SWEN 422 Lecture 5
Evaluation 1

Dr Jennifer Ferreira

13 March 2024

Agenda

- Review of previous lecture
- Summative research in HCI
- Measuring usability
- Participant considerations

Formative Research in HCI

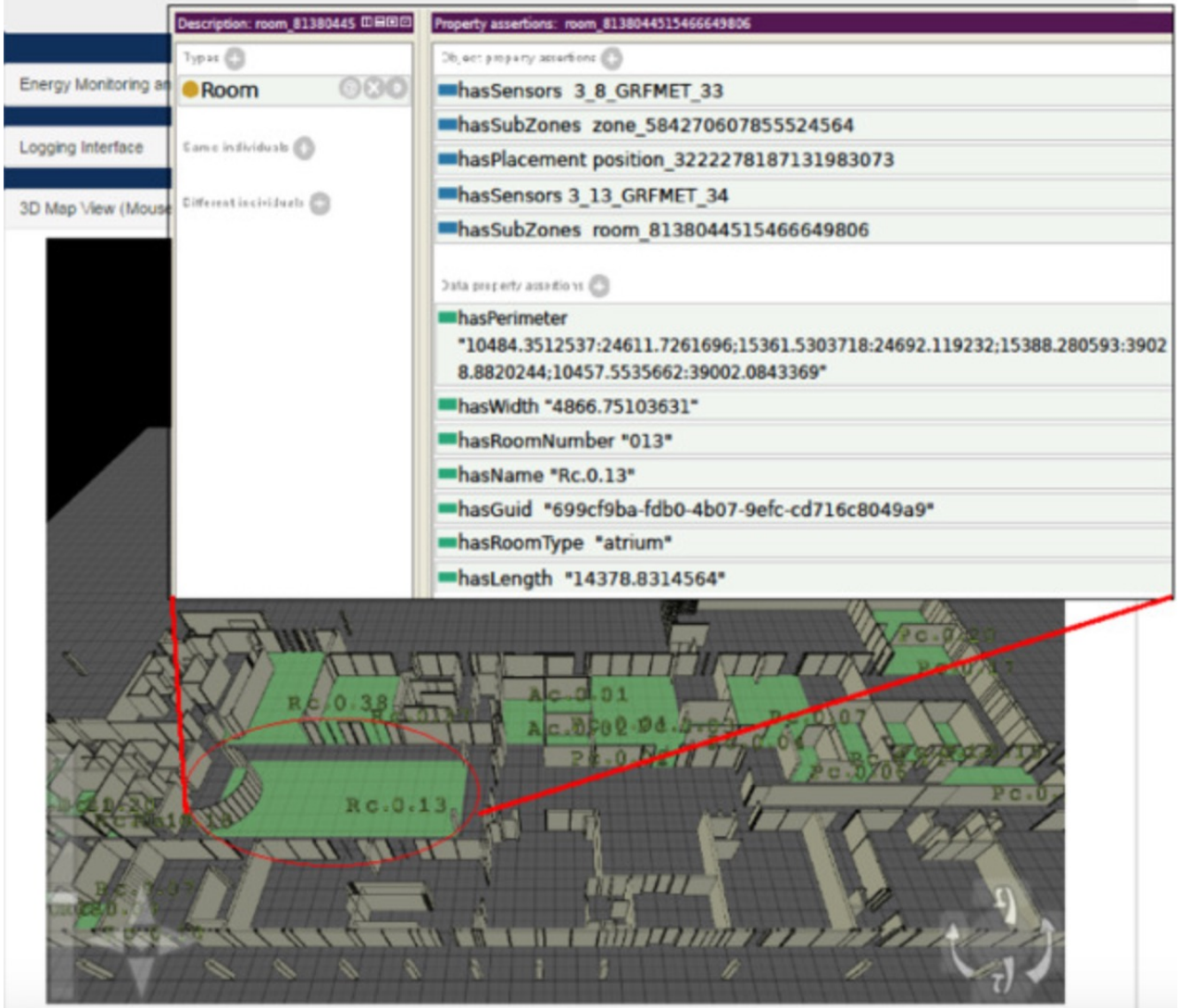
- Establish **relationships** between researchers and participants
  - Collaborative Research Stories: Whakawhanaungatanga
- Understand a **problem**
  - Why are people abandoning their online shopping carts?
- Understand **behaviour** in a **setting**
  - What are the experiences of emergency medical dispatchers?
- Understand **attitudes** to existing systems or situations
  - How do people view conversational agents (such as Siri)?
- Test an early (lo-fi) **prototype**
- Based on our findings we may want to
  - Formulate requirements for a system
  - Conduct further studies

