Placement Testing in Computer-Assisted Language Testing: Validating Elicited Imitation as a measure of language proficiency

Riska Risdiani

s4582829

RiskaRisdiani@student.ru.nl

Master's Program in General Linguistics

Radboud University Nijmegen

2015-2016

Supervised by

Dr. Frans van der Slik

ACKNOWLEDGEMENTS

All praise to Allah SWT who granted me the courage and capability to enjoy the process of writing this master thesis.

I owe my deepest gratitude to my supervisor Dr. Frans van der Slik for his countless times of supervision, valuable, valid, and reliable guidance. I also thank Dr. J.J.M. Rob Schoonen for being the second reader of this thesis. Their feedback and supervision have been essential in achieving my master's degree.

I sincerely thank Edith Schouten and Novolanguage team for making it possible to gather data for my thesis and for the willingness to share all knowledge that enriches my way of thinking.

I will take this opportunity to thank Indonesia Endowment Fund for Education (LPDP) for providing me with the financial support to finish my master study. Finally, I am indebted to my family and my friends in Indonesia and Nijmegen. I highly appreciate their support and kindness. I can just say thanks for everything and I dedicate this thesis to them.

Abstract

Advances in computer-assisted language testing is motivating the enhancement and variation in the testing of language skills. Especially tests for placement purposes have seen a major development. This paper uses an Elicited Imitation task intended as a computer-assisted placement module for NovoLanguage. NovoLanguage finds it important to correctly assess the level of their learners, in this case hotel staff, before they start using their application. The study aims to contribute to the knowledge base regarding the efficacy, validity and reliability of the use of Elicited Imitation (EI) as a way of language assessment. The paper describes the correlation between a multiple-choice placement test form, Elicited Imitation, and video interviews. The data were all gathered in three hotels located in Indonesia and Vietnam on the same day. To investigate the validity and reliability of the multiple-choice placement test, 59 participants from Indonesia and Vietnam were tested and the data was analysed with TIAPLUS ©. To evaluate the accuracy of the test, the results were then correlated with video interviews with 54 participants. Afterwards, the Elicited Imitation was also correlated with the video analysis. The data were analyzed using descriptive statistics, differential item functioning analysis, inter-rater reliability analysis, and correlational analysis. Most analyses indicated that there is a strong positive correlation between the three different types of assessment to assess the level of English as a foreign language. This might be an insufficient evidence that Elicited Imitation is the most suitable assessment for placement purposes. However, positive and strong correlations are important evidence for the close relationship between the tests and indicate that EI is sufficient to replace the other assessment formats.

CHAPTER I

INTRODUCTION

Along with global modernization comes international communication in English as the lingua franca (Sun, 2012; Prabhu & Wani, 2015). International hotels are one such place where international communication is inevitable. Henley (2016) notes that hotels do not solely offer a bed, but also hospitality and information services to their international guests. In addition, if hotels want to stay full, they have to ensure high quality facilities and make sure that the staff is fully prepared to communicate in English (Blue & Harun, 2003; Selke, 2013). It is thus essential that hotel staff have a sufficient command of English.

Many international hotels in popular destinations like Bali, Indonesia and Vietnam (Pham & Thirumaran, 2016) try to provide an English learning course because the use of English in a hotel is vital for the hotel's quality and reputation among the guests. This calls for a type of (English) language learning tailored to the hotel industry (Moore, 2013), which means hotels rely on assistance from language learning providers or companies to help their employees learn English sufficiently. Hotels can also use Computer-Assisted Language Learning (CALL), which is a self-study program that enables employees to study at home without the need of a classroom (Widyastuti, 2015). Chapelle (2008) states that Computer- Assisted Language Learning assists people who want to learn a language independently, with minimal or even no aid at all from their teacher or instructor.

NovoLanguage is a Computer-Assisted Language Learning company with a speech technology platform based in the Netherlands (NovoLanguage, 2016). They support several hotels in Asia by providing tailor-made courses for learning English for specific purposes. NovoLanguage's courses use gamification techniques and Automatic Speech Recognition (ASR) to learn English with a specific focus on hospitality (Widyastuti, 2015). The modules are based on real situations from daily hotel life. Gaillard (2014) shows that language learning providers are appointed to construct course materials, develop lesson plans, and design various forms of language tests. The tests not only track the progress of the student, but also give an overview of the student's language comprehension before and after the learning process.

The language placement test is a type of test that is commonly used before the learning process. Carr (2011) states that the placement test is used to determine the appropriate level of

the students. This shows that NovoLanguage finds it important to correctly evaluate the level of their learners, in this case the hotel staff, before they start using their application. The level of the learners can vary because of differences in educational background and exposure to English as a foreign language. Thus, NovoLanguage is trying to develop a short, affordable and convincing computer-assisted placement test that automatically assigns new learners to the correct courses or modules. In addition, NovoLanguage wants to possess a valid and reliable automated assessment, so that they do not have to rewrite and pretest new items for every new target group. The advantage of having an automated assessment has also been stated in the literature, i.e. "the development of automated systems promises to significantly lower costs and increase accessibility" (Cook & McGhee, 2011, p. 30).

NovoLanguage has therefore designed a computer-assisted placement module in multiple-choice format. Aside from the multiple-choice format which consists of listening, speaking/conversation and focus on form subtests, the placement test also includes a sentence repetition task named Elicited Imitation (EI) as one of the subtests. EI is an assessment method instructing test takers to repeat or imitate a series of stimuli (sentences, phrases, words, and even sounds) (Yan, Maeda, Lv & Ginther, 2015). NovoLanguage has included EI as a placement purpose, because Computer-Assisted Language Testing (CALT) is usually only available in multiple-choice format (Carr, 2011). In addition, EI is well-matched with NovoLanguage's main feature; automated speech recognition (ASR). Automated speech recognition is a system that allows users to utter responses instead of pressing a dial pad (Rouse, 2007). ASR is expected to improve oral skills of the users of the NovoLanguage application. Rahim (2011) notes that the providers or learners of English in the hospitality industry have to realize that oral proficiency is the most important for hotel staff. In addition, EI also incorporates automated scoring (Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008) to save time checking answers. Ashwell (2014) argues that scoring oral assessments is time consuming and labor intensive work.

Using experienced assessors to evaluate learners' oral proficiency on the spot is also an expensive solution. If EI automatic scoring becomes plausible, immediate feedback becomes a reality, and the usefulness of the tests will greatly increase. The cost benefits will increase, because EI is less expensive. Therefore, the NovoLanguage team is willing to develop the EI task as placement test in order to replace the other subtests which use the multiple choice format.

Even though many scholars argue that partakers in an EI test will only be able to repeat utterances from memory, without actual knowledge of the meaning of an utterance (Yan et al, 2015), EI is still promising for the following reasons, according to Mozgalina (2015):

(1) EI is able to assess core language knowledge, such as grammar, vocabulary, and phonology. In addition, EI can test this in a relatively short time.
(2) EI provides an inexpensive proxy measure of second language speaking proficiency.
(3) The proficiency of partakers of the EI test is tested independent of literacy, because the participants have to repeat sentences, rather than read them. Therefore, EI is able to asses illiterate test takers, such as children or blind people.

This study will test whether the Elicited Imitation is sufficient for placement purposes, considering the beneficiary assertions stated above.

This thesis is structured as follows; the first chapter presents an introduction to the study. It will discuss current literature on English language learning for hotel or hospitality purposes, placement testing, and Elicited Imitation. The second chapter will review the relevant literature on Computer-assisted Language Testing (CALT) and validating Elicited Imitation. Chapter two will also include the research questions and hypothesis. Chapter three elaborates on the design of the study in chronological phases, the research methodology, participants, and materials. Chapter four presents the result. Chapter five will summarize and conclude this thesis, as well as suggest further research.

CHAPTER II
LITERATURE REVIEW

This section provides an overview of the theoretical perspectives on Computer-Assisted Language Testing (CALT). The various academic models for the use and the validation studies of EI will also be reviewed. This chapter also thoroughly discusses previous studies and EI approaches that have been developed.

2.1 Computer-Assisted Language Testing (CALT)

Since the 1960s technology has been employed to make language assessment more efficient. (Chapelle & Voss, 2016). Computer-assisted language testing (CALT) has been a prominent player in the language assessment field since the mid-1980s (Chalhoub-Deville, 2001; Carr, 2011). Chapelle (2010) differentiates three important motives why technology is used for assessing language. The first is efficiency. Automated writing evaluation (AWE) or automated speech evaluation (ASE) systems are employed in computer adaptive testing and analysis-based assessment for reasons of efficiency. The second motive is equivalence, which means that CALT is of the same standard as paper-and-pencil techniques. The third reason is that CALT is flexible and an appropriate assessment medium in many different situations. Embretson and Reise (2000) state that CALT is developed to provide stimuli that are optimally efficient for assessing the true ability of every single test taker.

Ockey (2009) notes that computer-assisted language testing is available in many forms, but that the producer of a CALT system has to assess four different skills: reading, writing, speaking and listening. The placement test is such a CALT. Even though not many studies have addressed the notion of placement (Green & Weir, 2004), the appropriate starting level is a prerequisite for successful language learning.

In 2003, a new variety of placement assessment, The Quick Placement Test (QPT), was created. It is an adaptive assessment to measure English language proficiency of the test takers designed by Oxford University Press and Cambridge ESOL in an attempt to provide teachers or instructors with a reliable and efficient method of investigating a student's level of English. The test is aimed at learners of all levels and all ages. The computer-based QPT uses multiple choice questions to assess students in listening, reading, and structure, including grammar and vocabulary. The test has been validated in three phases, in 20 countries by more than 6.000

students. After analysis of the reliability of the scores, the final QPT was created. The reliabilities reported for approach .90 for the 60 item test. (Geranpayeh, 2003). Unfortunately, this CALT test does not have a speaking component.

Another CALT provider is Duolingo . This is a web-based language learning startup tool which became publicly available in 2012 (Vesselinov & Grego, 2012). Duolingo has attempted to develop a Duolingo English Test (DET) (Duolingo, 2014). "Duolingo conveys to test proficiency in daily English for all four skills" (Ye, 2014, p.4). The DET also aims at assessing the general language proficiency and giving an indication of the level appropriate level of proficiency in English (Oxford University Press, 2016). The DET is also adaptive. People can simply take this test using their computer, smartphone or tablet. Duolingo plans to use the DET as a university admission test in the near future. However, Wagner and Kunnan (2015) believe it is not a sufficient measure of academic English profiency and is hence not suitable for university admissions.

However, none of these tests assess the user's speaking ability, but are more focused on receptive skills instead. Most of the CALT placement tests are multiple-choice. It is possible that luck plays a role. The user can simply guess or select the answer by ruling out implausible options. Therefore, a new, and valid and reliable computer-assisted language placement test is called for, which can replace the multiple choice format and can hence avoid bias. Any bias existing in the placement tests can result in a small number of false negatives (i.e. learners assigned to the wrong level). These are not high-stake tests and do not have a large impact on the test takers' lives, so it does not cause undue worry. However, the tests are still useful to help the test takers with exercises and keeping track.

2.2 Elicited imitation

Elicited Imitation (EI) is an oral skill assessment method that has been employed over the latest few decades in various contexts, including normal native language development, abnormal language development, and second language development (Graham et al., 2010). According to Gaillard (2014), Elicited Imitation (EI) is a psycholinguistic method of assessment during which the testees are asked to demonstrate their speaking abilities in more condensed way by, for instance, repeating one sentence in one attempt. Vinther (2002) mentions that Elicited Imitation is used in three different fields: child language, neuropsychological and second

language studies. In recent years, many scholars have taken an interest in Elicited Imitation as a way of testing oral skills in second language learners (Graham, McGhee, & Millard, 2010).

Gaillard (2014) found that there are two EI versions, which are used depending on the aim of the study. The first version, called naturalistic design, demands that the testees (mostly children) directly echo the preceding utterance by another speaker in a natural setting, without receiving specific instructions. The other variety is used in an experimental situation and uses a default set of sentences. This more structured application of the EI technique, which calls for validity and reliability evidence, asks the participants to repeat items which are constructed to test specific structures, such as grammar, vocabulary, and/or syntax, depending on the research focus.

Ortega, Iwashita, Norris and Rabie (2002) use EI to analyze second language proficiency in English, Spanish, German, and Japanese. Their objective is to test the validity, reliability, and usefulness of EI to test syntactically complex structures (Gaillard, 2014). Similarly, Chaudron, Ngyuen, and Prior (2005a) and Chaudron, Prior, and Kozok (2005b) use EI to measure adult language proficiency in Vietnamese and Indonesian. The results of this pilot research illustrate the typical and successful characteristics of EI performance, which can serve as the basis for future studies (Gaillard, 2014).

Chaudron et al. (2005a) developed two assignments (assignment A and assignment B) for Vietnamese non-native speakers of English. Each assignment comprises 48 different sentences that include various grammatical structures from standard Vietnamese speech. The test was administered as follows: the participant was allowed to listen to each sentence once. The participants were asked to imitate the sentence. The participant were assigned either assignment A or assignment B. Two Vietnamese native speakers rated the test on an adapted version of Ortega (2000)'s holistic scale (0-4). The outcomes demonstrated a Cronbach Alpha of .99 for both assignments. "Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group" (IDRE UCLA, 2016).

Chaudron et al. (2005a) substantiated the evidence of concurrent validity between the participants' mean scores on the Vietnamese EIT and their self-report adjusting the Common European Framework of Reference for languages (CEFR) scale (Council of Europe, 2001). The results proved evidence of concurrent validity, with a high correlation between the Vietnamese EIT scores and the self-report on assignment A (.80 for listening and .72 for speaking). In contrast, participants from assignment B presented less variability in the self-

reports. The correlation was low and accounted for .20 for listening and .14 for speaking. The authors suggest that this is because of the lack of variability in the participants' self-assessments. In addition, Chaudron et al. (2005a) indicate that the more familiar the given sentence was, the better participants imitated it. The proficiency level also played an important role, because a higher proficiency led to better repetition of the items, since the L2 grammar corresponds to the grammar used in the test sentences.

However, Chaudron et al. (2005a) cannot satisfactorily claim that EI is a useful measure of L2 proficiency, because the test results are not always straightforward. They note that further research is necessary to substantiate the claim. They therefore designed a baseline which includes an EI design for Vietnamese and concluded that the ability of a participant to imitate a sentence in a foreign languages depends on the knowledge of that foreign language. To sum up, EI is a rational measure of global proficiency (Chaudron, 2005b; Gaillard, 2014).

Zhou (2012) reports on a synthesis of 24 researchers using EI on adult second language learners and claimed that EI is overall a reliable measure (the internal consistency coefficient ranged from .78 to .96, p. 90). In addition, the correlation between EI scores and other measures of language proficiency was higher than .50 in the majority of the studies reviewed (p. 90), which includes several pieces of evidence for the construct-related validity for EI as a measure of language proficiency (Yan et al., 2015). "EI was conceptually classified as a measure of implicit grammatical knowledge owing to four features: (1) respond according to feel; (2) respond under time pressure; (3) focus on meaning; and (4) requires no metalinguistic knowledge" (Bowles, 2011, p. 157).

