SMU Journal of Undergraduate Research

Volume 7 | Issue 2 Article 3

May 2022

Conceptualization of an Assessment System to Measure Vocabulary in Science (Project MELVA-S)

Sisi Kang
sisik@smu.edu

Recommended Citation
Kang, Sisi (2022) "Conceptualization of an Assessment System to Measure Vocabulary in Science (Project MELVA-S)," *SMU Journal of Undergraduate Research*: Vol. 7: Iss. 2, Article 3. DOI: https://doi.org/10.25172/jour.7.2.2
Available at: https://scholar.smu.edu/jour/vol7/iss2/3

Conceptualization of an Assessment System to Measure Vocabulary in Science (Project MELVA-S)

Sisi Kang

sisik@smu.edu

Mentor: Dr. Eric C. Larson[1]

ABSTRACT

This paper aims to report the conceptualization of a web-based Automated Speech Recognition Scoring System, project MELVA-S (Measuring the English Language Vocabulary Acquisition of Latinx Bilingual Students), to measure the science vocabulary of second- and third-grade Latinx students. ELVA (English Learner Vocabulary Acquisition First Iteration) and ELVA-2 (English Learner Vocabulary Acquisition Second Iteration) focused on student's learning and comprehension on science vocabularies. Both of the iterations are the foundation to build MELVA-S, which intends to measure and evaluate student's answers with greater accuracy with Machine Learning. As a web-based agent, this system increases satisfaction for both teachers' and students' User Experience (UX) from content, design, and engineering perspectives. The project utilized a design-thinking approach and prototyped both the algorithm and the automated system interfaces. Future iterations of ELVA-2 and MELVA-S could consider adopting a Human-Centered Machine Learning approach, implemented with incremental improvements that include evaluation and testing with users, to keep enhancing both usability and functionality of the system for better UX.

1. INTRODUCTION

A. Background of ELVA, ELVA-2, and MELVA-S

There has been a rising trend in machine learning and Child-Computer Interaction. It is noticeable that the English learner population, and specifically, the Spanish-speaking bilingual student population, has vastly increased over the last decade in the U.S. [1] Considering high risks of young Hispanic learners dropping out of school due to failure to learn effectively[1] and the iterative nature of technology, systems must have a cost-effective, efficient, minimally invasive tool to enhance and satisfy the ever-changing student and teacher needs.

ELVA (English Learner Vocabulary Acquisition First Iteration) was developed at SMU in 2014. The purpose was to promote and increase the vocabulary knowledge, text comprehension and English language proficiency of Spanish-speaking English Learners (ELs) in second grade. The application developed taught students scientific terminologies and assessed whether they understood them. Researchers designed this intelligent tutoring system where student audible interactions were recorded since it is advantageous for engaging second-grade Latinx students in describing and explaining phenomena using expanded comprehension from students' existing vocabularies [1]. This assessment involved asking students to accurately synthesize and utilize the words by speaking in front of a camera. Previous work has shown that by involving video and voice recordings, second-grade students participate more, practice well, can be assessed with individualized modules, and achieve high levels of enjoyment compared to typical content instruction [1].

ELVA-2 (English Learner Vocabulary Acquisition Second Iteration) was the extention of ELVA, promoting monolingual and bilingual second-grade students' scientific vocabulary and language proficiency. Specifically, ELVA-2 focused on testing a science vocabulary intervention delivered to second-grade students through the Canvas Learning Management System.

In addition to developing a science vocabulary program that supports student learning of science, MELVA-S intend to build a scoring system that automatically scores student answers to questions. This approach would help teachers understand the language and vocabulary students need to understand science and reduce the time teachers spend giving and grading tests. Thus MELVA-S incorporates the same concept with ELVA and ELVA-2 and evaluate students through the automated scoring system in the program. To accomplish this, project MELVA-S, a new web-based product, will include a speech recognition system and a user-friendly computerized scoring system for the target users [2].

The project is based on construction learning theory, which acknowledges that students can become participants in co-creating their learning journeys [16]. Their knowledge bases cannot be obtained by just listening and watching, but also by involving their overall educational experiences through interactive learning. ELVA-2 embraces construction learning theory by utilizing technology to create an embodied reality. Students can learn through audio interaction; they talk to the interface about scientific word definitions, and provide example sentences using the words learned through the system.

[1] Dr. Larson is an Associate Professor in Computer Science in SMU's Lyle School of Engineering.

B. The MELVA-S System as a Web-based Product

Technology has been advancing at an exponential rate; children grow up in an environment that is immersed in technology, such that they are considered digital natives and frequent users and owners of electronic devices. Research involving Child-Computer Interaction, including its design and methodology, according to Read and Markopoulos, is in urgent need of scientific investigation [3].

