Developing an English Summative Computer-Based Test (CBT) for Tenth Grade Students

[1]Putro Bagus Wahida Machtowil, [2]Sintha Tresnadewi, [3]Francisca Maria Ivone

[1,2,3]English Education-Universitas Negeri Malang

ARTICLE INFORMATION	ABSTRACT
	The use of ICT in language testing has recently expanded to more diverse computer-based exams. This includes achievement test like UNBK (*Ujian Nasional Berbasis Komputer*) and university entrance test such as SBMPTN *(Seleksi Nasional Masuk Perguruan Tinggi Negeri)*. thus, students need to get used to using this kind of test before taking a competency exam. The method employed in this study is Research and Development (R&D). Based on the process and results of the development of this English Computer Based Test, it has many features that can cover the difficulties in conducting online tests such as timers, anti-cheating, and other features that are always a challenge in online tests. Some features are not working normally like audio and also some bugs in the product. Furthermore, these products can provide insight into the effectiveness of online tests and can be adjusted to the needs of test makers.

Correspondence Address:

Putro Bagus Wahida Machtowil
Graduate Program in English Language Teaching
Universitas Negeri Malang
Semarang St, Number 5 Malang
E-mail: putro.bagus.1802218@students.um.ac.id

Many language tests are now administered via computer. However, widespread adoption of computer-based language test has been hampered by factors such as the difficulty of gaining access to computers and the associated costs. In language testing, the use of a computer as a testing medium is referred to as Computer-Assisted Language Testing (CALT), Computer-Based Testing (CBT), or Computer-Based Language Assessment (CALA) (Winke & Fei, 2008). Along with the advancement of Computer-Based Assessment (CBA), there will always be a place in the classroom for CALT. Especially since UNBK (*Ujian Nasional Berbasis Komputer*) and SBMPTN (*Seleksi Nasional Masuk Perguruan Tinggi Negeri*) have been computerized, students must be prepared for the transition before taking a language test in this new format.

Since computers and electronic devices have become popular in evaluating the language skills of English learners, Computer-based assessment techniques (CBA) provide a valuable resource for students who want to assess their academic progress through instant personalized feedback. According to High (2010), Computer-Based Assessment (CBA) techniques are facilitator of formative assessment. It can provide feedback for each student with appropriate mechanisms for analyzing data on students' strengths and weaknesses related to their response to assessment items which enables them to engage in self-assessment.

In addition, the utilization of Computer-Based Assessment (CBA) programs results in many advantages both for students and teacher, such as the teacher can provide feedback easily, the teacher can do monitoring or tracking in term of students' difficulty in English and time-saving because an assessment can be created using software (Haryani:2018). For students, Nikou & Economides (2016) examine the positive effect of computer-based assessment and mobile devices on students' learning motivation and achievement.

Moreover, CBA can also promote positive attitudes and awareness of learning autonomy. It is in line with the research from Bull (2006), who stated that Computer-based formative assessment to promotes reflection and learner autonomy. In the case of positive attitudes, taking part in Computer-based Assessment (CBA) programs are intrinsically motivating for students. They see it as a modern and valuable tool, enabling them to connect with the world. Computer-Based Assessment (CBA) programs provide software applications such as browsers (i.e., Mozilla Firefox, Google Chrome, Opera, etc.), and students are free to access with internet connection. Since English is the most widely used language online, students quickly see its practicality. The internet may be a motivational tool since it provides students the freedom to study at their speed and in whatever way best suits them. Students' awareness of learning autonomy helps them promote independence in learning and the development of learning strategies.

Computer-based testing (CBT) and the use of independent study, both advocated by the Ministry of Education and Culture before the epidemic, are now being implemented. COVID-19 (Circular Letter of Education and Culture Ministry, 2020). The problems of globalization need that today's students be digitally literate, creative, highly motivated, and able to produce an invention with the explicit goal of preparing tomorrow's Millennial for those challenges. However, some teachers did not implement computer-based assessment maximally because they cannot integrate technology into their teaching-learning process.

Kurt (2013) investigates some teachers who believe that they cannot use technology efficiently because they do not have adequate support and do not have enough time. Similarly, Jaber (1997) conducted a study about factors that influence teachers' use of CBT; he mentioned that teachers react negatively to CBT implementation because they spend too much time learning how to use CBT. Recently, Lesly (2021) stated that the implementation of CBA is not maximal because of some weaknesses such as electricity problems, time (session) delay, server error, and computer errors. These factors cause many teachers to tend to revert to traditional instruction, whereas the students are expected to be well prepared for the implementation of Computer-Based Assessment (CBA).

Regarding the problem mentioned above, needs analysis to know the obstacles and challenges of Computer-Based Test (CBT). It is important to see students' transition readiness and monitor the effectiveness of implementing a Computer-Based Assessment (CBA), such as at a high school level. SMAN 20 Surabaya is one of the high schools that has to transition to the development and use of computer-based tests.

Based on a preliminary review conducted by the author through questionnaires and interviews, indicates some information: As a first issue, some students suffer from a general fear of computers. However, for others, the degree of their computer literacy is of paramount importance. Second, the usage of the mouse, the clarity of the screen, the screen size, the quality of the display, the pace of the show, and the ability to scroll are all technical concerns in CBTs that might impact students' answers. Third, based on the interview with an English teacher, the training in Computer-Based Test is still needed. The teacher also said that it would be better if the government gave an opportunity to develop a test media based on the needs of each school so that the media can be effective to be used.

Furthermore, if Computer-Based Test (CBT) are used for a long term, it would be better if the students were given clear instructions about this new kind of test media because some students said that they still choose to use the traditional test media rather than Computer-Based Test due to the general anxiety of using computer and level of computer experience. So it needs more support in managing the implementation of Computer-Based Test (CBT) to achieve the success of learning targets through best practices and facing emerging challenges. These factors are the challenges for Computer-based Assessment (CBA) designers and administrators to construct fair, reliable, and valid test scores.

