I, not Robot: A Human-Centered Redesign of CAPTCHA

Tristan Daniel
tdaniel60@gatech.edu

***Abstract*—**Maintaining efficient cybersecurity has become a crucial element in ensuring safe Internet access across a diverse pool of users. The Completely Automatic Public Turing test to tell Computer and Human Apart (CAPTCHA) has become a widely adopted tool in pruning artificial users from human users. However, with the ongoing advancements of artificial intelligence, artificial agents have become progressively adept at solving these test. In conjunction, human users face increasing cognitive strain while interacting with CAPTCHA evaluations due to their rising complexity as a means to counteract the growth of A.I. This project aims to redesign the current CAPTCHA system by addressing its violation of fundamental Human Computer Interaction (HCI) principles. By incorporating cognitive assessments inspired by developmental psychology, specifically those used to gauge cognitive milestones in toddlers, this project proposes a more human-centered approach in the task of verifying human users. The redesign focuses on enhancing user experience and reducing cognitive load while keeping security paramount; thus situating CAPTCHA within the essential heuristics and principles of HCI.

1 INTRODUCTION

Technology has advanced society in ways never imagined by previous generations. Among one of the most significant technological innovations of the 21st century has been the rapid development and expansion of the Internet. The Internet was initially utilized as a way for government researchers to relay information to each other and could only be accessed through large, immobile computer structures (Georgia, 2020). However, the Internet has evolved into a platform intertwined within daily life; facilitating e-commerce, education, entertainment, and interactions with others through social media networks.

The exponential growth of Internet users since its inception in the early 1960's has

been paralleled by the rise in automation. The evolution of autonomous agents have allowed humans to offload repetitive tasks. However, while automation has streamlined efficiency, malicious users have exploited these agents as tools to breach online systems to compromise privacy and security. The malicious abuse of automation have been linked to various cybercrimes such as identity theft, cryptojacking, and other forms of exploitative online behavior.

1.1 CAPTCHA

In response to these threats and a growing concern for user safety, cybersecurity has become an essential component in the implementation of protective safeguards to mitigate vulnerabilities. One effective measure has been the development of Completely Automatic Public Turing test to tell Computer and Human Apart (CAPTCHA), which serves as a buffer against the aforementioned attacks by differentiating between automated and human interactions.

CAPTCHA exist in various forms, relying on users to engage their cognitive abilities to solve challenges such as image recognition, distorted text interpretation, and object orientation matching. These task are designed to verify users are human and not automated agents attempting to exploit online servers. While CAPTCHA serves as an effective security measure, it can become taxing for legitimate users who must complete these increasingly complex assessments to access websites.

1.2 Current issues & Frustrations

The conception of analyzing CAPTCHA stems from personal experiences interacting with its interface on a daily basis. Although helpful in the sense of maintaining security and preventing malicious access, I often found myself frustrated by the unnecessary cognitive burden needed to pass these assessments. This observation reflects a broader concern: as CAPTCHA test become more challenging to counteract advances in artificial intelligence, human users bear the cost.Von Ahln defines CAPTCHA as "a cryptographic protocol whose underlying hardness assumption is based on an AI problem." ((**vonAhn2003**, **vonAhn2003**)). As of 2024, researchers at *ETH Zurich* developed an automated system capable of solving 100% of the presented CAPTCHAs, while other models can only solve 68-71% of the presented CAPTCHAs from reCAPTCHAv2(Plesner, Vontobel, and Wattenhofer, 2024). In light of this development, it's reasonable to assume the current implementation of CAPTCHA will need to be revised to keep up with

AI advancements.

However, the repeated revision of these CAPTCHA systems cause users to report a growing dissatisfaction with the system as a whole. These refactored assessments of CAPTCHA have caused users to experience inconsistency, inequity, and an increased intolerance for user errors in what should be a simple test of human verification. This project aims to address these issues by redesigning CAPTCHA through a human-centered lens, improving both user experience while not compromising the prioritization of security.

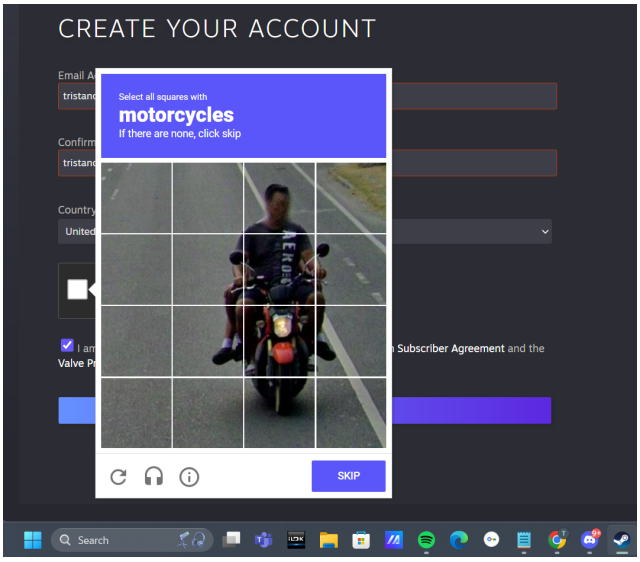


Figure 1—Example of image recognition CAPTCHA

2 NEEDFINDING PLAN

My initial step in the needfinding plan involves the recruitment of participants whom have had experience with taking any variation of CAPTCHA assessments. This way, we gather feedback from a diverse collection of novice and expert users on the task of human verification. The only requirement a participant must meet is that they have readily available access to the internet and web applications. Although this narrows our margin of participants by excluding those without internet or internet accessible devices, these are considered necessary components for CAPTCHA interactions. These are ideal participants for gathering feedback, as CAPTCHA systems are commonly integrated in everyday web access task. Any attempts for logging into personal accounts and participating in e-commerce typically triggers a CAPTCHA assessment.

Recruiting participants will occur by virtual consent provided through *PeerSurvey*. In addition, I plan to recruit a small sample frame of individuals, particularly close friends and family, to conduct in-person interviews regarding their user satisfaction with the current state of CAPTCHA.

The questions asked through this needfinding stage are designed with a focus on the data inventory pertinent to the task of human verification (see Appendix 14.1). By following these guidelines, we can minimize irrelevant data collections and enhance the accuracy of our feedback. The implementation of anonymous survey distribution aids in the mitigation of commonly found biases while conducting user experience research. However, certain questions, such as questions 3-5, may still be susceptible to recall bias; as participants are asked to give an estimate on how many attempts it takes to solve different variations of CAPTCHA assessments. To reduce the impact of recall bias, I opted for terms such as "typically", allowing participants to provide general estimates without specifying a particular date or instance.

