CS-639 Building User Interfaces, Fall 2020, Professor Mutlu

Dialogflow γ (3 Points)

Usability Testing

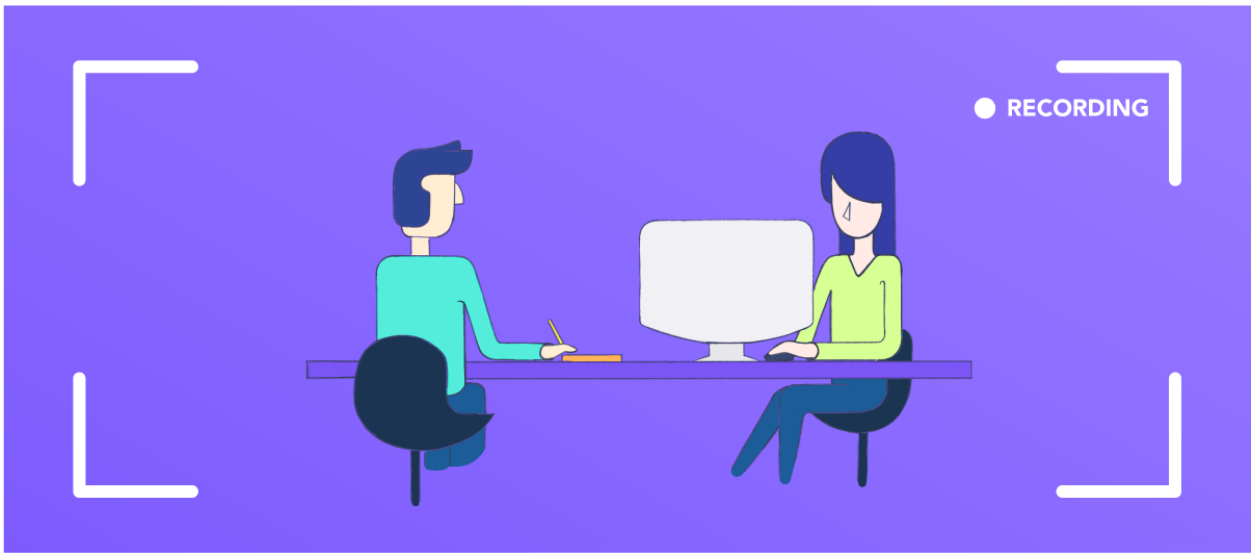


Image source

In this assignment, you will design and carry out a *mini* usability test of your Module 3 deliverable, *the shopping assistant*, in three parts:

Part 1—Designing A "Mini" Usability Test (0.8 Point): In the first part, you will make some decisions on the *why*, *what*, *how*, and *whos* of the study and write a two-page test plan that reflects your decisions.

Part 2—Executing Test Plan (1.4 Points): Next, you will recruit two volunteers from among classmates, family, and friends who can help you with your testing, and you will execute your test plan, over videoconferencing, to collect quantitative and qualitative data on the use and experience of the shopping assistant.

Part 3—Analyzing & Reporting Findings (0.8 Point): Finally, you will analyze your data and translate your findings into design insight.

Submission Details

Your deliverables for the assignment will be your test plan from Part 1, the data you collected in Part 2, and a report of your findings and a discussion of their design implications in Part 3, all as a single PDF document submitted to Canvas.

Note: Your assignment will be graded on the contents of this report and not the usability of your system. If you find that your agent is hard-to-use or unintuitive, you can be honest with your outcomes.

Part 1: Designing A "Mini" Usability Test (0.8 Point)

In this part, you will make some decisions about the format and design of a brief *formative* usability test and develop a *test plan*. First, you will determine two desired outcomes for your study. You can choose from five Es we have discussed in class (*effective*, *efficient*, *engaging*, *error tolerant*, and *easy to learn)*, the three dimensions of the ISO definition of usability (*effective*, *efficient*, *satisfactory*), or related concepts or outcomes (e.g., desirability, learnability, discoverability) that best fit to what you would like to evaluate. These will serve as your desired outcomes. Next, for each outcome, you will develop *questions*, *tasks*, and *scenarios* that will guide your testing. Then, you will choose two metrics: one performance, one self-report. Your deliverable will be a test plan that communicates these decisions and serves as a guide for the moderator (you) to run the test. Your study should be in the form of a remote *moderated* usability test conducted over videoconferencing, e.g., Zoom. The steps in the checklist below will help you in your decision-making and writing of your test plan and the form below that will help you draft your test plan. Your test plan should not exceed two pages.