Summative Research in HCI

- Understand **behaviour** with the new system
  - Is the web-based tool for supporting holistic building energy management usable?
- Understand **attitudes** to the new system
  - Will clinicians adopt the “CanRisk” tool for CanRisk tool for predicting risk of breast and ovarian cancer?
- Evaluate performance of a working **prototype**
  - Usability Assessments of STAR-Vote
- Based on our findings we may want to
  - Refine requirements for the new system
  - Establish that the new system is “better” / “usable”
  - Establish that the new system is “ready” for release / “fit-for purpose”
  - Establish that users will accept & use the new system
  - Conduct further studies

BuildVis Tool

The BuildVis tool is a multi-purpose tool for providing facility managers the capabilities to monitor the energy consumption of different building zones. The tool also provides suggestions on ways to improve the energy efficiency of the building.



Summative Research in HCI: Understand **behaviour** with the new system

Example Question: **What is the level of usability of BEMS BuildVis?**

- time to complete tasks,
- number of errors,
- whether a task is completed
- average satisfaction of users (using the System Usability Scale (SUS))

  (Based on previous work by J. Sauro, E. Kindlund. **A method to standardize usability metrics into a single score**. In Proceedings of the Conference in Human Factors in Computing Systems, April 2–7, 2005, ACM Publication, Oregon (2005), pp. 401-409)

McGlinn, K., Yuce, B., Wicaksono, H., Howell, S., & Rezgui, Y. (2017). Usability evaluation of a web-based tool for supporting holistic building energy management. *Automation in Construction*, *84*, 154-165.

Summative Research in HCI: Understand **behaviour** with the new system (McGlinn et al., 2017)

Evaluation in 2 parts:

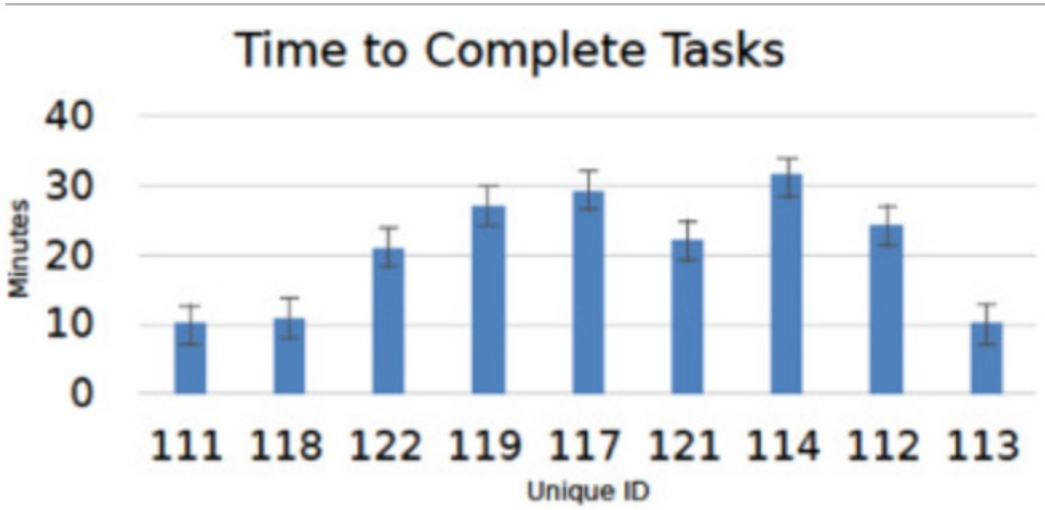
1. Participants with backgrounds in computer science, engineering, and related fields – 9
2. Participants who are facility managers – 5

Task-based

- Navigating the 3D building floor plan.
- Selecting a zone in the building and monitoring sensors related to energy consumption metering.
- Enacting suggestions from the real-time controller.
- Logging information regarding changes made to the building configuration related to those suggestions.

