

Hoorcollege 3

Issue-based metrics, performance metrics & self-reported metrics

Judith Masthoff & Jeroen Ooge

(Credits: gebaseerd op slides Christof Van Nimwegen, Imke de Jong)



Waar zitten we in de cursus?

Methodes voor (vooral) usability engineering

HC 1: Introductie usability engineering & UX
HC 2: Usability testing & inspection
HC 3: Performance & self-reported metrics

Methodes voor (vooral) user experience

HC 4: UX & working with human participants
HC 5: Surveys & interviews
HC 6: Focus groups & diary studies
HC 7: Ethnography & case studies
HC 9: Experiments

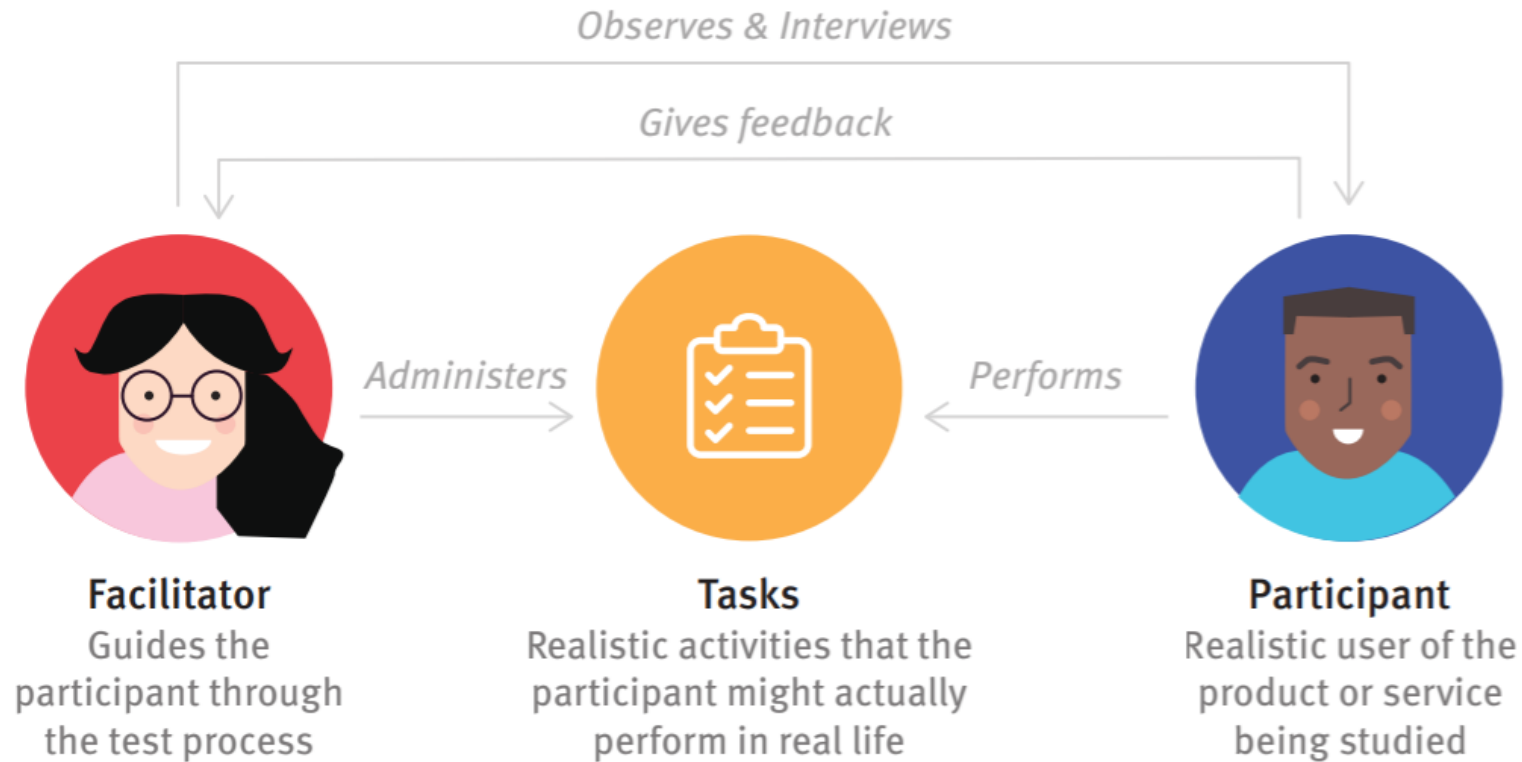
Methodes voor dataverzameling in UX-onderzoek

HC 10: Automated data collection & online research
HC 11: Ubiquitous research & measuring the human

Concreet usability tests inrichten

Wat meet je?

zie hoorcollege 2



Wanneer ingrijpen?

Hoe analyseren en rapporteren?

Welke taken?

Waar testen?

Hoe rekruteren?

Hoeveel deelnemers?

Opbouw van dit hoorcollege

Deel 1: usabilityproblemen

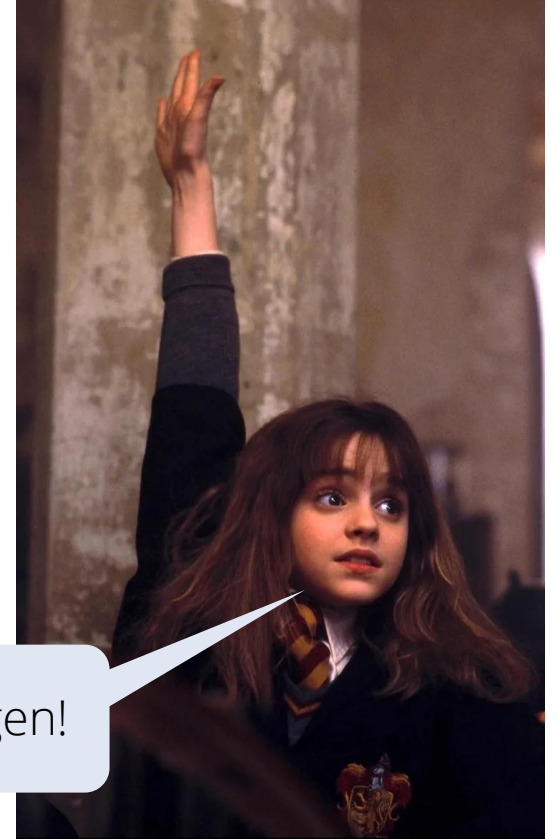
Deel 2: performantiemetrieken

Deel 3: zelfgerapporteerde metrieken

Literatuur:

Tullis & Albert: H4, H5, H6

Stel gerust vragen!