Usability Test Design Checklist

- ❑ Choose two intended **outcomes**, e.g., effective, efficient, engaging, error tolerant, easy to learn, usable, satisfactory, etc.
- ❑ For each outcome, formulate a **question**, e.g., "To what extent are users satisfied with the shopping assistant" or "What is the overall usability of the shopping assistant?"
- ❑ For each question, devise a **task** using your shopping assistant that can help you assess how well your design meets the outcome. The task description should capture what you expect the users to do to successfully perform the task.
- ❑ For each task, develop a **scenario** that will provide context and guidance to the user. The scenario should prompt the user to perform the task you developed.
- ❑ Choose two **metrics** for measurement: one performance, one self-report. Examples of performance measures include task success (e.g., number of task substeps completed), time (e.g., seconds), or errors (e.g., number of deviations from expected use). For self-report measures, you can use the SUS questionnaire or all or part of the USE questionnaire.
    - ❑ Templates for SUS and USE.
- ❑ Write out your **test plan** using the form on the next page. Your plan should have three sections: (1) overview, (2) study design, and (3) test procedure. The overview section will briefly describe the context (including the "what" of the usability test, i.e., the scope of your interim or final design), the general goals for the testing, and the intended outcomes of the test. The study design section will outline your questions, tasks, and scenarios and your metrics. In test procedure, you will provide a step-by-step plan for the test in the form of a checklist.
    - ❑ You can see an example usability test plan from Barnum (2011) here. Your plan will not be as detailed as this example and should be *at most* two pages.

Usability Test Plan[1]

Overview

The test focus on effective and error tolerant outcomes. Our test would test on how our features of the dialog flow agent functioning. Currently, our dialog flow agents was just implemented with based features to help users log in, get queries, conduct actions such as add/remove tags, add to/remove from cart, review and checkout cart, and navigate. However, although the agent is trained with some sample inputs, it sometimes still does not understand users' inputs and would behave inaccurately. Therefore, during this test, we want to find the possible functioning problems for our main features as mentioned above. It would be remote and conducted through Zoom. With this study, we hope to collect information that would help us improve and refine the functioning for our features to provide more accurate responses or actions. We hope to answer the question: "Can our dialog flow agent complete the tasks accurately based on users' inputs?" (effective) and "Can our dialog flow agent let users recover from their errors easily and accurately?" (error tolerant.)

Study Design

Effective: Can our dialog flow agent complete the tasks accurately based on users' inputs?
Tasks: Find and purchase a product. More specifically, log in, narrow down the search, get information about a product (products) that the user narrows down to, add to cart, review and/or modify the cart, and confirm the cart at last. This task would cover most of our main features of the agent. So, we can check whether our dialog flow agent can function accurately for its main features.
Scenario: This is the users' first time to our shop. He/she want to purchase a short but do not have any ideas. He/she only has some requirements about the short in mind but do not have any specific information about a specific product. Therefore, he/she need the shop assistant to help him/her find the product to purchase.

Error tolerant: Can our dialog flow agent let users recover from their errors easily and accurately?
Task: Understand and recover from the mistake by providing wrong information when asking and finding information about a product. More specifically, ask for info about a product but does not know that the user himself/herself provides the wrong info, understand that he/she provided the wrong product name, want to see products for a category but give wrong category name (e.g. want to check out shorts, but the categories that have product for short is bottoms), search with wrong tags to narrow down the products, understand that provide the wrong tags, recover and search again with correct ones. In this way, we can understand whether our agent could detect, identify, and help user recover from errors.
Scenario: The user was introduced to our shop by his friend and know that we provide a good short. So, he/she want to check out the short recommended by his/her friend. However, his/her friend mixed up the

[1] Or use the Usability Test Plan template

short with other products and provide him/her with the wrong name of short. Therefore, the shop assistant need to help him/her recover from the wrong information and find the product as needed.

metrics for measurement:
Performance: number of times that the dialog flow perform actions or gives responses that deviate from users' intents and in what situation does the agent fails(error)
Self-report: SUS questionnaire with template given in the url.

Test Procedure

As in the actual test, we want to have a step-by-step plan. It would be better for us to test for both effective and error tolerant together. As our two scenarios are not many conflicts with each other, we can put them together and provide a smooth plan in one iteration for the shopping process.

In the below: texts in blue would meet the tasks and scenarios for the error tolerant and texts in black are for those of effective outcome

New Scenario:
This is the users' first time to our shop. He/she was introduced to our shop by his friend and know that we provide a good short. So, he want to check out he short recommended by his friend. However, his/her friend mixed up the short with other products and provide him/her with the wrong name of short. He/she want to purchase a short but do not have any ideas as the name of the product from his friend is wrong. He/she only has some requirements about the short in mind but do not have any specific information about a specific product.