There are three main perspectives to consider in conceptualizing, designing, and developing a user-friendly product for both teachers and students. First, the content needs to be essential for teachers to teach and students to learn. Second, the usability and aesthetic aspects of the system need to be taken into account. Third, the application's technical facets and functionality need to be considered [14].

Seymour Papert, the first researcher who inspired Child-Computer Interaction, suggested that constructionist learning theory emphasizes the importance of children becoming authors and creators [4]. This seamlessly resonates with the concept of ELVA-2 and MELVA-S, which is to assess students' understanding and knowledge of a solid vocabulary as a building block of English Proficiency [1][2].

Next, the design must nurture children's critical thinking skills [5]. Building a better UX is more than having improved usability or simply creating more aesthetically pleasing user interfaces. It entails the social, cultural, and emotional experiences that are rewarding, fun, and aesthetically satisfying to use [6].

According to Evanini et al. [8], use of automated scoring technology within K-12 English Language Proficiency assessment has surged in recent years while limitations remain. Functionality-related issues such as the timeliness, quantity, and quality of feedback received from the students can be challenging [9].

This paper provides the conceptualization, design, and development of an automated scoring system that can support and expand the capabilities of ELVA-2 and MELVA-S. To build on these projects, I will consider the content, engineering, and design of ELVA and MELVA-S. I also include ideas on how to test the feasibility and usability of the programs.

2. METHODOLOGY

A User-Centered Design Thinking approach [15] was chosen to put user needs first and pull the constraints for automated scoring from the field of Child-Computer Interaction. Design thinking is a human-centric process that embraces a user-centered framework. Specifically, designers must first understand the stakeholders' requirements, the users, and their problems by empathizing, defining, refining, and only then exploring what could be done by ideation and prototyping. The last step is materialization by testing, implementing, and re-iterating for the best result.

From the empathy component of Design Thinking, it is crucial to understand (1) the initial problem ELVA-2 is solving, (2) the target users, and (3) the functions and features the design needs. Having better UX ensures both students and teachers have fewer distractions when using the system, which improves on the entire experience. Understanding users so as to empathize with them is the first step to find the root cause and pain points from the users' perspectives. While some background information of the program was shared through stakeholders, the team also evaluated various videos students took from two learning modules, about Rocks and Volcanoes, on Canvas.

The previous program, ELVA, utilized a software tutoring system, allowing students to practice scientific terminology using video and voice recording. However, this raises concerns about accessibility because different system requirements for software created a burden for teachers in assessing each student; it is far easier to assess using the form of multiple-choice only questions. Furthermore, students also do not have immediate feedback on the answers. Feedback is vital to enhance child engagement and satisfaction [3]. To solve the problem of immediate feedback, we explored the reliability and feasibility of using an automated scoring system.

Projects ELVA-2 and MELVA-S can also be conceptualized through Debra Levin Gelman's theories, principles, and real-life practices [5]. One key idea includes child communication through play. Designers are responsible for understanding the users so that they are able to complete a task with ease. We know that children love visual and audio feedback [5]; utilizing and including these features in the MELVA-S system will help to make children's experience more fun and interactive for better involvement. In this way, scoring and evaluating can be more effective and practical for validating their answers.

Five fundamental methodology principles of voice user interface design are utilized, including end-user input, integrated business, user needs, conversational designing, and context [10]. Tools used for prototyping and wireframing include Balsamiq and Adobe Creative Cloud.

The Automated Speech Recognition System for MELVA-S intends to utilize the following to recognize, examine, and score responses: ELVA research project secondary data and content, speech-to-text technology, services on Amazon Web Services, and transcribed audio files that are generated by existing students' recordings. To ensure that both ELVA-2 and MELVA-S are accessible, cost-effective, and practical learning tools, the Automated Speech Recognition System will be a web-based application. Its feasibility and reliability will be assessed by surveying students and teachers after the system is developed [2].

We started by assessing a total of 906 student video entries. The video entries were converted to transcripts and JSON files (Figure 1). All the data obtained was from two Earth Science modules: Rocks and Volcanoes.

The questions were categorized into three general types from the student entry data. The first type is open questions, for which the answers could vary and are very hard to predict. Questions include, "*Can you think of a school event that could be released in the news?*" and, "*Think of a sentence where you can use the word heat as a verb.*" The second type is semi-open questions, including, "*What do you know about volcanoes?*" and, "*What does erupt mean?*". ELVA-2 already has a set of expected answers provided for these questions. The last category is repeating questions; some examples include reading a word aloud, repeating a paragraph, and spelling a word's syllables.

Finding an excellent algorithm to match the keywords is crucial because it is essential to improve scoring accuracy.