Russell et al. (2003) state that Computer-Based Test (CBT) In most cases, test takers should have at least as much access to editorial feedback and control over their replies as they would on a Paper-Based Test (PBT). Many factors must be considered when deciding whether to use computer-based assessment (CBA) instead of more conventional, paper-and-pencil testing methods. These include software quality, secure delivery, a reliable network (if Internet-based), capacities, support, maintenance, software costs for development and test delivery, licenses, and software fees (Scheuermann & Pereira, 2008). The ability to reliably automate data processing activities during test administration (such as data collection, scoring, report generation, etc.) provides the foundation for the various benefits of computer-based administration over paper-and-pencil testing.

To develop a good test, several criteria need to be not only known but also fulfilled satisfactorily as a test is a set of data collection instruments that should function appropriately if accurate information about the learners is to observe optimally to avoid the so-called gi-go effects – garbage in garbage out impacts. The first is validity. Gronlund and Linn (1990:7) Describe validity as the conformity of historical interpretations of test scores and other assessment data to a specified application. This implies that the outcomes of the exam must have significance, be relevant, instructive, and helpful.

Consistency is the second. According to Brown (2004:20), a trustworthy test produces results that are stable and consistent throughout time. Theoretically, we should get consistent results by simultaneously administering the same exam to the same students. Being objective is the third. According to Borg and Gall (1983), test results are considered aim if they are not affected by the prejudices of the people administering and scoring the exam. The Rorschach Inkblot Test is an excellent example of a highly subjective test; its administration and scoring procedures are open to interpretation. According to Brownell, M. H. (1947), Rorschach tests are not as reliable as other types of tests. In this setting, tester bias might quickly emerge. On the other hand, multiple-choice exams are more impartial since anybody may administer them, and everyone who takes them can use the same scoring key without any disagreement.

These are the minimum requirements for an effective testing environment and scoring procedure. According to Borg and Gall (1983), a quality test will include an accompanying handbook that details the protocols to be followed by the test-maker in case of any contingency that might impact test performance. The developers will decide how long the test should take, whether or not participants may retake it, how questions should be answered, and how much contact between the tester and participants should be allowed. Finally, we have justice. If the test is valid, two groups of equal ability should get the same score on each test item (Borg & Gall, 1983).

Aptitude tests, proficiency tests, placement tests, diagnostic tests, and accomplishment tests are all examples of the many sorts of tests listed by Brown (2004: 43). An accomplishment exam is a standard tool for evaluators in the classroom. A placement exam, diagnostic test, and accomplishment test are the "three sorts that you will almost surely construct," as stated by Brown (2004). Achievement exams may be either formative or summative. A formative assessment aims to determine whether or not students have learned the content of a particular lesson, unit, or textbook section (Gronlund & Waugh, 2013). On the other hand, the summative examination is administered at the conclusion of a course or instructional unit, designed to determine how much a

student has learned. (Brown, 2004). Indonesian teachers popularly use summative test. This test is standard and always use to know students' mastery of the courses bases on the grading standard from National Education Standard Board (BSNP). To simplify the types of assessment, the figure below will give more assessment information:
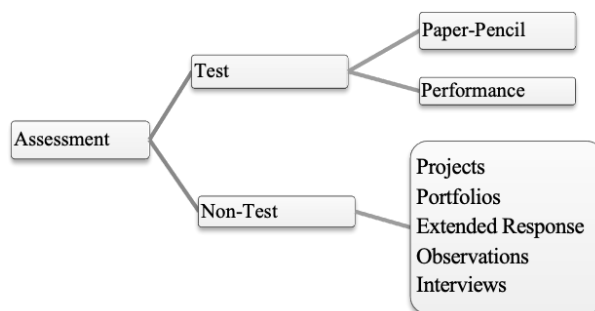

Figure 1. Scope of Assessment based on the Procedure (Brown, 2004)

From the explanation above, it can be concluded that there are some types of assessment to improve the understanding of what students know and understand, which can be used to consider before developing computer-based assessment. Numerous studies have attempted to explain the use of technology in testing. First, a study was conducted by Noyes and Garland (2008) that investigated whether computer and paper-based tasks are equivalent. A survey design was conducted by reviewing literature and research. In the research, it is indicated that in some cases, paper and computerized tests were equivalent, but in some cases, they were not, for example, the form of the test. In addition to this finding, the achievement of equivalence in computer-based and paper-based tasks poses a difficult problem. It is probably influenced by the test takers' confidence in using the computer and other psychological factors. The second study was conducted by Sofa and Sulistyo (2017), that developed online reading comprehension summative test for college students by using the research and development method. The research indicates that the product can be a model of an online reading summative test in the college level.

Based on the previous studies, the advantages and disadvantages of using computer as a testing media can be known. The advantages are, first, a lot of interfaces. For example, it has many dynamic presentation features as a test content. The second is the user population. Online Computer-Based Test (CBT) by using the Internet allows many people to participate in diverse areas, as long as they connect to the Internet. It also lets participants take part in the test from anywhere. Now comes the time for online scoring. These outcomes allowed quicker responses, improved accuracy, and fewer human mistakes. The quantity and quality of the mixture is the last consideration. According to a meta-analysis conducted over 10 years by Goldberg, Russell, and Cook (2003), computers outperformed paper in terms of both word count and quality of writing.

However, there are a few drawbacks to relying on computers for testing. The first has to do with the medium itself while also referring to a controlled environment in which reactions are made at different times and places, and maybe not even by the planned person (Noyes & Garland, 2008). When taking longer exams, working on a computer may be more tiring on the eyes than writing on paper. Some exam participants may not have much experience with computers. Hardware concerns are a consideration for computer-delivered content but not for paper-based exams. Displays on screens, for instance, must be standardized so that they're of the same quality for all test takers. Examinees who took the reading test on a big, high-resolution display outperformed those who took the test on a smaller, low-resolution monitor, according to research by Bridgeman, Lennon, and Jackenthal (2003). The main issue was that, although the complete passage was viewable on a high-resolution screen when answering questions, the entire text had to be scrolled through on a low-resolution device. The researcher may tell that the benefits are more difficult to get than the negatives by comparing the references above.