2.1 Survey Strategy

By utilizing surveys to gather user feedback, viable information can be gathered at a low cost. For classmate recruitment, I will be utilizing *PeerSurvey* to reach out to classmates who have experience with taking CAPTCHA surveys. The overall aim for the survey is to focus on user's experience and cognitive processes involved in proving human verification.The goal for participation sits at a minimum of 25 respondents.

2.2 Interview Strategy

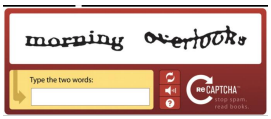
The use of interviews will provide a more dynamic and rich environment in investigating the cognitive processes involved in taking CAPTCHA assessments. The interview's questions will follow the same contextual format found in the survey questions. The goal for participation sits at a minimum of 5 respondents.

2.3 Heuristic Evaluation

In the context of heuristics, I aim to examine the the CAPCTHA assessment interface. For identifying potential problems with the interface, I plan to analysis the following heuristics, which derive from a combination of Nielsen's ten heuristics (Nielsen, 1994) and the principle discussed in *Lesson 2.5*: **error prevention, flexibility and efficiency of use, and equity**.

3 NEEDFINDING RESULTS

Concluding the needfinding process, I successfully reached my expected goal of 25 survey respondents. All respondents were collected through PeerSurvey, with 100% of the participant population consisting of classmates within CS 6750. Although the initial plan included distributing surveys via social media, my survey reached it's target number of participants before I circulated it outside of Georgia Tech's PeerSurvey interface. I was only able to conduct three interviews, falling short of the preferred goal of five, but they still provided valuable insights.



(a) CAPTCHA Flexibility and Efficiency Violation

(b) CAPTCHA Equity Violation

Figure 2—Example of CAPCTHA representations that violate heuristic and design principles

3.1 Heuristic Evaluation Results

For heuristic evaluation, the CAPCTHA interface violated the three key heuristics examined. First, in the area of error prevention, CAPCTHA violates this heuristic through its lack of constraints and support of "undo" functionalities. For example, in CAPTCHA image recognition assessments where users must select squares on a grid that make up an objects representation, there is no option to undo an incorrect selection, making the first selection the final answer. This failure to support correction violates the basic design principle of tolerance.

For flexibility and efficiency, the interface offers both audio and visual cues for some distorted text assessments, as seen in *Figure 2-a*. However, this dual-feature functionality is inconsistently implemented across all CAPCTHA variations, limiting its overall efficiency.

Lastly,regarding equity, the current implementation of CAPCTHA inadequately serves individuals with disabilities. For example, as illustrated in *Figure 2-b*, closely packed alphanumeric characters can become a challenge for users with learning disabilities such as dyslexia to interpret. CAPTCHA also fails to take cultural differences into account with their assessments as well; opting out in utilizing instinctive identifiable elements that's consistent across global societies.

3.2 Insights

This needfinding exercise provided several key insights into areas that should prioritized in future brainstorming and prototyping. I found participants highlighted the time consumption taken to complete CAPTCHA challenges as an unavoidable annoyance, with over 60% of respondents admitting to abandoning a task due to the time required to complete a CAPTCHA assessment (see Appendix 14.4 12). One participant explicitly noted their desire for a more efficient system "that does not eat into my time". Although the overall response from question 9 regarding satisfaction with CAPCTHA's error tolerance formed a bell curve. Around 87% of respondents reported refreshing their CAPTCHA assessment more times than often due to difficulty (see Appendix 14.7 17).

Through short answer responses, it was surprising to see image recognition CAPTCHAs, not text-based ones, were often cited as being most challenging due to low resolution and blurry images. My initial hypothesis suggested that text based assessment would have proved to be more problematic due to its infinite range of alphanumeric strings paired with differing fonts. However, only 18 of the 25 respondents had encountered them, compared to full participation in image recognition challenges(see Appendix 14.2 11).This suggest that image recognition CAPTCHAs deserve more priority in terms of usability in future improvements.

4 BRAINSTORMING

The brainstorming process will begin by further examining the heuristics the CAPTCHA interface fails to adhere to, along with the ones that it demonstrates compliance under. This is an effort to highlight the areas where the system's usability can be improved. Heuristics such as error tolerance, equity, and flexibility will especially be considered in order to gain a clearer understanding on how the interface can better serve it's users.

From my own intuition and experiences, I will also generate ideas that focus on improvements that can be made to the existing variations of CAPCTHA interfaces. I plan to take the context in which CAPCTHA assessments are completed into consideration as well, reflecting how different context can reflect the user's overall experience. As a personal observation, I will put emphasis on design alternatives that serve diverse audiences; putting equitable and accessible design

decisions at the forefront of my brainstorming process.

Lastly, my brainstorming process will be heavily informed by the insights gathered by respondents during the needfinding stage. By taking into account what real users experience when interacting with the interface, I can ensure a the redesign will be more catered in improving the frustrating and unsatisfactory aspects of the CAPCTHA interface through the user's perspective. With this in mind, time consumption and unclear prompts given by CAPCTHA will be considered while brainstorming new alternatives.

4.1 Structure

To keep this process structured, the initial plan starts with the examination of the core problem: "How can we redesign online human verification process to increase user experience while maintaining security against artificial users?" I plan to come up with a target of at least 20 ideas for a possible solutions over several brainstorming sessions spanning several days. These ideas will be free of constraints and ignore factors of feasibility, as this is just an exercise to encourage out-of-the-box thinking and create a broad spectrum of possible solutions. To mitigate bias's in this individual brainstorming activity, I plan to incorporate ideas from fellow classmates and peers, to eliminate blind spots in my ideologies. To further eliminate biases, I will take occasional breaks from brainstorming as a way to refresh my perspective on the problem.

5 BRAINSTORMING RESULTS

The brainstorming activity concluded after generating 20 ideas over the span of several days. A comprehensive list of these ideas is provided in *Appendix 14.5*. In an effort in facilitating a good brainstorming activity, the brainstorming document emphasizes the main task as a title:"User wants to verify they are a human while using encountering captcha test interface". This emphasizing of the task helped regain focus after occasional breaks to ensure clarity and proper idea formation direction. The central theme among my proposed ideas revolved around the integration of cognitive test commonly used in the field of developmental psychology. These cognitive test align well the CAPCTHA objectives of verifying human intelligence while pruning artificial users. Furthermore, these test promote equity by leveraging assessments that most competent individuals can easily complete. This address the inequitable practices deployed in CAPC-

THA's current system relating to cultural differences, learning disabilities, or age-related cognitive decline.

After careful evaluating all proposed approaches, I selected three alternatives to explore further in the prototyping phase:: A-not-B Test, Facial Expression Recognition, and the Violation of Expectation Paradigm.