Wat zijn usability-problemen?

Er is geen formele definitie, maar typische vbn. zijn:

- Gedrag dat voltooiing van taken hindert
- Deelnemer uit frustratie
- Iets niet zien wat opgemerkt had moeten worden
- Deelnemer denkt foutief dat een taak voltooid is
- Een actie uitvoeren die weg leidt van taaksucces
- Een onderdeel verkeerd interpreteren

Ernst van usability-problemen

Vraag: Als slechts 1 van de 20 deelnemers een bepaald probleem ervaart, is het dan echt een probleem?



Ernst van usability-problemen

Vraag: Als slechts 1 van de 20 deelnemers een bepaald probleem ervaart, is het dan echt een probleem?

Antwoord:

- Kijk naar de oorzaak van het probleem, het gedrag en het gedachteproces van de deelnemer
- Als de deelnemer niet kan uitleggen waarom die iets deed of afgeleid was, dan kun je het probleem negeren

Ernst van usability-problemen

“Severity ratings” highlighten de belangrijkste usability-problemen

Leg op voorhand vast wat de ratings precies betekenen

Om de ernst te bepalen kan je meerdere factoren beschouwen

	Few users experiencing a problem	Many users experiencing a problem
Small impact on the user experience	Low severity	Medium severity
Large impact on the user experience	Medium severity	High severity

Impact on user experience

(0 = low, 1 = medium, 2 = high)

Predicted frequency of occurrence

(0 = low, 1 = medium, 2 = high)

Impact on business goals

(0 = low, 1 = medium, 2 = high)

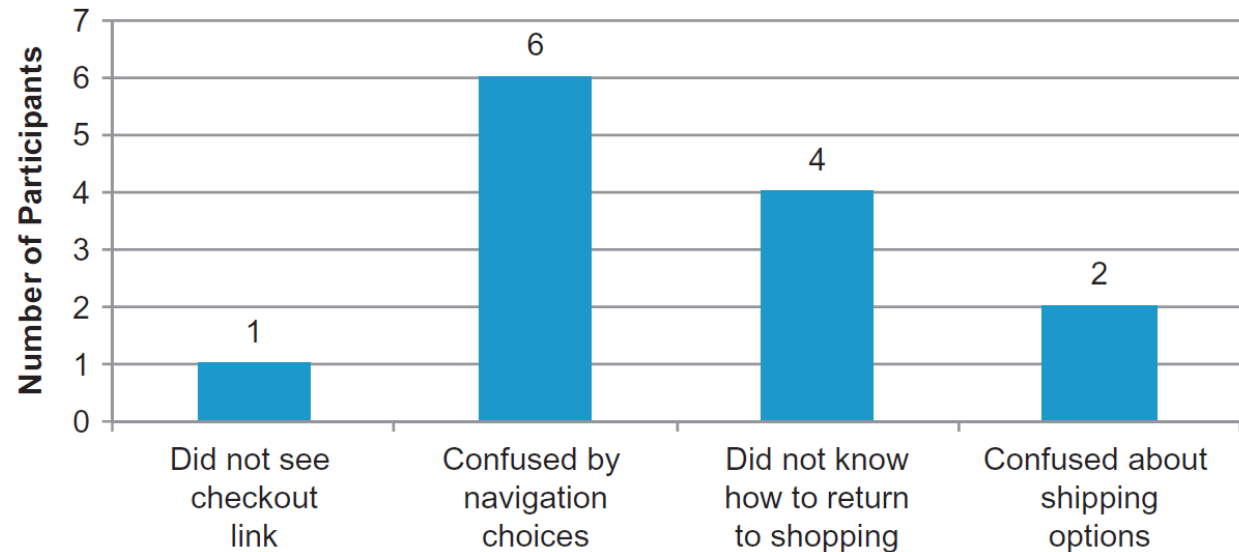
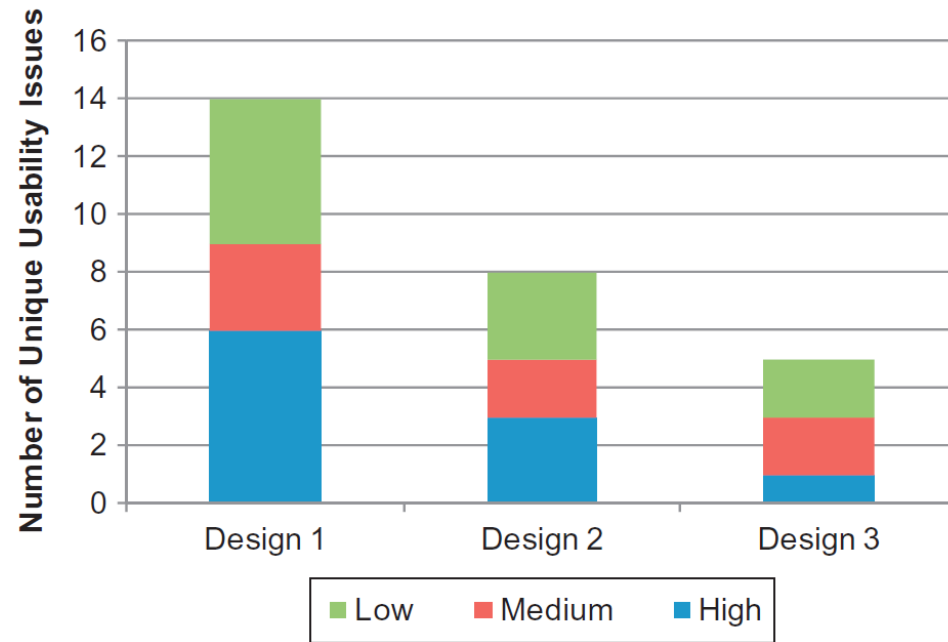
Implementation costs