Meanwhile, educational research and development is needed to fulfill the gap caused by the transition of student learning and testing methods, from the conventional (Paper-Based Test) to computerized (Computer-Based Test), where the website can be useful as a support to develop a summative computer-based test for them by online test because it is user friendly. Thus, this study offers the importance of developing a summative test using a website as potential media that may be effective by providing a different package of test questions for tenth-grade students, to prepare for their English test graduation as a part of university entrance until job requirement in the future.

The research object here is the tenth-grade students of SMAN 20 Surabaya in Academic Year 2021/2022. The reason for selecting the object caused by the transition of education level, from Junior High School (SMP) to a higher level; Senior High School (SMA) required to develop an effective English summative assessment through Computer-Based Test, where students need to be familiar with such occasions before encountering a real English proficiency test modes in the future. Each student takes a test by getting the development, the different package questions through a website as their learning tool. Thus, a thesis based on

research and development entitled "Developing an English Summative Computer-Based Test for Tenth Grade Students" will be conducted. Based on the background of the research above, this research aims to develop an English summative Computer-Based Test (CBT) for tenth-grade students of SMAN 20 Surabaya and increase their students' achievement. The developed online test is used at the high school level for X grade as a summative test in the odd and even semesters.

Based on the research objective, the theoretical framework here is used to discuss and review relevant theories as to the basis of the study. This part discusses ideas related to the English summative Computer-Based Test (CBT) and Computer-Based Assessment (CBA). In CBT involves language test, achievement test, diagnostic test, aptitude test, and proficiency test variables. While CBA involves CBT model itself.

Summative means "additive-cumulative" or something that is added as an element to change or improve the quality of a general trait or to fight an undesirable trait. After each semester, teachers administer summative assessments to students to evaluate how much progress they've made toward learning objectives. The indicators included in the scope are meant to be representative of the semester-long standard competencies. Periodically, students are evaluated on their knowledge and skills via summative tests (Garrison & Ehringhaus, 2010:1). a lot is riding on the results of a Summative exam because of its high point value. Some argue that high-stakes testing is ineffective because it relies on the fallacy that "motivating the unmotivated" to learn can be achieved via the promise of rewards and the threat of (Orfield & Kornhaber, 2001).

End-of-year exams meant to demonstrate proficiency in a language are examples of accomplishment tests; these exams, also known as attainment or summative tests, examine how much a student has learned about a specific curriculum or set of lessons. In more pragmatic and pedagogical terms, an accomplishment exam is defined by Brown (1994) as "tests that are confined to certain topics presented in a curriculum within a given time period." In other words, their primary purpose is not to encourage or reinforce language but rather to assess individual growth. Teacher-made exams of student success should be used very seldom.

Specific language characteristics are the primary focus of a diagnostic exam. Pronunciation diagnostics might, for instance, reveal which phonological aspects of the English language are most likely to be problematic for a particular student body. Diagnostic exams should ideally assess students' linguistic knowledge (knowledge of and about the language) and linguistic abilities (hearing, speaking, reading, and writing) before beginning a course.

Students' performance in a language is evaluated using a language aptitude exam, which may be used for a second or foreign language. That's why it's employed: to zero in on students who have the best shot at making it. The Modern Language Aptitude Test (Carroll & Sapon, 1958) and the Pimsleur Language Aptitude Battery are two widely used standardized language aptitude assessments in the United States (Primsleur, 1966). These exams are in English and include memorization, counting, listening, seeing spelling hints and grammatical patterns, and identifying grammatical errors (Brown, 1994).

A language competency test is one way to evaluate a learner's progress. It is not tied to any curriculum but gauges students' overall linguistic competence. To gauge a test taker's ability to understand and use the language, most standardized exams focus on familiar, high-frequency vocabulary. The study conducted by Meyling et al. (2018) shows that in the English Proficiency Test (EPT), at least three main indicators can be analyzed to see achievement, namely listening, writing structure, or expression, and reading. Reading skills are the main problem faced by the first year of university students,

The widespread use of computer-based assessment (CBA) is being heralded as a revolutionary change. It's commonly known that language instruction has evolved thanks to technological advancements (Zhao & Liu, 2011) significantly. To meet the needs of today's pupils, educators are increasingly turning to computer-based assessment (CBA) tools, which may provide both theoretical and applied data. Unlike conventional classrooms, online learning environments provide students with a more laid-back and stress-free setting, which is likely the primary motivation. Because of the increased chances for group work afforded by CBA, students devote much more class time to collaborating than they did in teacher-centered classrooms. The pupils' increased confidence in their language-learning abilities is another striking effect of computer-based assessment (CBA). They believe that training their minds on the computer may rapidly advance their linguistic skills.

Two primary models for computerized-based assessments (CBT) are the linear or fixed model and the item-adaptive model. As described by Ejim (2017), the linear model is the most straightforward kind of CBT since it mimics the methodology behind the distribution of traditional paper examinations. Each learner is shown the same collection of things, either in the same sequence or randomly jumbled. Scores on fixed form CBTs are calculated similarly to traditional examinations, either by adding up the number of correct responses or using item response theory (IRT) techniques.

In line with the objective of the research, the final product of this research is a Computer-based Test (CBT) media for Senior High School students of SMAN 20 Surabaya with a focus on listening and reading tests. Two main aspects focused of this CBT development study. First, the test format to be used is in the form of multiple-choice. The importance of using the multiple-choice type of test as stated by Borg & Gall (1983) is that multiple-choice is generally more objective This is due to the existence of an automated machine system, which minimize human error and fraud.

