# Usability Testing of the Germinator Task

**Purpose:** We tested the extent to which participants understood task demands given code for the "Germinator Task". Data collected from this task will hopefully result in two studies for my dissertation, and the project involves running thousands of participants. Thus, having the most optimal code and procedure is important.

**Germinator Task Detail**: Participants first undergo a practice task where they categorize the color of color-words by pressing keyboard buttons. In the practice, the color matches the meaning of the color-words, and participants must get above 70% accuracy to succeed. Thus, in the practice, participants must learn the six keyboard buttons that match their respective color-words, but in the main task blocks ("LWPC"), only 4/6 color-words are shown per block, and the colors in which color-words are printed do not always match the meaning of the color-words. On 2 main blocks, the meaning will match the color 75% of the time, and on the other 2 blocks, they will match only 25% of the time. After participants finish the color-word categorization, they categorize digits and letters based on a trial-by-trial cue (Alphabet/Letter vs. Digit/Number), using only two keyboard buttons (e.g., d for consonant/odd and k for vowel/even). This task ("LWPS") also involves a practice task with a 70% accuracy threshold criterion, and on 4 main blocks, participants have to switch tasks only 30% of the time, but have to switch 70% of the time on the other 4 blocks. Finally, after finishing both categorization tasks, participants filled out demographics questions, questions about the categorization tasks, and cognitive puzzles aimed at identifying their "intelligence". They also self-reported mental health symptoms.

**Usability Test Method:** I ran three students on the usability test. Participant 4 was a no-show. I add in my intuitions as a user tester, and when I asked for guidance on recommendation #4, my colleague Nick added his insight. I asked participants questions about the task and instructions at various break points.

## *Recommendation #1: Add greater clarity to the instructions*

**Problem**: At least two participants expressed confusion as to whether they needed to memorize the color-word to keyboard button mappings. One participant explicitly said that he wondered whether the mappings would stay on screen. The other participant was startled by the task timing, which is not explicitly expressed in the instructions beyond telling the participant that they should respond before the color-word/letter/digit disappears from the screen. Yet another participant did not understand what the cues in the LWPS meant: specifically, 2/3 were confused by the fact that a digit and letter were shown next to each other on the screen, but that they'd only categorize one (either digit or letter). Another participant felt that the task was really hard: keeping in mind the cue, finding the letter/number, and then processing all together.

**Solution**:

1. Add a gif or picture that expresses the task-timing for both tasks. The gif would be more ideal because it shows what the timing and sequence would look like. This may also clarify the confusion with the digit and letter next to each other.
2. Clarify that participants need to memorize these mappings. It may also reduce participant frustration if we explicitly state the purpose of the practice rounds again.

We do not want to explicitly post the mappings on the screen during the practice, because we need participants to learn these mappings before moving onto the main task. Nonetheless, here, the goal is to ensure that participants don't waste their time having to redo the practice task because the instructions were not clear enough.

***Recommendation #2: Reconsider practice criterion or lower threshold to proceed onto tasks***

As a reminder, we added a practice criterion, whereby participants cannot move past the practice task before achieving at least 70% accuracy. This was meant to ensure that participants did not accept the "HIT" on Amazon Mechanical Turk and then let the task run without trying. However, we have to make sure the solution is not a problem of its own.

**Problem**: One participant had to repeat the practice LWPC 5 times and the practice LWPS 8 times. One participant had to repeat the practice LWPC 2 times. Another participant told me she had done the LWPC before (likely my other SONA study) but took a picture of the color-word mappings to refer to. She tried the LWPS practice twice before admitting to increased anxiety and asking if she could stop the task, saying that she had learning disabilities and she "couldn't do it." Nick (mentioned below in #4), a colleague, also tried the LWPC practice twice and said that he couldn't advance and thus gave up. Whenever I do the LWPC practice, I basically do a sing-song rehearsal of the color-word mappings ("yellow red orange, blue green purple") for about twenty trials until I've learned the mappings, and I repeat the cues out loud for the LWPS practice. Most participants are not like me and are more like Nick and the students I tested.