(0 = high, 1 = medium, 2 = low)

Usability-problemen analyseren en rapporteren

- Aantal unieke problemen →
- Aantal unieke problemen per deelnemer
- Aantal deelnemers met een specifiek probleem →
- Problemen per categorie
- Problemen per taak

Onthoud: je wilt benadrukken wat cruciaal is voor jouw doelpubliek



Focus niet alleen op problemen

Belangrijk: rapporteer ook de sterktes van het prototype dat je evalueert, zodat gebruiksvriendelijke elementen behouden worden in volgende iteraties



Als je usability-problemen
“oplost” kun je er per ongeluk
nieuwe introduceren

Wijsneus



Voorbeeld van een eenvoudige overzichtstabel

Usability-aspect	Frequentie	Te ondernemen actie
Onderwerpen selecteren Probeert te klikken op de titel van het onderwerp ipv de checkbox Denkt initieel dat niveau het niveau van de oefeningen betekent	 4/6  1/6	Maak de titels ook aanklikbaar Hernoem “niveau” naar “level”
Visualisatie van de oefeningen Begrijpt aanvankelijk de bolletjes op de assen niet <i>“Zijn het reeds gemaakte oefeningen?” (P1)</i> <i>“Zijn het de pogingen voor eenzelfde oefening?” (P2)</i> Legende verduidelijkt dat de bolletjes oefeningen voorstellen <i>“Misschien kun je de legende bij de uitleg zetten zodat je deze sneller leest” (P4)</i>	 6/6  4/6	Maak een korte tutorial om de visualisatie stapsgewijs te introduceren en uit te leggen; voeg infobutton toe Legende behouden en samenvoegen met uitleg

Je kunt de tabel uitbreiden met “severity ratings”, screenshots ...



Opbouw van dit hoorcollege

Deel 1: usabilityproblemen

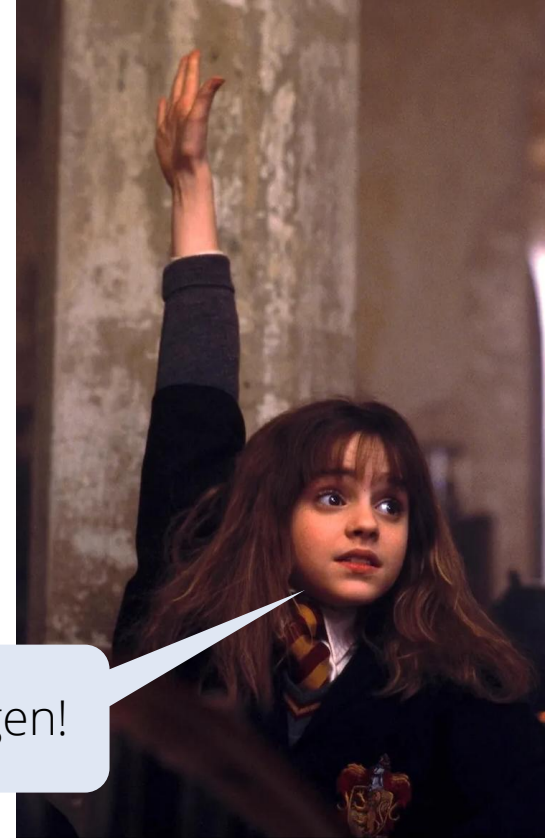
Deel 2: performantiemetrieken

Deel 3: zelfgerapporteerde metrieken

Literatuur:

Tullis & Albert: H4, H5, H6

Stel gerust vragen!



Wat zijn performantiemetrieken?

Algemeen idee: in plaats van werken met subjectieve “problemen” en “severity rates”, leg je **concrete taken** of doelen vast en je **meet de performantie** (performance) van mensen die een applicatie gebruiken

Aanname: het gebruikersgedrag waarin je geïnteresseerd bent is meetbaar met taken

Welke performantiemetrieken bestaan er?

1. Succesvol voltooide taken
2. Tijd om taken te voltooien
3. Fouten
4. Efficiëntie
5. Leerbaarheid

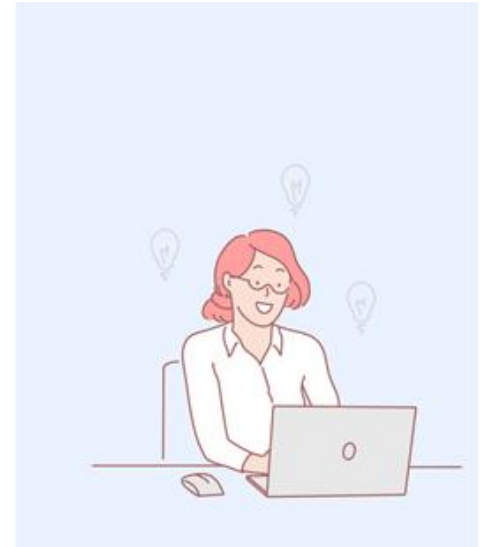


Succesvol voltooide taken

- Meest voorkomende performantiemetriek
- *Nodig*: taken met concrete einddoelen en een definitie voor “succesvol”

Goed voorbeeld: zoek op de UU-website wie het vak INFOUE doceren (succes = Jeroen en Judith)

Slecht voorbeeld: zoek op de UU-website een leuke docent (wat is “succes”? Jeroen? Judith? Niemand?)



Goed voorbeeld: zoek op de UU-website wie het vak INFOUE doceren (succes = Jeroen en Judith)



Wat als deelnemers slechts 1 van de 2 namen vermelden?

Is het dan volledig onsuccesvol of gedeeltelijk succesvol?

