Hello class, and welcome to our first of several voice-over-slides video lesson sets.

This set will provide an overview of some ways in which we can evaluate interfaces with our users.

I have divided it into smaller sub-videos in an attempt to make it easier to view them and refer back to them.

A later lesson set will explore a technique for HCI experts to evaluate interfaces.

As with many things, there are trade-offs between approaches and we look for the best fit in a certain situation.
Recall that evaluation is an integral part of the iterative design lifecycle, and this is part of an overall software engineering lifecycle.

We can, and should, evaluate on low-fidelity prototypes and on our medium and high fidelity versions as we implement a new vertical drop.

Evaluation continues to happen all the way through to acceptance testing, and then really keeps going as we think about a “2.0” version.
Please take a moment to look at this real snippet from a real grading platform.

Pause now. Welcome back. Did you notice that it says submissions 1 through 25 of 22 are unassigned?

How would you express your feelings about what is wrong here and why?
This is a photo of the Iribe first floor map from last Spring. The red areas were still under construction as classes were being held in the green areas.

What usability issues do you see?

Please pause to look.

Welcome back. Even without being in the building at the time, you should be confused by why the arrows for 1116 and the restrooms seem to contradict the map image. The reason is that the arrows were correct for where the map was located, the map orientation wasn’t aligned with reality.

Again, how would you categorize and explain this issue?
An overarching question we have hinted at is that of defining success from an HCI point of view. We don’t want to restrict ourselves to hindsight of “well, it made a great profit and is still around” for two reasons. First, we’d like more confidence before releasing a product. Second, software like Windows and Canvas would be booming HCI successes by that measure ;)

These five metrics are common ones by which we can judge either a system as a whole, or a feature within a system. You can imagine that tradeoffs occur. For example, a system that takes longer to learn how to use might also have a low error rate. Similarly, maybe a low error rate is achieved by slowing the user down as they undertake a task to avoid errors.

Take a moment to try to come up with some examples of where a trade-off might happen in a product or feature’s design.

Pause now. Welcome back. Launching a nuclear weapon might take checks and rechecks that slows down the performance time of the task, but you’re really looking for low error rate there.
In general, when looking at ways to evaluate a system, there are two broad categories; qualitative and quantitative.

Qualitative is sometimes referred to as being more naturalistic. It can do a better job of describing ongoing processes rather than atomic elements of a process. Consider the “real life” process of filling out your mailing address in a form and saving it versus the atomic elements of that such as “clicking in a box to type” or “clicking in a button to save” and which matters to you in your evaluation context.

These naturalistic/qualitative approaches tend to be seen as more authentic for real situations.
**Approaches: Experimental/Quantitative**

**Experimental**
- study relations by manipulating one or more *independent* variables
  - experimenter controls all environmental factors
- observe effect on one or more *dependent* variables

**Internal validity**
- confidence that we have in our explanation of experimental results

**Trade-off: Natural vs Experimental**
  precision and direct control over experimental design
  *versus*
  desire for maximum generalizability in real life situations

However, there are times when you want to undertake more precise experiments, such as when deciding what the best default height of a GUI button would be.

This is where we start to think about how we can manipulate something and observe the effect of that manipulation.

A big “pro” here is that if we get conclusive results, we know exactly what lead to them.

In reality, both families of approaches have their place, and some research employs what is know as a “mixed methods” model where elements of both are used.

We might categorize our users into three age groups, such as 7, 9, and 11 year olds and then look to see whether any of the things we are observing, even if they are qualitative, have a correlation to that age variable.
One question that often comes up is whether reported results are reliable. Would the same results be found if the tests were run again with a different group of users?

There are techniques that you might have learned about in a statistics course that come into play here.

We will not employ this in our projects this semester, but it is something worth thinking about in general.
Another question that often comes up is whether the reported results have relevance to real-world uses. The qualitative side of things can help with this, but not always.

Even if you have real people performing real or realistic tasks, the setting could be inauthentic. Take a moment to consider the list on this slide.

Pause now. Welcome back. We could imagine Dilbert’s colleague Wally testing noise-cancelling headphones in a library (well, maybe not McKeldin) and declaring success even if nobody figured out how to turn them on!

In the next video file, we will explore some qualitative methods for usability evaluation in more detail.
Hello class, and welcome to the continuation of our first of several voice-over-slides video lesson sets.

In this sub-part we will explore details and variations of two common qualitative methods used for usability evaluation; introspection and direct observation.
There are actually three method groups we will explore, but in this segment we will cover the first two.

The third, query, will be in the final segment of this video set.
First, let’s look at the idea of introspection and two variations on it.
Members of a software design team can carry out introspective evaluations on their prototypes.