Summative Research in HCI: Understand **behaviour** with the new system (McGlinn et al., 2017)

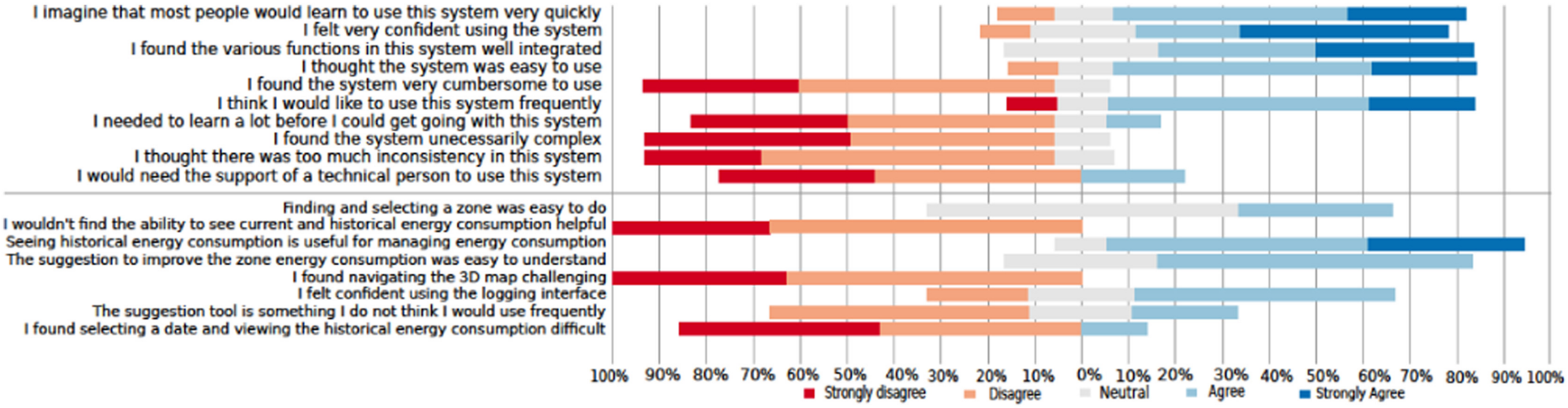
Results of technical users



Average time was 20.6 min with a standard deviation of 7.8 min.

Summative Research in HCI: Understand **behaviour** with the new system (McGlinn et al., 2017)

Results of technical users



The SUS scored 73.9

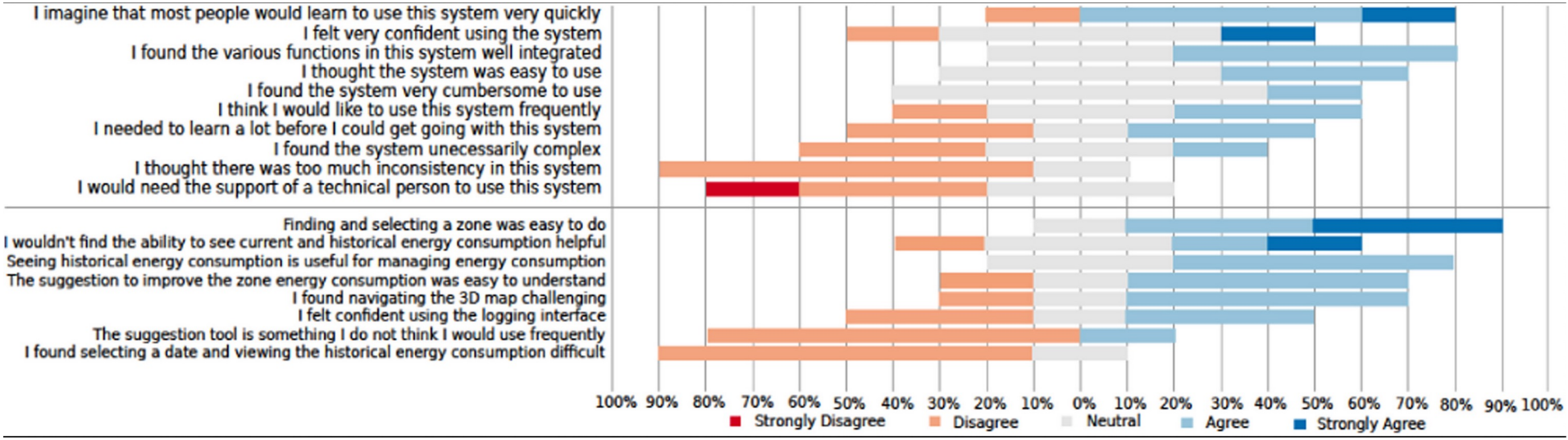
Summative Research in HCI: Understand **behaviour** with the new system (McGlinn et al., 2017)

Results of FM users

Average time was 13 min with a standard deviation of 4.24 min.