The second is the layout. The layout in the Computer-Based Test (CBT) format is similar to Paper-Based Test (PBT), so the students will not be confused with the format. The layout also will be designed as close to the layout of SBMPTN (*Seleksi Bersama Masuk Perguruan Tinggi* Negeri), so the students will be familiar with the new test media. The media itself will be

designed for one question, one page. The media can be used in an online and offline format. If in an online form, the test will need *Hosting* and *Domain* to administer, and if in an offline form, the test will be conducted through *Localhost*. Based on the background of the study, the font used will be Verdana with 11-12 points size. For the text length, Heaton (1988) states that the appropriate text length for pre-intermediate and intermediate is approximately 250-300 words with with a maximum of 5 questions on each text.

The Computer-Based Test (CBT) layout can be done by creating a Unified Modeling Language (UML) model, which consists of Use Case diagrams, activity diagrams, and sequence diagrams (Tedyyana, Danuri, & Lidyawati, 2017). As shown in a previous study conducted by Sofa and Sulistyo (2017), when designing the application process, it is necessary to pay attention creating content specifications/blueprints before they are displayed to participants. The layout diagram will start with the following parts: conducting needs analysis, formulating the test specification/blueprint, blueprint expert review, item writing, item writing review, test installing, test and ICT expert review, try-out, item revision, and finally publishing the final form, where the questions shown in the questions bank is randomized. The CBT application program uses the PHP programming language through MySQL as its storage database to be run on a Client-Server PC. The media used in this research is an open-source program of freeware types, which is available at e-ujian.com. Meanwhile, in the listening test process, the recording sound will be played two times at normal speed.

The result of the research is expected to be an alternative test media for the students and teachers of SMAN 20 Surabaya. More specifically, the product of this research would be beneficial for the teachers to conduct a test with a sophisticated media. For students it is expected to make them familiar with the new test media. For the test developers, this research is expected to consider and reference in conducting further tests. For the head of the school, it is expected to provide reference to conduct similar tests.

The research only developed a prototype of an online summative test to measure the students' achievement at the end of each semester. The prototype will focus on tenth-grade students so they will be familiar with a new trend of test media, not only for the summative test but also for further tests like SBMPTN (*Seleksi Bersama Masuk Perguruan Tinggi Negeri*). The test will also be limited only to listening and reading skills that have been taught in the school. Distinguishing between individual sounds in English, picking up on English stress patterns, and identifying grammatical word classes, systems, practices, rules, and forms are all that the micro skill of hearing will cover. Recognizing the communicative roles of speech, inferring contexts, participants, and aims based on preliminary information, and differentiating between literal and implied meanings will all be outside the scope of the listener's macro listening ability.

METHOD

This study follows the Research and Development stages suggested by Borg and Gall (1983). They generally recommend 10 stages, from initial research and information gathering through final product modification and implementation. These stages include designing and constructing a prototype of the product as well as conducting field testing.

Meanwhile, the design of the test development here was adapted from Sofa and Sulistyo (2017), in which the original stages are as follows: conducting needs analysis, formulating the test specification/blueprint, item writing, item writing review, test installing, test and ICT expert review, try-out, item revision, and finally publishing the final form. The adaptation model was modified and developed, but still sufficient with the original test form because it was carried out the essence research and development itself. This research proposed additional aspects; try-out to determine the reliability, item difficulty, and discriminating power. The model of CBT-test development in this study is presented in the following chart.
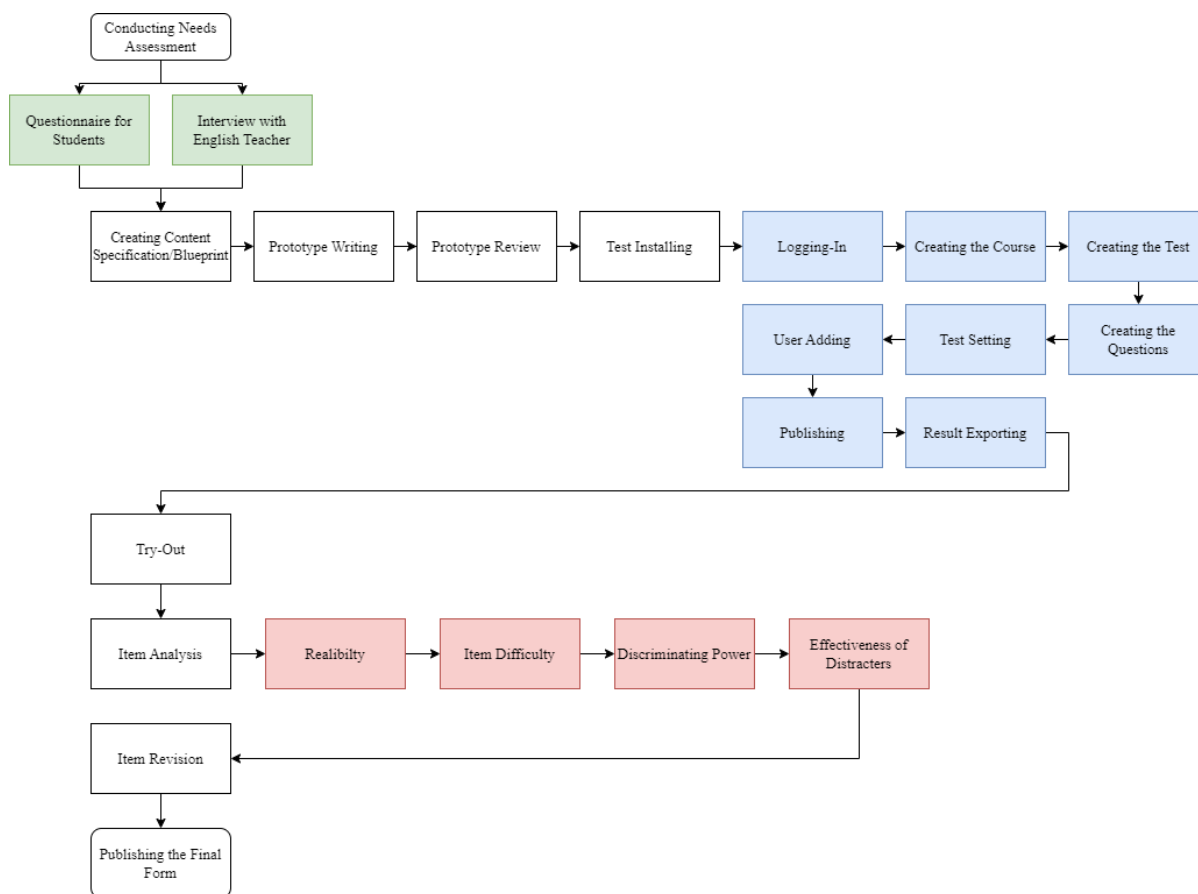
Figure 2. The Model of CBT Development (Adapted from Sofa & Sulistyo, 2017)