Vinther (2002) shows that there are four key task features that must be taken into account regarding the validity of Elicited Imitation: (a) length of sentence stimuli. "Sentence length has been frequently observed as a factor that influences the difficulty of EI tasks" (Yan et al., 2015, p.14). (b) delayed repetition. Time delay can be inserted after the test takers listen to the stimulus and before they imitate the sentences. The use of delay might cause intervention when eliciting the structure and meaning of the sentences (Vinther, 2002). Yan et al. (2015) report that EI tasks that applied delayed imitation (k = 13 (k stands for the number of studies), g = 1.25 (g stands for sensitivity of EI expressed by Hedges'g effect size), SE= .07) presented to be less discriminating than EI tasks that did not insert time delay (k = 11, g = 1.30, SE = .08), Q (1) = .31 (Q test examining the homogeneity of average effect sizes), p = .58; However, the 95% confidence intervals for the two groups largely overlapped, indicating that the use of delay

does not necessarily give much variation to the sensitivity of EI scores (Yan et al., 2015). (c) grammatical features of the sentence stimuli such as syntactic complexity, lexical difficulty, phonological structures and the employment of ungrammatical sentences (d) scoring rubrics (Yan et al., 2015). A further consideration that needs to be made during the development of an EI is how to score the elicited sentences. Various approaches have been applied to scoring EI performances: scoring based on the repetition of a certain structure, scoring based on the repetition of idea units, scoring that targeted different aspects of learners' proficiency, and finally automatic scoring (Yan et al., 2015).

2.3 Previous studies on Validating Elicited imitation

Validation is the process that legitimizes the test inferences made. This justification takes place through a compilation of pieces of evidence that motivate the proposed test interpretation and administration (Gaillard, 2014). According to Messick (1989), "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Messick, therefore, reduces the three types of validity (content-related, criterion-related and construct-related) to construct-related validity, because content- and criterion-related evidence contribute to score meaning. It is interesting to examine how his unitary view on validity has driven the reflection of test developers. The pieces of validity evidence are fundamental as they ensure that the test under development remains appropriate, suitable, and relevant over time.

The two most important objectives of a test are to assess language proficiency (Bachman and Cohen, 1998) and validate the interpretation and use of its results (Gaillard, 2014). A test is valid when it correctly measures the element that it is intended to assess (Hughes, 1989). Spolsky (1985) mentions that language test designers know that a test can never be completely accurate, because human language is too complex to be reduced to a single number. Therefore, scholars have to consider the limited information that testers receive, and should be more conscious of the social consequences of a test for test takers. As Davidson and Fulcher note, "validity theory occupies an uncomfortable philosophical space in which the relationship between theory and evidence is sometimes unclear and messy, because theory is always evolving, and new evidence is continually collected." (Davidson & Fulcher, 2007, p. 11).

According to Cronbach (1971), test validation is a multi-faceted notion that depends in part on the type of validity measured: content validity, construct validity or criterion validity containing concurrent validity and predictive validity. Content validity scrutinizes whether the assessment content is a valid measure of the skills that is assumed to measure. Construct validity is a way to test the validity of a test. It demonstrates that the test is actually measuring the construct it claims it measures. The inferences are made from test score interpretations and the construct being tested. "This type of validity examines the degree to which the test outcomes adequately reflect what the theory says about how that particular construct should operate" (Gaillard, 2014, p. 60). In addition, Cronbach describes that criterion validity needs to build an inference from test scores to performance. A high score on a valid test indicates that the testee has met the performance standard. For instance, language test 2 plays a role as a criterion against which the criterion of a new measurement (e.g. language test 1) has been created and is being validated. Thus, criterion validity examines the correlation of test grades with outside criteria.

> To assess criterion validity, two options are available. It is possible to establish either concurrent or predictive validity, which are two types of empirical validity that both require data to generate a numerical validity coefficient. The first one, concurrent validity, refers to the correlation that a test (e.g., language test A) has with another test (e.g., language test B) that is supposed to measure the same criterion (e.g., ability or language skill). The second one, predictive validity, indicates the extent to which a score on a test (e.g., language test A) predicts a score on another test (e.g., language test B). (Gaillard, 2014, p. 60-61)

Correlational analyses can be employed to test criterion validity. An independent variable could be used as a predictor variable and a dependent variable as a criterion variable. The correlation coefficient between them is called the validity coefficient (Cronbach, 1971).

As computer-based language testing proliferates, it is tempting to use EI as a placement indicator. EI requires testees to imitate stimuli in the target language, in this case English. The accuracy of the imitation is used as an indicator of language proficiency. Erlam (2006) tested the validity of EI as a measure of L2 implicit grammar knowledge. She used the participants' IELTS band for comparison. The correlations between EI scores and each IELTS score for

each skill, and their total IELTS score, were analyzed. The strongest positive correlation was found between EI scores and IELTS overall scores. Erlam (2006) suggests that IELTS measures the learner's implicit grammar knowledge and that overall EI performances represent it.

Cook & McGhee (2011) investigate the extent to which Oral Proficiency Interview (OPI) scores can be predicted using an EI test and analyze how to design an automated system to grade the EI. Gaillard (2014) investigates the possibility of implementing a French EI as a component of a language placement test. Her EI test comprised 50 items with at least 8 syllables and no more than 32 syllables. Furthermore, to reduce pressure on the participants, the experiments was self-paced. This allowed participant to take their time before moving on to the next item. She then designed a scoring rubric that targeted different aspects of a learner's proficiency. She developed six independent scoring rubrics to assess meaning, syntax, morphology, vocabulary, pronunciation, and fluency. For each criterion, the participants' responses were scored on a 7-point Likert scale. No score was given if a testee did not imitate the sentence or started repeating before the beep, violating EI directions. A maximum score of 6 was given when the testee imitated the stimulus perfectly. Participants were given a score from 1 to 3 if the utterance was up to 50% correct on a particular criterion and a score between 4 and 6 when their score was 50% or higher. Although the employment of different scales is advantageous for eliciting more specific information about test takers' second language ability, it is challenging to come up with a good description of scales and levels without any overlap. Gaillard (2014) notes that the descriptions of the vocabulary and the fluency scales were not sufficiently distinguishable, since the fluency scale included some elements of the vocabulary scale. Furthermore, the appropriateness of an EI for separately eliciting evidence about meaning, syntax, morphology, vocabulary, pronunciation, and fluency can be questioned in the first place. Gaillard concludes that EI as a measurement of French proficiency works well in the aural/oral modality. It is not difficult to operate and is reliable to rate, but still has several limitations such as the question whether EI is appropriate for separately eliciting evidence. Although placement tests are low-stake, Chapelle, Jamieson, and Hegelheimer (2003) argue that validating published computer-assisted tests is important, because test takers might see the tests as having a high-face validity, for instance in the case of a published placement test that is accessible to everyone.

Tracy-Ventura, McManus, Norris, & Ortega (2014) also investigated whether the scores on the French EI show any significant relationship to several points of oral language proficiency. They used Hulstijn's (2011) definition of oral proficiency as their baseline, which is restricted to the processing of oral language (listening and speaking) containing high frequency words, and grammatical, phonotactic and prosodic elements. They investigated the possible relationship between performance on the French EI and (a) the lexical diversity in a oral and written assignment, (b) vocabulary knowledge as measured by a vocabulary test, (c) the speech rate in a narrative task format, and (d) university final marks. A Pearson product-moment correlation was utilized in order to see the relationship. The Elicited Imitation materials consist of 30 test sentences ranging from 7 to 19 syllables. The EI items are presented in order from lowest to highest number of syllables. A native French speaker designed all the stimuli and they were checked by another native French speaker in terms of syllable length and naturalness. Tracy-Ventura et al. (2014) show that there is a significant and relatively large positive correlation between the EI scores and end-of-year grades (r =.78), lexical diversity in the oral interview data (r = .62), and speech rate in narratives (r =.67). These correlations provide evidence that the EI test yields scores that can be used as a tool for assessing French L2 oral proficiency. There were low and statistically not significant correlations between EI scores and lexical diversity in written essays (r = .32) and the scores for Meara and Milton's (2003) vocabulary test (r = .12). This is a kind of discriminant validity. The correlation of EI with final year grades and oral interview is logical, as both variables assess similar aspects of language proficiency, focusing on speaking and listening. Therefore, one would not expect very high correlations in different modalities and on relatively unrelated constructs such as vocabulary recognition and writing assessments.

Mozgalina (2015) applies an argument-based approach for validating the EI test for Russian. To investigate the accuracy of the score interpretation for the first use, Elicited Imitation was used with 97 Russian learners in Germany and the USA along with a background questionnaire. For the second use, EI was used with 67 Russian learners in the US together with the Russian Speaking Test, a listening comprehension test, and a C-test. Multiple descriptive, graphical, and inferential statistical techniques were used in the data analysis. She claims that EI is able to measure oral perception and production skills at the sentence level.

The results of Mozgalina's study demonstrate some counter-evidence to the assertion that EI addresses implicit knowledge. A higher correlation between EI scores with the length of the Russian study than with the length of residence in a Russian-speaking country indicates that it is less likely a measure of implicit knowledge than something else. If EI primarily measured implicit knowledge, there would be higher correlations with language learning in a natural setting, as in the case of residents in a Russian speaking country, who foster implicit knowledge and a large amount of exposure to Russian. This indicates that she restricts what is commonly believed about the use of EI.

2.3.1 Types of Validity Evidence

Standards for educational and psychological testing (AERA, 1996; APA, & NCME, 1966) distinguishes five types of evidence: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on the consequences of testing.

1. Evidence based on test content

   This first type of evidence is based on logical analyses and expert examination of the test, including items and task format. The evidence is also used to ensure that all parts of the test match the definition of the aim and the assessment is not biased against a particular gender, culture, or mother tongue.

2. Evidence based on response processes

   To gather this evidence, test designers observe test takers as they perform the required tasks or interview examinees to determine the reasons for providing their answers. Furthermore, this type of evidence targets the ways in which observers, judges, and raters use criteria to review and evaluate the behavior and performance of test takers without interference from irrelevant factors.

3. Evidence based on internal structure

   Evidence based on internal structure comes from the relationship between test items and test components. For instance, if the test items or tasks increase in difficulty, empirical evidence on the extent to which response patterns conform to this design is needed. Differential Item Functioning (DIF) is also included here to uncover if particular items function differently for identifiable subgroups of examinees.

Differential Item Functioning (DIF) is a statistical analysis in language testing that is able to evaluate whether test items perform similarly in different groups of testees. (Song, 2014). One example of DIF is the study on gender effects in the Pearson Test of English Academic (PTE Academic) (Song, 2014). PTE Academic is considered a recent language test. It engages an increasing number of testees and other pertinent stakeholders around the world (Song, 2014) and is used for admission, placement, and visa purposes. The study used statistical DIF methods combined with content analyses of test items and provided comprehensive and empirically-driven results regarding test validation and fairness.

DIF is a widely used method to detect bias items. DIF is possible in this case, because the participants come from two different countries: Indonesia and Vietnam. Although both countries are located on the same continent, their languages are very different. It is also interesting to see whether there are differences between males and females in performance on the computer-based test. "Test items exhibit DIF when testees with different background characteristics (such as gender, or cultural, social or linguistic) differ in their probability of answering these items correctly, after controlling for ability, or, formulated more accurately, overall test performance" (Van der Slik, 2009 p.278). Items showing DIF are carefully revised by test designers or simply omitted from the test.

4. Evidence based on relations to other variables

   This type of evidence is based on the dependence of test scores on external variables. Another way to test this is by means of a group separation study that can be used to test whether a particular instrument accurately predicts outcome variables.

5. Evidence based on the consequences of testing

   This type of evidence analyzes the extent to which expected or anticipated advantages occur, as well as to what extent unexpected or unanticipated disadvantages of testing occur.

Most studies in the field of validating elicited imitation have only focused on a formal educational setting. Thus, no previous study has given sufficient consideration to EI in English for specific purposes, such as in hospitality. This current study will address this gap by using an English Elicited Imitation test as a substitution of the placement test for the hotel staff.

Research questions

Based on the foregoing, two research questions have been formulated:

1. Is there evidence that Elicited Imitation could sufficiently replace multiple-choice testing for placement purposes?
2. Is there a positive relationship between performance on the multiple choice placement test, Elicited Imitation, and oral video interviews?

The hypothesis is that EI test is a sufficient placement test and that there are strong and positive correlations among the three different test. This study will gather essential knowledge about EI testing, including its evidence (i.e. validity, reliability, bias, access, administration) and build an argument for this language assessment. The study should prove that EI does not only increase the efficiency of test operation and grading, but also gives appropriate information about a learner's true ability. In addition, it is expected that this study will allow greater insight into the development of placement testing and test design.

CHAPTER III
METHODOLOGY

This study builds on previous research into EI as a test for placement purposes and aims to provide further evidence for EI as a holistic measure of English proficiency. This chapter will describe the design of the study, the materials, containing the specifications of three different tests including a placement and EI test, test takers, the data collection procedure in chronological order, different scoring criteria and the statistical analysis of the materials.

3.1 Design of the Study

The design of the study is based on correlation studies to investigate the construct-related validity of EI as a placement test. It included two groups of participants from two different countries (e.g. Indonesia and Vietnam) that were given the same tasks. They first did the placement test, followed by the EI task and finally, a video interview. This meant that this study consisted of the following four phases to validate EI as a placement test: (1) validating the Novolanguage placement test; (2) video rating; (3) Elicited Imitation rating; (4) correlation studies (investigating the relationship between scores on EI and other measures of language proficiency). Positive and strong correlations were expected between the multiple-choice placement test, the video interview and EI, since these three types of assessment addressed a wide range of language abilities. Correlations are important evidence for the close relationship between the tests. However, they do not inform about the distribution of the covarying scores. This can only be done with a scatterplot.

3.2 Materials

The study was based on placement test created by NovoLanguage. The NovoLanguage placement test aims at automatically placing test takers into the right courses or modules. It was expected that the result of the placement test will also provide information to hotel managers so they can assign staff to certain courses or modules. This requires a test that is able to assess language comprehension in a way that matches NovoLanguage's automated speech recognition (ASR) programme. The Novolanguage's choice of this format of the placement was motived by external reasons suited to the needs of the hotel industry, such as the practicability, effectiveness and the financial costs of the test. This placement module was designed based on the CEFR (*Common European Framework of Reference for Languages:*

Learning, Teaching, Assessment), which aids language test developers describing, creating, and reviewing language tests. There are six different levels: A1 and A2 (Basic User), B1 and B2 (Independent User), C1 and C2 (Proficient User) (Council of Europe, 2001). There were two different test formats. The first was in multiple-choice format and is often used in listening, conversation, and focus on form subtests. The second test format was the EI task that required the test takers to repeat each English sentence they hear. The sentence items increased in difficulty. There were no specific rules for delay time in the Elicited Imitation tasks. The placement test pilot was built up as follows: A1 listening = 10 items, A1 conversation = 8 items, A1 focus on form = 9 items, A2 listening = 8 items, A2 conversation = 8 items and A2 focus on form = 10 items. Furthermore, there are 10 sentence repetition/EI items. Altogether there were 63 relevant items.

3.2.1 Listening

In the listening section, the user heard the situation description and the question in audio format. In addition, they read the question on the screen. This was followed by a short pause to answer the question. The users had to click 'next' to go to the next question. The user could listen to the fragment twice. Ockey (2009) notes that in listening assessments in CALT should contain authentic items. In other words, the items should occur in everyday speech situations. Thus, the test takers were exposed to items/questions which represent the foreign language they were studying (Lewkowicz, 2007). The items in the NovoLanguage placement listening section were consistent with this view.

