Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature
||Volume||4||Issue||2||Pages|| 147 – 174 ||2018||
Available online at:http://ejournal.kopertais4.or.id/mataraman/index.php/efi |P-ISSN: 2460-0938; E-ISSN: 2460-2604|

ASSESSING ENGLISH SPEAKING AND LISTENING SKILLS WITH THE MOBILE APPLICATION TELEGRAM

Agus Wardhono
aguswardhono@unirow.ac.id

Stephen Spanos
spanosstephen@gmail.com

Universitas PGRI Ronggolawe Tuban, Indonesia
Jl. Manunggal No.61, Wire, Gedongombo, Semanding, Tuban, Indonesia

Middlebury Institute of International Studies at Monterey, USA
Middlebury Institute of International Studies 460 Pierce Street Monterey, USA

Article History:
Received:
September 12, 2018

Revised:
November 15, 2018

Accepted:
December 21, 2018

Corresponding Author:
aguswardhono@unirow.ac.id

Abstract: This paper describes an English as a Foreign Language (EFL) speaking and listening test designed and piloted through the mobile application Telegram. The test was designed to diagnose the speaking and listening skills of first-year students at the Indonesian university UNIROW Tuban before they arrive on campus. The mobile delivery of the test, which can be accessed by smart phone and computer, was aimed to evaluate if learning apps can increase communicative learner interaction with authentic materials in English. The mobile test takes advantage of the vast amount of multimedia that can be transferred online, such as video and podcasts. Test-takers respond to Speaking Section prompts by recording themselves and answer Listening Section questions by clicking on the correct multiple choice options. All test-taker responses are sent to the administrator of the test, created through an inline Telegram bot. After piloting the test with students in a first-year conversation class at UNIROW Tuban, the test underwent a series of analyses, including item facility, item discrimination, split-half reliability, inter-rater reliability, and subtest relationships. These analyses are important for verifying the test's validity, reliability, and practicality. Overall, the test seems to be an important tool not only for diagnosing students listening and speaking needs, but also to increase interest in learning and practicing conversational English. Although the test

was designed for one university in particular, it (along with variations) can be used in similar contexts throughout the world.

Keywords – Assessment, Speaking, Listening, Telegram

INTRODUCTION

The kind of informal language learning found today was a common occurrence in former times and raises some interesting questions on the relationship between technology and language learning. Language teaching and learning is moving towards a new direction (mobile-assisted language learning/MALL), it is becoming more and more learner-centered and autonomous (Lixun, 2017). Since the mid-1990s, MALL has focused on the exploitation of five mobile technologies: pocket electronic dictionaries, personal digital assistants (PDAs), mobile phones, MP3 players, and most recently ultra-portable tablet PCs (Burston, 2013). A study conducted by Deng and Shao (2011) indicated that there was a high readiness of students to undertake mobile learning in their everyday life (Guo, 2015). Social networking is one tool which can assist teachers and learners to access information and facilitate the learning of English (Srinivas, 2010). According to Heidar & Kaviani (2016), one of the technologies that can be used to help learner in learning a foreign language is Telegram.

Telegram is now considered as one of the most famous platform online social networks among media university students (Heidar & Kaviani, 2016). Telegram has channels and bots to access information with the teacher. According to Omidi & Fooladgar (2015), Telegram intermediary server handles all encryption and communication with the Telegram API for the users. The users communicate with this server via the Telegram API. The server calls that interface as Bot API (https://core.telegram.org/bots/api).

In the Telegram Messenger official webpage for its bot, https://core.telegram.org/bots/api, the Bot API is an HTTP-based interface created for developers keen on building bots for Telegram. In authorizing a particular bot, each bot is given a unique authentication token when it is created. The token looks something like 123456:ABC-DEF1234ghIkl-zyx57W2v1u123ew11, but we'll use simply <token> in this document instead. All queries to the Telegram Bot API must be served over HTTPS and need to be presented in this form: https://api.telegram.org/bot<token>/METHOD_NAME. In creating a bot, it will be guided by the BotFather and we just need to follow a few simple steps. Once we've created a bot and received our authorization token, head down to the Bot API manual to see what we can teach our bot to do.

The terms assessment, test and evaluation are interchangeably refer to the same activity of collecting information for making decisions about the students through observation, self report and tests in order to improve their learning process; therefore, assessment plays a great role in identifying the student's areas of strength and weaknesses (Nadia, 2013). Bachman and Palmer (2010) state that assessment is used to provide a description of the progress of individuals such as language use. Sarosdy *et al* (2006) argue that assessment focuses on testing, measuring or judging the progress and the achievement or the language proficiency of the learners. So, student test scores can measure learning (Haertel: 2013).

Based on the second author experience, before moving to Indonesia, where the second author expected to teach EFL courses at a teacher's college for his Peace Corps service, he developed, piloted, and validated a diagnostic test that he believed would help identify his future students' strengths and weaknesses. He did not have access to students in Indonesia while developing the original test, but was able to pilot the assessment with sixteen university students in Nicaragua. The original assessment, which measured listening, reading, and writing skills, was criterion-referenced, grading students' responses against a preset goal or objective rather than against the performances of other test-takers (Bailey & Curtis, 2015).

After two years of teaching EFL in Indonesia, he decided to create a new diagnostic test for university English learners in the country, revising his original work to reflect what he learned. In conjunction with the first author, the English Teaching Department Dean at UNIROW Tuban, a teachers college in East Java, Indonesia, he altered the test to measure speaking and listening, two areas on which the university wishes to focus instruction. The English Department requested that he make an oral communication assessment for first-year students that can be accessed by mobile device. Accordingly, he changed the medium of distributing the test from paper to the mobile application Telegram.