A needs analysis was conducted to identify senior high school students' problems when doing a CBT. The design of this assessment was qualitative and quantitative. The qualitative data was elaborated from the interview with the English teacher of SMAN 20 Surabaya. Before doing the interview, several questions were drafted as an interview guide to the English teacher of SMAN 20 Surabaya. The interview was about the procedure of conducting an English test in the school, the response of Computer-Based Test (CBT) in senior high school area, and the response about how students feel about this kind of new test media. The interview went for about 30 minutes. The interview guide detail is presented in the interview form.

The quantitative data was elaborated from the questionnaire with students of SMAN 20 Surabaya. The questionnaire is about the experience of students in Computer-Based Assessment (CBA), especially in English class. The goal of the questionnaire is to know general information about the students, students' experience in using computer, and students' experience in the use of computer as an assessment media, as presented in the questionnaire form.
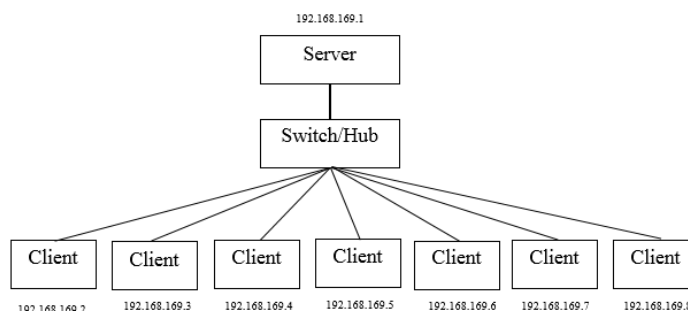
Figure 3. Topological Model of CBT Online Exam

Based on the chart above, the CBT online application program used here is the PHP programming language through MySQL as its storage database so that media can run on a Client-Server PC. The website provider used here is open-source programs of freeware types, which is available at e-ujian.com.

RESULTS

This section presents the research results from the process of validation stages (items expert and ICT expert), students' try-out and students' questionnaire. The data were presented both quantitatively and qualitatively. Past CBT data from the institution may be accessed at this point. As a first issue, some students suffer from a general fear of computers. However, for others, the degree of their computer literacy is of paramount importance. Second, the usage of the mouse, the clarity of the screen, the screen size, the quality of the display, the pace of the show, and the ability to scroll are all technical concerns in CBTs that might impact students' answers. Third, based on the interview with an English teacher, the training in Computer-Based Test is still needed. The teacher also said that it would be better if the government gave an opportunity to develop a test media based on the needs of each school so that the media can be effective to be used.

Furthermore, if Computer-Based Test (CBT) are used for a long term, it would be better if the students were given clear instructions about this new kind of test media because some students said that they still choose to use the traditional test media rather than Computer-Based Test due to the general anxiety of using computer and level of computer experience. So, it needs more support in managing the implementation of Computer-Based Test (CBT) to achieve the success of learning targets through best practices and facing emerging challenges.

The Results of Test Items Expert Validation

There were 3 specific aspects of the test items adapted from Fuhrman:1996. The aspects are material aspect, construction aspect, and language/culture aspect. From the test item expert validation, there were some results and feedback that suitable to be used in the try-out. From beginning, there were 70 items of English test, but after validation, 50 items are suitable to be used in the try-out. The validator said that for high school students, it was more suitable if the test has 50 items and it should be carried out within 120 minutes. If the students do it in one-time seat, it is worried that the subjects are tired and it will give effect to the result.

Other suggestion was about the length of the passage. The test should be ranged from easy to difficult. The validator suggested that it is better to arrange the questions in order of paragraph, so then the questions were re-arranged based on the paragraphs order. The length of each passage must be adjusted to the level of the students. Since it was used for high-school students, only passage with 350-500 words be used in try-out.

For the listening section, the validator suggested that the audio should be played twice with clear instructions. The time for listening section also must be considered, so it is not too short or too long. For the reading section, the validator suggested that the options should be written in order from short to long, so it did not give a hint for the students. For options that has numbers, the order must be from small numbers to big numbers and not in random order.

The Results of ICT Expert Validation

The second validation was ICT expert validation. There were 4 specific aspects of the ICT media development. The aspects are user interface, user control, layout/display, and feature. The checklist result was good but there were some suggestions in some aspects. The validator suggested that it should be better if the instruction for listening section is clear on each part (there were 4 listening parts). The audio in listening section should be played maximum twice, then the students cannot play the same audio again. The audio size should be considered more since if the test is conducted online, the internet speed of each student may result in different audio loading.