Summative Research in HCI: Understand **behaviour** with the new system (McGlinn et al., 2017)

Results of FM users

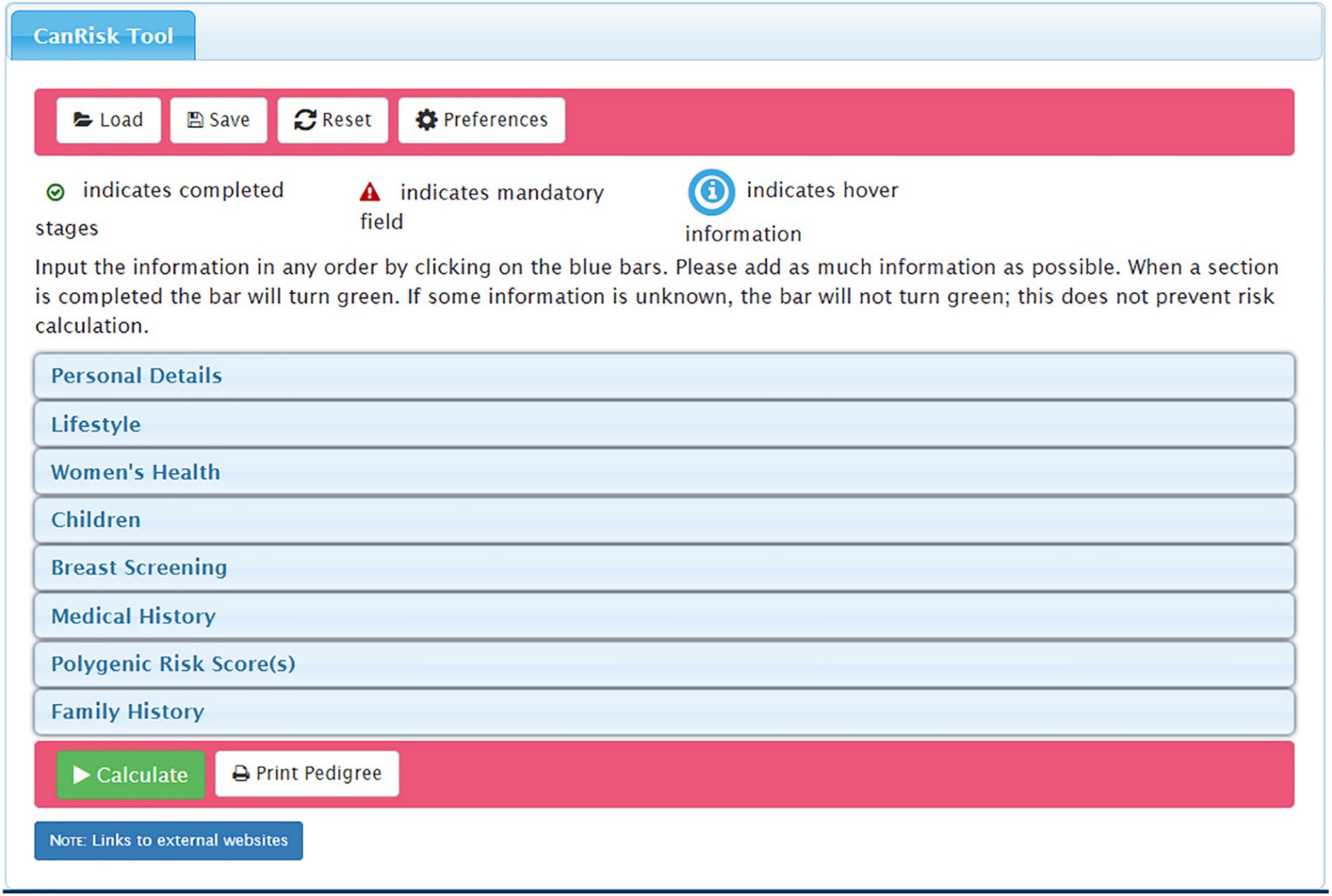


The SUS scored **59.5**

Summative Research in HCI: Understand **behaviour** with the new system (McGlinn et al., 2017)