3. RESULT

The ideation and prototyping phase evaluated student answer inputs using the handcrafted feature, matching keywords of the answers manually. A high-level algorithm was designed accordingly (Appendix B). First, the software will overlook the video input and focus on the audio input. If the audio entry is valid, the software will proceed to Amazon S3, utilizing the API through AWS Transcribe. Second, based on the three types of scoring, the software gives users real-time feedback; if a question type is undecided, then a manual evaluation will be required and queued to the teacher's system.

To further make this automated scoring system viable, feasible, and desirable, the result of correctness is compared using a confusion matrix (Figure 2). This matrix contains information about actual and predicted classifications in terms of attribute selection in these existing data [12]. The ground truth results are the data that has been manually evaluated, which can be compared with the text-to-speech results. Another way to improve the scoring model is to manually modify keywords for more accurate outcomes based on the performance and data over time.

The two figures below show two different ways of presenting the accuracy values of the current data. True positive indicates the scoring model positively predicted the correct answer; false positive represents data that should not be true, but the scoring model incorrectly predicted the response toward the positive class; false negative means the incorrect answers were not correctly predicted; and true negative indicates the data accurately predicted the wrong answers.

After utilizing AWS (Amazon Web Services) to transcribe without a custom setting, which is the 'before' result. There were 29 true negatives, 51 true positives, 28 false negatives, and 13 false positives.

The 'after' results using custom vocabularies indicate slight improvements. There were no changes to true negatives. However, true positives increased by 7, false positives decreased to 12, and false negatives decreased by 6.

After two iterations of running the confusion matrix model, improvements in the automated scoring procedure are shown in Figure 3. To indicate a better comparison algorithm, two different swarm plots were established to showcase the number of words in the students' response that overlap with the ideal answer, which is the ground truth for each question.

This figure has two sections: 'before' and 'after.' The 'before' section on top indicates the first iteration results performed manually (intersection_manual) and through AWS (intersection_AWS) before the automated scoring. The 'after' section on the bottom indicates the second iteration results and the associated improvements.

Correct answers are displayed in blue and incorrect answers are in yellow. The left side (Manual transcription) is the ground truth result, with the student answered evaluation by the teacher manually. This is the target result for the most accuracy. The right-side plots are the AWS Transcribe results (AWS Transcription), with the algorithm consistently overlapping with the ground truth. The more identical manual transcription and AWS graphs look, the more accurate the result is.

Besides the speech recognition from the engineering perspective, a product vision document was developed to communicate better and identify precise requirements and objectives (Appendix A). Low-fidelity sketches, mockups, and three versions of typical and high-fidelity prototypes were re-iterated and discussed among different stakeholders below.

4. DISCUSSION

The preliminary findings indicate that an automated speech recognition scoring system to enhance Child-Computer Interaction has potential and should improve over time as a web-based agent. It is reliable for measuring and assessing ELVA-2 (English Learner Vocabulary Acquisition) and MELVA-S students. The system can help increase satisfaction related to both teachers' and students' UX.

A frequent challenge in developing computer-based tools and technologies is the design of effective user interfaces. The MELVA-S system design focuses on one primary task, which is the voice interface, to ensure good usability. Weinerth stated that a good usability interface is easy to understand, easy to use, and easy to learn, making a system that enables the user to focus on the primary task of the assessment [13]. Similarly, Harms and Adams also mentioned that a central requirement for computer-based assessment delivery tools is ease-of-use with minimal training [14]. These concepts have been incorporated during the design process. However, UX, at its core, is about empathizing with the users.

From an eight- to ten-year-old student's perspective, the following are desired characteristics of programs: easy navigation, a button or other specified place to press a key to start recording, and ensuring that the project is saved.

From a teacher's or administrator's perspective, desired characteristics are an automated scoring system and centralized management system, including customized reports of individual students. Another desired feature is that the computerized system accurately adjusts to unique teacher curricula over time.

From the functionality aspect, the Automated Scoring System indicates better comparison algorithms would provide more margin between the ground truth and the AWS transcribe results. Incorporating students' input to improve automated scoring over time based on the constructionism theory will assist teachers' and administrators' manual grading process. One constraint to consider is the accuracy perspective with a better outlook in future iterations of the system development, which is currently in progress. Another constraint is the nature of users being children; quite clearly, they will sometimes not answer questions like adults.

However, adults and children share some similarities in interaction design. They demand consistency, knowing the purpose, not being surprised, and *lagniappe*, meaning a little 'add on easter egg' such as an animation. Children between the ages of eight and ten want to be

considered authorities on their favorite topics. They would like to share content and videos with their friends. When it comes to interaction design, specifically for children ages eight through ten, they are extremely curious and able to think logically about concrete ideas or events while failing to realize abstract concepts. [5] This means they do not like to read descriptions as much as six- through eight-year-olds, and they attempt to skip instruction material whenever they can. Providing feedback or motivation for this age group proved to be extremely helpful [5].