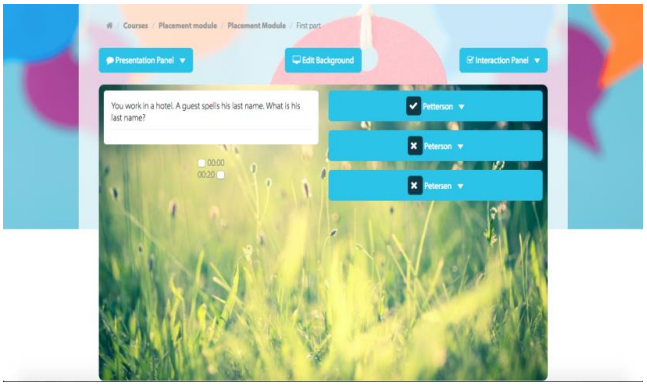


Figure 1. NovoLanguage Placement Module Listening Section, dev.novolanguage.com (2016)

Figure 1 is an example of a question in the listening section. The hotel staff were asked to mention the last name of a guest that the guest has spelled himself. In the next item, the test taker was situated in a restaurant. A guest ordered food and the test taker had to answer a question about the food ordered by the guest. The scenarios in both fragments were real life situations. These two items from the NovoLanguage placement test were not only authentic, they also confirmed one of the categories of the Language for Specific Purpose (LSP) assessment stated by Bachman and Palmer (1996). Carr (2011) notes, as do Bachman and Palmer, that branch-specific knowledge addressed in specific test items can be relevant to show that the test taker meets the job standard.

A. A1 Listening

In the A1 listening section, the test takers listened to a dialogue. This subtest was designed with the expectation that the test takers:

- Can follow speech that is very slow and carefully articulated, with long pauses to understand what is said.
- Can understand careful and slow instructions and follow short, simple directions.
- Can understand short, simple descriptions of people, situations and things.
- Can understand numbers, prices, the alphabet, dates, days of the week and other simple topics. (CEFR, 2001)

B. A2 Listening

In the A2 listening section, the test taker heard a fragment that is developed with the expectation that the test taker:

- Can understand short, clear, simple messages and announcements.
- Can understand enough to be able to understand provided speech that is clearly and slowly articulated.
- Can understand important phrases and expressions (e.g. very basic personal and family information, shopping, local geography, employment).
- Can identify the topic in slow and clear speech
- Can understand simple directions on how to get from X to Y, by foot or public transport.

- Can understand and extract essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly.
- Can follow changes of topic in factual TV news items, and form an idea of the main content (CEFR, 2001)

3.2.2 Conversation/speaking

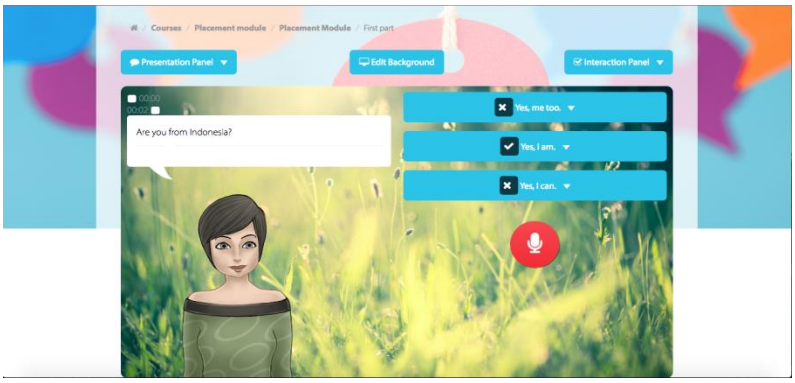


Figure 2. Avatar in the NovoLanguage Conversation Section, dev.novolanguage.com (2016)

In this conversation or speaking part, test takers were given a specific situation (prompt) by the avatar. The avatar is a virtual guide that gives instructions and asks the questions. The test takers listened to the avatar's question and later they were asked to respond by uttering one of the three available options. This kind of assessment followed the CALT trend of integrating listening and speaking skills in one single section (Ockey, 2009). The goal was to create an interactive situation for the test takers so they could produce a correct utterance based on the relevant input (Sawaki, Stricker, & Oranje, 2008).

The A1 speaking framework (CEFR, 2001) for the test is:

- Greetings, asking how people are and common standard expressions
- Describing age, address, hobbies, numbers, quantities, cost and time
- Describing people's appearances

According to the A2 speaking objectives test takers should be able to:

- use a series of simple phrases and sentences to describe their *family* and other people, *living conditions*, their *educational background* and their present or most recent *job*.
- construct *very short understandable utterances*, even though pauses, false starts and reformulation are evident.
- communicate in simple and routine tasks requiring a simple and direct exchange of information to do with work and free time.
- handle very short social exchanges, but do not have to be able to keep the conversation going.
- use simple everyday polite forms of greeting and address.
- handle invitations and apologies.
- say what they like and dislikes.
- discuss their plans in the evening or during the weekend.
- make and respond to suggestions.
- agree and disagree with others.
- discuss what to do, where to go and make arrangements to meet.
- ask for repetition when they do not understand what is being said.
- ask for and provide everyday goods and services.
- get simple information about traveling and the use public transport: buses, trains, and taxis. They should be able to ask and give directions, and to buy tickets.
- ask about things and make simple transactions in shops, post offices or banks.
- give and receive information about quantities, numbers, prices etc.
- make simple purchases by stating they want and asking the price.
- order a meal.
- deal with common aspects of everyday living.ask and answer questions about habits and routines.
- ask and answer questions about pastimes and past activities.

3.2.3 Focus on Form (FoF)

Focus on form means that the subtest comprises a compilation of grammar and vocabulary for the A1 and A2 levels. In this placement module, A1 FoF items were designed to measure the command of highly frequent isolated words and phrases related to everyday communicative situations (CEFR, 2001). In terms of grammar, A1 FoF items

expected a limited control of simple grammatical structures and sentence patterns from a learned repertoire (CEFR, 2001). Aside from A1 FoF items, there were also A2 FoF items. In the A2 FoF test the test takers were expected to be able to use simple constructions to fulfil basic communicative needs, but they could still make basic mistakes, for example mixing up tenses and forgetting to mark agreement. Nevertheless, it should be clear what they are trying to say.

3.2.4 Elicited Imitation

The study contained 10 EI items that vary in length. There were no EI researchers who mentioned the typical length of a sentence. It is important for the design of EI test that the quality of the items is high and that the range of syntax and vocabulary items vary in their complexity and that they represent the construction that the researcher wants to analyze, especially when aural and oral proficiency in a second language are tested (Bley-Vroman and Chaudron, 1994; Gaillard, 2014).

Table 1

Stimulus Elicited Imitation

NUMBER	SENTENCES	WORDS AMOUNT/ SENTENCE LENGTH
1	I work here	3
2	Her son is four years old	6
3	That computer is broken	4
4	My sister is afraid of spiders	6
5	Playing tennis is my favourite hobby	6
6	I am afraid I cannot remember your name	8
7	After the meeting had finished they all went to a nice restaurant	12

8	You should never have allowed him to go to that awful museum	12
9	I cannot believe you never told him you used to live in the city	14
10	She finally admitted that it was her father who had stolen the famous painting	14

Table 1 shows that the length of the stimuli varies between three and fourteen words. The difficulty of the items increases and, vocabulary and semantic plausibility were used as a basis for item construction, so that they can be applied and scored automatically in an ASR device. After the EI items were created, the stimuli were audio-recorded and digitized using the high-quality ASR-system from Novolanguage. All stimuli were recorded by a native speaker of English, on the same day, under the same conditions, and at a consistent volume.

Graham, McGhee, & Millard (2010) report that although the test takers are not familiar with the sentences they have to imitate, they can still process them when the sentences are short. However, as the sentences become longer and complicated, the test taker must have a remote memory of such a sentence to understand its meaning. Furthermore, there are no specific rules on the delay time in this type of placement test. The use of delay might cause intervention when eliciting the structure and meaning of the sentences (Vinther, 2002).

Table 2

English Attributes Used for the Elicited Imitation Sentences Construction

Sentence Attributes	**Explanation & Examples**
Sentence type	(1) Declarative - I work here - Her son is four years old - That computer is broken - My sister is afraid of spiders

	- Playing tennis is my favourite hobby - After the meeting had finished they all went to a nice restaurant - She finally admitted that it was her father who had stolen the famous painting (2) Negative - I am afraid I cannot remember your name - I cannot believe you never told him you used to live in the city (3) Imperative - You should never have allowed him to go to that awful museum
Modifier presence feature	(1) Adjective (e.g. broken, favourite, afraid, nice, famous, awful) (2) Preposition (e.g. to,of) (3) Adverb (e.g. here, in the city) (4) Singular/Plural (e.g. hobby, spiders, painting)
Tense & mood	(1) Present , (2) Past, (3) Past continous (4) Present perfect
Length	Short, < 3 - 4 words> Medium, < 6 - 8 words > Long, < 12 - 14 words >

Table 2 provides information about the variety of stimuli created by the Novolanguage language expert. Following previous EI studies, the length of the stimuli was controlled, a variety of grammatical structures were targeted, frequent vocabulary was used, and all

sentences were grammatical. This reduces the risk of parroting or the role of working memory.

3.2.5 Video Interview

The interviews were held on the same day as the placement tests. Participants were asked to answer each question in English. The questions were developed by language experts from NovoLanguage and are mainly based on length, difficulty, and diction. The interview was designed to test the English speaking skills of the test takers. On average, the test takers took at least 15 minutes to answer all questions.

The questions can be used for speakers from A1 to B1 level. The amount of questions could vary between participants, depending on how well they understood and answered the questions.

Table 3

Video interview question lists

Number	Queries
1	What is your name
2	What is your full name?
3	Where do you live?
4	What is your address?
5	What is your job?
6	What do you usually do as... (depends on the job) ?
7	How long have you been working here?
8	Do you like your job?
9	Do you interact with guests?
10	What are the most frequent questions that the guests ask you?
11	Have you ever met difficult guests?
12	Have you ever encountered problems with guests?
13	Are there guests that complain a lot?
14	What do you do to solve the problem?
15	What do you like most about working here?

16	What is the most difficult part of your job?

After the video interview, the test takers participated in a short role play, which was based on their task in their daily job at the hotel, in which the interlocutor acted as a guest. The interlocutor is the same interviewer from the previous video interview. The guest role started the conversation by asking about a certain location, the facilities, tourist information or he would ask general questions directed at housekeeping, spa crew, concierge staff, engineers, the food and beverage team, and the security division.

3.3 Participants

All tests were conducted in Indonesia and Vietnam. There were 59 participants of which 54 participants work in three different hotels: 20 staff member (11 males and 9 females) of Alila Villas Soori Hotel and Resort in Indonesia, 17 staff members (all males) of Sanur Paradise Plaza in Indonesia and 17 staff member (7 males and 10 females) of Intercontinental Nha Trang, Vietnam. The other five participants were NovoLanguage employees who tested the placement test and the EI, but did not do the interview or role play. Hence, they were excluded from the video analysis. The mother tongue of the participants who did all three assessments (multiple-choice, Elicited Imitation, and video interview test) is either Bahasa Indonesia or Vietnamese. They come from various divisions in the hotel industry such as housekeeping, spa, concierge, engineering, food and beverage, and security. All of them are literate and have no problem with hearing. They are familiar with computer use but they have no experience taking tests based on the CALL system. Furthermore, they are not taking English courses outside of their self-study at work or during the study. There is no information about their current proficiency level in English.

Figure 3 presents an overview of demographic information of the participants. 33.9% of the participants work at Alila Villa Soori Indonesia, 28.8% at both Sanur Paradise Plaza Indonesia and Intercontinental Nha Trang Vietnam and 8.5% employees at Novolanguage. In addition, there is an uneven amount of male and female participants.

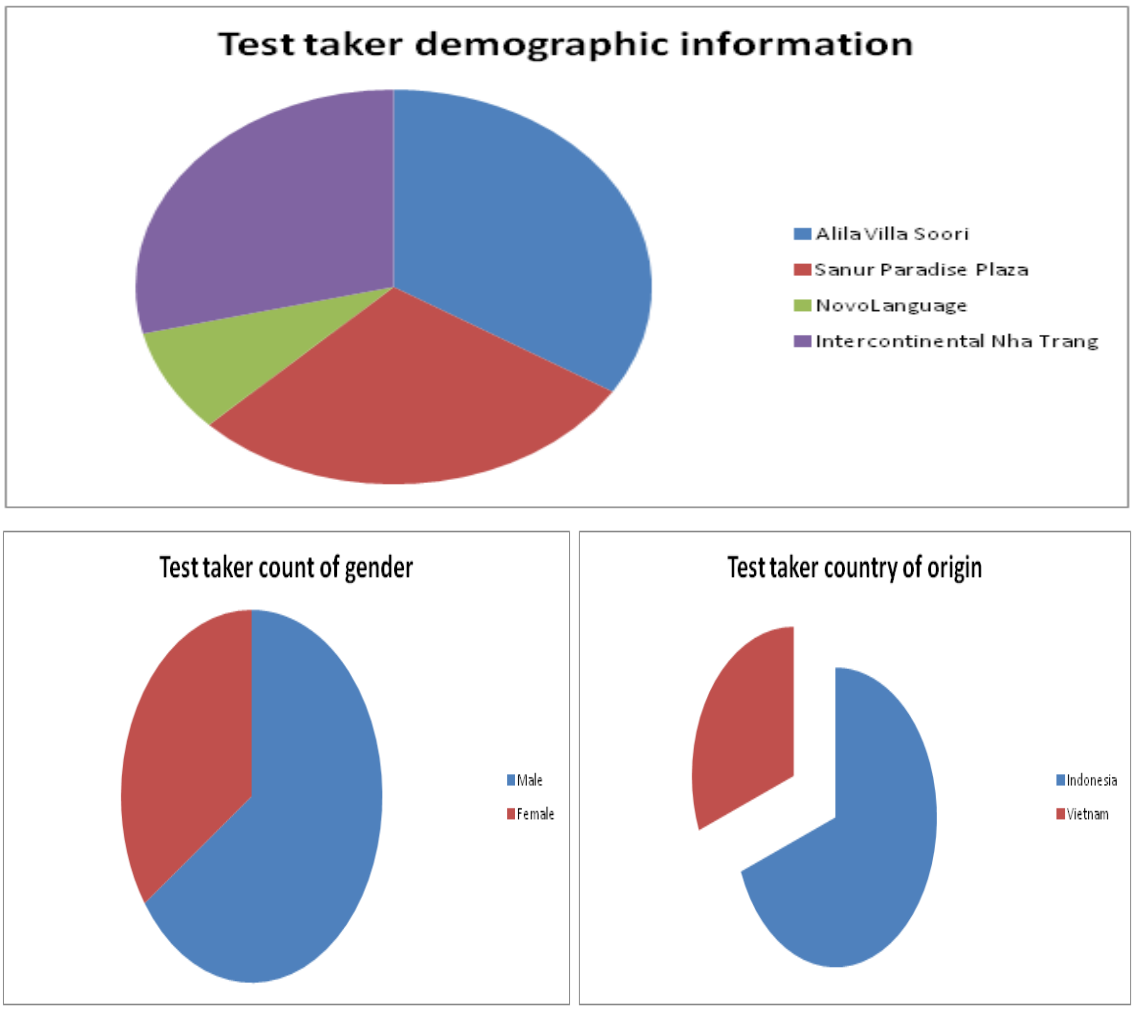


Figure 3. Demographic participant information

3.4 Procedure

The placement test and the interview were done on the same day. First, the test takers made the test in a quiet room with a laptop, mouse, headset with microphone, and fast internet connection. They signed in and were presented with a screen showing the test. The NovoLanguage placement pilot was administered in several parts. The first part started with the listening items. The test takers could play the audio twice. After the listening part, they were asked to answer a given question that assessed their conversation skills by using the record button on the laptop screen. In the Focus on Form (FoF) part, they simply had to click on the correct answer. During the Elicited Imitation task they had to listen to the sentences very carefully, pressed the record button and repeated the sentence in the microphone headset to the

best of their abilities. When the participants had imitated the sentence, they had to left-click to be able to continue to the next item. Each test taker received the sentences in the same order and only heard the sentences once. All sentence imitations were recorded and saved for later analysis.