It can be as simple as walking through some representative task scenarios, paying attention to whether the progression feels natural.

In reality, we do this all the time, but it has shortcomings.

The assessment is very subjective, similar to reading an essay you wrote aloud to see whether it sounds right to you. It might just be a case of confirmation bias.

Also, while you might think you can imagine being a typical user, in reality you are non-typical even if you are good at roleplaying users.
**Introspection Method: User**

Typically done as a user-centered walkthrough of a system. The idea here is typically one of conceptual model extraction by showing representative users prototypes or even screenshots of a mock-up.

- Can ask the user to explain what each screen element does or represents as well as how they would attempt to perform individual tasks.

This can allow us to gain insight as to a user’s initial perception of our interface and the mental model they might be constructing as they begin to use a system.

**NOTE:** Since we’re walking them through specific parts as their guide, we won’t really see how a user might explore the system on their own or their learning processes.

Having representative users carry out introspective evaluations on prototypes, whether low fidelity or higher, can provide valuable insights and is another aspect of including users in the iterative design process.

Even running a like, dislikes, and design ideas sticky note session with screenshots can provide useful insights regarding how the users will initially perceive things, and provide information about their anticipated use of the system being built.
This next qualitative approach set has more costs in both time and resources, but can provide very useful “holistic” evaluative information. Direct observations can be done either in a lab setting or “in situ” where a user would realistically undertake their tasks. The utility of the observations are often strongly tied to how the situations are designed and executed.

A study on how children search for information online done by going to the child’s house (with proper ethical review and parental permission of course) to watch them undertake searches on their own computer might reveal things that an in-lab trial would not. For example, in their room on their computer they are surrounded by their own typical distractions. In a lab setting the participants might all stay on-task, where in a more ethnographic setting, a child asked to find out what dolphins eat might go to a bookmarked SpongeBob game, not because they think he would know the answer but because they used up their game-playing quota for the week and wanted to sneak in one more game.
Within this category, there are three sub-approaches to running a session; simple observation, think-aloud, and constructive interaction.
With the simple-observation approach, you provide a user with tasks to perform, and watch what they do to assess the interface.

One challenge this presents is that you likely have to guess why they did certain things.

If that SpongeBob child had been in a simple observation trial, how would it have been understood why they went to that game. By coincidence the task was searching for what dolphins eat and the observer could have thought the child made a clever connection.
Instead, a slightly different approach is more commonly used where the user is asked to self-narrate what they are doing, and the observer can prompt them to speak if they are being silent. This is called the think-aloud approach.

While it can be awkward, a simple warm-up exercise can help. In a moment, pause the video and pick a place where you lived and count the number of windows the building had. Do this out loud, counting and describing as you count. For the full effect, do it with another person in the room, at a social distance from you of course.

Pause to try that now. Welcome back. Did it get less awkward. Did thinking it through aloud make the process any easier or harder?

This might decrease realism with tasks having a high cognitive load, by distracting the user, but it could also help the user in an inauthentic way.

Consider the “tell it to the duck” method of debugging, and whether you do that normally or not.
One final sub-approach here is to utilize constructive interaction where you ask two people to work together on a task, hoping to get the conversational aspect of think-aloud for free and more naturally.

Of course, this can introduce interpersonal issues, and can detract from realism when the tasks are normally solitary ones.
Co-Discovery

A variant of constructive interaction is to have co-discovery learning take place, where the pair working together are:

- a semi-knowledgeable “coach”
- a beginner (who is actually using the system)

Ideally, this results in
- the “naïve” beginner participant asking questions
- the semi-knowledgeable “coach” responding
- insights into thinking process of both beginner and intermediate users

One variation on the constructive interaction model is to create a setting of co-discovery.

While this still might not be realistic, it can provide insights into both confusions that a novice might have and how a more advanced user might describe things. That last part might be useful for providing help or rewording parts of the interface.
Recording Observations

Make sure you get permission!

Make sure you are mindful of privacy!

Whichever of the direct observation approaches you might use, it is common to want to record the session’s audio and/or video so that you can review it later, possibly to transcribe dialog and possibly to codify the behaviors you see in detail and have another person confirm your observations.

It is crucial to get permission to do this and to consider the user’s privacy in what you do with the recording.
When AV recording is not allowed, you might resort to designing a coding sheet to make live annotation more feasible.

You might also have a team of observers split the job by focusing on different things about which you want to keep a record.
The tools available to you are wide-ranging and changing constantly. Recording technically covers any record-keeping, including simply taking notes as the session is in progress. We more commonly think of AV recordings and there are many technologies available for that.