Elspeth McKay suggested that the automated scoring process is complicated and will remain a goal for education assessment for immediate feedback [11]. The challenge remains that speaking in a spontaneous context is difficult to assess unless the task is more restricted [8]. This, in turn, contradicts the educator's purpose to allow children to express themselves freely for encouragement and a sense of motivation. Therefore, some future implications could include inviting the second and third-grade students as stakeholders to participate and design the system or interfaces, incorporating them into the subsequent development iterations with evaluation and testing of the content program, usability, and functionality.

In conclusion, this project has great potential to make incremental improvements utilizing the Human-Centered Machine Learning approach. With opportunities to test and validate second and third-grade students, fine-tuning the output of the transcriptions will assist greatly in getting the system with better accuracy by combining UX with Machine Learning Automatic Scoring System. In this way, the project can adopt an Agile process that will make incremental success to keep enhancing both students and teachers with better experiences.

5. ACKNOWLEDGEMENTS

First and foremost I am extremely grateful to Prof. Doris Luft Baker and Dr. Eric Larson for their invaluable feedback, continuous support, and patience. Their immense knowledge has encouraged me to continue my learning journey. I am indebted to SMU Ph.D. student Yihao Wang, who helped me find my footing in Machine Learning and provided advice for this paper. I am also fortunate to have been a part of the Engaged Learning and Grand Scholars Program. Thank you, Adam Scott Neal and Beryl Hellinghausen for the opportunity and invaluable assistance during my time in the program. Finally, I would like to express my gratitude to my parents. Without their unconditional support and encouragement, it would be impossible for me to complete my study.

6. REFERENCES

[1] Baker, D.L., Otaiba, S. A., Cole, R., Ward, w., (2014). English Learner Vocabulary Acquisition (ELVA): Promoting the Vocabulary and Language Proficiency of Spanish Speaking English Learners in Second Grade.

[2] Baker, D. L., Ma, H., Polanco, P., Conry, J. M., Kamata, A., Al Otaiba, S., Ward, W., & Cole, R. (2020). Development and promise of a vocabulary intelligent tutoring system for second-grade Latinx English learners. *Journal of Research on Technology in Education*. Advance online publication. https://doi.org/10.1080/15391523.2020.1762519.

[3] Read, J. C., & Markopoulos, P. (2013). Child–computer interaction. *International Journal of Child-Computer Interaction*, 1(1), 2–6. https://doi.org/10.1016/j.ijcci.2012.09.001.

[4] Papert, S., (1988). The conservation of Piaget: The computer as a grist to the constructivist mill. In Forman, G. & Pufall, P., (Eds.), *Constructivism in the computer age* (pp. 3-13). Hillsdale, NJ.

[5] Debra Levin Gelman. 2014. Design for kids: digital products for playing and learning, Brooklyn, NY: Rosenfeld Media.

[6] H.Rex Hartson and Pardha S. Pyla. 2018. The UX book: process and guidelines for ensuring a quality user experience, Amsterdam: Morgan Kaufmann.

[7] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. 2009. Understanding, scoping and defining user experience: a survey approach. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). *Association for Computing Machinery*, New York, NY, USA, 719–728. DOI: https://doi.org/10.1145/1518701.1518813.

[8] Evanini, K., Hauck, M. C., & Hakuta, K. (2017). Approaches to Automated Scoring of Speaking for K-12 English Language Proficiency Assessments. *ETS Research Report* Series, 2017(1), 1–11. https://doi.org/10.1002/ets2.12147.

[9] Crook, A., Mauchline, A., Maw, S., Lawson, C., Drinkwater, R., Lundqvist, K., … Park, J. (2012). The use of video technology for providing feedback to students: Can it enhance the feedback experience for staff and students? *Computers & Education*, 58(1), 386–396. https://doi.org/10.1016/j.compedu.2011.08.025.

[10] Michael H. Cohen, James P. Giangola, and Jennifer Balogh. 2004. Voice User Interface Design. Addison Wesley Longman Publishing Co., Inc., USA.

[11] Elspeth McKay. 2013. EPedagogy in online learning: new developments in web mediated human computer interaction, Hershey, PA: Information Science Reference.

[12] Kohavi, R., and Provost, F. 1998. On Applied Research in Machine Learning. In Editorial for the Special Issue on *Applications of Machine Learning and the Knowledge Discovery Process*, Columbia University, New York, volume 30.