Furthermore, there was an invigilator present during every test, who supplied information and instructions about the experiment, so that the participants would feel comfortable and not fear for their jobs. During the placement test the participants only used the computer mouse. After the placement test, the participants did a short interview and role play section with the same invigilator. There was no information about their proficiency level in English. The interview was filmed and transcriptions and global impressions (based on the CEFR characteristics) were made to determine the participant's proficiency level. The placement test, the EI task, and the interviews were done in a soundproof room. This resulted in recordings with good sound quality with as little noise as possible.

3.5 Scoring rubric

The audio-recorded sentences (10 sentences per test taker) needed to be assessed by different raters to ensure the validity and reliability of the scoring system. Scoring methods can be either holistic/global or analytical. Each scoring method has advantages and disadvantages. The holistic method is faster but less reliable, because of the lack of information and because the scales do not always apply to one single skill. One inevitable drawback of the analytical method is that analytical marking is time-consuming. The scoring method that should be employed depends on the assessment condition. If time is not a major obstacle, analytical scoring is the better choice (Carr, 2011).

The video analysis is scored with a holistic scoring guide based on CEFR categories, which means that raters must make decisions about the students' abilities, but can only evaluate one level score from their oral production. Aside from that, scoring the EI test takers' productions was an interesting and challenging task to conduct. However, there is no standardized scoring method for rating EI items (Vinther, 2002). Ortega et al. (2002), Chaudron et al. (2005), and Graham et al. (2008) employed holistic scoring where descriptors for each level were used to evaluate the test takers' oral productions (see Table 4 and Table 5).

Table 4

Scoring grid used by Ortega (Gaillard 2014)

Score	**Score Description for the Holistic Rating Scale**
4	Perfect imitation
3	Changes in content or changes in form that affect content
2	Changes in content or changes in form that affect content
1	Imitation of half or less of the stimulus
0	Silence, only one word repeated, or unintelligible repetition

Table 5

Scoring Rubric by Chaudron et al. and by Graham et al. (Gaillard, 2014)

Score	**Score Description for the Holistic Rating Scale**
4	Test takers produces perfect imitation
3	The original, complete meaning is preserved as in the stimulus
2	The content of the repetition string preserves at least more than half of the idea units in the original stimulus string
1	Only about half of the idea units are represented in the string but a lot of important information in the original stimulus is left out
0	Repetition that produces nothing (testee is silent)

In the forthcoming NovoLanguage automated scoring system, utterances can receive a score ranging from A to M. A will always be the correct option and gives the maximum score. The test taker will get this maximum score when they repeat the stimuli perfectly (e.g. "You should never have allowed him to go to that awful museum", see Table 1). Score B means that speakers only produced part of the utterance (i.e. you should never have allowed him to go to that awful). However, it is still necessary to think about the best way to assign a score to the sentence repetition task. For a sentence such as "you should never have allowed him to go to that awful museum," A would be the perfect score and M would be the lowest score (you). F

(you should never have allowed him to go) is better, however. It might be interesting to look at the number of correct words produced/recognized in the right sequence. The study's manual rater scoring results can be adapted for future automated scoring.

The EI items were rated by analyzing audio files from the application. Raters will have access to the audio files and judge the correctness of the sentence items. Raters were allowed to play the audio file several times, in its entirety or in part. Raters used two different types of scoring guides. Following the first scoring guide (see Table 6), the raters would give "1" point for every single word in the correct order. If the test taker failed to imitate the original word order, the scoring was stopped. For the second scoring type (see Table 7), the raters would count every correct word in the sentence regardless of the order. Small lapsus linguae (slips of tongue) and mispronunciation were still counted as long as they did not deviate from the original meaning. The maximum score possible for both scoring rubric is 85. See the tables below for examples of the scoring method.

Table 6

Scoring guide based on the number of correct words in correct order

EI production	**Score**
I work here.	3
I … here.	1
I work	2
After the meeting had finished they all went to a nice restaurant.	12
After the meeting had finished they .... to .... restaurant.	6
After the meeting **have** finished	3

Table 7

Scoring guide based on the number of correct words

EI production	Score
I work here.	3
I … here.	2
I work	2
After the meeting had finished they all went to a nice restaurant.	12
After the meeting had finished they to restaurant.	8
After the meeting **have** finished they	5

3.6 Raters

Raters can be worthy and reliable in assessing a student's language ability, but not every rater's rating is worthy and reliable. Raters have to be instructed about the scoring grid and behave accordingly, for instance. Three raters scored the test takers performance in the Elicited Imitation task in this study; one experienced female rater who is also a senior trainer in the English Department of Radboud in'to languages, one native speaker of English from Nottingham and the author of this thesis. The experienced female rater and the author of this thesis scored the video interview based on CEFR frameworks. By drawing on rating experience, combined with teaching experience and mother tongue skill, the raters provide a critical and reliable interpretation of the students' language comprehension. Furthermore, gathering the perception of the raters on the rating process may help reveal information that would be lost if only using scales.

3.7 Analysis

The data was analyzed quantitatively and qualitatively. For the quantitative statistical analysis TIAPLUS © and SPSS 21 were used. The qualitative method comprised a study of the video interviews and a global impression based on CEFR framework. All data sets were compared to assure the validity and reliability by correlating the placement test results from each participant with the rating results from the video interview. Afterwards, the score

from EI and placement test were also correlated. Furthermore, inter-rater reliability was checked using Cohen-Kappa analysis.

To establish content validity, test items were examined for their representativeness of the target domain (Doe, 2013; American Educational Research Association, 1999). The test results from the placement module of each participant were calculated and analyzed using TIAPLUS ©.

> TIAPLUS © is a Windows program for test and item analysis within the framework of Classical Test Theory (CTT). The program offers flexibility in scoring the item answers and can handle missing values, subgroups and subtests, and tests that have mixed types of items (item formats, answer formats). TIAPLUS © also allows for DIF analysis, creates both numerical and graphical analysis results, and can report numerous item- and test characteristics, among which the GLB (Greatest Lower Bound coefficient) that gives the optimal estimate for test reliability (CITO, 2005).

The software is developed by CITO *(Centraal Instituut voor Toetsontwikkeling* "Central Institute for Test Development"*)*, a global company specializing in tests and assessments (CITO, 2005). TIAPLUS © software is available free of charge and is solely designed for scientific purposes. Since it provides the whole score of test and a subtest score, a comparison of both scores has to be made to get the overview and evidence for the test validation.

Furthermore, the overall scores on the sentence repetition task must be analyzed in order to see if there is a correlation between the score on the EI task (once for sentence order and once for words) and the score on the placement test. Correlation means two or more phenomena occur together and are hence dependent. Field (2013) notes that a correlation expresses the strength of linkage or co-occurrence between two variables in a single value between -1 and +1. This is called the *correlation coefficient*.

All data were analyzed in SPSS 21. The statistical analysis also included DIF analyses to check if female and male test takers performed differently and whether the DIF existed between Indonesian and Vietnamese employees. In addition, DIF analyses were performed by means of the Mantel-Haenszel statistics in the TIAPLUS © package (CITO, 2005). Afterwards, an inter-rater reliability analysis using the Kappa statistic was performed to determine consistency among raters.

CHAPTER IV

ANALYSIS AND RESULTS

Cronbach (1971) highlighted the need for various sources of evidence in order to adequately interpret the test takers' abilities. According to Messick (1989), test validation is an inquiry-based process that provides evidence and arguments in consistent support of, or against interpretations and uses of test scores. Therefore, several validation analyses were conducted to justify the use of the tests.

4.1 TIAPLUS © analysis placement test

Doe (2013) mentions that the purpose of placement assessment is to assign pupils to a group with identical learning demands. Therefore, this kind of assessment has to test four different skills: reading, listening, speaking, and writing (Ockey, 2009). TIAPLUS © is a program owned by CITO that can be used to analyze a language proficiency test (or any test for that matter) in terms of internal consistency in a very detailed manner. 53 items were investigated in this TIAPLUS © analysis. Ten EI items were excluded and separately analysed. 54 employees of three different hotels took the test. They received 1 point for each correct answer, so the maximum test score equals 53. The main results are as follows: the total scale is highly reliable, Coefficient Alpha = .88. Coefficient Alpha is a measure for the (lower bound of the) reliability of the test scores. It can also be interpreted as a measure of internal consistency. The reliability of a test has consequences for the decisions made based on the cut-off score. The less reliable a test is, the larger the likelihood that test takers will undeservedly pass or fail the test.

The test seems to be easy, since the average P-value = .78. The item P-value (multiple-choices items) or P'-value (non multiple-choice items) represents the difficulty of the item in the population (sample) tested. It is calculated by summing all available item scores for the item and dividing this sum by the item maximum score times the number of participants. High values indicate that the item was easy. P-values of '0' and '1' imply that the item was superfluous. Aside from P-values, there is also the A-value indicating the distractor value. It is the proportion of test takers that opted for one of the incorrect answers. A-values can also be used to check for coding errors or ambiguous items.

The analysis showed that 6 items had a 100% correct score (16, 17, 18, 23, 25, and 26). This means these items were easy and should be revised. Since there was no variation for these items, a factor analysis cannot be performed. A two-factor analysis is used to test if the test is either one-dimensional (homogeneous) or multidimensional (heterogeneous). The mean test score was 41.41. This means that the mean score did not differ significantly from the maximum score, which is 53. In addition, the mean Rit-value was .40. Rit is the correlation of the item with the total test (*item included*). It provides information on the discriminative power of the item. TIAPLUS © reports the arithmetic mean of the Rit coefficients in a test, because one cannot simply sum the Rit values (being Pearson product moment correlation coefficients) and divide by their number (CITO, 2005). Other statistic analyses at the item level yield the Rir and Rar value. Rir shows the correlation of the item with the rest of the test (*item excluded*). It is an alternative measure of the discriminative power of the item. Rar measures the correlation of an incorrect answer with the rest of the test. Hence, test developers aim at items that show strong and *positive* Rirs and Rits and *negative* Rars.

AR stands for Alfa-rest coefficient ('alpha if item deleted'). It provides information on the reliability of the test in case that particular item is removed. If alpha decreases, the item behaves well psychometrically, while it performs poorly when alpha increases. Beside the entire scale, TIAPLUS © also provides subscales containing information about the subtests (listening, speaking/conversation, Focus on Form) characteristics. From the subscales analyses, it is clearly that the reliabilities of the subscales is rather low, ranging from .24 to .68. This, at least partly, can be attributed to the low number of items in each scale.

Subtest 1: A1 Listening

In the A1 listening section, there were no items that the majority of the participants answered wrongly or correctly. Item 1 has a relatively low Rit and Rir value, so a relatively low correlation with the test in- and excluding item 1 itself compared to other items. The average Rit was .45, while for item 1 it is .30. The Rir for item 1 is .17. Furthermore, it was remarkable that the coefficient alpha increased when item 1 is removed. The p- and a-values show that 93% of the students chose answer 1, 7% chose B, and nobody chose C. This item was probably too easy, because it asked about the guest last name ("You work in a hotel. A

guest spells his last name. What is his last name?"). Although the question was not difficult, the answer alternatives were still plausible (response options: Petterson / Peterson / Petersen).

The Rit and Rir values for for item 7 with the pictures as alternatives (see *Figure 4*) had a negative value, -.2 and -.26 respectively. However, the p- and a-values for item 7 were not remarkable (69% chose the correct answer, 10% and 20% for the two alternatives), so the only possibility was that those students who scored high on other items scored low on this item, and vice versa.

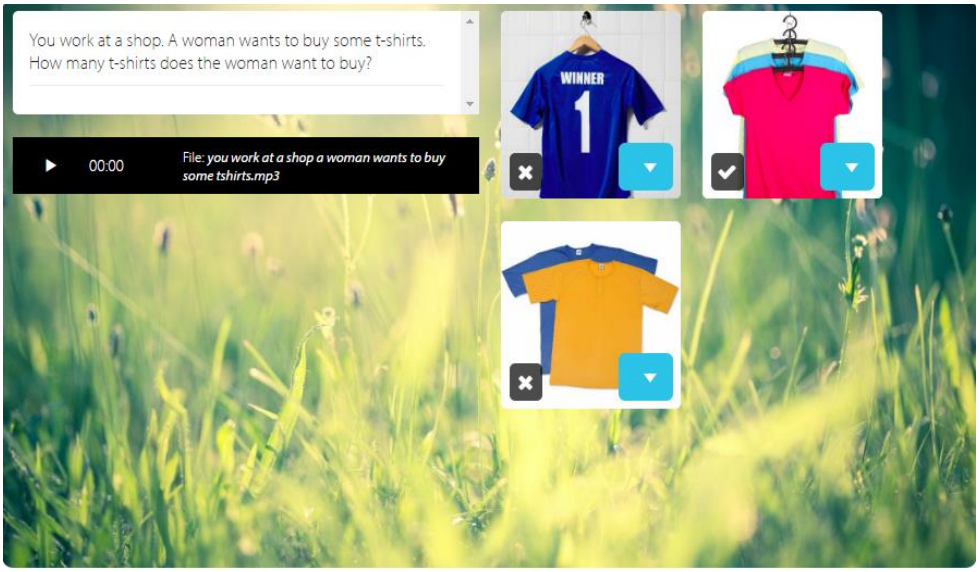


Figure 4. Item number 7 in Novolanguage multiple-choice placement test, dev.novolanguage.com (2016)

Subtest 2: A1 conversation

Item numbers 11 until 18 tested the conversation skills at A1 level. The maximum test score was 8 (because there were 8 items, and each correct item gave 1 point). The average test score was 7.2, which means that the test was quite easy. This was also reflected in the P-value, which was .90. Because the subtest only consisted of 8 items, the overall scores did not give a detailed explanation. The best way to judge the specific items was to look at the AR score: the alpha (= reliability) improved when the item was removed. The overall alpha was .49, which is quite low.

Item 11 asked about hobbies. Two participants gave more than one answer, while 1 person did not answer the item at all. This also holds for item 12. Item 13, 14, and 15 seem to

cause no problems because more than 80% of the test takers gave the correct answer. Item 16, 17, and 18 all have a 100% p-value (see Table 8), which means that everyone answered this question correctly. This means these items were quite easy. The three items were asking about personal information such as country of origin, age, and nationality.

Table 8

100% correct Items in the Placement Test

Item Number	Question	Alternatives
#16	Where are you from?	From England. / At 3 o'clock. / It's Monday.
#17	How old are you?	In America. / I'm 35. / Yes, I am.
#18	Are you from Indonesia?	I can't. / Me too. / Yes, I am.

Subtest 3: A1 Focus on Form

This subtest contained 9 Focus on Form items. The test-takers gained 1 point for each correct answer, so the maximum test score was 9. The mean test score was 8.1 and the average P-value equals .90, which means that the Focus on Form subtest was considered easy. The coefficient Alpha is .24, which means that the test is not reliable. However, this subtest comprised only 9 items, so it is unfair to be compared to the whole test.

Looking at the details, the average Rit is .44. All participants are able to correctly answer items number 23, 25 and 26 (see Table 9), which have a Rir value of 0. Furthermore, it can be concluded that these three items are probably too easy, or the alternatives are not plausible enough.

Table 9

100% correct Items in the Placement Test

Item Number	Question	Alternatives
#23	My name ... Susy.	are / is / called
#25	What is ... last name?	choose / all / your
#26	... are you from?	Where / Who / What

The Rir value of item 19 is -.17 and only three percent of the participants chose alternative B. There were two test takers who gave more than one answer or left it open and one person who did not answer the question for item 27.