5.1 A-not-B Test

The A not B test is a widely used assessment in infant psychology that relies on simple short term memory. The test involves hiding an object in location A and then moving it to location B. This test then prompts the user to track the movement of this hidden object. Constructed with a limited amount of choices and quick animations, the assessment addresses the issue of time complexity and unclear instructions identified earlier as user frustrations in solving current CAPCTHA designs. A proposed prototype to demonstrate this test within the CAPCTHA system would be a box-and-prize setup. The user is shown a "prize" (a distinct, identifiable object) that is placed into one of three boxes. The boxes undergo a brief shuffle, and the user is then asked to identify the box containing the "prize" object. The task implements concepts of object permanence from developmental psychology; while ensuring the assessment remains quick and intuitive to address aforementioned user complaints regarding time consumption and unclear prompts.

5.2 Facial Expression Recognition

Recognizing facial expressions is a deeply instinctive human ability that relies on emotional context, but has proven to be difficult for artificial agents to mimic accurately without a large amount of training data. This alternative promotes accessibility, as facial expressions are universally understood across diverse backgrounds and cognitive abilities. For prototyping, the user will be presented with an image of a facial expression. The user is then asked to identify the emotion that corresponds to image from a set of options (e.g., happy, sad, surprised). Another variation of this proposed alternative could prompt users to sketch a facial expression based on a given emotion. Seeing that this task leverages innate human abilities, the issue of time consumption and equity concerns are mitigated.

5.3 Violation of Expectation Paradigm

The violation of expectation paradigm is a infant assessment test that takes "two test events, an unexpected event that violates a particular expectation and an expected event that accords with it."Margoni, Surian, and Baillargeon, 2024.By presenting users with events that either align with or violate expectations, human intuition is capitalized for interpreting logically correct events. This inherently makes selecting correct solutions instinctive, thus lowering the cognitive load required to correctly pass CAPTCHA assessments. Additional, this task is designed to be difficult for artificial users to predict correct outcomes.For a potential prototype, users will be given an initial image of a real-world event, such as an individual holding a food item near their face. The user will then be prompted to select the next logical step from a collection of possible events. The correct response in this scenario would be an image showing the individual eating the food, while incorrect options might include unrealistic actions such as placing the individual placing the food near their ear or on top of their head. This intuitive design lowers cognitive load for human users while maintaining security through the use of unpredictable events.

Overall, the brainstorming process yielded a diverse range of design alternatives that address the key components found during needfinding exercises: time consumption, equity for all users, and unclear prompts. The selection of the aforementioned approaches draw inspiration from developmental psychology assessments; relying on human cognitive abilities.

6 INITIAL PROTOTYPING

For each of the aforementioned design alternatives, I developed low fidelity pen and paper prototypes. These prototype designs allow for changes to be easily be made based on user feedback and provide flexibility during the early stages of evaluation. By using paper and pen prototypes potential users can interact and add functionalities to the system themselves which can later be leveraged during more higher fidelity prototype designs

6.1 A-not-B Test Prototype

The prototype (*Figure 3*) follows the flow of a card prototype, considering the interaction with the CAPTCHA will involve multiple state changes. The user is presented with three numbered boxes (1-3). Positioned above these boxes is a

distinct object, in this case, a circle with added emphasis to attract attention. The prototype is designed to guide user through the task of human verification by tracking an object as it's shuffled between boxes.

The next card mimics an animation of the circle being placed in one of the three boxes at random (but for our prototype's case, the circle is visibly placed box 2).

The next phase will shuffle the boxes in a sequential manner. In the case of the prototype, boxes 1 and box 2's positions are switched, then box 2 and box 3's positions are switched as well. For simplicity, the numbers that label the boxes are removed during the shuffling process to minimize distractions. The interface gives feedback to the user, notifying them that the boxes are undergoing a shuffling sequence, which is indicated by the *shuffling* message displayed in the top-left corner of the interface. This added feedback from the system aids in the overall perceptibility of the system, allowing the users to easily perceive the state of the system.

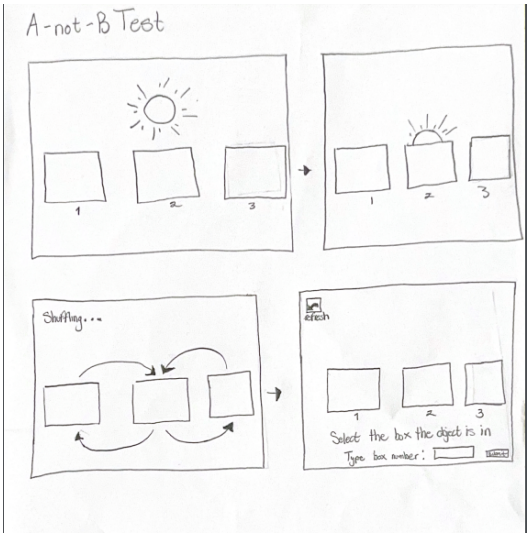


Figure 3—A-not-B pen and paper prototype

After the shuffling is complete, the boxes are displayed in a new arrangement. The user is then prompted to select the box where the object (circle) is inside of. The original number labeling of the boxes are again made visible on the interface. The selection by the user can be initiated through two options:

- The user clicks/taps the box they believe the object is in. This action is device dependent, as computer of desktop users can click on their box selection while

mobile users can select a box by tapping on the corresponding box.
- Users can enter the number of their box selection in a designated text field.

The implementation of multiple methods for answer selection supports the concept of flexibility in solving CAPCTHA assessments; allowing users to interact with the interface in multiple ways while addressing time consumption concerns by decreasing the complexity of the CAPTCHA test. In the top left corner in this display, an "redo" button is provided as an affordance, in case users need to restart the test if they lost track of the target box during the shuffle process. Once users are confident with their answer selection (in our prototype's case, box 1 is the correct answer), they may submit their answer for further evaluation.

6.2 Facial Expression Recognition Prototype

The CAPTCHA interface test the user's ability to recognize human emotions from facial expressions. The interface presents the user with a series of three facial expressions that represent the same emotion across different models(including a collection of diverse ethnicity, gender, and emoticon characters) to promote equity and avoid cultural biases (*Figure 4*). As seen in the prototype, all faces express a "smiling" facial expression. The user is then prompted to select from one of the following emotions to positively associated the displayed facial expressions with (based on the most common human emotions):

- Fear
- Happy
- Anger
- Disgust
- Sad
- Tired

Users are also given a "redo" button at the top right corner of the screen as an affordance. If users wish, they may refresh the set of facial expressions displayed that correspond with another random emotion. Along with initiating this refresh, emotion answer choices are shuffled from its previous arrangement.The absence of a text field for users to define their answers adds constraints to the CAPTCHA assessment that accounts for user errors that is not seen in current text-based CAPTCHA assessments. Once users are confident with their answer selection (in our prototype's case,"happy" is the correct answer), they may submit their answer for further evaluation.