Steps:

1. Log in to user's account (can just use a test account provided by us)
2. Ask for information of a product but provide a wrong product name
3. Want to check out products with wrong category provided
4. (Recover from the error), go to the page for the correct category
5. Narrow down his/her search by adding tag(s) but provide wrong tags
6. (Recover from the error), narrow down with correct tag(s)
7. View the page(s) for the product(s) meet his search criteria
8. Add the product(s) to cart
9. Review the cart and confirm the cart

Part 2: Executing Test Plan (1.4 Points)

In this part, you will identify two volunteers to help you test your shopping assistant over videoconferencing, e.g., Zoom, Microsoft Teams, Webex, etc., choosing a system that allows remote control of your computer (see documentation on conducting remote sessions where you give control of your computer to your partner for Zoom, Teams, Webex). They can be your classmates, friends, or family members. It is acceptable to pair up with a classmate and trade taking each other's test. You can use any version of your shopping assistant as long as you have a working prototype and choose to focus on any aspect of it. You can capture performance measures during the test, e.g., by timing them, counting errors, taking notes, or by recording them and watching later. You can present self-report measures on paper or on a computer screen after they perform all scenarios. Finally, be sure to make qualitative observations and ask questions, e.g., "you seemed surprised by that response, what were you expecting," to your participant where appropriate during and/or after the study. The deliverable for this part will be your data in table and/or text format pasted below. For performance, questionnaire, and qualitative data, provide the raw numbers or text that you will later organize and analyze in Part 3.

I conduct the test with two of my friends, on is a Senior in Computer Sciences and another is a Junior in Math and Economics. Both of them are Chinese and have experiences on "talking" to a dialog flow agent in some websites. My friend in CS said that the majority of her time talking with the agent is below her expectation. My friends in Math and Econ said that she was able to get the answer she want for about half of her time with agents.

I ask questions such as whether the user thinks the error was surprising when error happens; what seems to be confused and how do the user feel when the user seems to be stuck; and ask whether the user think the previous steps is burdensome when find that they seems to be unsatisfied from their facial expressions.

My friend in CS:

Errors: 2

1. When saying "I want to see shorts", agent called a default welcome intent.
2. When saying "go cart" mistakenly instead of "go to cart", the agent called the confirmCart intent.

SUS form[3], total score: 80

	Questions							
1	I think that I would like to use this system frequently.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
2	I found the system unnecessarily complex.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*

[3] Based on Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

Scoring Guide for SUS

1. For odd items, subtract one from the user response.For even-numbered items, subtract the user response from 5.
2. The subtractions scales all values from 0 to 4 where four is the most positive response.

Add up the converted responses for each user and multiply that total by 2.5, which will convert the range of possible values from 0 to 100 instead of from 0 to 40.

3	I thought the system was easy to use.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
4	I think that I would need the support of a technical person to be able to use this system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
5	I found the various functions in this system were well integrated.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
6	I thought there was too much inconsistency in this system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
7	I would imagine that most people would learn to use this system very quickly.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
8	I found the system very cumbersome to use.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
9	I felt very confident using the system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
10	I needed to learn a lot of things before I could get going with this system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*

Qualitative data:

My friend get confused and do not know what to do after ask the agent to find a wrong product. The agent gave response "Sorry, the category or product info about wisconsin short seems not to be exist. Please try again." after she said, "show me info about wisconsin short." She think the agent did not provide her with enough prompt or advise about what to do next and stuck for a while when know nothing to do next.

She said that similar situations would apply to some other intents. For example, when asking to navigate to a category that does not exist, the agent only let her try again but did not provide enough information about what she can do. (She then ask for what products that the shop have to know the categories).

When she tried to add a tag that does not in the category page (add white when at the bottoms page), the system does not tell her that white is not a tag, but it is successfully added. However, the products in the bottoms page all disappear. She was confused by this. (This is due to the problem that white is a tag, but no bottoms has the tag white, i.e., add a tag that is used to filter other categories)

When she ran into the errors as mentioned above, she was not really surprised as she went into similar situations in other websites and understand this kind of drawbacks. She said that there might be some probem such as giving some training phrases that seems not related to the intents for the confirm cart, but not enough training phrases given to the navigation intent.

My friend in Math and Econ:

Errors: 1

1. When saying "check out my cart", the agent called clearCart intent

SUS form[3], total score: 85

[3] Based on Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

Scoring Guide for SUS

3. For odd items, subtract one from the user response.For even-numbered items, subtract the user response from 5.

	Questions							
1	I think that I would like to use this system frequently.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
2	I found the system unnecessarily complex.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
3	I thought the system was easy to use.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
4	I think that I would need the support of a technical person to be able to use this system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
5	I found the various functions in this system were well integrated.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
6	I thought there was too much inconsistency in this system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
7	I would imagine that most people would learn to use this system very quickly.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
8	I found the system very cumbersome to use.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
9	I felt very confident using the system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*
10	I needed to learn a lot of things before I could get going with this system.	*Strongly disagree*	**1**	**2**	**3**	**4**	**5**	*Strongly agree*

Qualitative data:

My friend is not surprised about the error. She said that this is understandable that the agent would make suck kind of error, but she did say that mistaken “check out the cart” as clear the cart seems to be not usual and should be fixed.