While we did not know many details about the target population while designing the first version of this test, we were able to choose one university for the revision and learned a great deal about student strengths, weaknesses, career goals, and backgrounds. The intended participants for this test are Indonesian university students enrolled in a four-year English teaching program at UNIROW Tuban. The university, which opened in 2007, admits students who are enrolled in various programs, such as Mathematics, Fishery, and Indonesian. Although the English Teaching program aims to produce English as a Foreign Language (EFL) teachers for primary through secondary schools, at least a third of the majors plan to pursue careers outside of education, such as entrepreneurship, international

business, post-graduate study, and hospitality. As mentioned previously, the overwhelming majority are stronger in reading and writing than listening and speaking.

The designed test is intended to measure English speaking and listening proficiency of students entering a required introductory course entitled Speaking and Listening for Daily Conversation at *Universitas PGRI Ronggolawe* (UNIROW) Tuban, a teachers college in Tuban, Indonesia. As a diagnostic assessment, its results will allow administrators and professors to "more closely identify students' particular strengths and weaknesses" (Bailey & Curtis, 2015, p. 23). Although universities often view diagnostic tests as a way to place students in their appropriate levels (Bachman, 1990; Alderson, 2005), almost all UNIROW Tuban students attend the same classes as their peers regardless of proficiency level, so this test will probably not be used for course placement purposes. The main goal of this test will be similar to most diagnostic tests — to guide instruction so that it best addresses learners' needs (Alderson, 2005; Alderson, Clapham & Wall, 1995; Bachman, 1990; Bachman & Palmer, 2010). Instructors in UNIROW Tuban's English teaching department will be able to use the test to identify student strengths and weaknesses before the first day of class, and therefore will be able to better design their syllabi accordingly.

The long-term vision for this test, the first author expressed to the second author, would be a low-stakes, annual assessment of each cohort as they return after summer break. By using a version of this test each year, or perhaps twice a year, the first author would create a great opportunity for positive *washback*, which occurs when a test promotes desired teaching and learning outcomes (Bailey & Curtis, 2015). Instructors and students might be more inclined to improve their speaking and listening skills throughout the semester and summer break if they were aware that they would be tested each year, and that those scores would be compared to those of previous years.

Students of all levels consistently desired more practice with listening comprehension, spoken fluency, pronunciation, debate, academic article writing, slang, and idioms. The university staff's desires matched those of the students in all these areas except learning slang and idioms. Note that almost all of the skills mentioned by the students and staff are related to speaking and listening. Weaknesses in oral language mirror what we saw while teaching throughout Indonesia, where a heavy emphasis on grammar-translation and teacher-centered classrooms, as well as a lack of interaction with English speakers of any level, led students to feel much more comfortable with reading and writing skills than listening and especially speaking. Therefore, we hope that the washback from this test might address student concerns by shifting instructional focus to those areas.

METHOD

Identifying and defining specific test constructs is essential to designing an assessment that is valid. Without knowing what one hopes to assess, a test designer will likely create a product that is ultimately aimless and meaningless. Buck (2001) writes that "an understanding of what we are trying to measure is the starting point for test construction" (p. 1). Thus, the first step that we took in creating this test was to define the constructs to be measured. Then, keeping those constructs in mind, we created a test to measure those constructs, following Alderson, Clapham, and Wall's (1995) framework. According to Alderson *et al*. (1995), "the test specifications are the blueprint to be followed by test and item writers, and they are also essential in the establishment of the test's construct validity" (p. 9). While designing this test, it was essential to remind myself of the two constructs — listening and speaking — that we had chosen to assess.

The reason why we have chosen to assess speaking and listening, when the previous test measured reading, writing, and listening, is due to the needs and desires of the target population. While leading monthly workshops at UNIROW Tuban during the second author's two-year service, we conducted periodic needs assessments through questionnaires and interviews with students and staff. We also conducted the second author own observations during classes.

Most incoming students are seventeen or eighteen years old and have studied English in the Indonesian school system for at least six years, though the quality of instruction varies tremendously. First-year students generally range in their English language levels from intermediate to advanced. Tuban, a city of about 1,200,000 people on the northern coast of Java, is known as a fishing town and producer of hardwood teak. The university attracts students from the city and surrounding rural areas, many of whom have rarely interacted with fluent English speakers. While this test was designed for UNIROW Tuban, there are numerous universities throughout Java with similar student populations and English departments that might be interested in such a test.

Listening

The first construct that the revised test aims to measure is listening comprehension. In order to demonstrate their listening skills, students will need to listen to, comprehend, and write essential information about a spoken text. Within the context of L2 acquisition, Richards, Platt and Platt (1992) define *listening* as "the process of understanding speech in a second or foreign language… [that] focuses on the role of individual linguistic units, as well as the role of the listener's expectations, the situation and context, background

knowledge and the topic" (p. 344). Further, listening comprehension exists only when learners utilize multiple comprehension tools to engage, process, and understand the speaker (Buck, 2001). For example, in order to succeed on the listening section of this test, learners must rely on their knowledge of linguistic units while also considering contextual and background information.

Speaking

Speaking skills include numerous factors, but as noted by Iwashita, Brown, McNamara, and O'Hagan (2008) while examining speech samples of the TOEFL iBT, those which are measurable include linguistic resources (grammatical accuracy, grammatical complexity, and vocabulary), phonology (pronunciation, intonation, rhythm), and fluency (filled and unfilled pauses, rewording, total pausing time, speech rate, and mean length of run). Following the functional perspective of language use defined by Brown and Yule (1989), most speaking tasks are either transactional or interactional. The purpose of transactional language, which is used to convey factual information, is to give a message to someone (Brown & Yule, 1989). Interactional language, on the other hand, serves to express social relations, personal attitudes, or establish human relationships (Brown & Yule, 1989). In order to be competent in speaking, learners must develop not only linguistic resources, phonology, and fluency, but also be able to use those skills for transactional and interactional purposes.