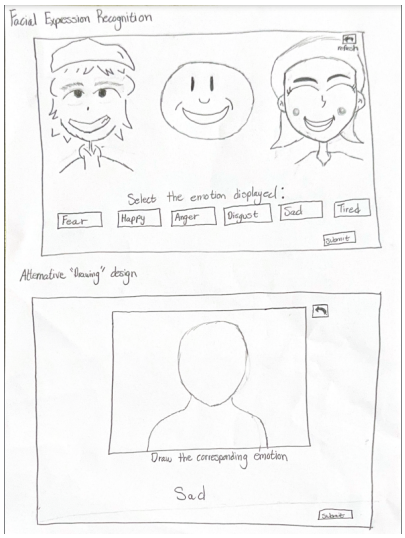


Figure 4—Facial Expression Recognition pen and paper prototype

An alternative to this concept of facial expression recognition, users could be asked to draw the facial expression corresponding to a given emotion on a blank human silhouette. In the prototype's example, the user is prompted to draw a "sad" facial expression onto the silhouette. The drawing can be submitted through drag-and-drop interactions or touch gestures, depending on the user's device. Once satisfied, users can elect to submit their drawing as their final answer. The design aims to address time consumption by applying a quick recognition task that leverages an innate human ability.

6.3 Violation of Expectation Paradigm Prototype

The prototype design is based on a popular cognitive paradigm within the field of developmental psychology, prompting users to identify event sequences based on common sense reasoning. The interface displays a series of images representing an action that is partially completed, and the user is tested on their selection of the most logically expected outcome. Initially two images are shown, labeled under the heading "prior actions"(*Figure 5*).

In the prototype, the first image depicts a person holding a fruit, while the second depicts the fruit moving towards their in preparation for eating. Users are shown four selections of possible events to choose from. The answer choices depict three incorrect actions (such as in the prototype, the fruit is placed near

the ear, on top of the head, and fruit further away from the mouth). The correct event would be for the user to select the fruit being eaten (which is the case for image 2). Users select their answer by entering the number associated with the image into a designated text field.

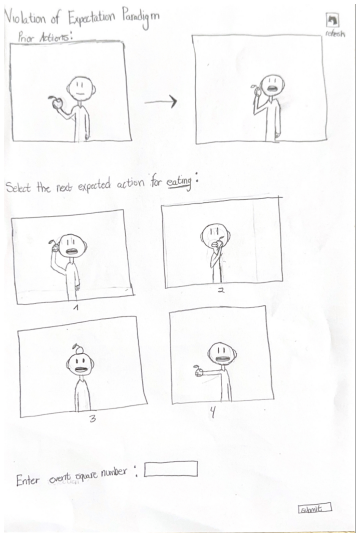


Figure 5—Violation of Expectation Paradigm pen and paper prototype

If a user is not satisfied with the current problem given, they have the ability to "refresh" the CAPCTHA assessment and are given a new problem. Once satisfied with their answer choice, users can submit their answer choice with the "submit" button. Having users interact with the submit button adds semantic constraints to the assessment, by allowing the user to change their answer selection before committing to their selection as their final submission. This element is often absent in current image recognition CAPTCHA assessments, as they force the user to commit to their first selection as their final answer. The prototype address concerns with unclear assessment instructions by making the task intuitive; aligning with human logic. By promoting cognition through human logic, we can create more accessible and usable test while simultaneously reducing the cognitive load required to solve the CAPTCHA test.

7 EVALUATION PLANNING

Considering this marks the early stages of the design life cycle, evaluation will be devised utilizing a qualitative approach. Qualitative approaches will use a mixture of a think aloud, synchronous sessions and asynchronous, survey delivery

to participants. Evaluation will test each of the proposed prototypes against each other; varying the order of the prototypes to each user. This decision aims to eliminate familiarity bias participants may have when encountering the already existing CAPTCHA interface. Another potential issue this mitigates is the inability to completely control the evaluation experiments when assessing against the current CAPTCHA test. Current CAPTCHA assessments are randomize, meaning each participant will not have the same experimental experience which in turn can pollute our feedback with inconsistency.

7.1 Participants and Recruitment

Participants will consist of family, friends, and in person classmates for think aloud activities. Online classmates will recruited from Ed Discussion and PeerSurvery for the asynchronous survey activity to gauge their satisfaction and concerns about the proposed design alternatives. Participants will be take part in these evaluation activities without incentives.

7.2 Procedures

Both activities will leverage the following general questions for all prototypes to gain valuable feedback and explore applicable variables:

1. How difficult is it to interpret what is being asked of you to accomplish?
2. As a first impression, what do you think the CAPCTHA assessment is testing you on?
3. In regards to learning what to do, was the length of the interaction too long or just right?
4. What would you change about the assessment to make it easier to complete?
5. What would you change about the assessment to make it faster to complete?
6. Were there any specific aspects of the assessment that frustrated you?

More prototype specific questions will be assessed as well, which can be viewed in *Appendix 14.3*. Many of the general questions yield qualitative results, which will be evaluated through note taking for think alouds and software logging for online surveys. The online survey will have the ability to enumerate many of these questions into empirical data, primarily yielding nominal and ordinal data. These data sets will be tested using the Chi-squared statistical test. The findings released from these test will give insight on which interface is best for reducing time consumption, the most satisfying interface to participants overall, and the

difficulty of interpretation on first impressions.

8 EVALUATION RESULTS

Concluding the evaluation process, I fell short of my expected goal of 25 survey respondents; by only completing the survey portion with 13 survey respondents. This shortfall is largely attributed to the massive influx of surveys currently circulating the PeerSurvey platform, thus reducing the accessibility of my survey. Survey respondents were collected through PeerSurvey, with the participant pool mixed with CS 6750 classmates and participants from think aloud activities. In addition, I was able to conduct five think aloud activities; successfully reaching the set goal and garnering valuable insights. These think aloud participants consisted of a mix between in person classmates and friends outside my academia cohorts. As a means to better structure the think aloud activities, I asked participants to complete the PeerSurvey survey and allowed them to ask questions about the proposed prototypes in real time.

In terms of quantitative analysis, qualitative questions were enumerated into testable data, the chi-square scores are reported in (*figure 6*). Contrary to hypothesized beliefs, the ch-square test indicated no significant association between the prototypes and how respondents rated their understanding of the task (x2̂ (3, N = 13) = 9.40, p = .31).