Subtest 4: A2 Listening

Looking at the items from the A2 listening test in detail, item 28 is quite remarkable, as there was one missing answer, and it happened twice that a student left out this item or gave two answers. The p- and a-values show that 79% of the answers is correct for this item, while 7% and 14% are wrong, which means that this item should be reconsidered.

Subtest 5: A2 Conversation

The alpha for the entire test is fairly good, with .68, but because there are only 8 items, this is hard to judge by itself. Out of a maximum score of 8, the average score was 6.05 and the average P-value .75, indicating that this subtest was made well. Looking at the individual items, item 41 has a good Rir-score (.66), but the alpha decreases (.56), which indicates that leaving this item out decreases the reliability. However, the P-value of this item is 53%, meaning that just over half of the people had this item correct, which could mean that this item was difficult. This also holds for item 39, which has a P-value of .54, but an AR of .60, which is fine. In general, these items seem to be doing fine.

Subtest 6: A2 Focus on Form

There were no missing answers in this A2 Focus on Form subtest. The coefficient alpha was .66. There was no item where the majority of the students the correct or wrong answers. The maximum test score is 10 for this part. The mean test score was 7.25 and the average P-value is .72, which means that the A2 Focus on Form subtest is not really difficult (since the lowest average P-value was .05 for the A2 Listening test). Looking at the details, there were no values which differed significantly or which were negative. The average Rit is .51. The factor analysis suggests that there were no major issues with this subtest.

Multiple-choice placement test cut-off score

After the tests, a cut-off score was established by two language testing experts from NovoLanguage to group the participants. The scores are illustrated in Table 10. There are 35 participants on A1, 8 participants on A2, and 11 participants on B1 level.

Table 10

Cut-off score based on TiaPlus result

Number Participants	**Items correctly answered**	**Score**	**Modules/Lessons**
35	0-21	0-75%	A1
8	22-27	>75-90%	A2
11	>27	>90%	B1

4.2 Differential Item Functioning (DIF) Analysis

Differential Item Functioning means that if there are subgroups within a population (gender, ethnicity), there is a possibility that items function differently, as a result of cultural bias. This study looked at the overall differences in performance between males and females and between participants from Indonesia and Vietnam. Female employees (M = 45.10, SD = 5.45) scored higher than males (M =39.37 , SD = 7.69) on average, $T(56) = 3.32$, $p < .01$.

DIF analyses reveal that none of the items behaved differently for male and female employees, since the Z-scores of the Mantel-Haenszel DIF statistics did not exceed the absolute value of 2.58 in any of the cases. Indonesian employees' scores (M = 40.21, SD = 7.84) did not

differ significantly ($T(50) = 1.90$, $p > .05$) from Vietnamese employees' scores ($M = 43.75$, $SD = 6.15$).

4.3 Analysis Video

The CEFR scale is based on an action-oriented approach and embedded in a fundamental hierarchy. The CEFR considers language users participants within society who want to accomplish tasks in certain circumstances, environments or fields of action. The scales describe competence along two broad dimensions: the **quantity** dimension (the number of tasks a person can perform successfully by using language, in what number of contexts, in relation to what number of themes, domains etc.) and a **quality** dimension (how effectively and efficiently a person can achieve their goals through language use (Council of Europe, 2001).

The results from the video interview allocate the test takes in three different CEFR levels; A1 or A2 (basic user) and B1 (independent user). The reason why this specific scale has been chosen is that an overall scale is not sufficient to evaluate specific oral productions in specific conditions. For instance, Hulstijn (2007) identifies that in overall oral production, a B1 speaker can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest. In this case, that impression is considerable global. There are no specific characteristics about the extent to which the speaker meet the requirements and how fluent he is as a B1 user. He might be good at producing the vocabulary, but have difficulty in using the correct B1 grammar. Considering this issue, CEFR interview and conversation scales were used to be able to obtain the more concrete, suitable and valid interpretation. Table 11, 12, and 13 shows excerpts of the video interviews from three different levels.

Table 11

Test taker A1 Level sample analysis

Participants	**Transcription**	**Impression**
A1-1	My name is xxx. I don't understand. Job is housekeeping. I'm job in the room. Cleaning in the room. Role play: Cleaning bathroom? Okay. Extra	This participant did not understand the utterance of his interlocutor and he repeated the question often. He did not use many words, hesitates often and thinks long and hard. He was not

	towel? Okay. Shower? Okay turn shower *eer*..(showing hesitation) bathroom *eer*.. check the shower.	very fluent and had a limited repertoire.
A1-2	My name is xxx. I'm job is security. I'm security. I am sorry. I security and responsible to hotel and save area hotel, and guest, building. Role play: The spa is this way and near. The beach is near *yaa* five minutes. Yes, this way is the close.	There was good use of vocabulary, but the repertoire was limited. The participant often did not understand the question.
A1-3	My name is xxx. I'm job in laundry. Pick up from room to the room in hotel. Yes. Greeting. Good morning sir, how are you? Yes, little bit. The complain about the. ... Broke the ... With the cloth. Role play: Good afternoon! Laundry! Your laundry finish. Can you check with me together here? Can I see the stain? Can I bring the item to laundry I will be try to be remove it.	Pronunciation was difficult to understand. He used good vocabulary to talk about his job, but made a lot of grammar mistakes.
A1-4	My name is xxx. *Di*(at) Wayang. Roomboy. Clean bed, bathroom, all of them. Role play: Can I clean your room now? Cleaning bathroom now? One or two? Please you take it to the on the bathtub. Yes, Sir! I'm finished cleaning your room.	It seemed as if this participant did not understand the question and find it difficult to produce sentences.
A1-5	I job is security, where you are going? Do you have promise instead of do you	This participant was very hesitant. He produced short and standard phrases only.

	have an appointment.	He did not do very well in the role play, because he did not understand the instructions, even after the proctor tried to explain it in the Indonesian language.

Table 12

Test takers A2 Level sample analysis

Participants	**Transcription**	**Impression**
A2-1	My name is xxx. Waitress. Start in the morning, I prepare a food for the guest and then we starting at 6.30 and then close at 10.30 everyday prepare for breakfast, food for the guest. Sometimes the food is cold and also the food come lates and then coffee was cold. That's it. We say to the guest I'm sorry this first and I then talk to with the guest. And then I say like this if the guest have any complaint at the restaurant about the food and then i talk with the guest would you change the food? And then with the other food? Role play: Good morning Madam, maybe you want to sit inside or outside? Please sit down. So this morning we have a buffet in the restaurant. And also we *hep* (have) *kopi* (coffee) . If you want *kopi*' (coffee) this morning you can pick by the corner and then if you want omélete inside and then you just order request for	This participant made many pronunciation mistakes. However, she demonstrated a good range of vocabulary to explain her daily job. Sometimes it was hard to follow and there was also staccato.

	omelete. Yes can you show me the coupon? Your room number? Enjoy your breakfast this morning.	
A2-2	My name is xxx. I live in Borobudur. My job is mechanical electrical engineer. To maintenance and repair electronic plumbing in there. Sometimes, we can greeting like good morning and then we tell i'm sorry. Role play: Good morning. I'm sorry there are problem in the room. I'm sorry can I look your? The problem is the toilet is the tint is lost. For the use the hot water you can turn left the end of the valve and pull the valve. Okay I show you.	The vocabulary was quite good specifically when explaining daily tasks. The pronunciation was clear, but sometimes the sentence was not complete.
A2-3	My name is xxx. Job is housekeeping. Cleaning room and then cleaning the public area, restaurant, the corridor and then daily place that's all that. Role play: Good morning, housekeeping please. Yes mam I'm sorry I will wait outside the room. I'm sorry mam are you ready to cleaning your room. Yes mam I will start from the bathroom. I'm sorry mam if you need extra towels, i will get you two towels just two towels for one day. House tenant hotel just two towels one day And then if you needs extra towels everyday you can confirm to reception	This participant showed a good range of vocabulary and could produce long sentences. He could generally understand clear, standard speech about familiar matters. He could express how he feels in simple terms, and express thanks.

	first and then I will complete your request towel to occupied in your room. How much you need extra mineral water mam? Yes mam. I'm sorry mam I was finish cleaning your room anything else you need mam? Have a nice day.	
A2-4	My name is xxx. Our job at spa therapist. I'm doing the treatment massage like a massage, facial, manicure, pedicure. Four years. They have complaint with the water, spa areas, no treatment. Water is no hot and the temperature no. Spa area is a dirty. About spider web. The best I ask to the guest and apologize with that and ask with the spa manager and then call housekeeping. And also our spa team also clean the areas. *Ya*, they still complain but our manager of course explain it for the guest. Role play: Good evening! How are you today? This is our spa! How may I serve you? Have a sit. *Ya*, we have massages, and also we have facial and manicure pedicure. What treatment do you like for today? Balinese massage is traditional massage and good for your body and brain. What time do you want to have spa? Wait a moment. I check first for the reservation. We have appointment for one person. Is it okay? and your treatment is Balinese massage.	This participant had a good range of vocabulary, clear pronunciation, and could make herself understood in an interview and communicated about familiar topics and, in particular, her daily job.

A2-5	My name is xxx. My job in the tanker of engineering. I usually the visit of the trouble in AC, electric car and the pantry. Yes I like to work here. I working here in the 20 years. I interesting of the Ac and I interesting of the my department and my staff and my friends the engineering. Not usually. Not so much. Yes, we can the guest with the meet and the park area, we call it good morning and good evening. but the complain is the handle it by villa house. Role play: Good morning! Yes the way is the from here you get the streetway and to the left. From the spa you got the room of the *'tangga'* (stairs) and you can to the left. And you got the restaurant. Open at the nine o'clock and close at the eleven o'clock at the night. Yes you can because the restaurant get info the name and you can get the order what do you want.	This participant could understand the question that are asked and showed a good range of vocabulary. Sometimes, he used incorrect words and conjunctions, but but realized this and corrected himself.

Table 13

Test takers B1 Level sample analysis

Participants	**Transcription**	**Impression**
B1-1	My name is xxx. I job is in rest area concierge. Actually I'm guide. We have many types of the journey here. So i become guide. For example a journey of Tanah lot. I'll	This participant had vocabulary range. He could produce long sentences and used many conjunctions. He was able to participate in unprepared conversations on

	take our client to many kinds of temples That we have around here. To give them explanation what history behind. The story that we have In our religion every day. Of course. Role play: Sure. Actually in Bali there are so many temples. 2000 temples and it can be classified in four different types of temple. The third one is associated with our occupation in Bali.	familiar topics.
B1-2	My name is xxx. My job is as a frontdesk. Actually I'm working for frontdesk department. So i need to be standby upstairs in the lobby as a reception. I need to do check in process and check out process and also I make a deal with the guest for some activity, some request. Yes, I did but not always like not always not every day I mean. But that is our challenge in front office department because the first complain will be coming to us so we need to resolve the problem and then make the guest stay comfortable with us here. So trying to speak nicely to them first and then try to find the good solution for them so just make like everything running smooth, calm and then don't try to argue with them. I have been working here for two years. So far, I like it but it's normal when you are working you have a lot of challenge but while you are working you have to be	The vocabulary was good, especially when talking about daily tasks. The pronunciation was clear and she could clearly explain familiar topics with pronunciation and grammar.

	professional. Maybe there are a lot of challenge here. Like you are the first person directly with the guest. Everything they need is goes to you. Sometimes, friendly guest is good right? We have to make. Be yourself and you have to be part of the guest. Lobby, the challenge is with the bill. You need to find solution. Before the guest leave. Role play: Good morning! Certainly, Miss. Just give me a second with the room department. Normally the bill here should be signed by the guest. Yes certainly. Let me check first. You are consuming some beverages. If you don't mind I will double check for you.	
B1-3	My name is xxx. You can call me xxx. My job is a as a butler. So it front office department. I do serve the guest for the dining in villa Arrange activity, and mostly meet the guest and also organize housekeeping and another department to do for the cleaning room and maybe dining for the other like activities. I do. I've been here for three and half years. Sure. Mostly the problem with the building, I mean the facility but sometimes you get the problem with service as well like the food and beverages timing. Sometimes comes late sometimes wrong maybe misunderstanding with the guest order then receipt from	He had a wide vocabulary range and produces longer utterances with a small amount of errors. His pronunciation was good and easy to understand. He could also understand clear, standard speech on familiar matters. He could express how he feels in simple terms and express emotions.

	staff and also sometimes we have a problem as well with the cleanliness in the villa. Sometimes the guest is really concern with the cleanliness. they are really upset when see hair or something else. Ya, Try to understand first what is the problem with the guest and of course to say apologize first then we follow up to a clean what the problem solve the problem then try to promise don't will happen again for the next time and also if the guest still unsatisfied maybe you can give them a bit complimentary with drink of the food. But if I can't handle it mess for you should be calm down. The last chance. Yes, because sometimes mostly from the phone because you know our villa is private. That means how they can find us by call. When we come to their villa. Mostly they come from the phone the control operator or our cell phone. So this number 24 hours for everyone. We have three shift here. In the morning start from seven until three. Then three until eleven for another person. And continue again for another shift. And also we have another service like intelligent It means we handle guest i mean one butler for one villa. Sometimes it depends on the guest. Role play: Good afternoon.	

	How are you today? Sure. Im sorry. Let me double check. Maybe. do you want enjoy something outside? Do you mean you want to lunch outside. Excursion. You can go to ubud. There is so many places like a shop. So you can walk away so you can see so many thing or you can do so many activity. You can see beautiful. Actually we do have so we can arrange. Half day tour. Half day will be like. Our travel will be accompany. Surrounded. Our travel is expert. Maybe six hours we can give you..	
B1-4	My name is xxx. I'm a GSA. As a GSA usually I do check in guest and check out specially for the billing. So if the guest need something so they come to me need transport need some and we will ask somebody to help them. Mostly because it is interesting and we found many characters of guest and let me. It will be great. We have problem. Air conditioner is leaking and to fix it soon with the . we will fix it and we will move them. When the restaurant wrong room. We have special programme. If the bill is not too much it is okay. But seven days. It's a bit difficult. We have limited power. But if the guest asking for the discount and our supervisor at the moment. Of course we call our manager or supervisor. Sometime it's interesting of course. Just	This participant could understand the main points of clear standard input on familiar situations regularly encountered in her work. She could also deal with uncommon situations that are likely to arise whilst she is at work. Additionally, she could produce simple connected text on topics, which are familiar, or of personal interest. She was able to describe experiences and events and briefly gave reasons and explanations for opinions.

	stay cool and keep smile. And then manager to explain. Because satisfaction of our guest is our priority. Role play: Good afternoon! Your room number is 271. Okay thi is your bill can you check. Oh really? I will check it first. This is for another room. Maybe it is repeated in our system. Sometimes internet is going down. I hope it's not disturbing. Certainly! For other expense. For your credit card. Not deposit. And then swap again. Can we process it?	
B1-5	My name is xxx. I'm working as a coordinator in the housekeeping department. Coordinator is a person that order taker. the order and request from guest and you transfer to the person in charge. I think that the most challenge is that the time because I only have limited time to receive and transfer any request. One year. I think that about ten years in secondary school. I like it. I think I like it. Busy everyday especially when the high occupancy. It's the part of our job. the guest usually complain about the make up room service. They might said that the room is still dirty or they want to the waste or they want special request. For example they want us to set up extra bed in the room everyday. Set up something like that. First, I can say someone to bring them to the	She had a wide range of vocabulary and produced longer utterances with a small amount of errors. Her pronunciation was good and easy to understand. She was able to provide concrete information required in an interview, but did so with limited precision. She could carry out a prepared interview, checking and confirming information, though she may occasionally have to ask for repetition if the proctor response is rapid or extended.

	guest immediately usually in five minute. It have to be offer by the front office. The guest have to pay to set up the extra bed in room. Food not. Food it belong to room service. We only about the.. we are housekeeping. Role play: hallo! Housekeeping! Okay i got it! I will sent the room attendant for you in a few second! So i will sent you our room attendant to clean your room and do anything you want. Yes and I stay in the office. I really sorry for this inconvenience. So now we will bring for you two complimentary water. So, The fruit is only set up when is the an offer from the guest reservation manager. Okay I have to explain something, because our job. Because It is the duty of the front office. Do you want some fruit? Okay i will talk the problem to the duty manager. You want to free fruit or you want to pay fruit? Okay i got it. So the fruit will be set up in the room. It is our hotel standard. In case you still want fruit and you have to pay them. Is not free. I see your problem. It's our mistake about the room attendant behaviour.	