[13] Weinerth, K. (2015). How does usability improve computer-based knowledge assessment? Doctoral thesis, 2015, Luxembourg. Supervisors: Martin R., Koenig V. and Brunner M.

[14] Niamh McNamara and Jurek Kirakowski. 2006. Functionality, usability, and user experience. Interactions 13, 6 (2006), 26–28. DOI: http://dx.doi.org/10.1145/1167948.1167972.

[15] Hasso Plattner, Christoph Meinel, and Larry J. Leifer. 2011. *Design thinking: understand, improve, apply*, Heidelberg: Springer.

[16] N. Brooks, Language and language learning: Theory and practice. 1964.

7. APPENDIX 1: FIGURES AND TABLES

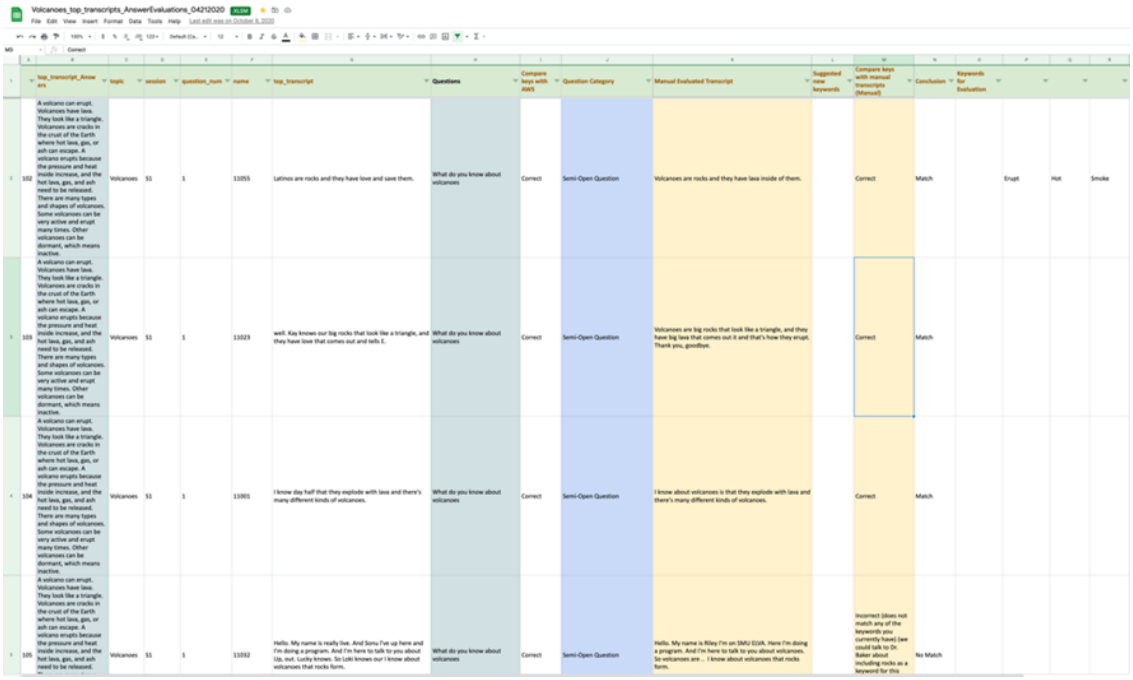

Figure 1. Excel sheet example of the raw data input of the ELVA-2 scoring system.