Results						
	1	2	3	4	5	*Row Totals*
ANB	2 (1.67) [0.07]	2 (1.67) [0.07]	9 (5.00) [3.20]	1 (3.33) [1.63]	1 (3.33) [1.63]	15
FET	1 (1.67) [0.27]	1 (1.67) [0.27]	2 (5.00) [1.80]	4 (3.33) [0.13]	7 (3.33) [4.03]	15
VOE	2 (1.67) [0.07]	2 (1.67) [0.07]	4 (5.00) [0.20]	5 (3.33) [0.83]	2 (3.33) [0.53]	15
Column Totals	5	5	15	10	10	45 (Grand Total)

Figure 6—CHI-test for understanding user opinions

Similarly, for user confidence (Question 5), ease of use (Question 11) and perceived cognitive load (Question 15), the chi-squared analysis (*figure 7*) indicated no significant association between different prototypes and respondent's opinions on the aforementioned topics.(x2̂ (3, N = 13) = 14.8, p = .06).

Results						
	CC	MC	N	MU	CU	*Row Totals*
ANB	2 (1.27) [0.42]	4 (2.22) [1.43]	3 (3.17) [0.01]	2 (1.90) [0.01]	2 (4.44) [1.34]	13
FET	1 (1.46) [0.15]	1 (2.56) [0.95]	3 (3.66) [0.12]	1 (2.20) [0.65]	9 (5.12) [2.94]	15
VOE	1 (1.27) [0.06]	2 (2.22) [0.02]	4 (3.17) [0.22]	3 (1.90) [0.63]	3 (4.44) [0.47]	13
Column Totals	4	7	10	6	14	41 (Grand Total)

Figure 7—CHI-test for understandability of each prototype

In regards to qualitative analysis, here are the some of the more insightful feedback's given by respondents in the open response portion for each prototype:

A-Not-B Test

- "It is an interesting idea, although I am not sure that it is any easier for the user than the existing options."
- "A-not-B takes too much mental effort"
- "Also another concern is the speed of the shuffle; is it going to be paced at the speed that a user can realistically capture the image but also fast/confusing enough for an AI to detect."

For the "A-Not-B" prototype, several respondents reported difficulties in understanding the shuffling sequence on the low-fidelity interface. This was partly due to an error on my side, as I did not clarify that the arrows were simply inductors of the shuffling movement, not a sequence to be followed to solve the assessment. In hindsight, even adding indicators to the arrow's order of sequence (e.g.,labeling the arrow sequence from 1-6) could improve user understanding.

Facial Expression Recognition Test

- "I think this is easy and maybe more pleasant than the existing captcha questions."
- "People who are neurodivergent have more difficulty in recognizing facial emotions. They may see a face and not understand that they are happy or sad."
- "Simple and immediately able to understand and find an answer"

Regarding the "Facial Recognition" prototype, most users found identifying facial expression intuitive. However, numerous participants cited concerns about equitability, noting that individuals with neurodivergent traits such as autism, might struggle with expression recognition. Additionally, other respondents pointed out that advanced AI models are becoming more increasingly accurate in classifying facial expressions, which could potentially compromise security.

Violation of Expectation Paradigm Test

- "I think this ones pretty good in terms of both natural intuition, and security."
- "cute! but not sure that there's always just "one" right answer"
- "I can tell action 2 is the correct one, but action 1 could be thought of as a continuation of the original sequence of events"

For the "Violation of Expectation Paradigm" prototype, respondents generally found the test intuitive while also maintaining security against artificial users. The main concerns stated by the users regarded the answer choices provided by the test, as they must be distinctive enough for assessment takers to easily distinguish between correct and incorrect actions sequences.

9 SECOND ITERATION PLANNING

Moving forward into the second iteration of development, insights from the previous evaluation phase has provided essential guidance for refining the design alternatives. The evaluation added more emphasis to recurring user concerns; particularity regarding the A-not-B prototype around time consumption and equitable use for diverse users in regards to the Facial Expression Recognition prototype. In addition, new concerns emerged as well after analyzing user feedback. Many respondents stated their desire for more auditory options for more flexibility when interacting with the assessments along with its visual representation.

Overall, the evaluation process gave me a more profound appreciation for those that design these CAPCTHA assessments, as it can be a challenge to balance user equitability, security, and time consumption without compromising one for another.

9.1 Design Alternative Finalization

The evaluation results confirmed that modeling assessments after cognitive development principles made users feel more capable of answering correctly by leveraging more intuitive tasks. Under careful deliberation, I've decided to move forward with the Violation of Expectation prototype. Although this model did not receive the highest overall rating among the other three prototypes, the concerns raised by survey respondents about the other prototypes were inherent in their design. For example, concerns centered around the A-not-B prototype contradicted reducing time consumption and the Facial Expression recognition contradicted accessibility efforts for neurodivergent individuals; the Violation of Expectation prototype aligned well with user expectations and simply required adjustments in regards to answer choice clarity.

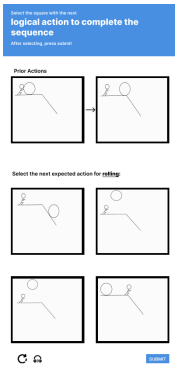
To address the newly identified need for auditory flexibility, I plan to incorporate an audio descriptive feature within the VOE design alternative, allowing for

narrations of the presented problem and answer choices.

10 FINAL PROTOTYPE

The final prototype of design of the Violation of Expectation CAPCTHA assessment can be seen in *Figure 8*. To display good practice of interface design, I decided to keep the VOE prototype design relatively consistent with the current CAPCTHA's design (i.e., text font, color, picture formatting). To address user's needs for equitability catered to visually impaired individuals, the interface design implements an audio narration function to allow for more flexibility, depicted by the headphone icon in the bottom left corner. Beside the audio narration icon resides a "refresh" button; enabling users to request a new CAPCTHA assessment if they potentially become overwhelmed or confused by the current test. As a way to improve tolerance, user's are asked to press the "submit" button (bottom right corner) to complete their human verification task. This way, if user's suddenly want to change their answer choice or make a slip when making their answer selection, they are allowed to make any changes before the task is completed.

(a) VOE CAPCTHA Assessment Example 1(Eating)



(b) VOE CAPCTHA Assessment Example 2 (Rolling)

Figure 8—Medium Fidelity Prototype of VOE CAPCTHA Assessment

10.1 Interaction Sequence

The sequence of expected user interaction goes as follows:

1. Users are given the prompt to select the next logical action to complete the prior sequence of actions. Two images are used as "Prior Action" squares and arrows are used as signifiers to aid in the inherent affordance of the design;

notifying users the sequence of the prior actions(14,15).

2. After reviewing these images, users direct their attention to the below prompt asking to select the next logical action that completes the explicitly named task (e.g., eating, rolling, etc.). This is implemented to give users clear directions as to what next action sequence the user should be looking for, in case it was not evident by previewing the "Prior action" sequence.
3. Once users are confident in their answer selection, the user selects the image square; indicated by a blue highlighted border(9).