Table 14

Distribution of participants across CEFR levels (video impression)

CEFR Level	A1	A2	B1	Total

Rater 1	8	7	3	18
Rater 2	24	23	7	54
Total	32	30	10	

The data in Table 14 show that rater 1 agreed to rate only 18 participants upon specific circumstances, while rater 2 determined the CEFR level for 54 participants.

4.4 Inter-rater reliability video analysis

The analysis of the participant interviews and role plays (conversations) is the result of the work of two female raters. Rater 1 is a highly proficient near-native speaker of English with Dutch as her native language. Rater 2 is a master student with Bahasa Indonesia as her mother tongue. They rated the test takers one by one by watching the video and then rated the oral production from participants based on the CEFR interview and conversation scale. The CEFR levels were then coded by number (A1=1, A2=2, and B1=3). The inter-rater reliability analysis was performed using Cohen's Kappa. Cohen's Kappa (κ) is a statistic of inter-rater agreement for categorical scales when two raters or observers measure a variable on a categorical scale (Laerd statistics, 2016).

Cohen's kappa (κ) is **.82**. This is the proportion of agreement above chance agreement. Cohen's kappa (κ) can range from -1 to +1. Based on the guidelines adapted from Landis & Koch (1977), a kappa (κ) of .82 can be characterized as substantial to almost perfect agreement. Furthermore, since p = .000 (which actually means p < .0005), our kappa (κ) coefficient is significantly different from zero.

4.5 Correlation between the placement test and video rating

If two sets of assessment scores are highly correlated, this means that if the scores on one scale increase, the scores on the other scale do so too. However, if the two tests are designed to test the same skill, but have a negative or no correlation, something is amiss (Carr, 2011). A bivariate correlation test was performed to see whether the total scores of the placement test and the video analysis are related. In addition, Pearson‘s r was calculated to provide statistical

evidence for a linear relationship among the same pairs of variables in the population, represented by a population correlation coefficient, ρ ("rho"). The bivariate Pearson correlation test does not say anything about case and effect, regardless of the correlation.

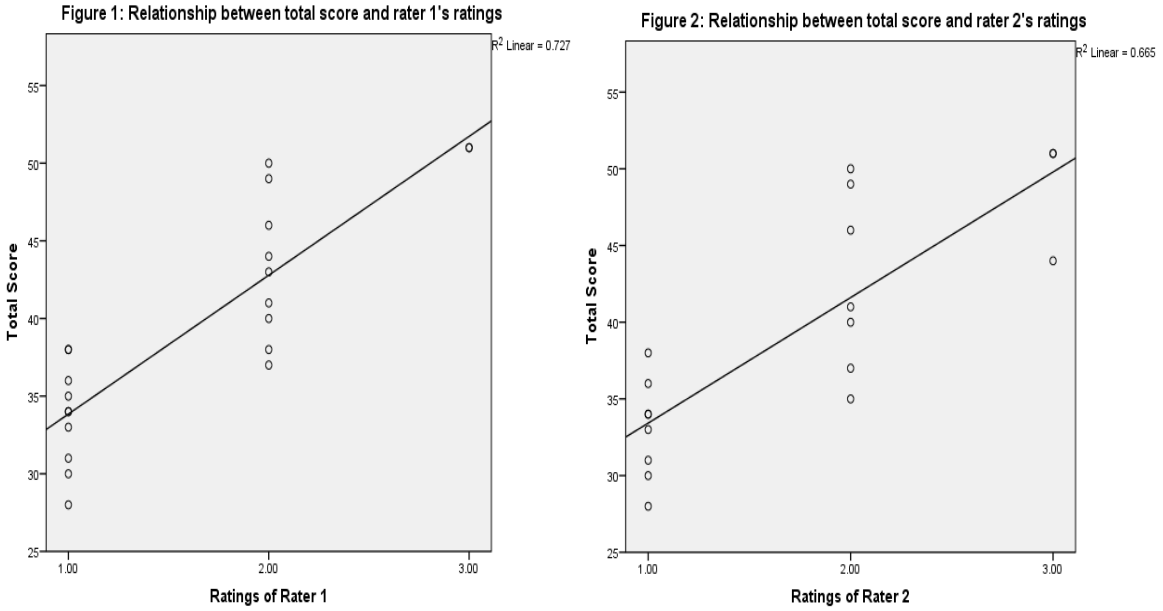


Figure 5. Correlation between the video analysis of two raters and placement test

The relationship or visual correlation between two variables (the total score of the multiple-choice placement test and ratings of the two raters) were presented in Figure 5. Figure 5.1 plots the data points for the video rating of rater 1 (x-axis) and total score of multiple-choice placement test (y-axis). In addition, figure 5.2 shows the ratings by rater 2 on the horizontal axis (x-axis) and the total score of the multiple-choice placement test on the vertical axis (y-axis). Both figures have a positive correlation, meaning that the variables in each figure move in the same direction. In particular, the rating of both raters and the total score were positively correlated. The two figures showed that there is almost no overlap between A1 (coded as 1,00) and A2 level (coded as 2,00). This means that the ratings are rather reliable.

Table 15

Correlation between placement test total score and raters

	Total score	Rater 1 (n=18)	Rater 2 (n=54)
Total score	-	.816**	.789**
Rater 1 (*n*=18)	.816**	-	.903**

Rater 2 (*n*=54)	.789**	.903**	-

Note **. Correlation is significant at the p < 0.01 level (two-tailed)

Looking at the correlations presented in Table 15, all three variables were positively and strongly correlated. Rater 1 rated only 18 cases, but there was a strong, positive relation between the scores of the placement test and her ratings, (r = .82, n=18, p < .001). In addition, the Pearson correlation coefficient also showed a significant positive correlation between the total score from the placement test and rater 2, (r = .79, n=54, p < .0001).

4.6 Elicited Imitation analysis

The EI outcomes were rated by three female raters/observers. Rater 1 is the author of this thesis, Raters 2 is a highly proficient near-native speaker of English with Dutch as her native language. She works as a language expert and has much experience with foreign learners. The last rater is a native speaker of English who lives and studies in the Netherlands for one year and she is getting used to hearing an Asian accent. The order of test takers was randomized, so that raters did not consecutively hear test takers from the same level.

4.6.1 Correlations between EI scoring based on sentence and word order

Table 16

Correlation between three raters on sentence order

	Rater 1 (n=54)	Rater 2 (n=13)	Rater 3 (n=54)
Rater 1 (*n*=54)		.896**	.961**
Rater 2 (*n*=13)	.896**	-	.749**
Rater 3 (*n*=54)	.961**	.749**	-

Note **. Correlation is significant at the p < 0.01 level (two-tailed)

Table 16 illustrates the correlation between EI scores based on sentence order across three different raters. The raters gave 1 point for every single word in the correct order. If the test taker failed to imitate the original word order, the scoring was stopped. The correlation between

rater 1 (*n*=54) and 2 (*n*=13) is very strong (*r* =.89) and is significant, *p* <.001, meaning that the two raters rated the test similarly. Furthermore, the correlation between rater 1 and 3 is the strongest, *r* = .96. Both raters rated 54 participants. The correlation between raters 2 and 3 was the weakest, *r* = .75. However, it is still characterized as a strong positive correlation.

Table 17
Correlations between EI scoring based on correct words

	Rater 1 (n=54)	Rater 2 (n=13)	Rater 3 (n=54)
Rater 1 (*n*=54)		.945**	.992**
Rater 2 (*n*=13)	.945**	-	.938**
Rater 3 (*n*=54)	.992**	.938**	-

Note **: *p* < .001

For scoring based on words, the raters counted the correct words from each sentence item regardless of the order in which they appeared. Small lapsus linguae (slips of the tongue) and mispronunciation were counted as long as they did not deviate from the meaning. If the mistake or any other error led to imitation of a major part of the sentence, it was given a 0 score. Table 17 showed that the correlation between rater 1 and 2 was very strong, *r* =.95, and significant, *p* <.001, meaning that the two raters rated the test similarly. Furthermore, the correlation between raters 1 and 3 noted was almost perfect correlation, *r* = .99.

4.6.2 Elicited Imitation Stimulus analysis

The EI Stimuli analysis was based on sentence length and complexity. The stimuli length varied between 3 and 14 words. The design of the test did not control for syllable length and intrinsic aspects such as morphology, syntax, or grammatical structures. The sentences therefore varied in difficulty.

(1) I work here

This declarative sentence was the easiest stimulus. The ratings showed that most participants imitated the sentence perfectly. Rater 1 reported one participant who solely repeated the first two words (e.g. “I work”)

(2) Her son is four years old

The second stimulus was a declarative sentence in the present tense sentence, but only contains 6 words. Many participants were able to repeat the sentence, but there were some participants who repeated "Her son is four years" only. They did not produce the last word.

(3) My sister is afraid of spiders.

This sentence was the first stimulus containing an adjective (i.e. afraid). Many participants elicited the sentence nicely, but they tended to make a mistake in the last word. Instead of saying spiders, the participants used the singular form (i.e. spider).

(4) A. After the meeting had finished they all went to a nice restaurant

B. You should never have allowed him to go to that awful museum

Both sentences contained 12 words and are categorized as difficult, because they used past and perfect tense. Many participants only imitated the first five to six words of the sentences. They also had tendency to change 'that awful museum' to 'the awful museum'. In the B sentence, participants often omitted the words 'never' and 'have'.

(5) I cannot believe you never told him you used to live in the city.

This 14 words-sentence was considered the second most difficult stimulus. There were only a few participants who perfectly imitated the sentence.

(6) She finally admitted that it was her father who had stolen the famous painting

None of the participants was able to imitate this sentence.

4.7 Correlation between video analysis and elicited imitation word scoring

From all the positive correlations at the $p < .001$ level, the strongest correlation is between the scoring based on words. One possible interpretation for this relationship is that the test takers obtained better scores by imitating words regardless of the sentence order. This is possible, because the words are familiar and frequently heard. This correlation is an important piece of evidence that justifies the use of an EI task as a replacement for the placement test. The length

of the sentences in some cases forced the test taker to remember words to repeat the sentence. It is difficult for them to imitate the sentence in its exact form and order. Departing from that result, this section will show the relationship between the two other assessments and EI scores based on word order.

Table 18

Correlation between video analysis and EI word scores

	Video Rater 1	Video Rater 2
Words Rater 1	.743**	.601**
Words Rater 2	.719*	.498
Words Rater 3	.727**	.609**

Note *: $p < .05$, **: $p < .001$, two-tailed.

Table 18 shows that there was a moderately strong positive correlation (r = .498 to .743) between EI word scores and video ratings, which means that there was close relation between the constructs measured by the tests. However, the Pearson correlation values for rater 2 was weaker than the others. This is possibly because rater 2 only rated a small amount of participants in both assessment forms.

4.8 Correlation between Placement test and Elicited Imitation word scoring

Table 19

Correlation between overall placement test and EI word scores.

Correlation	
Words Rater 1(N=54)-Placement test	.63**
Words Rater 2 (N=13)-Placement test	.83**
Words Rater 3(N=54)-Placement test	.62**

**. Correlation is significant at the p < 0.01 level (2-tailed).

Table 19 demonstrates that there were moderate to strong positive correlations between the overall scores on the multiple-choice placement test and the EI scoring that was based on the correct words produced by the test takers. The strongest correlation ($r = .83$), significant at $p < .001$ level in this case, is by rater 2, who rated only 13 participants, while the other two raters scored all participants (N= 54).

CHAPTER V

5.1 Discussion

The purpose of this study was to investigate whether the scores on the Elicited Imitation test show any significant and positive relationships with the multiple-choice format assessments and oral video interviews. The EI test was done in Vietnam and Indonesia and followed the following procedure: the participant was allowed to listen to each sentence once and when prompted by the researcher the participant imitated the sentence. The participant imitated ten different stimuli ranging from 3 to 14 words. The results were rated by one native speaker of English, an experienced rater and the author of the study on a word scoring (receive a point when correctly imitating the word) and sentence order (receive a point when a word is imitated correctly in the right order) level. The result of the present study suggest that the EI reported here is a reasonably valid and reliable measure to assess language proficiency for placement purposes. The outcomes demonstrate that the strongest correlation (r = .96) is between scores on word level. One possible interpretation for this relationship is that the test takers obtained better scores by imitating words regardless the sentence order. It is claimed that EI is a short, affordable, and convincing oral assessment method and showed a positive correlation with the other assessment types (multiple choice assessment and video interview). Eventhough the multiple-choice format and video interview have less similarity with EI in terms of test contents and administration, the correlations between EI and the multiple choice test, and between EI and video interview are (r = .83) and (r = .74) respectively.

It is also seems to be the case that the EI test takers have the tendency to be able to repeat the short declarative sentences containing vocabulary that is related to common everyday things, such as *work, computer, playing, hobby, restaurant.* None of the participants was able to imitate the long and unfamiliar stimuli containing more complicated grammar. The reason for this might be that most participants are at A1 level. The results of multiple-choice placement test showed that there are 35 participants on A1, 8 participants on A2, and 11 participants on B1 level while the video rating distributes the participants as follows: 24 participants on A1 level, 23 participants on A2 level, and only 7 participants on B1 level. The results of this study support the notion that EI is an effective measure of global language proficiency and has many

benefits in terms of simple and economical administration procedures and the flexibility in task design.

The results support and add to the findings of Zhou (2012) and Tracy-Ventura, McManus, Norris, & Ortega (2014), who reported similar correlations between EI scores and other measures of language proficiency. However, it is currently unknown whether and how design and scoring affect the reliability and validity of EI tests. This needs further research. It is therefore crucial for EI test developers to explicitly state why a certain procedure was used. Furthermore, when EI is used for placement purposes, it is possible that there are serious mismatches between the results of the placement test and the actual proficiency of participants. There may be people who do well at listening, speaking and focus on form, but get a very poor result on repeating sentences.

In addition, it is also necessary to take into account that people may be able to function at A2 level during their jobs at the hotel, while they struggle in other areas. Placement test items should therefore be general, but should not be too far removed from what people do. Furthermore, while people are accustomed to reading from a computer screen, they might be less comfortable talking to a computer screen. The exam aims at A1 to B1 speakers and was done by a homogeneous group of staff (with two different mother tongues).

Finally, there are some limitations to the present study. First, there is no information on the participant details, such as age, education background, or other experience with the target language (e.g. going abroad and the frequency of interacting with foreign people). Additional investigations should also look into distinct populations of participants.