The Medium Chosen to Present the Test

This test was designed on the mobile application Telegram. As student test can measure learning (Haertel: 2013), especially language learning, this mobile app next can be called as Telegram Assisted Language Learning (TALL). Telegram is a downloadable, free messaging app which can be accessed by mobile phone, tablet, or computer. Every student in the English teaching program has access to the Internet via personal phones, laptops, or internet cafes. In fact, many students in Indonesia have smart phones, but lack error-free English textbooks with authentic materials. By using the Internet, instructors can gain access to a vast amount of testing stimulus materials that they can send to students, such as podcasts, videos, and website links. The use of Mobile Assisted Language Learning (MALL) has been credited with expanding multimedia use, particularly for listening and speaking activities in situations where learners may wish to collaborate. (Kukulska-Hulme & Shield, 2008). There have been numerous studies in recent years on the impact of mobile learning and mobile assisted language learning (MALL). Given the powerful

features of the smartphone, its connectivity, multimedia support, growing ubiquity, and communication capabilities, it may seem surprising that MALL remains as Burston (2014a) comments, "on the fringes" of instructed language learning (p. 115). He points out in this study—as well as in his meta-analysis from 2015—that most published studies of mobile devices in the service of language learning are experimental in nature (with often no follow-up), have short time frames (often four to six weeks), and tend to focus exclusively on vocabulary development. Most MALL projects emphasize drill-type exercises, rather than communicative activities. As Burston (2015) comments, "nearly all [studies] presuppose a behavioristic paradigm involving rote learning and structuralistic tutorial exercises" (p. 16). His extensive annotated bibliography of MALL studies (2013) provides ample evidence of his assertion. The possibility to easily incorporate multimedia into this test was a major reason why we decided to use a mobile application.

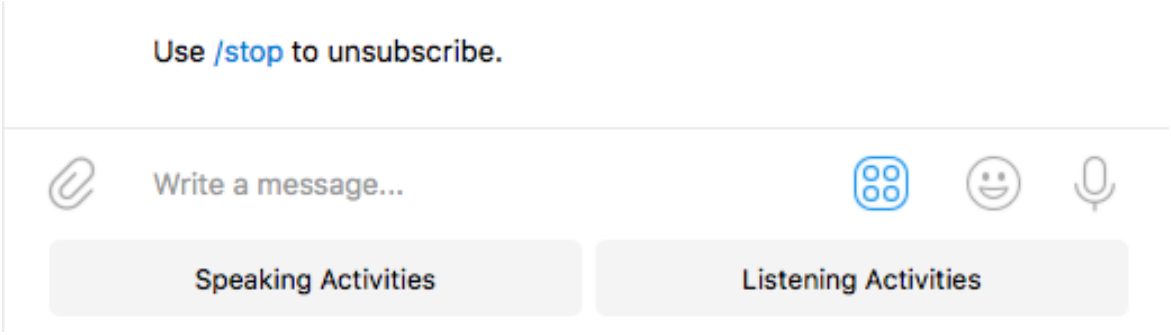
Aside from insufficient textual English resources, many professors in rural Indonesian universities are vastly outnumbered by students. With so many tests to grade, professors struggle to provide timely, qualitative, and thorough feedback, even though such feedback has been shown to increase student learning (Vitiene & Miciuliene, 2008). By using a process that ensures learners receive timely feedback after submitting responses, students will better be able to correct their errors. Even though this is a diagnostic test, students will likely want to receive feedback on their responses.

Another benefit of using a mobile application is that learners can access the test whenever and wherever they choose. Such remote access is especially important because this test will be administered before students arrive on campus. The professor does not need to deliver the test by mail, which is costly and takes time, and the students do not need to travel to the university to pick up the test. It is possible that students will look up the answers or talk to classmates, but because this is a low-stakes test that will not have any impact on their schooling, the temptation should not be too high.

We chose Telegram instead of other mobile applications for several reasons, the most important being that Telegram has the most attractive and supportive features. The first benefit is that Telegram is not only used for chat, it also has capability to send files of any type. Besides, Telegram also has channels so that professors can often send out announcements, audio files, and website links to students via "course channels" created in the application. A second major benefit of using Telegram is that its programmable bots allow for automatic delivery of testable items and feedback to learner responses. With the help of a program called Chatfuel, we programmed a bot to send instructions and testing items to students. Within the test, we were able to create folders, such as 'Speaking

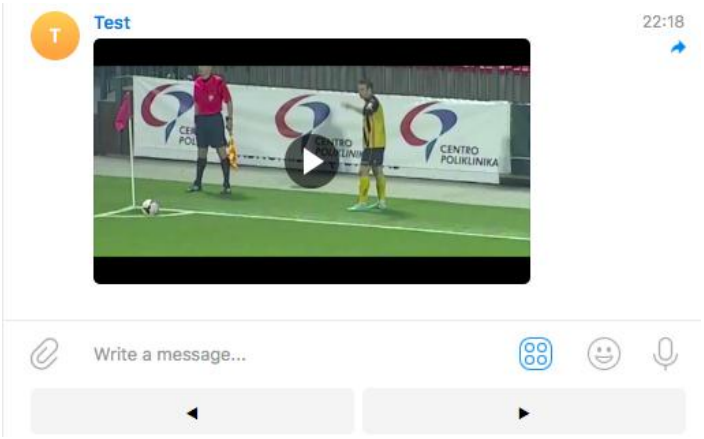
Activities', which students can click on. Once they open a folder, whatever text or media we choose will send itself to the students. Telegram's bots also allow for the creation of subfolders, such as 'Speaking Activity 1' and 'Speaking Activity 2,' which students can click on, leading the bot to send another set of text or multimedia. Students are not only able to navigate the test by clicking on folders and subfolders, but they also can respond to prompts by typing and recording themselves speaking. All student answers are sent to the administrator of the bot. We listed the steps to complete a sample listening activity below.

Step 1: Students click on "Speaking Activities" folder



Step 2: Students click on "Speaking Activity 4"

Step 3: Students receive prompt (video of a soccer player scoring a goal and celebrating).