She also think some parts are cumbersome. For example, when she want to check the tags, she did not input category to check tags from. In this case, the agent let her try again and input category. She said that it would be much better if agent ask her for category and she can just input category instead of asking for the whole sentence again.

Similar to my friends in CS, she also thinks that the some responses are not helpful enough to let her recover from her errors.

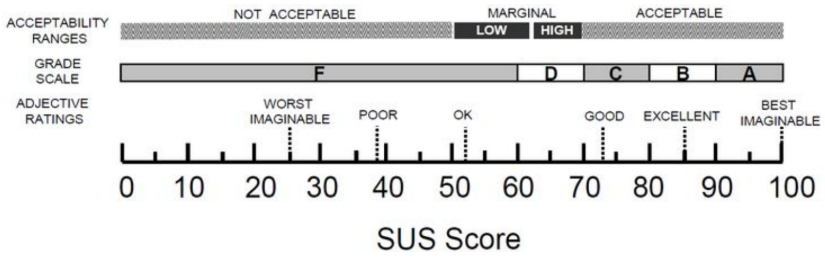
She also have trouble when telling agent to add a product to the cart. She want to add the Wisconsin Qualifier Woven Short but the product is so long that she did not spell it correctly for some time. In this case, the agent just tell her that the product is not found. This make her confused as she thought that the product is exist and did not realize that she spell it incorrectly. She said that it would be better to have the agent be able to find the product that match most instead of forcing the user to spell correctly everytime.

4. The subtractions scales all values from 0 to 4 where four is the most positive response.

Add up the converted responses for each user and multiply that total by 2.5, which will convert the range of possible values from 0 to 100 instead of from 0 to 40.

Part 3: Analyzing & Reporting Findings (0.8 Point)

In this part, you will clean, consolidate, and analyze your results and translate them into design insight. For your quantitative data, calculate the average values from your metrics and report the averages. For self-report data, if you used SUS, follow the scoring method included in the template and give your shopping assistant a grade (e.g., "D") and level of acceptability (e.g., "high marginal") using the guide below.[2] If you used a subscale of USE, such as "ease of use," average out the scores for all items to arrive at a single value and average out the values for both of your test participants. For qualitative data, categorize your notes and observations into a minimum of two high-level findings. If the quantitative data or the qualitative comments from your two participants vary significantly, you can also comment on these differing views. Report your findings in narrative form and end your report with high-level design insight and recommendations for how your shopping assistant might be improved. Your report should not exceed a page.



---

[2] Based on Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, 8(2), 29-40.

Usability Findings

Quantitative Summary

Performance: 2 errors and 1 error for 2 tests. 1.5 errors on average.
Self-report: 80 and 85 on SUS. 82.5 on average. This receives a grade B and is acceptable.
The quantitative data shows that the error rate is low, and the overall implement is acceptable with more improvements can be made. The errors mostly occurs in the navigation and confirm cart intent.

Qualitative Summary

1. Both users believe the error is not surprising but some of them seems to be unusual. Specifically, the training process to map users' input to intents seems to have the most problems and the problems lies mostly in the navigation and confirm cart intent. There might be problems such as giving some unrelated training phrases to the confirm cart intent but not enough or not related enough phrases to the navigation

2. Both users think that there are some responses are not helpful for them to recover from errors. (more details in the qualitative data sections)

3. One user thinks the system is a little cumbersome. She hope that when the agent found some missing info, she only need to provide the info needed instead of sending the whole input again. Also, the agent should be "cleverer" to tolerate some mistakes, such as spelling mistakes, in the inputs.

Conclusions

1. There might be some problem when defining the training phrases. We could refine the training phrases by checking again whether the phrases are all related, add more phrases, and add some phrases that could be confusing to the intent that it should be map to. In this way, the agent could be better trained and understand users' input better.

2. The responses to inputs with missing messages or wrong messages do not provide enough information for user to recover from their errors. We could refine those responses by adding related and useful information. For example, when the webhook does not find the categories that user want to navigate to, it could include the categories in the responses to let users know some possible navigation options.

3. The agent is not able to handle follow-on questions. We could let the webhook add context to the agent and generate responses or perform tasks with those contexts to let users have follow-on inputs to avoid entering the whole sentences again. Also, we could add more synonyms or have some algorithms to find the items that match the input most in the database in the future.