Binair succes meten: *pass or fail*



Je moet iemand
reanimeren



Succes (1)



Geen success (0)

Binair succes meten: *pass or fail*

Deelnemers falen voor een taak als:

- ze opgeven
- ze te veel tijd nodig hebben
- ze de taak fout uitvoerden
- onderzoekers hen onderbreken



Geen success (0)

Succesvol voltooide taken analyseren en presenteren

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Average
Participant 1	1	1	1	0	1	1	1	1	0	1	80%
Participant 2	1	0	1	0	1	0	1	0	0	1	50%
Participant 3	1	1	0	0	0	0	1	0	0	0	30%
Participant 4	1	0	0	0	1	0	1	1	0	0	40%
Participant 5	0	0	1	0	0	1	0	0	0	0	20%
Participant 6	1	1	1	1	1	0	1	1	1	1	90%
Participant 7	0	1	1	0	0	1	1	1	0	1	60%
Participant 8	0	0	0	0	1	0	0	0	0	1	20%
Participant 9	1	0	0	0	0	1	1	1	0	1	50%
Participant 10	1	1	0	1	1	1	1	1	0	1	80%
Average	70%	50%	50%	20%	60%	50%	80%	60%	10%	70%	52.0%

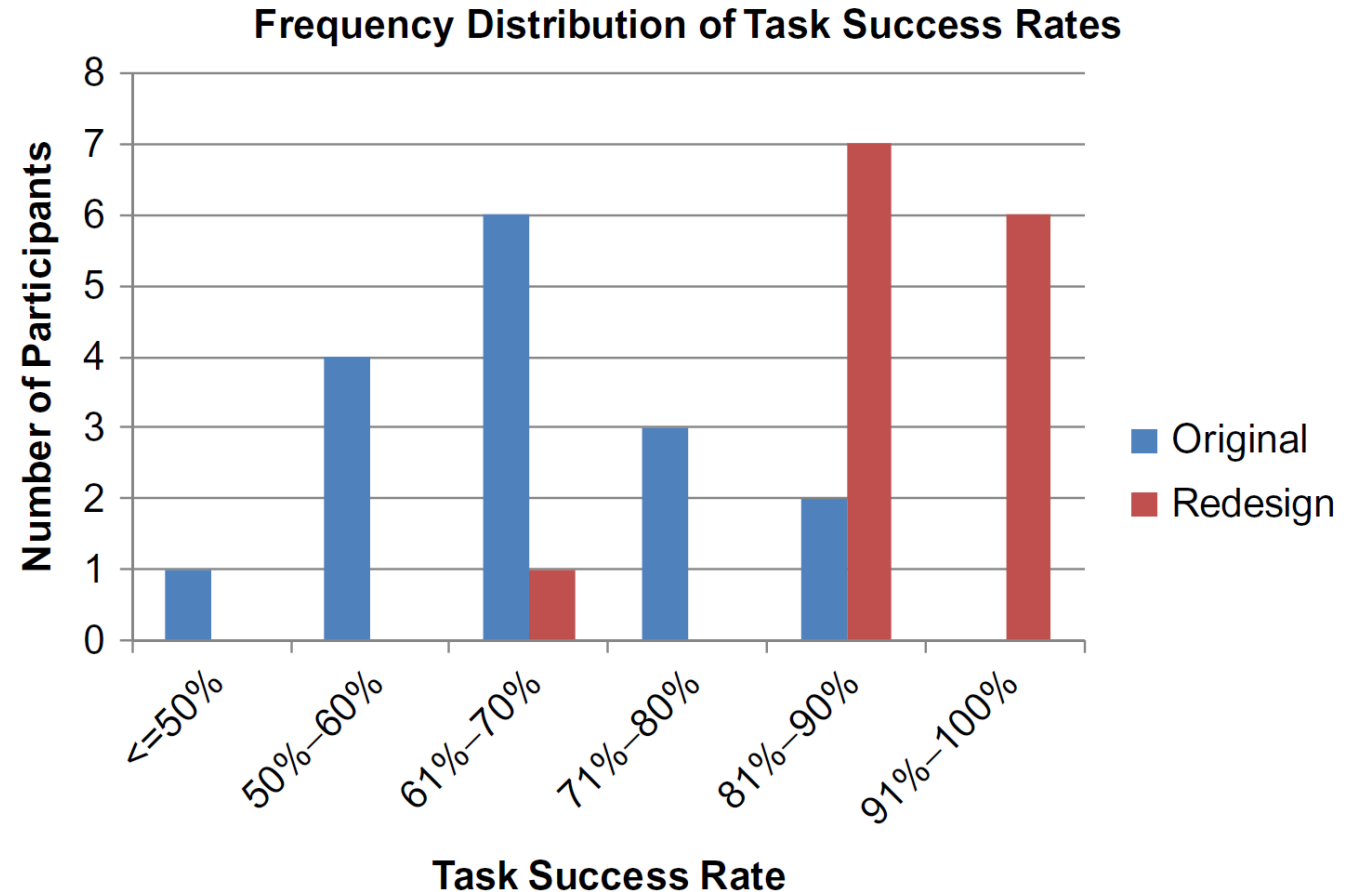
Focus op de
deelnemers ...