Overall results

- Still usability issues for target users (facility managers)
- Suggestions for improvement
- Low error rate and low task completion time -> system is robust


CanRisk Tool
Load
Save
Reset
Preferences
indicates completed stages
indicates mandatory field
indicates hover information
Input the information in any order by clicking on the blue bars. Please add as much information as possible. When a section is completed the bar will turn green. If some information is unknown, the bar will not turn green; this does not prevent risk calculation.
Personal Details
Lifestyle
Women's Health
Children
Breast Screening
Medical History
Polygenic Risk Score(s)
Family History
Calculate
Print Pedigree
NOTE: Links to external websites

Summative Research in HCI: Understand **attitudes** to the new system

Example Question: **Will clinicians adopt the "CanRisk" tool for predicting risk of breast and ovarian cancer?**

- demographics questionnaire,
- two test cases (either face-to-face with a simulated patient or via a written vignette)
- semi-structured interview or equivalent open-ended questionnaire

Archer, S., Babb de Villiers, C., Scheibl, F., Carver, T., Hartley, S., Lee, A., ... & Walter, F. M. (2020). Evaluating clinician acceptability of the prototype CanRisk tool for predicting risk of breast and ovarian cancer: A multi-methods study. *PLoS One*, *15*(3), e0229999.

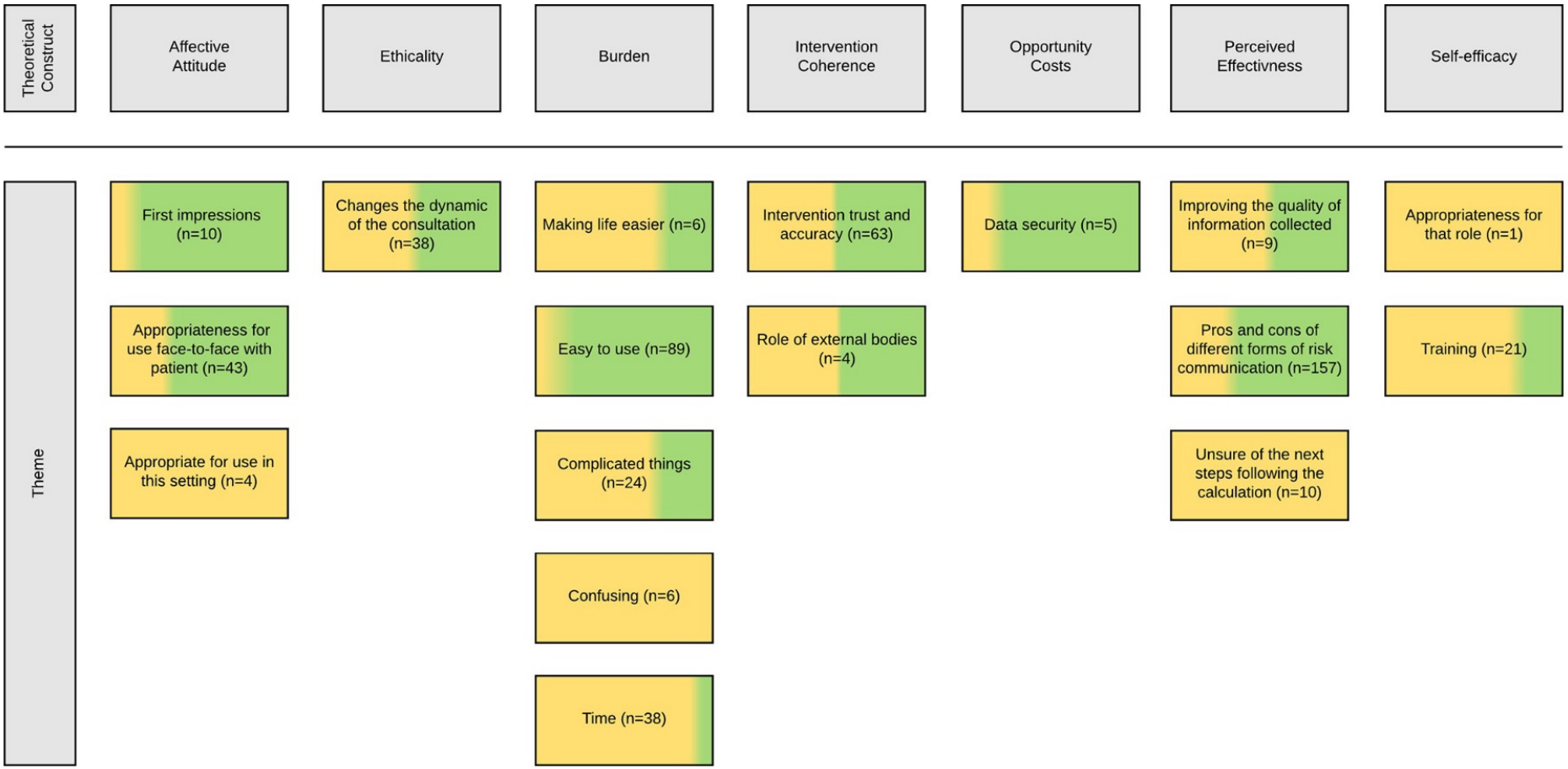