Figure 9—User answer selection highlighted

4. To confirm their final selection, the user presses the "Submit" button to have their selection evaluated by the system.
5. If the user's selection is the correct choice, a "successful" toast is displayed to the user along with a message notifying the user that they will be directed to the next webpage. Otherwise, if the user selection is wrong, the user is given a different CAPTCHA prompt to solve.(17,16)

More detailed images of these prototypes and frames can be further examined in *Appendix 14.8,14.9,14.10,14.11*

10.2 Video Prototype

The following the link to VOE medium-fidelity prototype will take you to a video detailing all the functionalities of the prototype in real time.

11 FINAL EVALUATION PLANNING

The final evaluation of the medium fidelity prototype will follow a similar approach as the aforementioned evaluation processes. To achieve meaningful insights within a limited time frame, a live demonstration will be conducted with five participants, each participating in a think-aloud interview.

11.1 Recruitment and Procedures

Participants in this survey will consist of in person classmates and other peers outside my academic cohorts to gather a diverse perspectives. Participants will participate voluntarily, however, classmates were incentivize with a "survey-for-survey" favor exchange. Participants will be asked to engage with the prototype and participate in a structured think-aloud activity; to allow for a space for participants to share their immediate thoughts and reactions as they navigate the interface.

During the think-aloud sessions, participants will be asked guiding questions to encourage feedback on prevalent aspects of the prototype, allowing them to freely expand on their options (see *Appendix 14.5* for full list of question). Each interview session aims to capture qualitative and quantitative feedback on the overall user experience and task performance.

11.2 Quantitative Analysis Planning

Quantitative analysis will cover the time consumption concerns address by users in earlier evaluation phases. To produce empirical results, a Chi-squared statistical test will be applied to user ratings on nominal data (e.g., task clarity, intuitiveness, and confidence in answer selection). Additionally, time taken to complete the CAPTCHA will be recorded, allowing for an interval comparison against standard completion times.

11.3 Qualitative Analysis Planning

Qualitative will be summarized and categorized based on common themes and user satisfaction, with a focus on aforementioned priorities: security, answer clar-

ity, general usability, and equitability for visually impaired users. This feedback will help determine the effectiveness of adjustments made to the prototype to enhanced usability, met user expectations, and improve accessibility.

12 FINAL EVALUATION RESULTS

The final evaluation was successfully executed with 5 participants: 2 classmates and 3 peers outside of my academic cohort, all recruited through voluntary response. My initial "survey-for-survey" recruitment tactic was only utilized for one respondent's participation. Interview notes and qualitative data can be further examined in *Appendix 14.6*.

12.1 Quantitative Results

As shown in *Figure 10*, the Chi-squared test shows overwhelming positive results in terms of the nominal data collected during the prototype interviews. In regards to understanding how clear the task was (Question 2) , the overall respondents score 23/25. Intuitiveness (Question 4), scored 20/25 and 23/25 for answer selection confidence (Question 4). These scores suggest the adjustments between iterations successfully enhanced the overall user experience.The chi-square statistic is 0.3563. The p-value is .999964, which does not correlate to a significance in association between participants and the categories they scored.

For time taken to complete the assessments, participants averaged 13.072 seconds in CAPTCHA verification completion, slightly outperforming the industry average of 13.6 seconds(Yilmaz, Zavrak, and Bodur, 2015). However, with a relatively small sample size, more participants would be required to accurately generalize these findings to a broader population.

Results						
	Question 2	Question 4	Question 5			*Row Totals*
Respondent 1	4 (4.53) [0.06]	4 (3.94) [0.00]	5 (4.53) [0.05]			13
Respondent 2	4 (3.83) [0.01]	3 (3.33) [0.03]	4 (3.83) [0.01]			11
Respondent 3	5 (5.23) [0.01]	5 (4.55) [0.05]	5 (5.23) [0.01]			15
Respondent 4	5 (4.53) [0.05]	4 (3.94) [0.00]	4 (4.53) [0.06]			13
Respondent 5	5 (4.88) [0.00]	4 (4.24) [0.01]	5 (4.88) [0.00]			14
Column Totals	23	20	23			**66 (Grand Total)**

Figure 10—Medium Fidelity Evaluation of Satisfaction Nominal Chi-Square Results

12.2 Qualitative Results

For a qualitative summary, respondents not only found the test easy to navigate, but also cited it's engaging nature, and uniqueness as contributed to an overall

positive user experience. The use of consistent elements that closely align with the standard CAPCTHA interface design was a key factor in respondent's ease of usability. Adding elements for tolerance control (refresh button and submit button) was generally accepted by the survey respondents However, one participant favoring an automatic answer submission, rather than a separate submission button. Respondents widely recognized the addition of audio narration as an ideal way to not only provide flexibility in completing human verification, but as an equitable choice to address the needs of users who may be visually impaired. In summary, a more refined and cleaner design of the prototype aided in an increased user usability and affordance, besides some components that are inherent in the system's design to handle the task of human verification (i.e,. time consumption for question answering and presence of cognitive load).

The overall aim for this iteration of the CAPTCHA redesign shows promising results that can be built upon with more detailed evaluations and higher fidelity prototypes. Modeling these human verification test off cognitive milestones that the generalized population has reached appears to increase equitability, prompts intuitive actions, and maintains security from artificial users.

13 REFERENCES

[1] Georgia, University System of (2020). "A Brief History of the Internet". In: *The Online Library Learning Center*.

[2] Margoni, Francesco, Surian, Luca, and Baillargeon, Renée (2024). "The violation-of-expectation paradigm: A conceptual overview." In: *Psychological Review* 131.3, p. 716.

[3] Nielsen, J. (1994). "Enhancing the explanatory power of usability heuristics". In: Proc. ACM CHI'94 Conf., pp. 152–158.

[4] Plesner, Andreas, Vontobel, Tobias, and Wattenhofer, Roger (July 2024). "Breaking reCAPTCHAv2". In: *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. Vol. 14. IEEE, pp. 1047–1056. DOI: 10.1109/compsac61105.2024.00142. URL: http://dx.doi.org/10.1109/COMPSAC61105.2024.00142.

[5] Yilmaz, Seyhmus, Zavrak, Sultan, and Bodur, Hüseyin (Dec. 2015). "Distinguishing Humans from Automated Programs by a novel Audio-based CAPTCHA". In: *International Journal of Computer Applications* 132, pp. 17–21. DOI: 10.5120/ijca2015907575.