Step 4: Students click on the microphone symbol to record themselves summarizing what happened in the video. The administrator receives the response that they send.

Test Methods and Organization

To describe the test, we will use Wesche's (1983) framework, which holds that "tests generally consist of a number of items, each composed of *stimulus material* and a related *task* which requires a *response* on the part of the examinee. Responses are then scored according to certain *criteria*" (Wesche, 1983, p. 43). When *stimulus material*, a *task posed to learners*, the *learners' response*, and *scoring criteria* are combined to create an assessment tool, they are "intended to reflect whether the examinee possesses certain knowledge, or to predict whether he or she can perform certain acts" (Wesche, 1983, p. 43). Although all four aspects are integrated during a test, we will explain each component separately so that readers can see each stage that the test-takers in this project have faced.

Stimulus Material

The first piece of Wesche's (1983) framework is *stimulus material*, or what Bailey and Curtis (2015) refer to as "whatever linguistic or nonlinguistic information is presented to the learners in a test to get them to demonstrate the skills or knowledge to be assessed" (p. 347). We chose the stimulus materials because they include content that might be relevant or interesting to students, which would increase investment and washback. For example, a speaking prompt asks students to discuss a pleasant experience that they had over the summer, presenting an opportunity for learners to share a personal experience. We attempted to reduce test anxiety by including a humorous video prompt of a cat trying to jump from a table towards a windowsill but falling (the cat was not injured). Not only is the video clip lighthearted, but it also provides good content for students to describe an event in English. The topics are also the same or similar to those that are covered in the Speaking and Listening for Daily Conversation course at UNIROW Tuban.

Following Swain's advice (1984) and *start from somewhere*, meaning that "test development should build from existing knowledge and examples" (p. 188), we made stimulus materials similar in structure and content to those of the TOEFL and IELTS. Those two exams test populations with similar English levels and goals, and low-stakes exposure to their item types will benefit UNIROW Tuban learners who plan to pursue postgraduate degrees or work abroad, as they often must take the TOEFL or IELTS. We also followed the advice of Kukulska-Hulme and Shield (2008), keeping each item's stimulus material, as well as its associated task and learner response, no longer than ten

minutes. If any sections took too long to complete, we found a shorter alternative or divided it into parts. The stimulus materials for each test section are described in Table 1.

Table 1: The stimulus materials for each test section

Test Section	**Stimulus Materials**
Speaking I and II	Two written prompts — the first asks students to describe a pleasant experience that they had during the summer; the second asks students to tell a tourist how to travel from Tuban to Surabaya (the capital of East Java).
Speaking III and IV	Two visual prompts — the first is a photograph of a group of people eating at a street food vendor; the second is a 28 second video of a soccer player scoring a goal and celebrating by running off the field, sitting in an empty stadium seat, and clapping.
Listening I and II	Two short (about one-minute) clips from conversations between two people; the first is informal, between friends; the second is formal, between a student and a professor.
Listening III and IV	Two visual prompts — the first is a photograph of a van so full of bananas that they are falling out of the open door; the second is a 17 second video of a cat trying to jump from a table to a windowsill but falling.

Task Posed to the Learner

The second component of Wesche's framework, the *task posed to the learner*, refers to the cognitive processes that test-takers use to understand the task and produce output (Bailey & Curtis, 2015). The tasks in this part of the test are described in Table 2.

Table 2: *Task Posed to the Learner*

Test Section	**Task Posed**
Speaking I and II	Understand the prompt; think of a suitable topic/travel option; organize their answer so that it makes sense to the listener.
Speaking III and IV	Recognize what is happening in the photograph/video.
Listening I and II	Understand the topic being discussed; remember pertinent details (students can replay the clip if they choose); understand two written multiple choice questions and four potential answers for each question about the audio clip.
Listening III and IV	Recognize what is happening in the photograph/video; understand four spoken, potential descriptions of the scene.

Although the section titles might make it seem so, none of the subtests completely isolate each language skill. For example, students will need to understand written prompts in order to produce a suitable spoken response. Still, the amount of language skill interference is minimal because the prompts are short and relatively simple compared to the rest to the language on the test.

According to Bachman and Palmer (2010), a test's language should match the language used outside of the test, and be both authentic and interactive. In this context, *authenticity* refers to the way that language is used in natural communication (Douglas, 2000). *Interactiveness* is the "extent and type of involvement of the test taker's… language ability (language knowledge and strategic competence of metacognitive strategies), topical knowledge, and affective schemata" (Bachman & Palmer, 2004, p. 25). Therefore, the tasks on this test mirror situations in which students could plausibly find themselves, such as listening to a professor's lecture or giving instructions to a tourist.

The tasks are intended to measure a range of language skills, including pronunciation, grammar, vocabulary, and cohesion, all of which the introductory Speaking and Listening course aims to cover during the semester. The tasks also allow for a fair amount of flexibility in the response of the learners. Students with advanced speaking skills can elaborate on their responses and use difficult grammatical structures, while those

possessing lower English abilities can still answer the questions at their own level, and thus provide instructors with information about their language skills.

Learners' Response

The *learners' response* is the physical answer that the test-taker produces (Wesche, 1983), whether by tapping a multiple-choice option on their phone's screen, typing a short answer, or speaking to a Telegram bot interviewer. Descriptions of learners' responses for the test are provided in Table 3.

Table 3: Descriptions of learners' responses for the test

Test Section	**Learners' Response**
Speaking I and II	Record their spoken answer for the Telegram bot; use appropriate vocabulary, grammar, and pronunciation.
Speaking III and IV	Record their spoken answer for the Telegram bot; describe the situation using appropriate vocabulary, grammar, and pronunciation.
Listening I and II	Tap on the correct multiple-choice options.
Listening III and IV	Tap on the correct multiple-choice options.