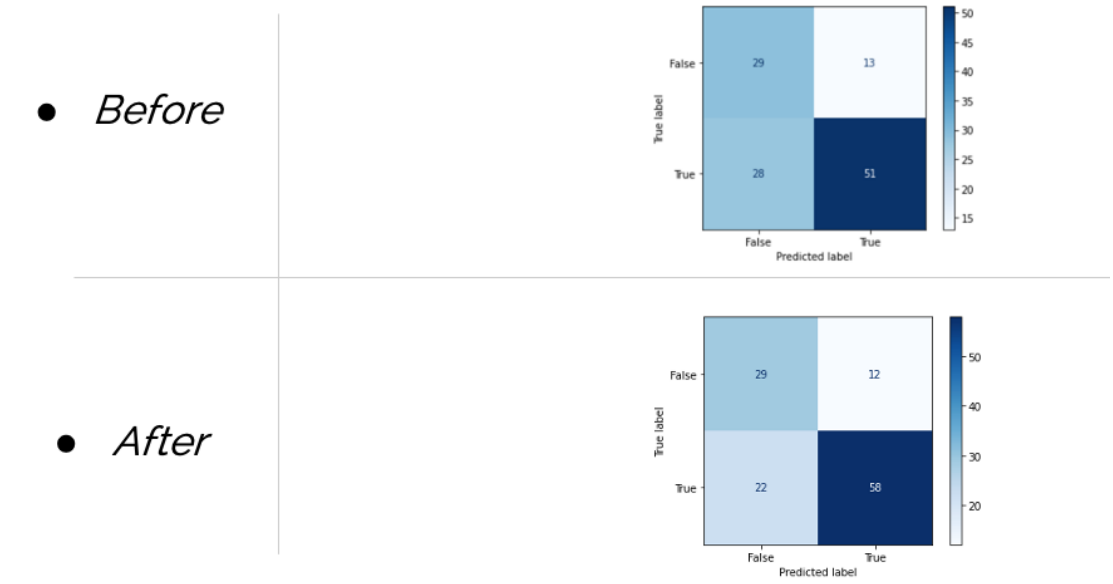

Figure 2. Confusion matrix indicating the accuracy values of the scoring model.

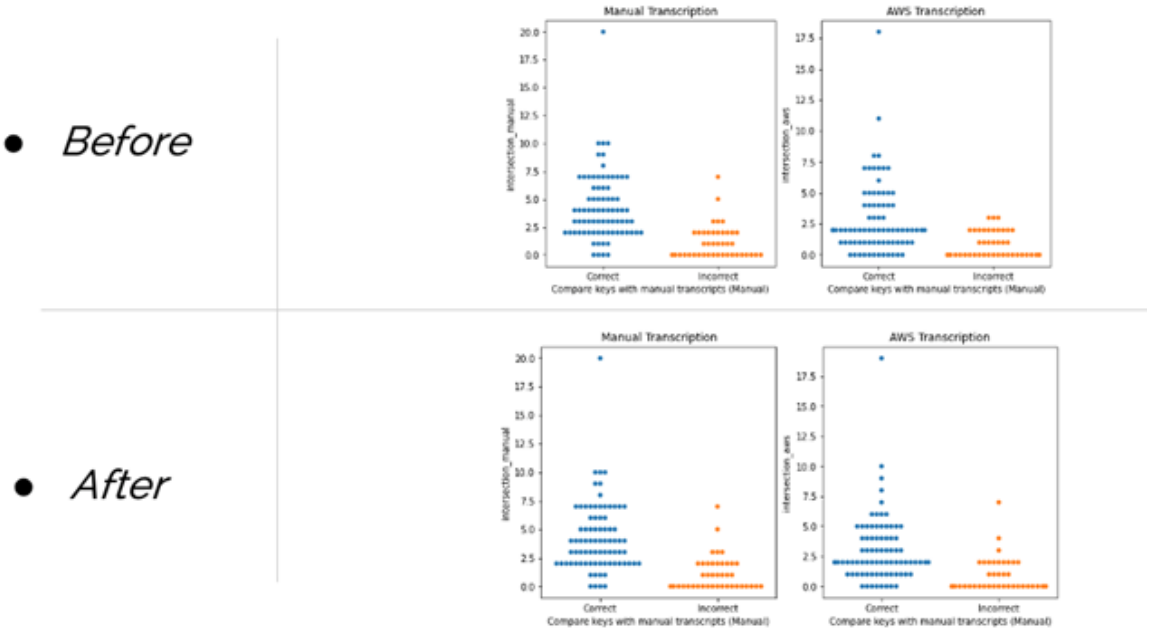

Figure 3. Confusion matrix swarm plots indicating the accuracy values of the scoring model.

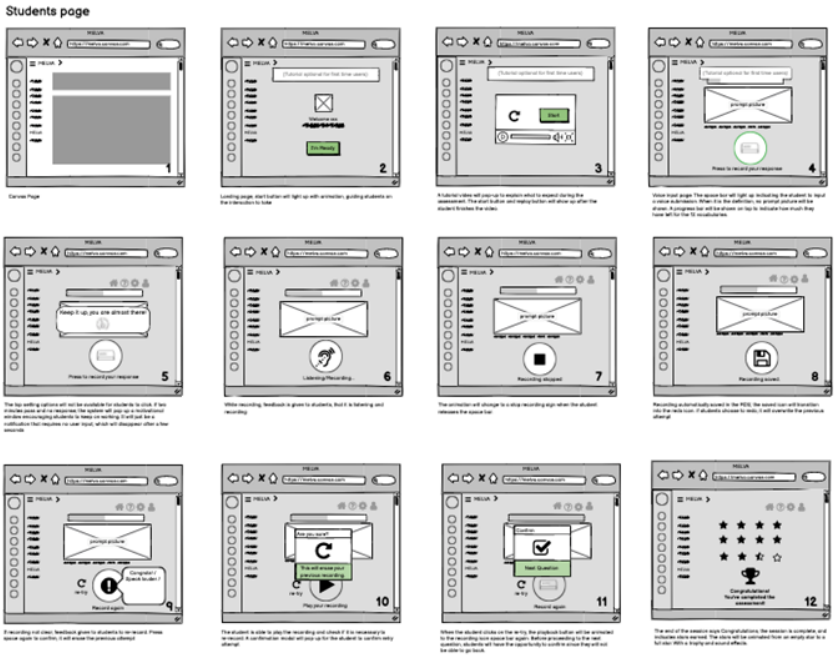

Figure 4. Low fidelity prototype of the student's voice scoring system.