14 APPENDICES

14.1 Needfinding Survey

1. In what context have you been prompted to solve a CAPTCHA assessment? (Select all that apply)
2. If 'Other' specify the platform/context:
3. How much do you agree with the following statement: I often encounter CAPTCHA test while performing time-sensitive task.
4. What CAPTCHA variations have you had experience in solving? (Select all that apply)
5. If "Other", please specify:
6. If you have taken Image Recognition CAPTCHA tests, how many attempts does it typically take you to successfully solve the challenge?
7. If you have taken Distorted Text CAPTCHA tests, how many attempts does it typically take you to successfully solve the challenge?
8. If you have taken Object Oriented CAPTCHA tests, how many attempts does it typically take you to successfully solve the challenge?
9. How much do you agree with the following statement: I am satisfied with the level of tolerance for user error in the current CAPTCHA system.
10. How do you usually feel after unsuccessfully failing a CAPTCHA test?
11. How often have you found the need to "refresh" a CAPTCHA prompt because the challenge was too difficult to solve or interpret?
12. What challenges, if any, do you face when solving CAPTCHA test?
13. Have you ever abandoned a task due to difficulties with a CAPTCHA test?
14. For you, how important is the usability of the CAPTCHA system compared to the security it provides?
15. How much do you agree with the following statement: The current CAPTCHA system effectively implements accessibility measures for individuals with disabilities, such as color blindness and dyslexia.
16. What features or changes, if any, would you like to see implemented in future CAPCTHA systems?

14.2

14.3 Prototype Evaluation Survey

1. Are you familiar with CAPCTHA assessments and its variations? (Image

Field1	Count of Field1
Image Recognition;Distorted Text Interpretation;Object Orientation Matching	14
Image Recognition	4
Image Recognition;Distorted Text Interpretation	4
Image Recognition;Object Orientation Matching	1
Grand Total	23

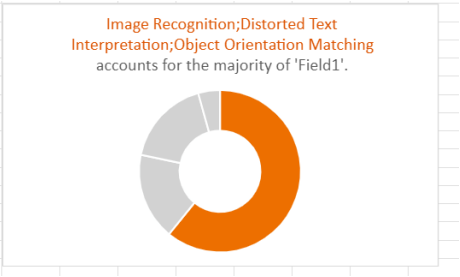


Figure 11—Q4 variation distribution

recognition, text-based, etc)

2. Take a look at the prototype via the link provided. From your first impressions, how difficult is it to understand what the "A-not-B" CAPCTHA prototype is testing?
3. In a short description, what do you think the prototype is testing the user to do?
4. The "A-not-B" prototype test user's ability to follow an object's box placement after shuffling the original arrangement. Were you able to deduce that from your initial viewing ?
5. Rate your opinion on the following statement: I feel confident in selecting the correct box the object is in, after the shuffle is done.
6. Do you have any positive takeaways from the "A-not-B" prototype, compared to current CAPTCHA assessments?
7. What concerns, if any, do you have regarding the "A-not-B" prototype?
8. Take a look at the prototype via the link provided. From your first impressions, how difficult is it to understand what the "Facial Expression Recognition" CAPCTHA prototype is testing?
9. In a short description, what do you think the "Facial Expression Recognition" prototype is testing the user to do?
10. The "Facial Expression Recognition" prototype test user's ability to correctly identify presented facial expressions. Were you able to deduce that from your initial viewing ?
11. Rate your opinion on the following statement: Recognizing facial expressions in the prototype was an intuitive (natural) task.
12. Do you have any positive takeaways from the "Facial Expression Recognition" prototype, compared to current CAPTCHA assessments?
13. What concerns, if any, do you have regarding the "Facial Expression Recogni-

tion" prototype?

14. Take a look at the prototype via the link provided. From your first impressions, how difficult is it to understand what the "Violation of Expectation Paradigm" CAPCTHA prototype is testing?
15. In a short description, what do you think the prototype is testing the user to do?
16. The "Violation of Expectation Paradigm" prototype test user's ability to identify expected events based off a prior sequence of other events. Were you able to deduce that from your initial viewing ?
17. Rate your opinion on the following statement: The alternative answer choices seemed clearly wrong compared to the correct except action.
18. Do you have any positive takeaways from the "Violation of Expectation Paradigm" prototype, compared to current CAPTCHA assessments?
19. What concerns, if any, do you have regarding the "Violation of Expectation Paradigm"prototype?
20. Out of the 3 presented prototypes, which did you prefer the most?
21. Out of the 3 presented prototypes, which did you prefer the least?

14.4

% distribution of 'Q13'

Q13	Count of Q13
Yes	60.00%
No	40.00%
Grand Total	100.00%

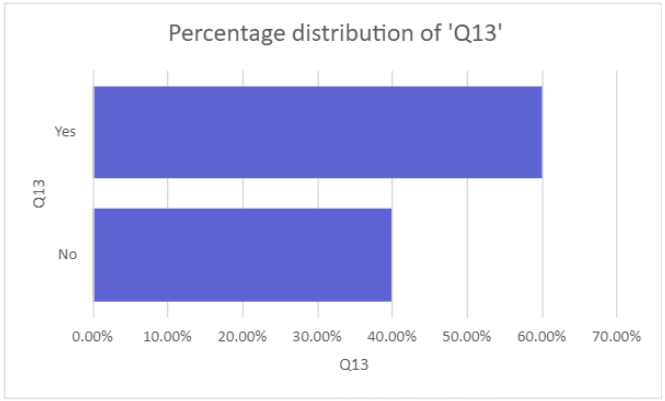


Figure 12—Q13 percentage distribution

14.5 **Medium Fidelity Prototype Interview Questions (after think aloud is completed**

1. What are your initial impressions of the task at hand?

2. On a scale from 1 to 5, how clear is it to know how to carry out the task, with 1 being "not clear at all" and 5 being "very clear"?
3. Does the sequence of events make sense to you?
4. On a scale from 1 to 5, how intuitive was the task, with one being "not intuitive at all" and 5 being "very intuitive"?
5. On a scale from 1 (not confident) to 5(very confident), how confident did you feel in selecting the correct answer choice?
6. Do you feel adding a "submit" button rather than an automatic answer submitting component (similar to the one's in current CAPCTHA systems) impacted your performance?
7. What are your opinions on the audio narration functionality ?
8. In terms of security, do you think the CAPTCHA assessment is secure enough to effectively prune artificial users? Why or why not?
9. What improvements, if any, would you like to see implemented to improve your experience?
10. How satisfied were you with your overall experience with this alternative CAPCTHA design ?

14.6 **Final Evaluation Interview Notes**

Participant 1:

Time to complete test: 11.25s

1. Found the task interesting; different from usual CAPTCHA assessments
2. 4
3. The sequence of events is easy to follow
4. 4
5. 5
6. Feels as like the submit button didn't affect performance negatively
7. Respondent wouldn't use it, but finds it as a fair implementation
8. Does not know much about AI capabilities, but assumes security is maintained
9. Expand on test questions
10. Overall satisfied; enjoyed the assessment!