Students will have the option of typing and recording their answers on whatever devices they choose — mobile phone, tablet, or computer. Although a time limit also allows for a more accurate measurement of students' real-world proficiency than if they had unlimited time to think, rerecord, and edit their answers, we have chosen not to include time constraints because they would be difficult to enforce. Additionally, by not imposing a time constraint, the test is more likely to elicit the learners' best performance, known as *bias for best* (Bailey & Curtis, 2015).

Scoring Criteria

Scoring on the test as a whole will be criterion-referenced, meaning that "a student's score is interpreted relative to a preset goal or objective — the criterion — rather than to the performances of other test-takers" (Bailey & Curtis, 2015, p. 56). We chose this scoring philosophy because the test aims to measure language competence, not to compare

learners against each other. As for individual subtests, each section has different scoring criteria. The listening section, composed of multiple-choice items, will be graded objectively, as the items have one correct answer. With an objective scoring procedure, subjectivity involved in rater judgement is reduced. And because test-takers' answers are sent to the professors' phone or computer, they can be scored by machine.

The speaking subtest, in which learners provide oral responses, requires a subjective scoring procedure. For each exercise, we scored responses against an analytic rubric, a rating scale in which "the teacher scores separate, individual parts of the product or performance first, then sums the individual scores to obtain a total score" (Mertler, 2001, p. 1). We chose to adopt a TOEFL (2017) speaking rubric because the items in my test were similar to those on that high-stakes test; students and teachers also expressed the view that they wanted practice with TOEFL scoring. While neither us nor any professors at UNIROW Tuban are official TOEFL scorers, using the rubric provided a framework for what students should produce. The original TOEFL rubric relied on a holistic grading procedure, with responses receiving a total score from 0-4. To make the rubric analytic, we kept the four sections (general description, delivery, language use, and topic development), but scored each section from 0-4. By aggregating the four section scores, responses on my test could receive a total of 0-16 points. To increase reliability, the concept that results of a test should be consistent (Brown, 2005), we used two raters to compare scores. By averaging each rater's scores, students are given more reliable score that is less prone to one person's view. An analytic rubric not only provided structured scoring and feedback, but was also optimal because the speaking responses involved integrated hierarchical components of language. By dividing scoring into sections, we could assess specific pieces of language ability within each answer.

Test Piloting

Piloting this test has been instrumental in creating a reliable, valid, and accurate test. As Alderson, Clapham, and Wall (1995) write, "However well designed an examination may be, and however carefully it has been edited, it is not possible to know how it will work until it has been tried out on students" (p. 73). Stephen first made sure that the technology worked properly by pre-piloting the test with Dr. Agus. After receiving feedback from Dr. Agus, he made a few changes to the test. Most notably, he provided more explicit instructions for test items and desired student responses. We also changed the programming of the bot so that learner responses were automatically sent to the bot — previously, the students would have had to manually forward their answers to the bot. With

the help of the English Department at UNIROW Tuban, twenty first-year college students enrolled in English Speaking and Listening for Daily Conversation were asked to complete the test. The students completed the test on their own time (mostly on campus after class) and all twenty students responded.

FINDINGS AND DISCUSSIONS

Using the test piloting results, we will conduct several statistical analyses consisting of item facility, item discrimination, distractor analysis, response frequency distribution, split-half reliability, inter-rater reliability, and subtest relationships. These statistical analyses will allow me to improve my test by measuring and critiquing individual test items, internal consistency, as well as the exam's overall reliability and validity.

Findings

According to Bailey and Curtis (2015), item facility (IF) is "an index of how easy an individual item was for the people who took it" (p. 198). To calculate the IF for the objectively scored items, which comprise the entire Listening Section, we divided the number of test takers who answered the item correctly by the total number of test takers (Bailey & Curtis, 2015). Item facility statistics are listed in Table 4. An IF of 1.00 means that all test takers chose the correct answer; an IF of 0.00 means that no one answered correctly (Bailey & Curtis, 2015). Additionally, Oller (1979) writes that "items falling somewhere between about 0.15 and 0.85 are usually preferred" (p. 247). Therefore, item 8 was too easy for test takers, with items 3, 4, and 6 close to being too easy. My average IF was within Oller's aforementioned desirable range at 0.7, though there was a gap between the relatively difficult items (1 and 2) and the easier items (3, 4, 5, and 6). Upon revision of this test, we will consider making the items listed above more difficult or adding other items that are more challenging. Because my test is diagnostic and criterion-referenced, it is not necessarily problematic that some test items yielded a high IF. If all students answered an item correctly, it simply means that they knew the information being tested — or that they were able to guess correctly.

Table 4: *Listening Comprehension Subtest Item Facility (n=20)*

Item	Students who answered the item correctly	Item Facility (IF)
1	8	0.4
2	9	0.45
3	16	0.8
4	16	0.8
5	18	0.9
6	17	0.85
		Average IF = 0.7

Item Discrimination

Item discrimination (ID) provides an analysis similar to item facility, although the information is more detailed because it shows how the higher and lower scorers did on each item (Bailey & Curtis, 2015). To calculate item discrimination for the objectively scored items, We selected the high and low scorers by ranking students from highest to lowest based on their total score. Flanagan's (YEAR - SOURCE) method for estimating item discrimination recommends selecting the top 27.5% and bottom 27.5% of the total number of students tested, and several authors (see, *e.g*., Bailey & Curtis, 2015; Mertler, 2003; Nitko, 2001) write that between 25 and 33 percent of test takers can be used. We selected the top five (25%) and the bottom five (25%) of the total test takers. ID values range from +1 to -1, with +1 indicating perfect, desirable discrimination between high and low scorers, and -1 showing a complete but Wong discrimination (Bailey & Curtis, 2015). Table 5 presents our test's Item Discrimination analysis.