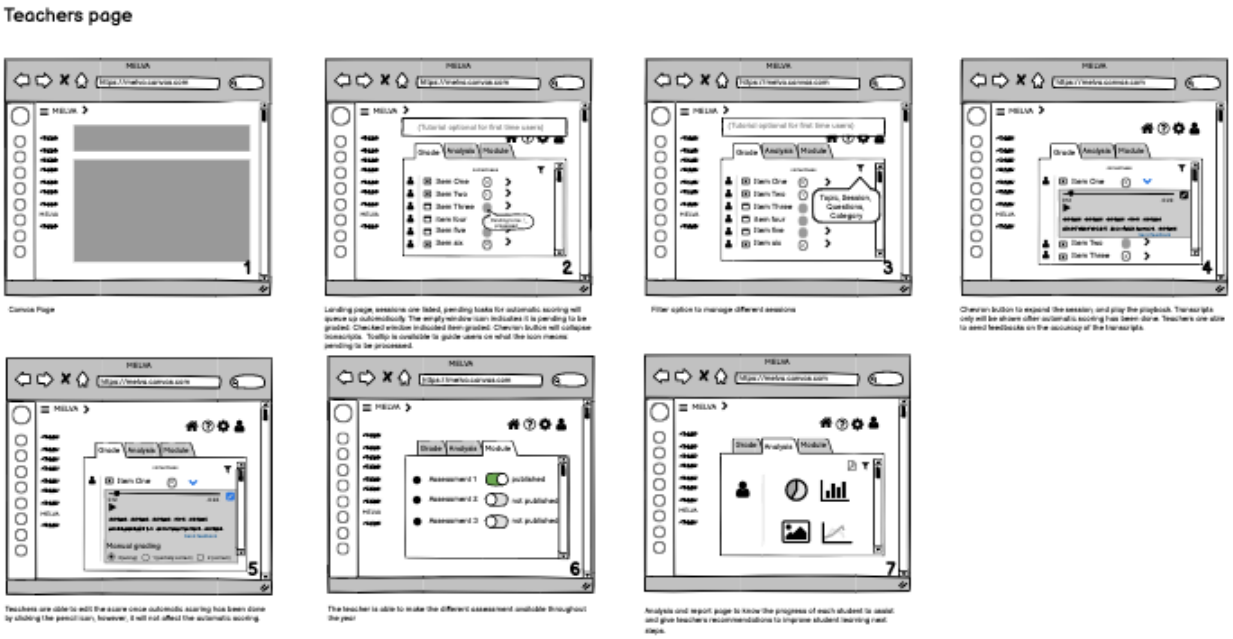

Figure 5. Low fidelity prototype of the teacher's voice scoring system.