Succes vergelijken volgens deelnemers

Vergelijk groepen deelnemers:

- Ervaring met het onderzochte systeem (vb. weinig vs veel ervaring)
- Domeinexpertise (vb. leken vs experten)
- Andere persoonlijke kenmerken (vb. leeftijd)



Succesvol voltooide taken analyseren en presenteren

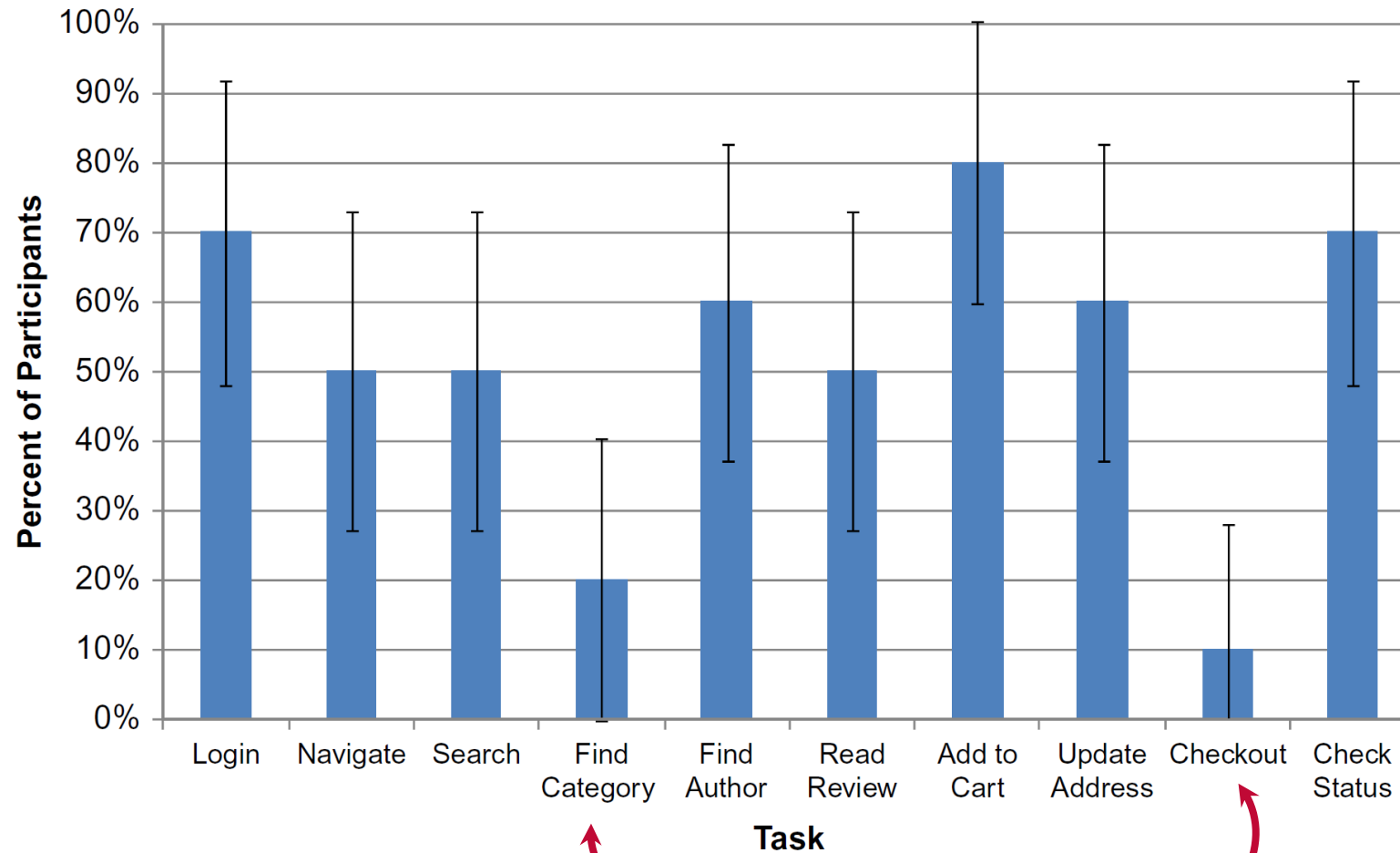
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Average
Participant 1	1	1	1	0	1	1	1	1	0	1	80%
Participant 2	1	0	1	0	1	0	1	0	0	1	50%
Participant 3	1	1	0	0	0	0	1	0	0	0	30%
Participant 4	1	0	0	0	1	0	1	1	0	0	40%
Participant 5	0	0	1	0	0	1	0	0	0	0	20%
Participant 6	1	1	1	1	1	0	1	1	1	1	90%
Participant 7	0	1	1	0	0	1	1	1	0	1	60%
Participant 8	0	0	0	0	1	0	0	0	0	1	20%
Participant 9	1	0	0	0	0	1	1	1	0	1	50%
Participant 10	1	1	0	1	1	1	1	1	0	1	80%
Average	70%	50%	50%	20%	60%	50%	80%	60%	10%	70%	52.0%

Focus op de
deelnemers ...



... of op de taken

Percent Correct, by Task
(Error bars represent the 90% confidence interval)



Problematisch

Fijnmazig succes meten met verschillende graden

Definieer wat de verschillende graden zijn en laat eventueel meerdere onderzoekers de graad bepalen

vb. Graad op basis van correctheid (3 niveaus):

- Volledig correct → zonder hulp / met hulp
- Gedeeltelijk correct → zonder hulp / met hulp
- Gefaald → fout voltooid / opgegeven