Other suggestions, the image in the test must clear and have the best resolution. In some questions, the images were too small and have low resolution, so it made the students hard to see the images. For passages that has 3-4 questions, the validator suggested that it should be in one page, so the students can read and answer the questions in the same page. The next is about the feature and interface. Since the test was used online, the validator suggested that it would be better if the test has anti-cheating feature like open-cam or face-cam to reduce cheating. The auto-save feature also should be considered if students have connection or gadget problem. Below is the product before and after the feedback from the ICT experts.
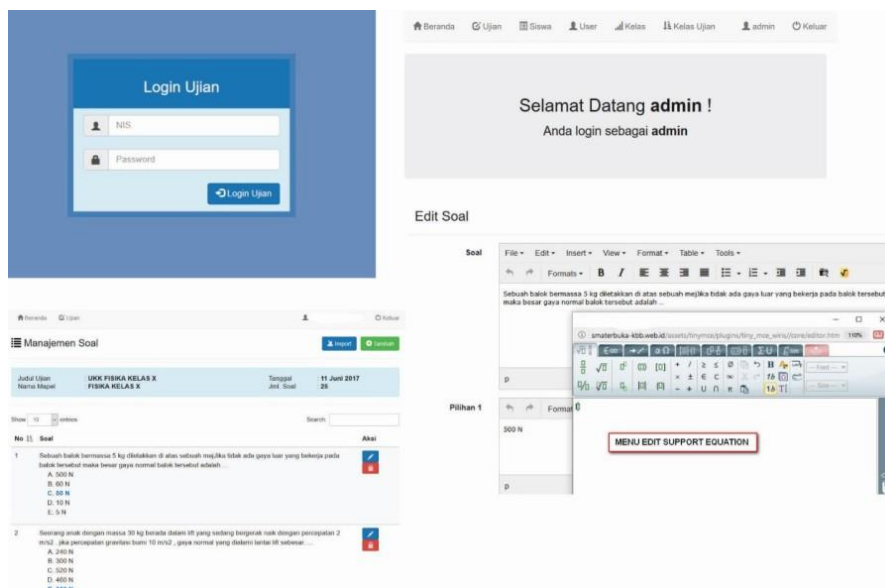
Figure 4. The Product Before the Feedback from Test Items and ICT Experts

The Results of Students' Try-Out

The try-out was carried out on July 12 and 13, 2022 to 34 students of SMAN 20 Surabaya. The item revision was used *Anates* 4.0 software that covers reliability, level of difficulty, item discrimination, and effectiveness of distractors.

The first step in the States analysis is to determine whether or not the test can be trusted. Minimally acceptable levels for test-score reliability must be defined in the context of scores that were utilized. According to Ebel and Frisbie (1991:86), "Experts in educational measurement have agreed informally that the reliability coefficient should be at least 0.85 if the scores are the only available useful information."

The reliability test score of English summative test for tenth grade students of SMAN 20 Surabaya was 0.94. It is categorized as acceptable and fair reliability because the standard reliability of multiple-choice items is 0.94. The next analysis was the level of difficulty. Level of difficulty is to examine the test questions in terms of test difficulty so can be obtained which questions are including, easy, moderate, and difficult questions. From the result, it can be seen that the English summative test for tenth grade students of SMAN 20 Surabaya has moderate arrangement of level of difficulty since it has 60% of the test that has good difficulty level. The result is shown in the Table 1.

Table 1. The Result of Level of Difficulty Analysis

No	Level of Difficulty	Items Number	Total	Percentage
1	0.00 - 0.25 (Difficult)	6, 7, 13, 14, 15, 28, 29, 30, 32, 34, 41, 42, 45, 49, 50	15	30%
2	0.26 - 0.70 (Moderate)	1, 2, 3, 5, 8, 11, 12, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 31, 33, 35, 36, 37, 38, 39, 40, 44, 47, 48	30	60%
3	0.71 - 1.00 (Easy)	4, 9, 10, 43, 46	5	10%

Based on the result shown in the Table 1, there are 5 easy items, 30 moderate items, and 15 difficult items. In order to know how good the item in discriminating the low and high ability students, the analysis of item discrimination was carried out. From the *Anates*, the result is presented in the Table 2.

Table 2. The Result of Item Discrimination Analysis

No	Item Discrimination	Items Number	Total	Percentage
1	Negative	7, 12, 42, 50	4	8%
2	Non-Discriminating	4, 14, 23, 34, 41	5	10%
3	Poorly-Discriminating	3, 5, 6, 13, 16, 18, 19, 24, 28, 32, 33	11	22%
4	Fairly-Discriminating	1, 2, 8, 9, 10, 11, 15, 17, 22, 25, 26, 27, 29, 30, 35, 38, 45, 46, 47, 48, 49	21	42%
5	Discriminating	20, 21, 31, 36, 37, 39, 40, 43, 44	9	18%

Ebel and Frisbie (1986) state that the smaller the sample of answer sheets used in the analysis, the larger the sampling errors. An item that appears highly discriminating in one small sample may appear weak or even negative in discrimination in another small sample. The values obtained for achievement-test items are also sensitive to the kind of instruction the students received relative to the item hence the use of refined statistics to measure item discrimination seldom seems to be warranted. The criteria of item discrimination are if negative (-) there is no discrimination, 0.00 – 0.20 also categorized as non-discriminating and should be discarded, 0.21 – 0.40 categorized as poorly discriminating and should be revised, 0.41 – 0.70 categorized as fairly discriminating and should be accepted with revision, and 0.71 – 1.00 categorized as discriminating and should be accepted.

English summative test for tenth grade students of SMAN 20 Surabaya has good quality because 82% of the items have good item discrimination. Then, items that already have good item discrimination (fairly discriminating and discriminating) should be included in items bank and items that have bad item discrimination still can be revised. For items that have negative item discrimination, preferably no longer used for test. The last analysis was the effectiveness of distractors. Summative test items using multiple choice, where the students choose the answers based on alternative options that already reserved. From 5 options, there is only one correct answer and 4 options are distracters. The answers pattern obtained by summing the number of students who choose answer a, b, c, d, or e and students who are not answered the questions.