Table 5: *Listening Comprehension Subtest Item Discrimination (n=20)*

Item	Top 5 scorers with correct answers	Bottom 5 scorers with correct answers	Item Discrimination (ID)
1	5	1	0.80
2	5	0	1.00
3	5	2	0.60
4	4	2	0.50
5	5	4	0.20
6	5	4	0.20
			Average ID = 0.55

According to Mertler (2003, working with ideas offered by Chase, 1999), ID values of 0.50 and above are optimal and should be kept (p. 187). Fortunately, four of the six items have IF values of 0.50 or higher. On the other hand, two of the test items — five and six — had ID values of 0.20. For a criterion-referenced test, low ID values are not necessarily problematic. Bailey and Curtis (2015) write that "if all the test-takers got an item right on a progress test or an achievement test after instruction, the ID value would be 0.00, but this result could indicate their mastery of the item's content" (p. 205). High IF values for items five and six (0.9 and 0.85, respectively) indicate that a high ID for those values is probably due to most students knowing the content. If a teacher were to instruct the students who took this diagnostic test, they would likely choose to spend little time practicing the skills tested in those items. It is also worth noting that because we worked with twenty students, a small test population, each scorer's choice had a large effect on the data. we would need to run this test again with a larger population to determine if items five and six did not discriminate effectively between high and low scorers.

Distractor Analysis

Bailey and Curtis (2015) write that "a 'Distractor Analysis' is a procedure specifically related to the multiple-choice format" (p. 199). They further write that it is important to analyze the effectiveness of each individual item in order to improve a multiple-choice test (Bailey & Curtis, 2015). Table 6 presents the Listening Comprehension Subtest Distractor Analysis. Correct answers are marked by an asterisk.

Table 6: *Listening Comprehension Subtest Distractor Analysis (n=20)*

Item	A	B	C	D	Omitted Response
1	2	3	8*	7	0
2	1	8	2	9*	0
3	3	16*	0	1	0
4	16*	2	0	2	0
5	2	0	18*	0	0
6	17*	1	1	1	0

Table 6 shows that several distractors did not sway many test takers. Of the 24 total distractors, four were not chosen by any test taker, and five managed to convince only one person. Mostly, the unevenly distributed answers occurred in items that most test takers chose the correct answer, an issue related to item facility. Of course, a lack of distractor selection does not meant that distractors were poorly designed. The students could have known the material well enough to sift through all potential answers to select the correct one.

Upon first calculating the distractor answer distributions, we considered revising items one and two due to the high number of learners selecting another distractor. For item one, seven students chose distractor D, compared to eight who chose the correct answer, C. Similarly, for item two, eight students chose distractor B, while nine students picked the correct answer, D. After looking at these items, we decided that the questions were not misleading or confusing. With IFs of 0.4 and 0.45, the reason students were misled was probably because the items were difficult. Further, the items' IDs of 0.80 and 1.00 suggest that they were difficult for many of the low scoring test takers, while high scorers were able to understand the content.

Response frequency distribution

According to Bailey and Curtis (2015), "The response frequency distribution combines information from both the distractor analysis and the item discrimination analysis" (p. 208). Just as item discrimination analysis showed us a more detailed view of item facility by looking at the responses from the highest and lowest scorers, the response frequency distribution allows us to examine individual distractor strength from only the top

four and bottom four scorers. Examining the response frequency distribution allows us to examine which specific distractors are selected by high and low scorers. The response frequency distribution is listed in Table 7.

Table 7: *Listening Comprehension Subtest Response Frequency (n=20)*

Item	Scorers	A	B	C	D	Omitted Response
1	High	0	0	5*	0	0
	Low	1	0	1*	3	0
2	High	0	0	0	5*	0
	Low	0	3	2	0*	0
3	High	0	5*	0	0	0
	Low	2	2*	0	1	0
4	High	4*	1	0	0	0
	Low	2*	1	0	2	0
5	High	0	0	5*	0	0
	Low	1	0	4*	0	0
6	High	5*	0	0	0	0
	Low	4*	0	0	1	0

Reliability

Due to time constraints, we were not able to administer the test twice, so we calculated the reliability of the objectively scored section using the split-half method, a method of internal consistency. First, we split the listening comprehension questions in half by even- and odd- numbered items. We then recorded the scores as shown in Appendix B. To correlate the scores, we used the raw score formula for Pearson's correlation coefficient, or Pearson's r. After this initial calculation, we adjusted r in order to present an accurate value of r for the entire subtest. Bailey and Curtis (2015) note that the split-half reliability estimate will likely be lower when the test is halved compared to an entire test. Fortunately, a formula exists that allows us to accurately adjust and raise the

coefficient. Hatch and Farhady (1982) write that, "When we have obtained the reliability of half of the test, we can then use Spearman Brown's prophecy formula to determine the reliability of the full test" (p. 246). The values that we calculated are shown in the two left columns of Table 8.

Table 8: *Listening Comprehension Internal Consistency Measures*

Split-half reliability	Reliability after using Spearman Brown Prophecy Formula	Standard Deviation	Standard Error of Measurement (SEM)	Points Possible
r = 0.309	0.472	1.259	0.915	12.00

After adjusting r with the Spearman Brown Prophecy Formula, the correlation between the scores is just under 0.5, weaker than we would have liked. However, considering that there were only six total items, a low r value is not surprising. Each item's facility weighed heavily, so even a small difference in IF, such as 0.05, heavily impacted the correlation between the three even items and the three odds.

To examine the consistency of the test scores, we calculated the Standard Error of Measurement (SEM), also shown in Table 8. Brown (2005) writes that SEM "is used to determine a band around a student's score within which that student's score would probably fall if the test were administered repeatedly to the same person" (p. 188). Therefore, a student earning a five on the listening comprehension subtest would likely score between a six and four upon repeating the test. While we would have liked an SEM as low as possible, we are not disappointed by the result because this test was designed for diagnostic means. Additionally, the SEM is low because of the fairly low r value, which we believe was due to having only six test items.