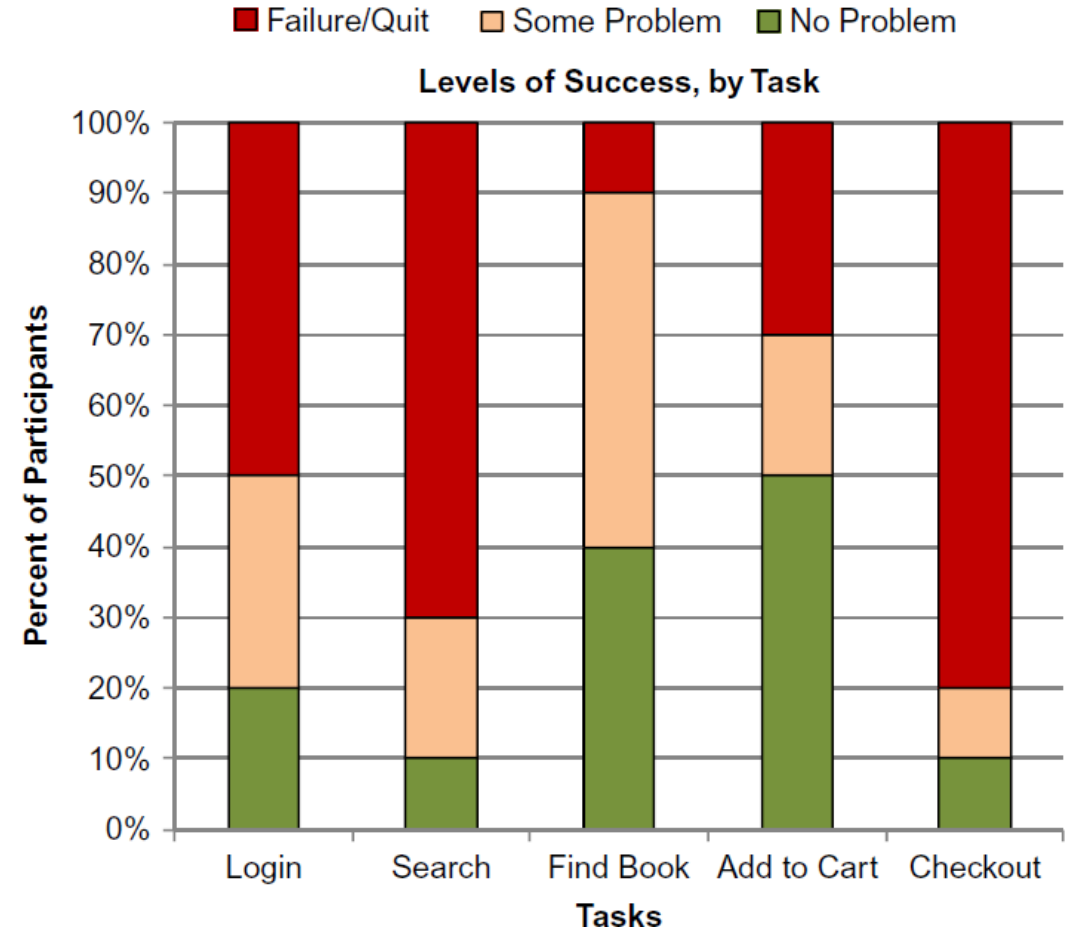
X aantal foute
antwoorden

Zelf aangegeven of
te veel tijd nodig

Succesvol voltooide taken analyseren en presenteren

vb. Graad op basis van ervaren problemen tijdens het voltooien van taken

- Correct voltooid zonder problemen
- Correct voltooid met kleine problemen
- Correct voltooid met grote problemen
- Niet correct voltooid of opgegeven



Welke performantiemetrieken bestaan er?

1. Succesvol voltooide taken
2. Tijd om taken te voltooien
3. Fouten
4. Efficiëntie
5. Leerbaarheid



Tijd om taken te voltooien

Tijd wordt vaak gebruikt als een metriek voor **efficiëntie**

"In most situations, the faster a user can complete a task, the better the experience. [...] users generally want to be spending *less* time on the site, not *more*."

Is sneller altijd wenselijk?

Pro snelheid

- Repetitieve taken moeten snel voltooid kunnen worden

Contra snelheid

- Doel = gebruikers laten nadenken (vb. e-learning)
- Doel = entertainment



Bespreek met je buur: uitdagingen bij tijdsmetingen

- Hoe kunnen we tijdsmetingen combineren met think-aloudstudies?
- Moeten we de tijd voor alle voltooide taken gebruiken of alleen succesvol voltooide?
- Hoe moeten we omgaan met outliers?
- Mogen deelnemers weten dat tijd wordt gemeten?



Hoe kunnen we tijdsmetingen combineren met think-aloudstudies?

Probleem: gedachten formuleren vergt tijd, vertekent de tijdsmetingen en kan leiden tot minder succesvolle taken

Mogelijke oplossing: retrospectieve think-aloud

- Testpersonen voltooien de taken in stilte
- Vraag achteraf waar testpersonen aan dachten



Moeten we de tijd voor alle voltooide taken gebruiken of alleen succesvol voltooide?

Probleem:

- Soms proberen deelnemers eindeloos lang tevergeefs een taak te voltooien
- Alleen succesvol voltooide taken bekijken is misleidend als weinig taken snel voltooid konden worden

Mogelijke oplossing: laat deelnemers zelf aangeven of ze een taak opgeven en gebruik alle tijdsmetingen voor een correctere weergave van performantie

Hoe moeten we omgaan met outliers?

Probleem:

- Tijdsmetingen zijn soms erg kort, omdat deelnemers bv. alleen uit zijn op de beloning voor deelname
- Tijdsmetingen zijn soms erg lang, omdat deelnemers bv. andere dingen doen tijdens het experiment

Mogelijke oplossing: verwijder onwaarschijnlijk korte of lange tijdsmetingen (typisch buiten 2σ - of 3σ -interval)

Mogen deelnemers weten dat tijd wordt gemeten?

Probleem:

- Deelnemers kunnen afdwalen van de taken als ze niet weten dat tijd wordt gemeten
- Deelnemers worden vaak nerveus als ze weten dat tijd wordt gemeten en kunnen zich beginnen haasten

Mogelijke oplossing: vraag deelnemers om taken zo snel mogelijk te voltooien zonder te zeggen dat je tijd meet



Welke performantiemetrieken bestaan er?