5.2 Conclusion

Having a sufficient command of English is an essential element for workers in the hotel industry. In an attempt to provide a successful English learning environment in hotels, many language learning providers assist their clients, international hotels, in assessing the workers' English proficiency before they send their staff to a module or course. In addition, it is necessary to have automated assessment in the form of oral assessment, since oral competence is the important for hotel staff. Elicited imitation (EI) is an active oral skill assessment method that has been used in the last few decades. It is chosen as an assessment method that is suitable to test English proficiency of hotel staff.

The following conclusions can be drawn from the present study. There is a highly positive correlation between the EI and the other two types of assessment, which suggests that the tests are dependent. However, the correlation analyses used in these studies cannot simply verify the assertions the researchers make on the basis of the EI scores, from which they argue that EI is a sufficient replacement for the multiple-choice placement test to measure English proficiency for the purpose of placement. The stastistical packages (TIAPLUS © and SPSS 21) and the video interview developed by language experts adopted in this study assisted in gathering evidence to present an interpretive and descriptive argument for validating EI. This study presented EI scoring based on words and sentences and has outlined the limitations of both strategies. EI is not difficult to perform, reliable to rate, and easy to design. Furthermore, DIF analyses revealed that items behaved similarly for male and female employees from both nationalities.

There is insufficient evidence to say that the Elicited Imitation is the most suitable assessment for placement purposes compared to the multiple-choice format test or even the video interview. However, positive and strong correlations are important evidence for the close relationship between the tests. In addition, the correlations give an indication of the desired scores and show that the word scoring method is the best to determine the proficiency of the test takers. The evidence of this study suggests that EI is sufficient to replace the other assessment formats.

Based on the findings, it is recommended that future studies using EI as a measure of language proficiency carefully design the test. In addition, further research efforts should be made to investigate how the design of key EI task features functions under specific assessment purposes and contexts. However, another important factors have to be taken into account when selecting the sentence length, the diction and grammatical features. Thus, it will be fair and more valid and reliable for participants from different levels.

References

Abbott, M.L. (2006). ESL reading strategies: differences in Arabic and Mandarin speaker test performance. *Language learning, 56*(4), 633–670.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education, American Psychological Association, & American Educational Research Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: APA.

Ashwell, T. (2014). Automated scoring for elicited imitation tests. *Journal of global media studies, 13*, 37-41.

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

Bachman, L. F., & Cohen, A. D. (1998). Strategies and processes in test-taking and SLA. In M. H. Long & J. C. Richards (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90-111). Cambridge, UK: Cambridge University Press.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass & A. D. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 245-261). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Blue, G. B., & Harun, M. (2003). Hospitality languages as a professional skill. *English for specific purposes, 22* , 73-91.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in second language acquisition, 33*(2), 247– 271.

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

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language learning and technology, 5*(2), 95-98.

Chapelle, C.A. (2008). Computer assisted language learning. In B. Spolsky & F.M. Hult (Eds.), *The handbook of educational linguistics* (pp. 585-595). Oxford: Blackwell Publishing Ltd.

Chapelle, C.A. (2010). *Technology in language testing. [Video]*. Retrieved from http: / /languagetesting.info/video/ main.html.

Chapelle, C.A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409–439.

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology*, *20*(2), 116–128.

Chaudron, C., Ngyuen, H., & Prior, M. (2005a). Manual for the Vietnamese elicited imitation test. *NFLRC Research Note 41* (pp. 1–33). Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i.

Chaudron, C., Prior, M., & Kozok, U. (2005b). *Elicited imitation as an oral proficiency measure.* The 14th World Congress of Applied Linguistics, Madison, WI.

CITO. TiaPlus, *Classical Test and Item Analysis* © Arnhem: Cito M & R Department., 2005.

Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Cook, K., McGhee, J., & Lonsdale, D. (2011). Elicited imitation as prediction of OPI test scores. In *Proceedings of the sixth workshop on innovative use of NLP for building*

educational applications (pp. 30-37). Portland, OR: Association for Computational Linguistics.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Education Measurement* (pp. 443–507). Washington, DC: American Council on Education.

Davidson, F., & Fulcher, G. (2007). The common European framework of reference (CEFR) and the design of language tests: A matter of effect. *Language teaching*, *40*, 11.

Doe, C.D. (2013). *Validating the Canadian academic English language assessment for diagnostic purposes from three perspectives: Scoring, teaching, and learning.* Ontario: Queen's University Kingston, Canada.

Duolingo. *Duolingo English Test*. Pittsburgh, PA: Duolingo Test Center., 2014.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Erlam, R. (2006). Elicited imitation as a measure of L2 Implicit knowledge: An empirical validation study. *Applied linguistics*, *27*(3), 464-491.

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics (4th edition).* London: Sage Publications Ltd.

Gaillard, S. (2014). *The elicited imitation task as a method for French proficiency assessment in institutional and research settings* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

Geranpayeh, A. (2003). *A quick review of the English quick placement test*. Research Notes, published quarterly by University of Cambridge ESOL Examinations.

Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari, K. Choukri,

B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the 6th international language resources and evaluation conference* (pp. 1604–1610). ELRA.

Graham, C. R., McGhee, J., Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In M. T. Prior (Ed.), *Selected proceedings of the 2008 second language research forumed* (pp. 57–72). Somerville, Mass.: Cascadilla Proceedings Project.

Green, A.B. & Weir, C.J. (2004). Can placement tests inform instructional decisions? *Language Testing, 21*(4), 467-494. University of Roehampton.

Henley, P. (2015). Trends and innovation in the hospitality industry. Retrieved June 1, 2016 from Web site: http://www.traveldailynews.asia/columns/article/50093/peter-henley-onyx-hospitality-group.

Hughes, A. (1989). *Testing for language teachers (second edition)*. United Kingdom: Cambridge University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal*, *91*, 663–667. (doi:10.1111/j.1540-4781.2007.00627_5.x)

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*, 229-249. (doi:10.1080/15434303.2011.565844)

IDRE UCLA. Institute for Digital Research and Education Los Angeles: University of California., 2016.

Laerd statistics. *Cohen's kappa using SPSS statistics. Lund Research Ltd*., 2016.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lewkowicz, J. A. (2007). Authenticity in language testing: Some outstanding questions. English Centre, University of Hong Kong. *Language Testing 2000*, *17*(1), 43–64.

Meara, P. & Milton, J. (2003). *X-lex, the Swansea levels test.* Newbury: Express.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational Measurement (3rd edition)* (pp. 13–103). New York: American Council on Education/Macmillan.

Milton, J. (2007). Lexical profiles and learning styles: What do these show us about the construct validity of lexical tests? In Daller, H., Milton, J. and Treffers- Daller, J. (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234-244). Cambridge: Cambridge University Press.

Moore, B. (2013). Discovering the language needs of hotel workers in Thailand. Retrieved April 21, 2016, from Language Institute, Thammasat University Thailand Web site: http://tujournals.tu.ac.th/thammasatjournal/detailart.aspx?ArticleID=53.

Mozgalina, A. (2015). Applying an argument-based approach for validating language proficiency assessments in second language acquisition research: the elicited imitation test for Russian. Retrieved April 15, 2016, from Faculty of the Graduate School of Arts and Sciences of Georgetown University Web site: https://repository.library.georgetown.edu/bitstream/handle/10822/760901/Mozgalina_georgetown_0076D_12888.pdf?sequence=1&isAllowed=y.

NovoLanguage. *Interactive and Personalized Language Learning* Nijmegen., 2016.

Ockey, G.J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, *93*, 836-847. Published by: on behalf of the Wiley National Federation of Modern Language Teachers Associations.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners.* (Unpublished Ph.D.). University of Hawai'i at Manoa.

Ortega, L., Iwashita, N., Norris, J., & Rabie, S. (2002). *An investigation of elicited imitation in crosslinguistics SLA research*. Conference handout from paper presented at the Meeting of the Second Language Research Forum, Toronto, Canada.

Oxford University Press. Oxford: Online Placement Test., 2016.

Pham, V. P. Q., & Thirumaran, K. (2016). Managing development of resort destinations in Southeast Asia: Emerging and peripheral Phu Quoc island. *Tourism and hospitality and business management*. James Cook University, Singapore Campus, Townsville City, QLD, Australia.

Prabhu, A., & Wani, P. (2015). A study of importance of English language proficiency in hospitality industry and the role of hospitality educators in enhancing the same amongst the students. *A Journal of Hospitality, 1*, 56-63. Retrieved April 21, 2016, from Web site: http://www.publishingindia.com.

Rahim, S. (2011). Analyzing the training and internship needs assessment of verbal communication skills amongst hotel practitioners, *4*(3), 44-53. Retrieved June 12, 2016, from Web site: www.ccsenet.org/elt.

Rouse, M. (2007). Automated speech recognition (ASR). Retrieved June 12, 2016, from Web site: http://searchmobilecomputing.techtarget.com/definition/automated-speech-recognition.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). Factor structure of the TOEFL Internet-based Test (IBT): Exploration in a Field trial sample. Educational Testing Service. *TOEFL research report 08-09.* Retrieved June 2, 2016, from http://www.ets.org/Media/Research/pdf/RR08-09.pdf

Selke, R. (2013). The importance of foreign language skills in the tourism sector: A study of employees' perceptions in hotels in Malaysia. In: The 3rd Advances in Hospitality and Tourism Marketing & Management Conference, 25-30 June 2013, Taipei, Taiwan. Retrieved April 21, 2016 from Web site: http://www.academia.edu/6274649/_2013_The_Importance_of_Foreign_Language_Skills_in_the_Tourism_Sector_A_Study_of_Employees_Perceptions_in_Hotels_in_Malaysia.

Song, X. (2014). DIF investigations with Pearson test of English academic. Queens University Canada. Retrieved April 21, 2016 from Web site: http://pearsonpte.com/wp-content/uploads/2014/07/Song_X_2014.pdf

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing* 2, *1*, 31 - 40.

Sun, X. (2012). On interrelations between language teaching and speech teaching. *Theory and Practice in Language Studies*, *2*(1), 179-182.

Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. In A.J. Kunnan (Ed.), *The companion to language assessment*, (pp. 593-613). Malden, MA: Wiley-Blackwell.

Tracy-Ventura, N., McManus, K., Norris, J.M., & Ortega, L. (2014). Repeat as much as you can: Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Proficiency assessment issues in SLA research: Measures and practices* (pp. 143–166). Clevedon, UK: Multilingual Matters.

Van der Slik, F.W.P. (2009). Gender bias and gender differences in two South African tests of academic literacy. *Southern African Linguistics and Applied Language Studies*, *27*(3), 277-290.

Vesselinov, R., & Grego, J. (2012). *Duolingo effectiveness study*. City University of New York, USA.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73. (doi: 10.1111/1473-4192.00024)

Wagner, E. & Kunnan, A.J. (2015). The duolingo English test. *Language Assessment Quarterly*, *12,* 320-331. (doi: 10.1080/15434303.2015.1061530)

Widyastuti, I. O. (2015). *Attitudes towards specific computer-assisted language learning courseware in the workplace*. Radboud University Nijmegen.

Yan, X., Maeda, Y., Jing, L., Ginther, A. (2015). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language testing*,1-32. (doi: 10.1177/0265532215594643)

Ye, F. (2014). *Validity, reliability, and concordance of the Duolingo English test.* University of Pittsburgh School of Education.

Zhou, Y. (2012). *Willingness to communicate in learning Mandarin as a foreign and heritage language (doctoral dissertation).* Honolulu: University of Hawai'i at Manoa, Honolulu, HI.

APPENDIX

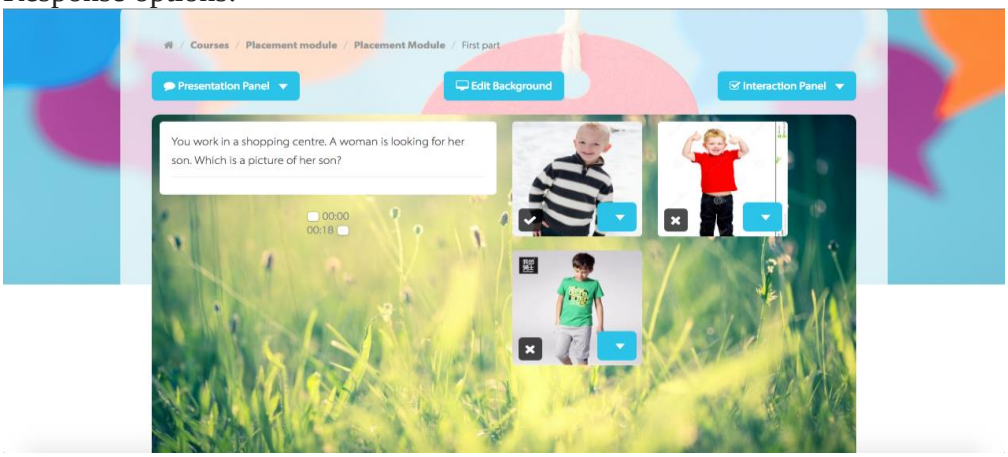
A. A list of the items Novolanguage placement test

A1 Listening

Item 1
Question: You work in a hotel. A guest spells his last name. What is his last name?
Response options: Petterson / Peterson / Petersen

Item 2
Question: You are working in a restaurant. A guest orders some food. What does the guest order?
Response options: chicken with rice / spaghetti and chicken with rice / spaghetti

Item 3
Question: You work in a shopping centre. A woman is looking for her son. Which is a picture of her son?
Response options:



Item 4
Question: You are listening to the weather forecast on the radio. What will the weather be like tomorrow?
Response options: pictures

Item 5
Question: A friend from England shows you a photograph of his children. Which is your friend's photograph?
Response options: pictures

Item 6
Question: Two friends want to go to the movies. When are they going to the movies?
Response options: Sunday morning / Sunday afternoon / Sunday evening

Item 7
Question: You work at a shop. A woman wants to buy some t-shirts. How many t-shirts does the woman want to buy?
Response options: pictures

Item 8
Question: You want to play tennis with your friend. When can your friend play tennis with you?
Response options: Monday / Tuesday / Friday

Item 9
Question: Listen to the conversation. What are these people talking about?
Response options: pictures

Item 10
Question: You work at a hotel. A guest wants to pick up her suitcase. Which suitcase is hers?
Response options: pictures

A1 Conversation

Item 11
Question: Do you have any hobbies?
Response options: Yes, I play tennis. / No, thank you. / Yes, at home.

Item 12
Question: What would you like to drink?
Response options: Yes, I would. / Tomorrow, please. / A coffee, please.

Item 13
Question: Excuse me. Is this your phone?
Response options: It's a phone. / No, it's not mine. / 360225

Item 14
Question: Have a nice weekend!
Response options: Thanks, you too. / Yes, please. / Yes, two nights.

Item 15
Question: Excuse me, how much is this pen?
Response options: It's one pen. / It's two dollars. / You're welcome.

Item 16
Question: Where are you from?
Response options: From England. / At 3 o'clock. / It's Monday.

Item 17
Question: How old are you?

Response options: In America. / I'm 35. / Yes, I am.

Item 18
Question: Are you from Indonesia?
Response options: I can't. / Me too. / Yes, I am.