Inter-rater reliability

We computed coefficient alpha for the subjectively scored Speaking subtest to determine inter-rater reliability between the two raters. We first found the variance for each of the raters on each subtest by entering given scores (found in Appendix C) on a calculator before plugging the numbers into the coefficient alpha formula. The resulting coefficient allowed us to examine the consistency of the two raters. As Bailey and Curtis write, "the closer the value is to the whole number 1.00, the greater the inter-rater

reliability" (p. 168). We were pleased with our test's inter-rater reliability, as the efficient alpha was 0.867. We attribute high inter-rater reliability to the detailed analytic rubric (Appendix A) as well as sufficient preparation and communication between raters before grading. Before scoring the test, the two raters discussed grading methods and to ensure that there was no confusion. Such efforts seemed to reduce what Bachman (1990) cites as the main cause of inconsistency among raters: "the application of different rating criteria to different samples of the inconsistent application of the rating criteria to different samples" (p. 178). Both raters not only applied the same criteria for each test but also graded consistently.

Subtest Relationships

We used Pearson's correlation coefficient to determine the correlation between scores on the two subtests and the total test. In order to find the extent to which the subtests measure the same construct, we also computed r-squared to determine whether there is overlap among the subtests. Bailey and Curtis ask, "Do the tests that are designed to measure the same construct correlate more highly than tests designed to measure different constructs?" (p. 273). If the answer is yes, we might have favorable *construct validation*, "the single, fundamental principle that subsumes various aspects of validation" (Cumming & Berwick, 1996, p.5). Although this test acknowledges that the subtests are integrative, and listening and speaking are oral communication skills, each section measures a distinct language skill. Therefore, we would expect some overlap between subtests but hopefully not too much. As the information presented in Table 10 suggests, we can be pleased with the results. There is a moderate amount of overlap (0.489) between Listening and Speaking scores, but that was to be expected because they are highly integrated language skills. Particularly noteworthy is the high rate of correlation between the two subtest scores and total test scores (0.923 Speaking and 0.912 Listening). Those who did well on each subtest did well on the entire test.

Table 10: *Subtest Relationships*

Test	**Correlation Coefficients (Pearson's r)**		
Total Test	0.923	0.912	-
Listening	0.699	-	0.912
Speaking	-	0.699	0.923
	Speaking	Listening	Total Test

Table 11: *R-squared for Subtest Relationships*

Test	**Correlation Coefficients (Pearson's r)**		
Total Test	0.852	0.832	-
Listening	0.489	-	0.832
Speaking	-	0.489	0.852
	Speaking	Listening	Total Test

Discussions

The traditional criterion for evaluating tests include reliability, validity, practicality, and washback (Bailey & Curtis, 2015). The reliability of a test, according to Oller (1979, "is a matter of how consistently it produces similar results on different occasions under similar circumstances" (p. 4). From the data analyzed in this essay, it is clear that inter-rater reliability on the subjectively scored section is high, though internal consistency on the objectively scored subtest is low. Internal consistency should be improved, even though Bachman (1990) admits that short tests are generally less reliable than long tests. Additionally, although an SEM of 0.915 on a subtest with twelve total points is not entirely problematic, this should be improved to create a more reliable test.

Validity, according to Oller (1979), is "how well the test does what it is supposed to do, namely, to inform us about the examinee's progress toward some goal in a curriculum… or to differentiate levels of ability among various examinees on some task" (p. 4). As a criterion-referenced, diagnostic English proficiency assessment, this test's validity is related to how well it measures test takers' English abilities. It is difficult to determine the test's validity at this point, since there has been no longitudinal analysis of test takers and their success in the classroom or a comparison of this test's results to other valid tests. After revising this test, we hope to determine its validity by comparing its

results with student scores in their classes. This test contains a fair amount of face validity (Cumming and Berwick, 1996) since it appears on the surface level to be valid for test takers. Much of the face validity stems from content validity, but the tasks of giving directions and describing a recent memorable experience reflect tasks learners might face in everyday life.

Bailey and Curtis note that, "developing, revising, administering, and scoring tests take time, money, and person-power" (p. 3). For these reasons, tests must be practical — a characteristic which includes "the preparation, administration, and interpretation of the test" (Oller, 1979, p. 4). This test is highly practical, as it should take no more than thirty minutes to complete, can be completed on mobile devices or computers, and can be sent, accessed, and scored from anywhere in the world (as long as there is Internet connection). If we wanted to further improve practicality, we could provide test questions on paper format, though we would need to replace the multimedia prompts with written descriptions. The scoring of the test is straightforward and quick, as an answer key is provided for the objectively scored items and an analytic rubric is presented for the subjectively scored subtest.

Washback, defined by Bailey and Curtis (2015) as "the effect a test has on teaching and learning" (p. 3), can be positive or negative. We believe that most of this test's washback will be influencing students and teachers to spend more time practicing oral language skills. Hopefully, the interactive content and mobile access encourages students to engage with their learning.

The analysis of subtests, test items, and item distractors will allow for improvement of this test. We have pinpointed weaknesses, whether from poor item distractors or items that do not sufficiently discriminate between high and low learners. We also have learned that internal reliability of the Listening Comprehension subtest is lower than desired and should be improved. We look forward to revising this test, and we hope that UNIROW Tuban will be able to use an improved version for their students.

CONCLUSION AND SUGGESTION

Technology is making it possible for people from all over the world to be able to communicate at the palm of their hand. Due to the enhancement of technology, digital learning allows people to learn in a more efficient and effective way. Language learning it evolving due to the usage of instant messaging applications like Telegram becoming a need for users. Telegram Assisted Language Learning (TALL) as technology progresses,

the importance of using Telegram alongside in education makes it even more vital to the overall success of a student ability to communicate internationally.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, UK: Continuum International.

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

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

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. New York, NY: Oxford University Press.

Burston, J. (2013). Mobile-assisted language learning: A selected annotated bibliography of implementation studies 1994–2012. *Language Learning & Technology*, *17*(3): 157–225. Retrieved from http://llt.msu.edu/issues/october2013/burston.pdf

Burston, J. (2014). The reality of MALL: Still on the fringes. *CALICO Journal*, *31*(1): 103–125.