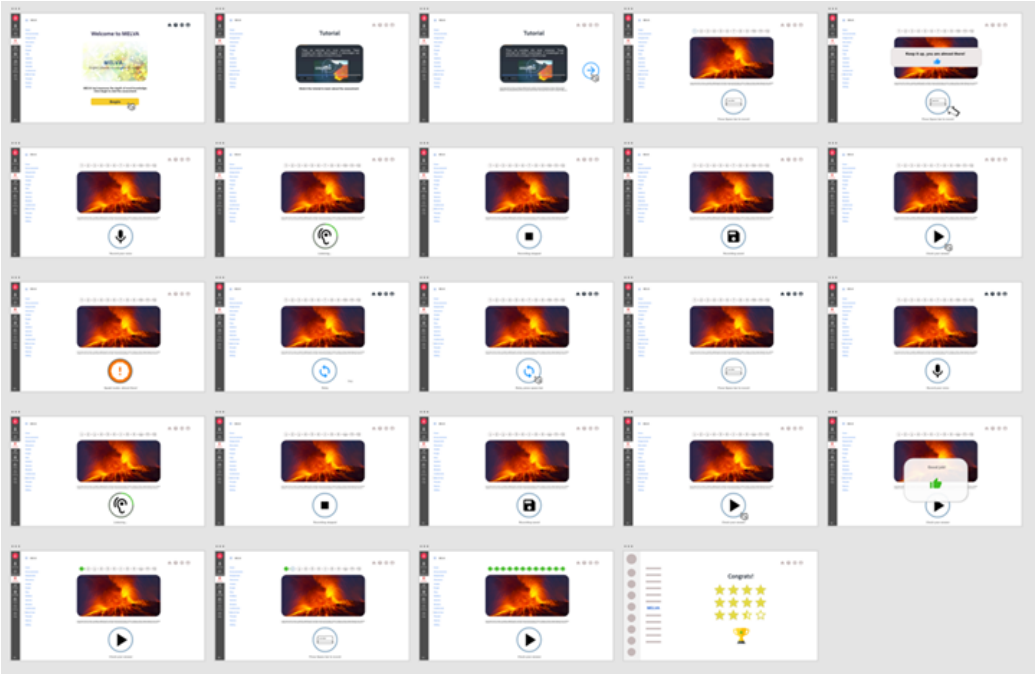

Figure 6. High fidelity prototype of the voice scoring system incorporated/integrated on Canvas.

8. APPENDIX 2: PRODUCT VISION DOCUMENT, MELVA-S, VISION 0.2, 02/2021

Version	Date	Description of Changes	Author
0.0	2/2021	Initial Release	Sisi Kang
0.1	2/3/2021	First Revision	Sisi Kang
0.2	3/5/2021	Proofing for spelling/grammar/style	Andy Rash

8.1. Executive Summary

MELVA-S is an online tool and management system that utilizes automatic scoring to measure the degree of students' vocabulary knowledge.

8.2 Approach

8.2.1 Key Requirements

Tutorial unction
Assignment submission function
Voice recording function
Playback function
Automated scoring function
Student progress monitoring
Student progress reporting function
Teacher reporting function

Nice to have:
Student stars
Encouraging/motivating animations

8.2.2 Current High-Level Flow

Here is the current high-level flow of the website experience.
Initial Flow:
Student clicks on the assignments tab on the left
Student selects the specific assignments option
Student plays the recording and understand the prompt
Student clicks on 'record and upload media'
Student grants microphone and camera permission from pop-up window
Student will start recording
Student clicks 'Finish' to end audio recording
Student may continue with the measurement or re-record their response when necessary
Student clicks on 'submit' to submit the assignment
Student clicks on the 'close' window icon

8.2.3 Ideal High-Level Flow

Here are the steps for ideal user flow: High-Level Stories.

8.2.4 Key Actors

Student
Teacher
MELVA-S Administrator (School Admin)

8.2.5 Student Stories

I need to turn on my microphone access when required.
I need to understand the purpose of this project and how it works by browsing the main page.
I need to record my answer with ease so that I have as few buttons to click as is possible.
I would like to receive motivational stars at the end of the session to feel more encouraged.
I need to have a notification to remind me to continue the session when I idle for two or more minutes to continue the assessment.
I need to see photos to have the proper context for answering the prompt.
I need to click on the following vocabulary quickly to complete the session as soon as possible.
I need the ability to re-record my response when necessary so that I don't submit a bad reaction.
I need to have a pop-up confirmation window before re-recording so I don't accidentally overwrite my previous recording.
I need to have the ability to resume if the page is refreshed or if the network is unstable to encourage me to continue my assessment without re-answering prompts.
I want to have an animated tutorial to help me grant permission to the browser to record audio to record responses to prompts adequately.
I want to hear my recordings to know how well I did.
I want to see a completion screen with stars.

It would be nice to have accessibility features to be inclusive.

8.2.6 Teacher Stories

I need the ability to toggle the availability of the assessments so that students can access the appropriate content at the proper time.
I need the ability to monitor student progress to decide whether to intervene.
I need to filter students' submissions by different categories.
I need to view or listen to student submissions' audio to provide my manual grade to students for feedback purposes.
I need an automated scoring system to score students to evaluate their responses.
I would like to send feedback to MELVA-S regarding the accuracy of the automated scoring system.
I would like to export a weekly and monthly report of student progress to be informed when it comes time to assist students.
I would like to see the transcripts of the students' responses to be informed about the accuracy of the students' responses.
I want to send feedback specific to individual transcriptions to help the developers improve the system.

8.2.7 MELVA-S Admin Stories

I want to store the data collected for research purposes.
I want to add and modify content intended for release to research subjects.
I want to release content to teachers so that they can then publish it for their students.
I need to have the ability to set curriculum keywords.

8.2.8 Website Development Stories

As for website development, I want to make sure recordings are saved in cases of network instability.
As for website development, I want to make sure the user interface/user experience is simplistic, functional, and feasible.

8.2.9 Security Stories

As a website administrator, I want to make sure students' data and information are secure—both at rest and in transit—and that privacy concerns (FERPA, etc.) are minimized.

8.3 Backend Automated Scoring System Design