**Solution**:

1. Lower the practice threshold. 70% is the goal for the whole task so that I can include the data in the analysis, and we ask participants to achieve above 80% (especially for the bonus), but they may not necessarily need to hit this threshold to advance in the task, especially since there is trial-by-trial feedback.
2. Nick suggested doing "blocks" within the practice. Given 120 total practice trials, we could, for instance, make the first 60 trials devoted to the 3 buttons for the left hand (z, x, c) and then the next 60 trials for the right hand (b, n, m). He said that he could only keep about 3 or so in mind and consistently get those right. Right now, I have all 120 trials randomized, with 20 per button.
3. Nick also suggested that we could do one block where participants can't move forward on the trial until they make the correct response. Then, we could do the timed practice block; that way, participants have a better sense of the mappings.
4. Remove the practice criterion altogether and hope that participants who let the task run without trying would get discouraged by the length of the study.
5. Putting the mappings on screen would create a false expectancy for the main task (when they won't be), but we could include the response mappings for the first half of the practice block and then make the mappings reminder disappear.

All participants reported that the practice task was effective at helping them memorize the mappings for other task blocks, but the goal here is not to remove the practice task, but to remove barriers that would impede its effectiveness.

## ***Recommendation #3: Change some of the post-task questions***

These questions assess how aware participants were of the fact that some color-words were more often shown in a conflicting color (RED shown in blue). The same happens with certain blocks of trials, and we also assessed awareness in the LWPS task. In the LWPS, participants switch task categorizations more often for certain letters/digits and in certain blocks of trials.

**Problem:** At least one participant was confused by what the post-task questions were asking. E.g., was the frequency question asking whether red was shown in purple? What is this hard vs. easy question asking? What does this task-switching stuff mean? Another said "she didn't really have an opinion" and didn't notice hard vs. easy in the task.

**Solution:**

1. We may want to play around with the phrasing more, because we have used these questions in a number of studies, and our previous results could be all noise.
    a. For instance, the participant who didn't understand what the frequency or the matching question was asking – we could, instead of the frequency question, ask what other color red was shown in (and the same of the other colors). That is basically probing the same issue: i.e., whether participants noticed how color-words were paired together.
    b. We could then ask how frequently each color was shown in its own color vs. another color (i.e., estimate of congruency), and then do the matching question.
    c. For the task-switching questions, we could ask about all the letters and digits for the matching question just to see if people think the letters that are unbiased are actually rated as neutral.
    d. We could rephrase all task-switching questions to ask about multitasking instead of defining task-repeat vs. task-switch.
    e. We could rephrase all the "yes/no" questions to essentially ask: were some color-words/letters/digits harder to categorize than others? Were some blocks of trials harder than others? While people might find certain stimuli or blocks harder for reasons that aren't related to the experiment, this is essentially what those questions are asking, and we don't need to define congruent/incongruent/task-switch/task-repeat.

## ***Recommendation #4: Change the JavaScript code***

**Problem:** Only one of the participants had slow wi-fi. For this participant, sometimes the feedback didn't show during the LWPC task, and sometimes the actual cue didn't show on the LWPS, just the letter/digit (G9). Notably, these issues did not happen when she stopped pressing buttons, suggesting that the problem stemmed from the event listener (i.e., pressing buttons / when data records) or, as Nick says, "not due to the browser getting overloaded but due to times not firing correctly." While we do not know what percentage of future online participants will have bad internet connectivity, if truly 1/3, that is way too many participants to lose.