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*(1): 4–20.

Bailey, K. M., & Curtis, A. (2015). *Learning about language assessment: Dilemmas, Decisions, and Directions* (2nd ed.). Boston, MA: National Geographic Learning

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* White Plains, NY: Pearson Education.

Brown, G., & Yule, G., (1989). *Discourse analysis.* Cambridge, UK: Cambridge University Press.

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

Chase, C. I. (1999). *Contemporary assessment for educators*. New York, NY: Longman.

Cumming, A., & Berwick, R., (Eds.) (1996). *Validation in language testing*. Clevedon, England: Multilingual Matters, Ltd.

Deng, H., & Shao, Y. (2011) Self-directed English vocabulary learning with a mobile application in everyday context. *Paper presented at the Proceedings 10th World Conference on Mobile and Contextual Learning (mLearn)*, Beijing, China: Beijing Normal University.

Douglas, D. (2000). *Assessing languages for specific purposes*. New York, NY: Cambridge University Press.

Guo, Hui (2015) *Analysing and Evaluating Current Mobile Applications for Learning English Speaking*. Birkbeck: University of London.

Haertel, Edward H. (2013) *Reliability and Validity of Inferences about Teachers based on Student Test Scores*. Princeton: ETS Research and Development.

Hatch, E. M. & Farhady, H. (1982). *Research design and statistics for applied linguistics.* Rowley, MA: Newbury House.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29,* 24-49.

Kukulska-Hulme, A., & Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, *20*(3): 271-289.

Lixun, Wang (2017) *Public Lecture Series 2017: The English You Didn't Learn in School V, Mobile Assisted Language Learning*. Retrieved from https://www.ied.edu.hk/ele/pls/spring 2017/seminar4.pdf on June 8th 2017.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, *7*(25). Retrieved from http://PAREonline.netgetvn.aspv=7&n=25

Mertler, C. A. (2003) *Classroom assessment: A practical guide for educators.* Los Angeles, CA: Pyrczak Publishers.

Nadia, Maarouf (2013) *The Importance of Continuous Assessment in Improving ESP Students' Performance*. Kasdi Merbah Ouargla University.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Oller, J. W. (1979). *Language tests at school*. London: Longman Group.

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics*. Essex, UK: Longman Group.

Sàrosdy, J. and Bence, T,F. Vadnay, M (2006) *Applied linguistics 1 for BA students in English*. Cambridge: Cambridge university press.

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. J. Savignon & M. Berns (Eds.), *Initiatives in communicative language teaching* (pp. 185–201). Reading, MA: Addison-Wesley.

Understanding your TOEFL iBT test scores. (2017). Retrieved from https://www.ets.org/toefl/ibt/scores/understand/

Vitiene, N., & Miciuliene, R. (2008). Application of criteria-referenced assessment and qualitative feedback to develop foreign language speaking skills in the context of e-teaching/learning. *Quality Of Higher Education*, *5*(7),:152-168.

Wesche, M. (1983). Communicative testing in a second language. *Modern Language Journal*, *67*(1): 41-55.

Appendix A

TOEFL iBT® Test

Independent SPEAKING Rubrics

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. ~~A response at this level is characterized by all of the following:~~	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. ~~A response at this level is characterized by at least two of the following:~~	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is some-what limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/ or overall coherence occur; meaning may be obscured in places. ~~A response at this level is characterized by at least two of the following:~~	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/ or coherence or is only minimally connected to the task, or speech is largely unintelligible. ~~A response at this level is characterized by at least two of the following:~~	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

Appendix B: Split-half Reliability Calculations

Learner	**X** = score on odd numbered items (3 points possible)	**Y** = score on even numbered items (3 points possible)	X^2	Y^2	XY
1	1	2	1	4	2
2	2	2	4	4	4
3	2	2	4	4	4
4	2	2	4	4	4
5	2	2	4	4	4
6	2	3	4	9	6
7	2	3	4	9	6
8	3	3	9	9	9
9	2	3	4	9	6
10	1	2	1	4	2
11	2	2	4	4	4
12	2	2	4	4	4
13	3	3	9	9	9
14	2	2	4	4	4
15	1	3	1	9	3
16	2	2	4	4	4
17	3	2	9	4	6
18	3	2	9	4	6
19	1	0	1	0	0
20	3	2	9	4	6
Σ	41	44	93	106	93

Appendix C

Individual Pilot Exam Scores by Section

Learner	Listening Section (12)	Speaking Section (16)	Total Scores (28)
1	2	R1 (4) R2 (5) = 4.5	6.5
2	10	R1 (12) R2 (11.5) = 11.75	21.75
3	12	R1 (13.25) R2 (12.5) = 12.86	24.86
4	8	R1 (7) R2 (7.5) = 7.25	15.25
5	8	R1 (11) R2 (11) = 11	19
6	8	R1 (10) R2 (9.5) = 9.75	17.75
7	12	R1 (11) R2 (12.5) = 11.75	23.75
8	8	R1 (13.74) R2 (15) = 14.38	22.38
9	8	R1 (11.75) R2 (12) = 11.88	19.88
10	6	R1 (12) R2 (11) = 11.5	17.5
11	6	R1 (11.5) R2 (11) = 11.25	17.25
12	12	R1 (15.75) R2 (14.75) = 15.25	27.25
13	12	R1 (15.25) R2 (14.75) = 15	27
14	10	R1 (15.5) R2 (13) = 14.25	24.25
15	10	R1 (15) R2 (14) = 14.5	24.5
16	10	R1 (13) R2 (14) = 13.5	23.5
17	8	R1 (12) R2 (11.75) = 11.88	19.88
18	8	R1 (13) R2 (12) = 12.5	20.5
19	6	R1 (11.75) R2 (12) = 11.88	17.88
20	8	R1 (14.25) R2 (13.5) = 13.63	21.63