Fig 4. Inductive themes mapped onto Sekhon et al's Theoretical Framework of Acceptability [12]. The total number of times the theme occurred is presented after each theme name (e.g. n = x). The yellow (primary care) and green (specialist genetics clinic) colouring shows the proportional contribution of each sample to the theme.

https://doi.org/10.1371/journal.pone.0229999.g004

Sekhon's Theoretical Framework of Acceptability (applies to the medical context)

Technology Acceptance Model (TAM)

Developed through research on adoption of email as a mechanism of communication and of a graphics software package with a menu interface (Davis, 1989) and (Davis et al., 1989).

"theory that models the decision-making process by which users may or may not adopt and implement a new technology"

Perceived Ease of Use: "the degree to which a person believes that using the system will be free of effort."

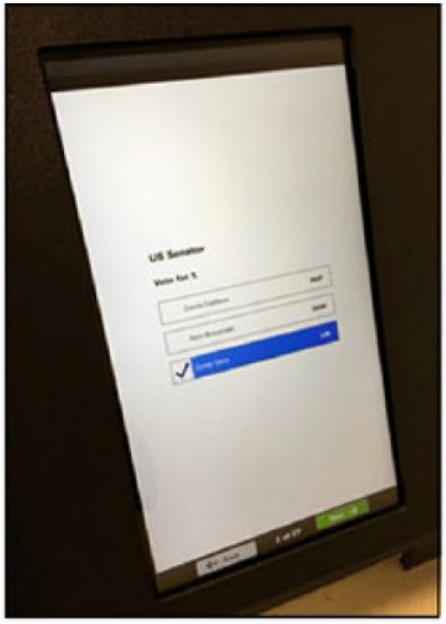
Perceived Usefulness: "the extent to which a person believes that using a particular technology will enhance her/his job performance."

Summative Research in HCI: Understand **attitudes** to the new system (Archer et al., 2020)

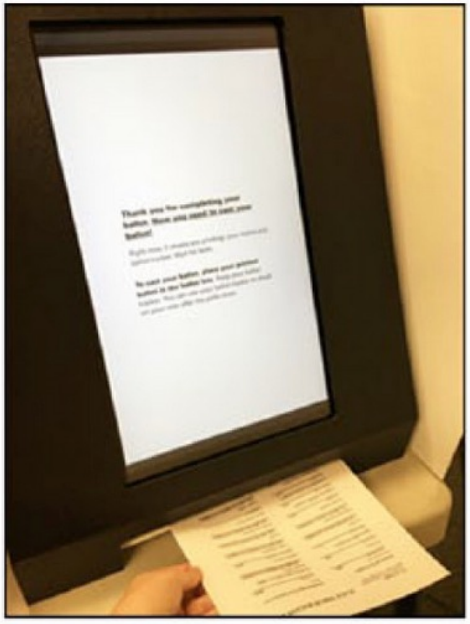
Found that

- the CanRisk tool was broadly acceptable, easy to use
- Potential barriers to use:
  - Amount of time needed to complete and interpret a risk calculation
  - Lack of opportunity to interpret the risk score before sharing with a patient
  - Requires changes to their consultation style
  - Required lots of additional information for the algorithm
  - Missing guidance on managing patients following a risk calculation