Participant 2:

Time to complete test: 12.03s

1. Thought the task was "funny", but understandable
2. 4

3. Thinks the sequence of events was easy for them; but might be a little slower for other test takers
4. 3
5. 4
6. Thinks the submit button is "a good practice of tolerance"
7. Respondent cites that it's "similar to the current CAPTCHA accessibility, but states never having a need to use it.
8. Respondent thinks AI would struggle in solving the test, but with enough training,can get better.
9. Has not explicit improvements; thinks users will get use to interface after multiple encounters
10. Respondent again states the assessments were funny, but kept them engaged.

Participant 3:
Time to complete test: 14.36s

1. Respondent states the instructions made it clear on how to complete the task
2. 5
3. Thinks the sequence of events was "straightforward"
4. 5
5. 5
6. Finds the submit button to "expected part of the interaction"
7. Respondent likes the idea, but is confused on how it would effectively be implemented (time it takes to narrate, how to describe images, etc.)
8. Respondent says AI could probably solve some but not all problems. However, the overall opinion is that security is maintained.
9. Suggestion to "add more prior action images"
10. Respondent were satisfied with assessment (more than a text entry or "find the bike type of question").

Participant 4:
Time to complete test: 18.43s

1. Respondent states the VOE prototype is similar to other CAPTCHA assessments
2. 5
3. Thinks the sequence of events are easy and makes enough sense to make answer choice

4. 4
5. 4
6. Thought interacting with the submit button "slightly slowed down their speed" but understands it's necessary
7. Respondent thinks the audio accessibility is a great feature, and has stated using it in current CAPTCHA test before ("probably once or twice; out of curiosity)
8. Respondent thinks AI can not solve these assessments as easily as the other current CAPTCHA assessment.
9. Respondents want to see more VOE examples in the test (states they understand this is just a "prototype" but are interested in the overall idea.)
10. Respondent "had fun taking the test".

Participant 5:

Time to complete test: 09.29s

1. Respondent can easily follow the instructions on screen (likes the underlined task to identify what next action to look for)
2. 5
3. Respondent agrees that the sequence of events are easy to follow
4. 4
5. 5
6. Respondent did find the submit button as an "annoyance", but would rather have automatic submission
7. Respondent thinks the audio accessibility adds "equity" but would like to see it in action to have a more "concrete opinion")
8. Respondent states current CAPTCHA assessment may do a better job with security in some of their test variants (i.e., mouse movement).
9. Respondents revisit suggestions to add automatic answer submission (although it lowers tolerance in the current design alternative).
10. Respondent found the overall test intriguing and very engaging with its "unique" assessments.

14.7

14.8

14.9

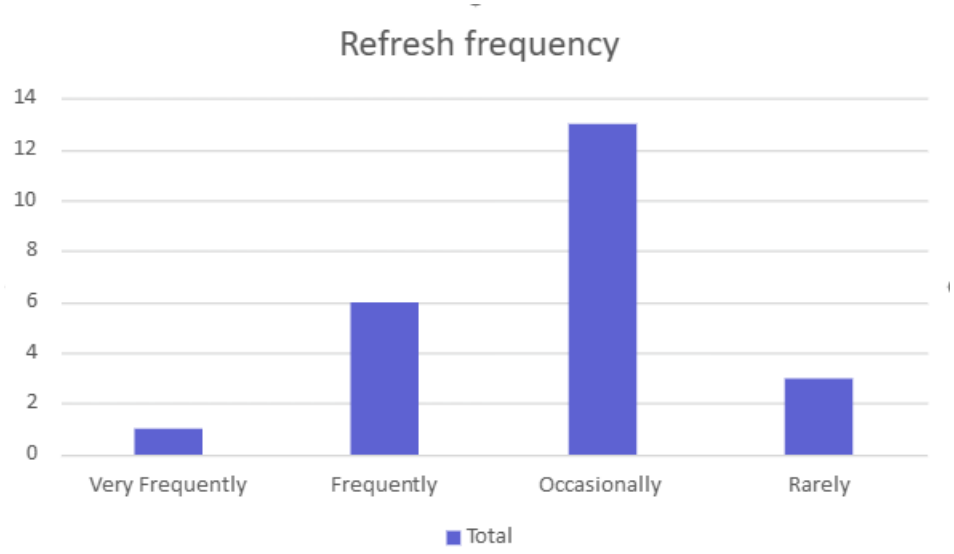


Figure 13—Q14 Refresh Frequency

14.10

14.11

Figure 14—Medium Fidelity Example 1 (Eating)

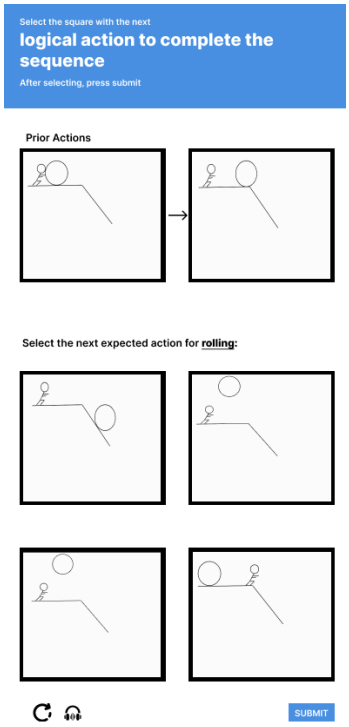


Figure 15—Medium Fidelity Example 2 (Rolling)

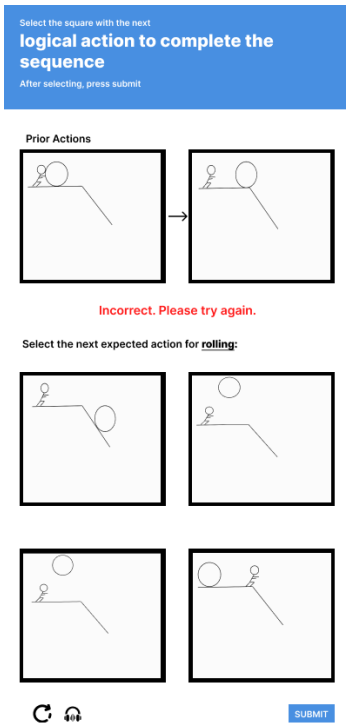


Figure 16—User answer choice incorrect, user must take another assessment

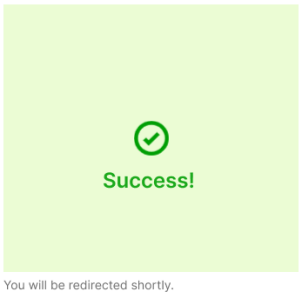


Figure 17—"Successful" Interface