**Solution**:

1. If not already evident, I contacted Nick to ask if he saw any place in my code that could be optimized. He suggested using a different task timer that he'd created and hasn't had issues with (he does not like the timer typically used in JavaScript).
    a. I also pointed out that it might only occur with low wifi issues, and he said that developers use some kind of tool to mimic poor internet connections. (I looked this up: there are settings you can change in the browser for this.) In other words, when testing, find a way to mimic the low internet connectivity issue, too.
2. He also said that he could sit down with me and recode the task probably in a day. I don't know if we should do this, because Audrey and I will have both published our tasks with our particular code. This also is probably a longer delay and may require some kind of credit acknowledgement for Nick. He also said, "There's certainly nothing that should be taxing for the browser and there weren't any obvious errors. It all seemed to run fine for me." So, if we recode the task, I don't know how different it will be.

***Recommendation #5: Increase the pay for the task***

**Problem:** Every single participant had to repeat either the LWPS or the LWPC practice at least once. That adds 4 minutes per repetition to the total time of the task. I calculated $11.50 based on a $0.13/minute and ~85-90 minute task estimate, with $3 bonus for >80% total accuracy.

**Solution:** My estimates are usually based on the exact timing, and as revealed by MTurk reviews, participants largely do not think this is fair. We should assume greater leniency, which would cover additional practice runs for participants. If we assume 2 hours for the pay, that means $15.60 (total cost per participant: $13.80 vs. $18.72).

***Minor:***

I remind participants of the color-word mappings before they start the practice the first time, but don't do this for the LWPS. I need to add that in.

The same participant who nearly had a panic attack doing the LWPS said that the LWPC was not "difficult, but really boring." Is it worth thinking about putting something "fun" between the LWPC and LWPS?

Some of the cognitive puzzle / "ICAR" questions confused participants. One question asks what's 1/5 of 1/4 of 1/9 of 900 —one participant asked if she could use a calculator. The questionnaire doesn't state "don't use a calculator," but perhaps the survey creators envisioned participants being run in-person. We may need to specify this instruction. Another participant was confused by what "alphanumeric" sequence means, and 4 of the questions say, "In the following alphanumeric sequence K L M P, what is the next letter?" If we are editing instructions, we could add a definition of alphanumeric. An ASEBA question also states: "I repeat certain acts over and over". The participant wanted to know what acts meant ("like taking a shower everyday?"). I imagine a clinician would clarify this question in person, but I think that explicitly clarifying the wording here might also interfere with what the creators meant.

Participants explicitly understand that the symptomatology questions are trying to assess whether

they have OCD, schizophrenia, etc. One participant said that the ICAR was like an IQ test. So, the questionnaire is largely *explicit* for folks. In other words, Part II (questionnaire) is explicit, while the Part I tasks (LWPC/LWPS) may not be – expectations issue?

I noticed some weird formatting issues with Qualtrics, so I asked Matt and Brenda about how they code some questions. In short, they suggested that I make all my questions of the "loop and merge" type (instead of matrix tables), which would also speed the questionnaire and make sure that things are smoother. Matt suggested putting a timer on each question too, which should hopefully prevent button mashing without thought. Moreover, I might add in an attention check or two. Even with questions/pauses from me, all participants took ~30 minutes.

The people who do better seem to notice things about the task. For example, the first participant who had to repeat both practice tasks didn't notice the "change" in contexts between blocks, but the second and third participants noticed that not all the colors were shown in each task block for the LWPC. Potentially interesting to consider from a research perspective.

## ***Limitations***

Because of COVID-19, we were not able to run participants in person. I ran participants by giving them a link to a Qualtrics consent form. After they consented, participants were told to contact the researcher to schedule a usability test via Zoom. They screen-shared with me while doing the task. Although I told participants that this was a usability test and that I wasn't judging their performance, I think that a "watching" experimenter presence is not the most ideal.

On one hand, this might actually reflect the scenarios that our online participants undergo (i.e., extra, unanticipated distractions in their environment). But, the video component, despite additional reassurance that the researcher was not judging task performance, likely created additional pressure for participants to do well. This might have inflated some of the task difficulties or resulted in better performance for some participants—it's hard to know.