Distracters can be said effective if the distracters are chosen by at least 5% of the test-takers. There are 34 students as test-takers, so the distracters are effective if 2 students choose the distracters. Based on the data from *Anates* result analysis, the test has good distracters because more than 80% of the options are effective. The making of items needs to consider the items difficulty with alternative answers. Multiple choice items that compiled without looking the homogeneity have a chance not to be functioned. Because the students are easily answer the questions. And also, items that have hints on the questions is also have a chance not to be functioned because the students are directly choose the hint answers and not looking the distracters.
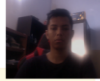
All the stages in conducting Research and Development have been carried out and 41 items remained after the analysis and put in the items bank of the product. The final product of this research covers the instructional media that includes the website, tutorial, blueprint, the English test in online form, and also screenshot of the try-out documentation. The final website of the product can be accessed in www.e-ujian.com/cebetein.

DISCUSSION

After some feedbacks and suggestions from test items experts and ICT experts, there are some adjustments and improvements for the product. The main adjustment and improvement are from the format of the test that follow the feedback from test items expert like the total of the items in the test, the text length, the time, and the options. From the media itself, the main adjustment and improvement are from the features like anti-cheating, face camera, live monitoring, and instant review. The newest update for the product, the test-maker and test-taker can use mobile application to reduce the limitation from the website version like cheating, lagging, and network problem. For now, test-maker and test-taker can download the mobile application only for Android version and can be downloaded from Google Play Store. For the product after the feedback, adjustment, and improvement, you can see from the figure 5.
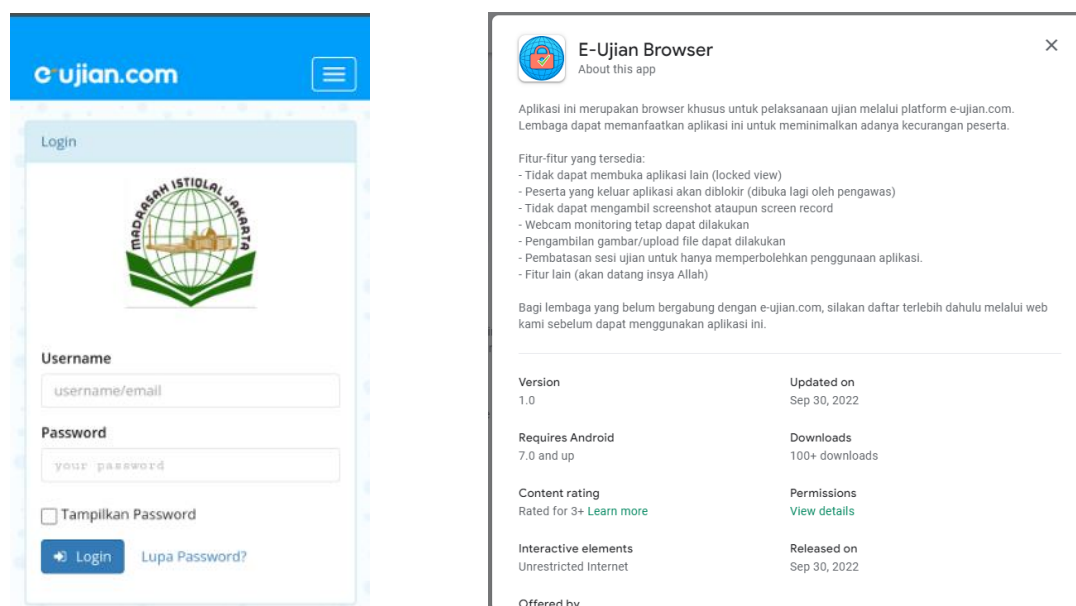
Figure 5. The Product After the Feedback from Test Items and ICT Experts

CONCLUSIONS

The conclusions were based on the whole process of the research and also the result of it, while the recommendations are there for the future perfection of the products. Based on the process and the result in developing this English Computer-Based Test, there are two domains of conclusions that can be drawn. The conclusions comprise the strengths and also the weaknesses of the product of the research.

Related to the strengths, first, the product of this research can be a model of English test as well as for other subjects in SMAN 20 Surabaya. Since it has many features that can cover the difficulties of conducting online test such as timer, anti-cheating, and other features that always be a challenge in online test. The product of this research is an online test which is not time consuming in administration as the previous test. This Computer-Based Test is also practical and cost effective since the maintenance of the product is still reasonably and the user do not need to use the paper time by time the test is used. Then it is also time effective since the product has the feature to automatically process the score after the students finish their test. This product is also a good model of the test in this 4.0 era.

As the product has strengths, it also has weaknesses. The final product of this test only consists of small group of students. Some of the features did not really applicable like the audio of the listening section and also some bugs in the product. Based on the strengths and the weaknesses of the product, it can be concluded that the developed Computer-Based Test can be used as a test model which replace the previous conventional, paper-based test, impractical, time and cost consuming.

Some suggestions are presented after completing the whole processes in conducting this research for future perfection. To SMAN 20 Surabaya, this online test can be a model for not only English test, but also other tests in other subjects. Especially in 4.0 era where paper-based test is no longer used and replaced with this kind of Computer-Based Test. As the maintenance of the online test is not as expensive as paper-based test, the online test can reduce the time consuming and cost consuming. The features of the product also can be adjusted based on the test-maker's need. As this online test needs high quality network and server if it is use larger of test-takers, it is hoped that the institution can provide larger and also stable network and server to reduce the trouble.

To the future researchers, this product can be an insight for the effectiveness of online test and can be adjusted based on what the test-maker's need. As this research had limited subject (only a class or about 30 subjects), it is suggested that the future researcher can have larger subjects to gain more reliable and valid result. Since the feature are not only for English test (listening and reading section), it is suggested that the future researcher can develop for other subjects like mathematic (using formula or numbers) or other language test like Arabic.