How to vote with **STAR-Vote:**



1 Complete your ballot



2 Print your ballot and ballot tracker

3 Cast your ballot by placing it in the ballot box

Summative Research in HCI: Evaluate performance of a working **prototype**

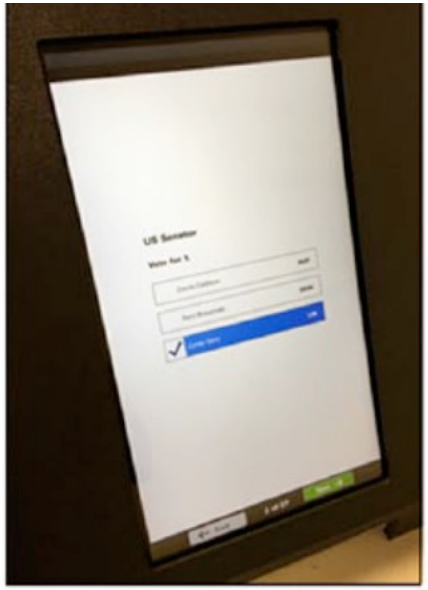
Example Question: **Usability Assessments of STAR-Vote**

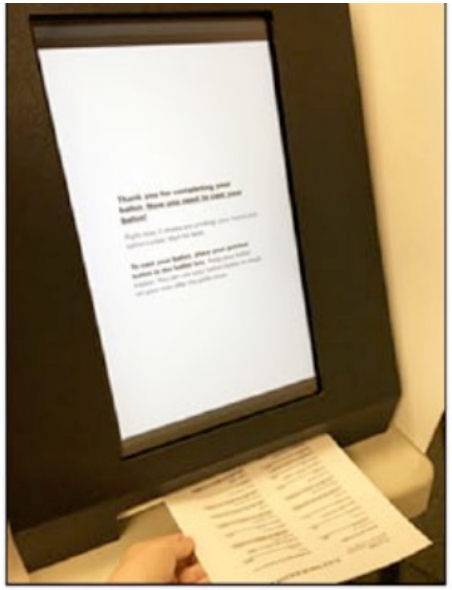
- Security – usability trade-off
  - Mismatches in mental models
  - Usability was not a priority in highly secure systems
- Known challenges in voting systems
  - Anonymity
  - Older people
  - Infrequent activity

Acemyan, C. Z., Kortum, P., Byrne, M. D., & Wallach, D. S. (2018). Summative usability assessments of STAR-Vote: a cryptographically secure e2e voting system that has been empirically proven to be easy to use. *Human factors*, 0018720818812586.

Summative Research in HCI: Evaluate performance of a working **prototype** (Acemyan et al., 2018)

Evaluation method (replication): 2 studies, mock election, randomly generated candidates