1. Succesvol voltooide taken
2. Tijd om taken te voltooien
3. Fouten
4. Efficiëntie
5. Leerbaarheid

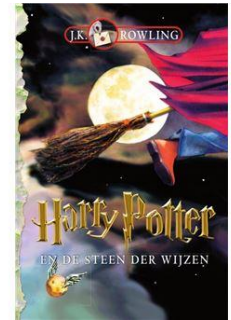


Fouten en usability-problemen zijn niet hetzelfde

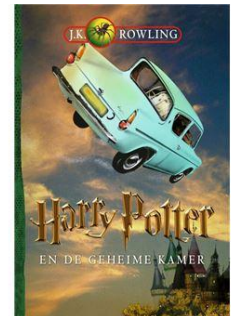
Fouten zijn het *gevolg* van usability-problemen
vb. Deelnemers bestellen een verkeerd product (fout);
de oorzaak is dat de positie van de bestelknop
verwarrend is (usability-probleem)

Er bestaan verschillende **types fouten**, bv.:

- Incorrecte data in een tekstveld invullen
- Een verkeerd item in een menu kiezen
- Foute acties uitvoeren

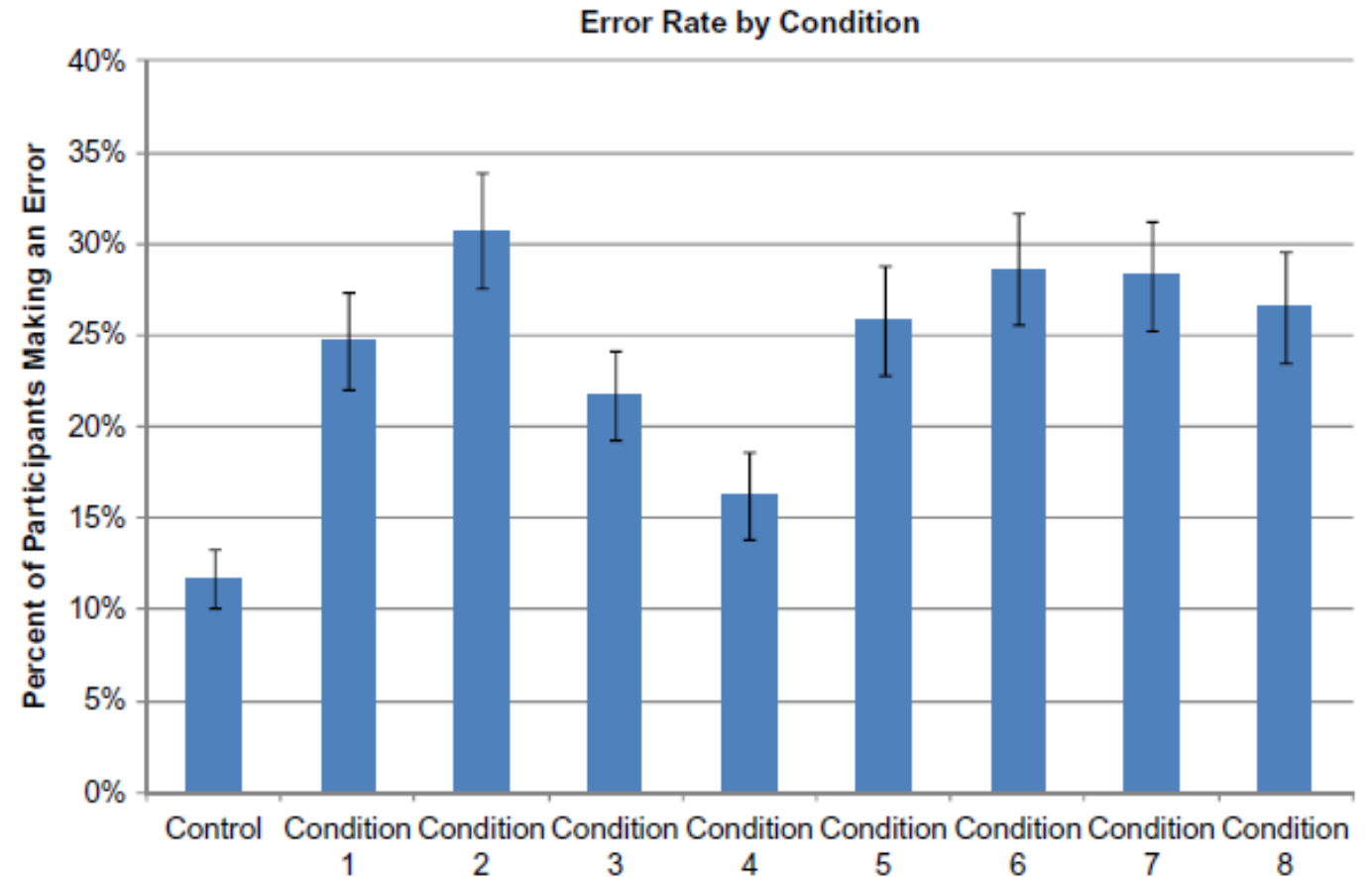


Koop nu!



Fouten analyseren en presenteren

- Gemiddeld aantal fouten per testpersoon
- Percentage testpersonen dat fouten maakt →
- Niveau van ernst aan fouten toekennen (vb. hoog, gemiddeld, laag)
- Terugkerende fouten



Welke performantiemetrieken bestaan er?

1. Succesvol voltooide taken
2. Tijd om taken te voltooien
3. Fouten
4. Efficiëntie
5. Leerbaarheid



Efficiëntie

Inspanning nodig om taken te voltooien

- *Cognitieve inspanning* = bepalen wanneer welke actie uitgevoerd moet worden en het resultaat interpreteren
- *Fysieke inspanning* = muis bewegen, toetsen indrukken ...

Efficiëntie wordt vaak gedefinieerd in termen van het aantal stappen en de tijd om taken te voltooien (of beide)



