In this lecture we discuss considerations in making questionnaires. Most human computer interaction experiments will use questionnaires. Typically they serve two main purposes:

1) To gather demographic information, in order to find if the participants are representative of the population you intend to study. The questions can pertain to age, gender, experience in certain areas, etc. You are trying to get to know your participant with these questions.

2) To gather responses about preferences, of the participant. Through the use of correctly designed questionnaires we can quantize a subject’s responses (as opposed to attempting to interpret free form style answers). These surveys are designed to minimize social desirability (meaning the tendency for people to want to say things that are considered to be more acceptable or closer to the norm). Minimizing social desirability can be especially useful when approaching a potentially sensitive topic.

Questionnaire Design Concerns

Here we will discuss practical design concerns in creating questionnaires.

1) Formatting of the Questionnaire

We want to focus on:

- Ease: Responses to questions should be easily recorded and simple to sift through in order to draw conclusions.
- Reliability: The format should bring about reliable data that is directly comparable to responses from other subjects
- Reaching the Population: The survey should be available to all parts of your population. Electronic communication can make this process easy but may also leave out certain populations thus potentially skewing data.

We may implement metrics to help determine which parts of the data are valid in comparison to outliers. As mentioned above when we put our questionnaire online we self-select the population by limiting access to only those that have online capabilities. Online surveys have the advantage that they are easy to handle and avoid legibility issues, while they maintain the disadvantage in that the parts of the target population who cannot get internet access will be left out. Using paper (instead of online surveys) will have a much larger population of ascertainable populations but still can suffer from legibility, distribution, and interviewer presence issues. As was mention above, social desirability (among other swaying factors) could cause a subject to respond differently than if they were alone.
An example of improving formatting concerns would be to implement the survey in multiple languages. If your desired population spans multiple languages, then this clearly helps to reach the population and another benefit would be easier for a participant to read if their first language is not that of the questionnaire creator.

2) **Page Layout**

The layout of the page, questions, and answer selections can affect how a participant may respond. Thus, the page layout can increase or decrease the reliability of the results. For example, some multiple choice answer selection pages will alternate between different background colors per row of answer selections, so that there is no confusion as to lining up the answer selection with the question number for which the answer was intended.

An example of a page layout which could bring about less reliable results would be one with inconsistent spacing among answer selections. A depiction of this is seen below.

![Example of inconsistent spacing among answer selections](image)

The spacing in the text may lead the participant to believe that Agree is closer in meaning to (or grouped together with) “Strongly Agree” as opposed to “Undecided”. To dissect this example even further we can pay attention to the terms used. The term “Undecided” could be interpreted differently than the term “Neutral”, this may affect the outcome. The use of capitalization for some terms as opposed to others could also be seen as a way to unintentionally emphasize a specific answer.

3) **Ordering of Questions**

The question presentation order can affect how a participant may respond. It is generally better to place more neutral questions in the beginning and more intrusive ones at the end. An example would be, asking someone about their income as the first question. Some may not answer at the far ends of the income amount spectrum (if they are rich or poor), while middle incomes may be more likely to answer. This could skew the results in that a subset may be more comfortable with answering truthfully (if at all). Also, if the question were intrusive enough to come off as offensive then this may lessen the willingness of a subject to participate or may cause them to answer differently.

4) **Open vs. Close Ended Questions**

Open ended questions (those having no preset answer selections) give the participant the freedom to express their thoughts as they see fit and may elicit some responses that the questionnaire designer(s) may not have thought of. The problem is that the open ended question responses are not standardized and thus are more of a challenge to compare with
other subjects. Close ended questions tend to be quicker to answer and lead to much more comparable and consistent (less mood based) results.

5) **Rating vs. Ranking**

Say you are asking about food preference. You could say rank these in terms of favorites (carrots peas, etc.) or you could ask the participant to rate each individually using a scale of enjoy ability.

A typical pitfall to the rating system is that participants tend to pick some median rating on an item and then rate the rest according to (relative to) that medium rated item. This can lead to less reliable results. Thus generally the ranking system works better.

Should a rating system be used in a questionnaire, there are two types of scales to choose from.

- **Unipolar scale**: A unipolar scale only moves in one direction, in that it will either go from 0 to 5 or from 0 to -5 but not in both directions. It has been shown that a 5 point Likert scale gives the best unipolar rating results.
- **Bipolar scale**: A bipolar scale is different in that it spans in both directions. For example a bipolar scale can span from -5 (disagreeing) to 5 (agreeing). It has been shown that a 10 point Likert scale works best for bipolar ratings.

For either scale it is useful to attach words to the numbers and make sure they are semantically evenly distributed.

Below we will depict and discuss an example of a unipolar scale:

The context of this answer selection could be asking a user to rate the usefulness of a tool in a specific learning task.

```
0  Not Useful  1  2  Moderately Useful  3  4  5  Apotheosis of Usefullness
```

While the designer may have wanted to emphasize how truly high a rating of 5 could be, this can skew results in that the use of diction that may be unfamiliar to some could cause them to avoid that answer. Thus “Apotheosis of Usefullness” may be better put as “Extremely Useful”. Also, there may be added utility to lining up the semantic meaning with numbers by having a one to one correlation of meaning (to a selection number).

Below we will depict and discuss an example of a bipolar scale:

```
0  Abysmal  0  Poor  0  Neutral  0  Good  0  Very Good
```
It is good to use words to quantify (instead of numbers) but the words should be symmetric, unlike the words above. It would have been better to use the term “Very Bad” in place of “Abysmal” (for symmetry with “Very Good”), and “Bad” in place of “Poor” (for symmetry with “Good”).

6) **Repeat The Concept**
   Even a carefully worded question still is up for interpretation by the listener. Thus it is useful to repeat the same concept in an attempt to drive out the possibility of ambiguity. An example question for which this is useful would be: “Does this avatar/character appear to be natural?”. This could be interpreted as “Does the rendering of the character look realistic?”, “Could this be a part of our real world?”, or “Given the context of the character, does this character seem realistic?”. Thus we want to repeat the concept several different ways in order to isolate our meaning. Another useful technique is to flip the way we ask the question later. For the given example, this could be done by asking “Do we not believe this character is realistic?”.

7) **Group Your Topics**
   If your questionnaire spans multiple topics, it can be useful to group the questions by their topics. That can make the cognitive load on the participant much more manageable while also lending to easier data management with related questions being adjacent to one another.