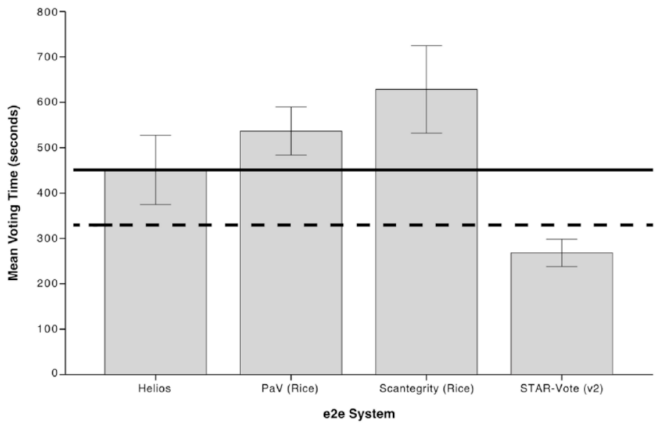


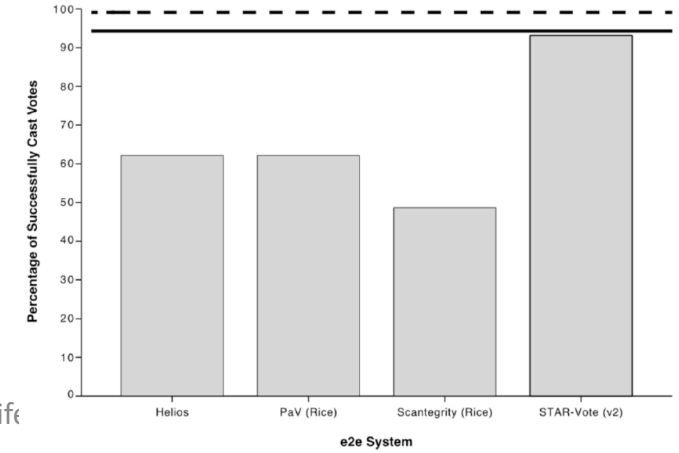
How to vote with **STAR-Vote:**

1. Complete your ballot
2. Print your ballot and ballot tracker
3. Cast your ballot by placing it in the ballot box

Summative Research in HCI: Evaluate performance of a working **prototype** (Acemyan et al., 2018)

- Evaluation method (replication) 2 studies mock election, randomly generated candidates
  - Version 1: ballot box accepted all votes
  - Version 2: ballot box could verify votes
- Measured time to vote, errors, completion of task, satisfaction
- Compared with measures from other voting systems





Summative Evaluation

- Focus is on
  - Establishing that the new system is “better” / “usable”
  - Establishing that the new system is “ready” for release / “fit-for purpose”
  - Establishing that users will accept & use the new system
- Need to define formal acceptance criteria

Measuring usability

- Measuring ***learnability***
  - Time to complete a set of tasks (by a novice)
  - Ability to improve performance
  - Learnability/efficiency trade-off
- Measuring ***effectiveness/ efficiency***
  - Ability to complete a task (pass or fail)
  - Extent to which a task is completed
  - Time to complete a set of tasks (by an expert)
  - How to define and locate “experienced” users

Measuring usability

- Measuring ***memorability/recall***
  - Ability to distinguish between visual elements
  - Ability to remember content
  - "recognition over recall" – NN/g
  - Users may be nervous
- Measuring user ***satisfaction***
  - Likert scale (agree or disagree)
  - Standardised questionnaires e.g. System Usability Scale (SUS)
  - Physiological measure of stress
  - Users are reluctant to be critical of a system – userfocus.co.uk

Measuring usability

- Measuring ***errors***
  - unintended actions, slips, mistakes or omissions that a user makes while attempting a task
  - Classification of minor vs. serious vs. critical
  - only about 10% of tasks are completed without any mistakes, and **the average number of errors per task is 0.66** – Jeff Sauro – "A Practical Guide to Measuring Usability".

Participant Considerations

- Characteristics of study participants should **match those of your end users**
  - Level of experience with computers
  - Level of experience with interface
  - Quality of domain knowledge
  - Experience with similar software

Tovi Grossman, George Fitzmaurice, and Ramtin Attar. 2009. A survey of software learnability: metrics, methodologies and guidelines. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). Association for Computing Machinery, New York, NY, USA, 649–658.

Participant Considerations

- Select participants who can provide **relevant insights**
  - General criteria (easy to find) – age, location, experience with a widely used product / software
  - Specific (harder to find) – specific medical treatment, member of community, disabilities
- Avoid testing with people you know (colleagues, friends, family, etc.) to avoid conflicts of interest and bias
- "Users who are invested in completing a task act very differently than those who are not" – The Myth of Usability testing

Further reading

- Usability metrics: tracking interface improvements (Nielsen) (https://ieeexplore-ieee-org.helicon.vuw.ac.nz/abstract/document/8740869)
- Bevan, N. (2006). Practical issues in usability measurement. *Interactions*, *13*(6), 42-43.
- Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, *64*(2), 79-102.
- Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.
- Drew, M. R., Falcone, B., & Baccus, W. L. (2018, July). What does the system usability scale (SUS) measure?. In *International Conference of Design, User Experience, and Usability* (pp. 356-366). Springer, Cham.
- Chapter 15 in Preece, Jenny, et al. *INTERACTION DESIGN : BEYOND HUMAN-COMPUTER INTERACTION*, Wiley, 2015. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/vuw/detail.action?docID=4901891