There are also services you can pay to transcribe your recordings, but again you need to consider user privacy and the conditions under which you were allowed to make the recording. This is a topic we will explore more later in the semester under research ethics.

Some researchers find that for every hour of recorded sessions they have, they might invest 5-10 hours in reviewing, transcribing, and codifying it, as well as looking for patterns and having others review the footage to confirm the observed patterns.

In the next video file, we will explore sways in which we can query users.
Hello class, and welcome to the conclusion of our first of several voice-over-slides video lesson sets.

In this sub-part I will discuss some of the ways to query users, along with some advantages and disadvantages of the approaches.
**Querying Users: Interviews**

**Excellent for pursuing specific issues**
- vary questions to suit the context
- probe more deeply on interesting issues as they arise
- good for exploratory studies via open-ended questioning
- often leads to specific constructive suggestions

**Problems:**
- accounts are subjective
- time consuming
- evaluator can easily bias the interview
- prone to rationalization of events/thoughts by user
  - user’s reconstruction may be wrong

One solid use of this approach, and one that you have already used to a degree, is the interview.

The questions you ask can easily be varied to match the context in which you are doing the evaluation.

As the participant replies, you can easily extend your questioning on specific aspects of the interface.

The one-to-one nature makes it more likely that the participant will offer constructive suggestions to you.

However, this is not a fast process and the one-to-one nature can lead to the evaluator introducing biases and can lead to the participant to try to rationalize their feelings in a non-critical way.

In the next video lesson set, I will introduce a variety of psychological effects that can manifest themselves in an interview-based evaluation, as well as in some of the other approaches this set has explored.
**Querying Users: Structured Interviews**

**Plan a set of central questions**
- could be based on results of user observations
- gets things started
- focuses the interview
- ensures a base of consistency

**Try not to ask leading questions!**
- “Now that was easy, wasn’t it?”
- “How hard would you say this task was?”

**Start with individual discussions to discover different perspectives, and continue with group discussions**
- the larger the group, the more the universality of comments can be ascertained
- also encourages discussion between users

When undertaking an interview, you should plan out a central set of questions and follow that plan as precisely as possible to maintain consistency.

It is common to write an actual script to follow, even if there is only one person conducting the interviews.

When constructing your questions, avoid leading questions like, “Do you agree this was an amazing video set on this topic?”

Interviews can also move from one-on-one settings to small groups settings if you want to see what comes out of spontaneous group discussions.
An interesting approach to querying the user can be to record them as they perform tasks, then have them watch the recording and explain to you what they were thinking as they did the tasks.

You potentially get a more natural flow than with the think-aloud approach, while still getting insight into the users thought process.

The user can ask you to pause the playback to discuss or explain their thoughts on an element of the UI, which can provide more time for concrete suggestions than think-aloud would.

However, even only a few minutes later, the user might not be able to exactly reconstruct their thoughts of a particular moment.
It is tempting to jump to surveys and questionnaires since you don’t need a person there to administer it, and can potentially reach a much wider audience.

It also provides more raw data in a form that can be statistically processed.

However, there are large challenges to getting a good response rate from a representative slice of the population that you want (as opposed to Redditors or people on mass survey answering sites looking to earn 5 cents a survey, hoping to do 30-40 surveys an hour).
## Querying Users: Surveys and Questionnaires Details

**Establish the purpose of the questionnaire**
- what information is sought?
- how would you analyze the results?
- what would you do with your analysis?

**Typically will not ask questions whose answers you will not use**
- this is unlike many other types of surveys you may have discussed in a psychology class

**Determine the audience you want to reach**
- typical survey: random sample of between 50 and 1000 users of the product

**Determine how would you will deliver and collect the questionnaire**
- on-line for computer users
- surface mail (with pre-addressed reply envelope for better response rate)

**Determine target demographics**
- e.g. level of experience, age, income, etc.

This slide tries to provide an overview of things to consider as you build and administer a survey or questionnaire.

Please pause to review it.

Welcome back. The next few slides show some of the types of questions you could ask.

For each slide, it is probably good to pause and consider the example.
Open-ended questions take us back to the qualitative side of things, but can provide useful freeform insights.

**Styles of Questions (I)**

**Open-ended questions**

- asks for unprompted opinions
- good for general subjective information
  - but difficult to analyze rigorously

**eg:** *Can you suggest any improvements to the interfaces?*
**Styles of Questions (II)**

**Closed questions**
- restricts the respondent’s responses by supplying alternative answers
- makes questionnaires a chore for respondent to fill in
- can be easily analyzed
- but watch out for hard to interpret responses!
  - alternative answers should be very specific

Do you use computers at work:
- often
- sometimes
- rarely

In your typical work day, do you use computers:
- over 4 hrs a day
- between 2 and 4 hrs daily
- between 1 and 2 hrs daily
- less than 1 hr a day

Multiple choice questions require a lot of thought to design well, and often benefit from a test-run on, for example, a dozen participants where you ask them to comment on any shortcomings of the options you provided.