A1 Focus on Form

Item 19
Sentence context: ... is your name?
Response options: What / Who / When

Item 20
Sentence context: They ... speak English.
Response options: not / don't / no

Item 21
Sentence context: The swimming pool opens ... 9 a.m.
Response options: between / in / at

Item 22
Sentence context: Would you like ... water?
Response options: some / order / drink

Item 23
Sentence context: My name ... Susy.
Response options: are / is / called

Item 24
Sentence context: Have you seen ... mobile phone?
Response options: my / me / mine

Item 25
Sentence context: What is ... last name?
Response options: choose / all / your

Item 26
Sentence context: ... are you from?
Response options: Where / Who / What

Item 27
Sentence context: I would like to buy ... ticket to Hong Kong.
Response options: two / a / not

Item 28-37 Elicited Imitation

A2 Listening

Item 38
Question: You want to go to the hospital to visit a friend. A man tells you how to get there. How should you go to the hospital?
Response options: pictures

Item 39
Question: You want to smoke a cigarette. A woman tells you where to go. Where should you go?
Response options: pictures

Item 40
Question: You are taking an English course. Your teacher talks about next week's lesson. What time will your English lesson start next week?
Response options: 12:30 / 2:30 / 2:45

Item 41
Question: You work at a restaurant. Two guests would like to order something to drink. What do they order?
Response options: pictures

Item 42
Question: You are on a train. You want to go to Oxford. You hear an announcement. What do you have to do at the next station?
Response options: Go back to London / Take a train to Oxford / Take a bus to Oxford

Item 43
Question: You are on the phone to your friend Mike. You are meeting him this evening to go to the movies. What does Mike want to change?
Response options: The place / The time / The day

Item 44
Question: You are having a problem with your computer. You call the service center. What do you need to do?
Response options: Make an appointment / Go to the service center / Buy a new computer

Item 45
Question: You are shopping. You hear an announcement. What should you do?
Response options: pictures

A2 Conversation

Item 46
Question: Do you speak Spanish?
Response options: Just a little bit / Last year / I listen

Item 47
Question: Do you like playing tennis?
Response options: No, next year. / I can't right now. / Yes, I play every week.

Item 48
Question: Let's go to the theater this Friday.
Response options: No, next week. / No, I can't. / How was it?

Item 49
Question: We had a party yesterday.
Response options: How was it? / Can I come? / Be my guest.

Item 50
Question: What are you reading?
Response options: A computer magazine / In the evening / For one hour

Item 51
Question: Good evening. I've got a reservation for 7 o'clock. The name is Garcia.
Response options: Yes, I know. 7 o'clock / Yes, you are. Thank you. / Yes, I see.
Table 4

Item 52
Question: Can I please speak to Mr. Edwards?
Response options: Yes, you're right, Mr. Edwards. / He's not here right now. / Mr. Edwards can speak.

Item 53
Question: How long have you been working here?
Response options: For about two years. / Probably next year. / Only during the summer.

A2 Focus on Form

Item 54
Question: Excuse me. Can I take a bus to the city center from here?
Sentence context: Yes, it ... in 10 minutes.
Response options: leaves / was leaving / has left

Item 55
Question: How do I get to the shopping mall from here?
Sentence context: Just ... please.
Response options: turn straight ahead / turn over / turn left here

Item 56
Question: Mummy, can I have some ice cream?
Sentence context: No, I didn't bring ... money.
Response options: nothing / any / a

Item 57
Question: It's our wedding anniversary next Wednesday.
Sentence context: How long ... married?
Response options: have you been / are you getting / will you get

Item 58
Question: We still need to buy Mark a gift for his birthday.
Sentence context: I will buy him ... .
Response options: somewhere / anywhere / something

Item 59
Question: Can I follow English lessons here?
Sentence context: Yes, the course ... next semester.
Response options: starts / comes / leaves

Item 60
Question: Is it very far to the city center?
Sentence context: No, it ... about 15 minutes.
Response options: can / only takes / arrives

Item 61
Question: I have a pain in my knee.
Sentence context: You ... to the doctor.
Response options: were talking / went / should go

Item 62
Question: Would you like me to take you to the station?
Sentence context: No, I ... .
Response options: will walk / went to the park / have not been there

Item 63
Question: Jim has moved to New York.
Sentence context: ... .
Response options: When is he going? / When did he move? / Where is he going?

B. TIAPLUS © analysis result

```
TiaPlus®  Test and Item Analysis  Build  314
Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2013.

Population          : hotel employees Bali Vietnam
Test                : Placvement test
Date                : Friday, July 01, 2016
Time                : 11:31
Data file           : U:\PROJECTEN\STUDENTEN\Riska Risdiani\Placement data.txt
Database            : C:\Cito\Tiaplus\TiaPlus2000.mdb
Missing handling    : Missing as Missing

Persons: All persons                                    Items: All items

Test and Item Analysis

                                                                  %   #
Item      Item             ------ P- and A- values ------
Mis-|-------------- Weighted ----------------
Label      nr. Weight  Key    A    B    C    D    E    F  O/D sing| Max Mean   P
Sd   RSK  Rit  Rir   AR
-------------------------------------------------------------------------------
---------------------------
           1     1    A     93*   7    0    0              0   0 |  1  0.93  93
0.25  0.25   32   29   88
           2     1    C     15   29   56*   0              0   0 |  1  0.56  56
0.50  0.50   49   43   87
           3     1    A     76*  12   12    0              0   0 |  1  0.76  76
0.43  0.43   45   40   87
           4     1    B      7   68*  25    0              0   0 |  1  0.68  68
0.47  0.47   47   42   87
           5     1    C      5   24   71*   0              0   0 |  1  0.71  71
0.45  0.45   43   38   87
           6     1    C      8    3   88*   0              0   0 |  1  0.88  88
0.32  0.32   52   48   87
           7     1    B     10   69*  20    0              0   0 |  1  0.69  69
0.46  0.46  -19  -25   89
           8     1    B     17   75*   8    0              0   0 |  1  0.75  75
0.44  0.44   57   53   87
           9     1    A     93*   2    5    0              0   0 |  1  0.93  93
0.25  0.25   38   35   88
          10     1    A     61*   7   32    0              0   0 |  1  0.61  61
0.49  0.49   45   40   87
          11     1    A     88*   7    5    0              2   1 |  1  0.88  88
0.33  0.33   15   11   88
          12     1    C     10    0   90*   0              2   1 |  1  0.90  90
0.30  0.30   42   39   87
          13     1    B     15   85*   0    0              0   0 |  1  0.85  85
0.36  0.36   48   44   87
          14     1    A     83*   5   12    0              0   0 |  1  0.83  83
0.38  0.38   49   46   87
          15     1    B     20   80*   0    0              0   0 |  1  0.80  80
0.40  0.40   52   48   87
          16     1    A    100*   0    0    0              0   0 |  1  1.00 100
0.00  0.00    0    0   88
          17     1    B      0  100*   0    0              0   0 |  1  1.00 100
0.00  0.00    0    0   88
          18     1    C      0    0  100*   0              0   0 |  1  1.00 100
0.00  0.00    0    0   88
          19     1    A     97*   3    0    0              0   0 |  1  0.97  97
0.18  0.18    4    1   88
          20     1    B     15   85*   0    0              0   0 |  1  0.85  85
0.36  0.36   -4   -9   88
          21     1    C      8    3   88*   0              0   0 |  1  0.88  88
0.32  0.32   49   45   87
```

```
           22     1    A     69*   7   24    0            0   0 |  1  0.69  69
0.46  0.46   17   11   88
           23     1    B      0  100*   0    0            0   0 |  1  1.00 100
0.00  0.00    0    0   88
           24     1    A     93*   2    5    0            0   0 |  1  0.93  93
0.25  0.25   20   17   88
           25     1    C      0    0  100*   0            0   0 |  1  1.00 100
0.00  0.00    0    0   88
           26     1    A    100*   0    0    0            0   0 |  1  1.00 100
0.00  0.00    0    0   88
           27     1    B     21   79*   0    0            2   1 |  1  0.79  79
0.41  0.41   40   35   87
           28     1    B      7   79*  14    0            2   1 |  1  0.79  79
0.41  0.41   31   26   88
           29     1    C     14    3   83*   0            0   0 |  1  0.83  83
0.38  0.38   32   27   88
           30     1    A     69*  19   12    0            0   0 |  1  0.69  69
0.46  0.46   45   40   87
           31     1    B     31   66*   3    0            0   0 |  1  0.66  66
0.47  0.47   48   43   87
           32     1    C     42   14   44*   0            0   0 |  1  0.44  44
0.50  0.50   58   53   87
           33     1    A     63*  32    5    0            0   0 |  1  0.63  63
0.48  0.48   39   34   87
           34     1    B     17   59*  24    0            0   0 |  1  0.59  59
0.49  0.49   21   15   88
           35     1    A     64*   3   32    0            0   0 |  1  0.64  64
0.48  0.48   43   38   87
           36     1    A     97*   2    2    0            0   0 |  1  0.97  97
0.18  0.18   26   24   88
           37     1    C      2    0   98*   0            0   0 |  1  0.98  98
0.13  0.13   23   22   88
           38     1    B     32   59*   8    0            0   0 |  1  0.59  59
0.49  0.49   51   46   87
           39     1    A     54*  31   15    0            0   0 |  1  0.54  54
0.50  0.50   54   50   87
           40     1    A     90*   7    3    0            0   0 |  1  0.90  90
0.30  0.30   47   44   87
           41     1    C     31   17   53*   0            0   0 |  1  0.53  53
0.50  0.50   71   68   87
           42     1    B     31   59*  10    0            0   0 |  1  0.59  59
0.49  0.49   76   73   87
           43     1    A     95*   2    3    0            0   0 |  1  0.95  95
0.22  0.22   14   11   88
           44     1    A     71*  17   12    0            0   0 |  1  0.71  71
0.45  0.45   16   10   88
           45     1    C     31    7   63*   0            0   0 |  1  0.63  63
0.48  0.48   38   33   88
           46     1    B      5   71*  24    0            0   0 |  1  0.71  71
0.45  0.45   33   27   88
           47     1    A     71*  12   17    0            0   0 |  1  0.71  71
0.45  0.45   48   44   87
           48     1    C      7    8   85*   0            0   0 |  1  0.85  85
0.36  0.36   41   37   87
           49     1    A     92*   3    5    0            0   0 |  1  0.92  92
0.28  0.28   25   22   88
           50     1    B     10   75*  15    0            0   0 |  1  0.75  75
0.44  0.44   49   44   87
           51     1    C     12   12   76*   0            0   0 |  1  0.76  76
0.43  0.43   63   59   87
           52     1    A     66*   7   27    0            0   0 |  1  0.66  66
0.47  0.47   59   55   87
           53     1    B     12   56*  32    0            0   0 |  1  0.56  56
0.50  0.50   36   30   88
```

```
--------------------------------------------------------------------------------
---------------------------
SubGroup number             : 0                 SubTest number            : 0
Number of persons in test   : 59                Number of selected items  : 53

Minimum test score          : 0                 Maximum test score        : 53

Average test score          :  41.41            Standard deviation        :
7.50
Average P-value             :  78.23            Std. Error of Measurement :
2.62
Average Rit                 :   0.40
Coefficient Alpha           :   0.88            SE Coeff. Alpha           :
0.02
Guttman's Lambda2           :   0.89

--------------------------------------------------------------------------------
------
90% Confidence limits for Coefficient Alpha: (0.84 =< 0.88 =< 0.91)
--------------------------------------------------------------------------------
------
Estimated Coefficient Alpha if this test had a standard
norm length of 40 items: 0.84 (Spearman-Brown)

Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2013.
```

C. Table CEFR

Global Proficiency scale

Independent User	**B1**	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics, which are familiar, or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	**A2**	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Table interviewing and being interviewed scale

B1	Can provide concrete information required in an interview/consultation (e.g. describe symptoms to a doctor) but does so with limited precision. Can carry out a prepared interview, checking and confirming information, though he/she may occasionally have to ask for repetition if the other person's response is rapid or extended.
	Can take some initiatives in an interview/consultation (e.g. to bring up a new subject) but is very dependent on interviewer in the interaction. Can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow up questions.
A2	Can make him/herself understood in an interview and communicating ideas and information on familiar topics, provided he/she can ask for clarification occasionally, and is given some help to express what he/she wants to.
	Can answer simple questions and respond to simple statements in an interview.
A1	Can reply in an interview to simple direct questions spoken very slowly and clearly in direct non-idiomatic speech about personal details.

Table conversation scale

B1	Can enter unprepared into conversations on familiar topics. Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases. Can maintain a conversation or discussion but may sometimes be difficult to follow when trying to say exactly what he/she would like to. Can express and respond to feelings such as surprise, happiness, sadness, interest and indifference.
A2	Can establish social contact: greetings and farewells; introductions; giving thanks. Can generally understand clear, standard speech on familiar matters directed at him/her, provided he/she can ask for repetition or reformulation from time to time. Can participate in short conversations in routine contexts on topics of interest. Can express how he/she feels in simple terms, and express thanks.
	Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord, though he/she can be made to understand if the speaker will take the trouble. Can use simple everyday polite forms of greeting and address Can make and respond to invitations, invitations and apologies. Can say what he/she likes and dislikes.
A1	Can make an introduction and use basic greeting and leave-taking expressions. Can ask how people are and react to news. Can understand everyday expressions aimed at the satisfaction of simple needs of a concrete type, delivered directly to him/her in clear, slow and repeated speech by a sympathetic speaker.

D. EI Scoring rubric sample

A	B	C	D	E	F	G	H	I	J
		I cannot believe you	4	4		Diffcult pronunciation			
		x	0	0					
			14	16					
7	6Fhw40CUdo	x	0	0					
		Her son is four years old	6	6					
		That computer is broken	4	4					
		My sister is afraid of spider	5	5					
		x	0	0					
		I'm afraid I cannot remember your name	8	8					
		After the meeting had finished they always go to the nice restaurant	6	9					
		x	0	0					
		I cannot believe you tell him is live in the city	4	9					
		x	0	0					
			33	41					
8	6mXcb7mqen	I work here	3	3					
		My son four year old	0	3					
		x	0	0					
		My sister of afraid ID	3	3					
		Play tennis is my flay hobby	0	3					
		x	0	0					
		After the finished the meeting they had go to the restaurant	2	7					
		x	0	0					
		I cannot believe you told him go to university	4	6					
		x	0	0					
			12	25					
9	7lO47medsJ	I work here	3	3					
		x	0	0					
		That computer is broken	4	4					
		My sister is afraid of spiders	6	6					
		Playing tennis is my favourite hobbies	5	5					
		I'm afraid I cannot remember your name	8	8					
		After the meeting when finished would like to go to the restaurant	3	6					
		You should never allow him to go to the museum	3	8					
		I never believe you never tell him to live in the city	1	10					
		x	0	0					
			33	50					
10	9RL5UyRJfU	I work here	3	3					
		Listen I'm four years old	0	3					
		That computer is broken	4	4					
		x	0	0					

A	B	C	D	E	F
	9 files	playing tennis is my favourite hobbies	5	5	
		you should never allowed him to go over to the museum	3	7	
	what do we do with hobbies instead of hobby? I did not count hobbies as	I work here	3	3	
		I'm afraid I cannot rememeber your name	8	8	
		that computer is broken			
		after a meeting has finish they all went to a restaurant	1	8	
		my sister is afraid of spiders	6	6	
		x	x	x	
		total score (1 file missing)	**32**	**43**	
	oQ7TBlBO9M	her son is old	2	4	
		xxxx	0	0	
	9 files	playing tennis is my favourite hobby	6	6	
		xxxx	0	0	
		I work here	3	3	
		I'm afraid	3	3	
		that computer is broken	4	4	
		after meeting is finish	1	2/3?	
		my sister is	3	3	
		x	x	x	
		total scores (1 file missing)	22	25	
	23gl602vOR	I cannot believe	3	3	
		playing tennis is my favourite hobby	6	6	
		xxx	0	0	
	9 files	I'm afraid I cannot remember your name	4	4	
		that computer is broken	3	5	
		after the meeting is over they going out the restaurant	4	5	
		My sister is afraid spiders	4	5	
		her son is four years old	6	6	
		i work here	3	3	
		x	x	x	
		total scores	33	37	