Further, although low, but as this test still open the chance for the students to cheat, so the researcher should work on the test in the same place and time between test-maker and test-takers. This attempt is hoped to not only reduce the cheating action but also increase the students' independence and self-esteem.

REFERENCES

Arikunto. (2013). *Prosedur Penelitian: Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.

Blackburn, M., Alexander, J., Legan, J. D., & Klabjan, D. (2017). Big Data and the Future of R&D Management. *Research-Technology Management*, *60*(5), 43–51. https://doi.org/10.1080/08956308.2017.1348135.

Borg, W. R., & Gall, M. D. (1989). Educational Research. *An Introduction*. New York: Longman. https://doi.org/10.1604/9780801303340.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. *Applied Measurement in Education*, *16*(3), 191–205. https://doi.org/10.1207/s15324818ame1603.

Brown, H. D., & Abeywickrama, P. (2010). Language Assessment: *Principles and Classroom Practices*. New York: White Plains.

Bull, S., Quigley, S., & Mabbott, A. (2006). Computer-based formative assessment to promote reflection and learner autonomy. *Engineering Education*, *1*(1), 8–18. https://doi.org/10.11120/ened.2006.01010008.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The Four Generations of Computerized Educational Measurement. *ETS Research Report Series*, *1988*(1), i–148. https://doi.org/10.1002/j.2330-8516.1988.tb00291.x.

Djiwandono, M. S. (1996). *Tes bahasa dalam pengajaran*. Bandung: ITB Bandung.

Fasold, R. W., & Connor-Linton, J. (2014). *An Introduction to Language and Linguistics*. UK: Cambridge University Press.

Fuhrman, M. (1996). Developing Good Multiple-Choice Tests and Test Questions. *Journal of Geoscience Education*, *44*(4), 379–384. https://doi.org/10.5408/1089-9995-44.4.379.

Gronlund, N., & Waugh, C. (2013). *Assessment of Student Achievement*. Pearson.

Haryani. (2018). The Use of Computer-Based Assessment (CBA) in Measuring Students' Achievement Test. *English Language and Literature International Conference (ELLiC) Proceedings*, *2*, 231–234.

Indonesian Ministry of Education and Culture. (2020). *Circular Letter of Merdeka Belajar*, (Online), *www.kemdikbud.go.id*, accessed on June 11th, 2020.

Kurt, S. (2012). Examining teachers' use of computer-based technologies: A case study. *Education and Information Technologies*, *18*(4), 557–570. https://doi.org/10.1007/s10639-012-9199-7.

Latief, M. A. (2012). *Research Methods on Language Learning: An Introduction*. Malang: UM Press.

Mukhallafi. (2014). *Computer Assisted Language Learning for Learning English in Saudi Arabia*. Sydney: University of Technology.

Ndun, L.N. (2021). The Implementation of Computer Based Test (Cbt) As Final Assesment and Its Relevance to The Teaching and Learning Process (A Case of Students and Teachers in SMA Negeri 1 Soe). *Wiralodra English Journal*, *5*(2), 15–27. https://doi.org/10.31943/wej.v5i2.132.

Nikou, S. A., & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior*, *55*(Part B), 1241-1248. https://doi.org/10.1016/j.chb.2015.09.025.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375. https://doi.org/10.1080/00140130802170387.

Russell, M., Bebell, D., O'Dwyer, L., & O'Connor, K. (2003). Examining Teacher Technology Use. *Journal of Teacher Education*, *54*(4), 297–310. https://doi.org/10.1177/0022487103255985.

Scheuermann, F. & Pereira, G. (2008). Towards a Research Agenda on Computer-based Assessment - Challenges and Needs for European Educational Measurement. *JRC Scientific and Technical Report*.

Setyaningsih. (2014). *Senior High School Teachers and Students' Utilization of and Attitudes Towards Computer-Assisted Language Learning and Their Awareness of Students' Learning Autonomy*.

Sofa, S., & Sulistyo, G. H. (2017). A Model of an Online Reading Comprehension Summative Test for College Students. *IJEE (Indonesian Journal of English Education)*, *4*(2), 168–187. https://doi.org/10.15408/ijee.v4i2.8344.

Sulistyo, H. (2015). *Assessment at Schools: An Introduction to its Basic Concepts and Principles*. Malang: CV. Bintang Sejahtera.

Talebinezhad, M. R., & Abarghoui, M. A. (2013). The Iranian High School Students' Attitude toward CALL and the Use of CALL for EFL Receptive Skills. *Theory and Practice in Language Studies*, *3*(2). https://doi.org/10.4304/tpls.3.2.329-337.

Tedyyana, A., Danuri, & Lidyawati. (2017). Design of Computer Based Test Using the Unified Modeling Language. *IOP Conference Series: Earth and Environmental Science*, *97*, 012006. https://doi.org/10.1088/1755-1315/97/1/012006.

UNESCO (*United Nations Educational, Scientific and Cultural Organization). 2020. Covid-19 Education Issue Notes, (Online), en.unesco.org, accessed on June 10th, 2020*.

Verderber, R. F. (1998). *Speech for Effective Communication*. New York: Holt Rinehart and Winston. https://doi.org/10.1604/9780030975257

Winke, & Fei. (2018). Computer-assisted language assessment. *Encyclopedia of Language and Education*, *4*.

World Bank Group Education. (2020). *Educational Policies in the Covid-19 Pandemic: What can Brazil Learn from the Rest of the World?*, (Online), *www.worldbank.org*, accessed on June 10th, 2020.

Yuyun, I., Meyling, M., Laksana, N. L., & Abenedgo, D. (2018). A Study of English Proficiency Test Among the First Year University Students. *Journal of Language and Literature*, *18*(1), 1–8. https://doi.org/10.24071/joll.2018.180101.