Consider the second example here. It might be better than the first since it provides more details in theory, but is “over 4 hours a day” really a good option when that range is likely to be picked often and could span from 5 to 12 hours quite easily?
The use of bipolar scaling might resolve some of the previous issues. This slide presents two examples.

The second is sometimes referred to as a Likert scale, but there are subtleties that differentiate the Likert approach from a general rating scale like these.

I’ll leave that to your PSYC 100 class if you take it.
Styles of Questions (IV)

Multiple choice (possibly multiple responses)
• respondent offered a choice of explicit responses

How do you most often get help with the system? (tick one)
☑ on-line manual
☐ paper manual
☐ ask a colleague

Which types of software have you used? (tick all that apply)
☑ word processor
☑ data base
☐ spreadsheet
☐ compiler

Here are some more examples of multiple-choice questions.
**Styles of Questions (V)**

**Ranked**
- respondent places an ordering on items in a list
- useful to indicate a user’s preferences
- forced choice

Rank the usefulness of these methods of issuing a command
(1 most useful, 2 next most useful..., 0 if not used)

- **2** command line
- **1** menu selection
- **3** control key accelerator

Something that can be very useful in situations is to ask participants to rank a set of choices.

I encourage having a “0” rank that can be used an unlimited number of times to avoid forcing users to rank options they just don’t want.
**Styles of Questions (VI)**

Combining open-ended and closed questions

- gets specific response, but allows room for user’s opinion

It is easy to recover from mistakes:

<table>
<thead>
<tr>
<th>disagree</th>
<th>agree</th>
<th>comment: the undo facility is really helpful</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3 4 5</td>
</tr>
</tbody>
</table>

Going with a hybrid approach can be quite effective at providing raw data that can be analyzed along with freeform responses that might provide useful contextual insights.
Survey Question Style/Wording “Risks”

When undertaking a survey, one of the challenges is often in the wording and ordering of questions.

Even a subtle difference that involves negations or the meaning of high and low numbers can have an impact.

This slide has the results of an in-class activity that I did on the 6Round exam cheating topic last year. These are authentic results.

Please pause to look at the two graphs.

Welcome back. Notice how it is close to the same, but not identical, even here when you throw a simple negation in.

This is one of the reasons that many surveys will ask the same question with different wording.
What might the future hold?

We live in a time where the use of AI is on the rise and chatbots are in the thick of things. We went from the Universal McCann agency’s Jill020306 to Microsoft’s Tay in the span of a decade. Where could they take us in the future? What are the ethical issues that could come out of this?

When you write a new library or program module, you can use unit testing tools to automatically assess the accuracy of various things. The tools continue to expand, even into the ability to automatically test GUI elements. Where might they go in the future?

Possible direction: “Chatbots: Your Ultimate Prototyping Tool”

https://medium.com/ideo-stories/chatbots-ultimate-prototyping-tool-c4e2831967f3#74ci9iy2n

The idea of replacing human participants with chatbots programmed to mimic humans feels like something out of a speculative fiction story, but there have been explorations in this already.

The linked article is an optional reading if you find this an interesting idea.
What you now know about...

Observing a range of users use your system for specific tasks can reveal successes and problems and qualitative observational tests can be quick (and somewhat easy) to do. Several methods can reveal what is in a person’s head as they are doing the test. Particular methods include:
- Conceptual model extraction
- Direct observation (simple observation, think-aloud, constructive interaction)
- Query via interviews, retrospective testing and questionnaires
- Continuous evaluation via user feedback and field studies

Evaluation is crucial for designing, debugging, and verifying interfaces

There is a tradeoff in naturalistic -vs- experimental approaches
- internal and external validity
- reliability
- precision
- generalizability

UP NEXT: ETHICS and EFFECTs!

One last pause-and-read slide for the set to try to wrap up some of the overall ideas of today, but the details matter so please review earlier slides to gain comfort with each method / approach.

Next up when we return virtually will be a video lecture set on ethics and psychology effects in human-subjects research.

These two sets have strong ties to each other.

I hope you are able to enjoy your break, and hope this and our upcoming virtual class sessions are successful at conveying the core ideas, even without me being able to interactively ask for your thoughts and examples in class.

The topics presented this way will set up Phase 3, and I’ll adjust Phase 3 if and as needed to reflect whatever the health protocols suggest at the time.