Voorbeeld 1: “verloren” zijn op een website

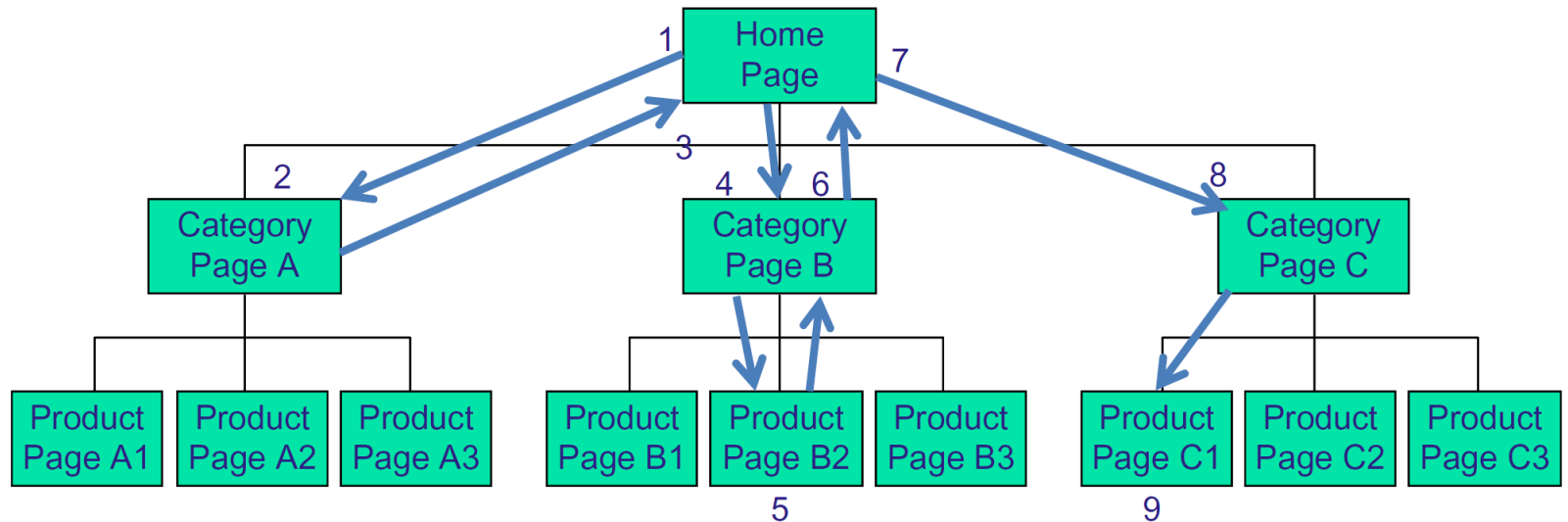
Efficiëntie uitgedrukt als een combinatie van:

- zo weinig mogelijk dezelfde webpagina's bezoeken
- het kortste pad naar een webpagina vinden

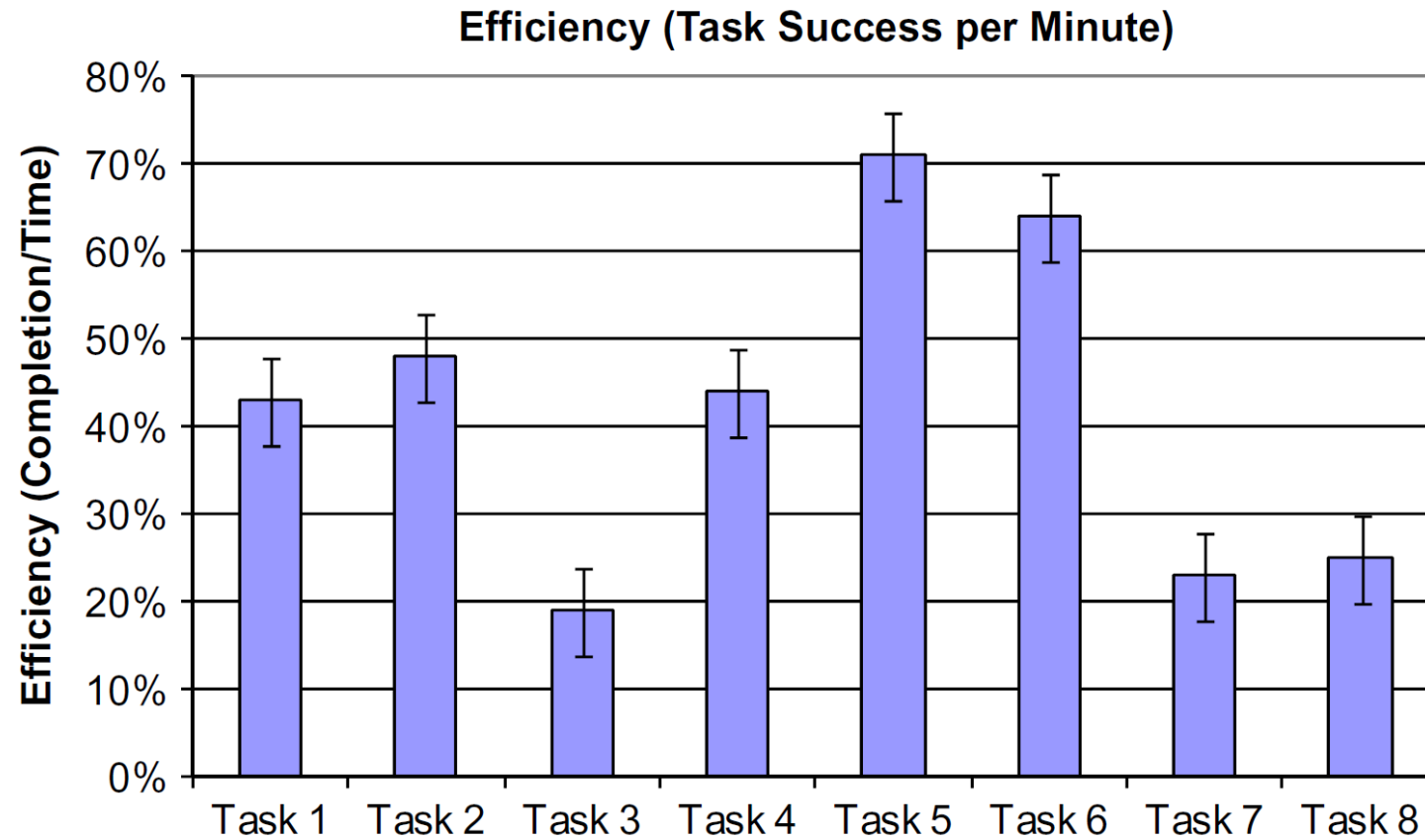
N : The number of *different* web pages visited while performing the task
 S : The *total* number of pages visited while performing the task, counting revisits to the same page
 R : The *minimum* (optimum) number of pages that must be visited to accomplish the task

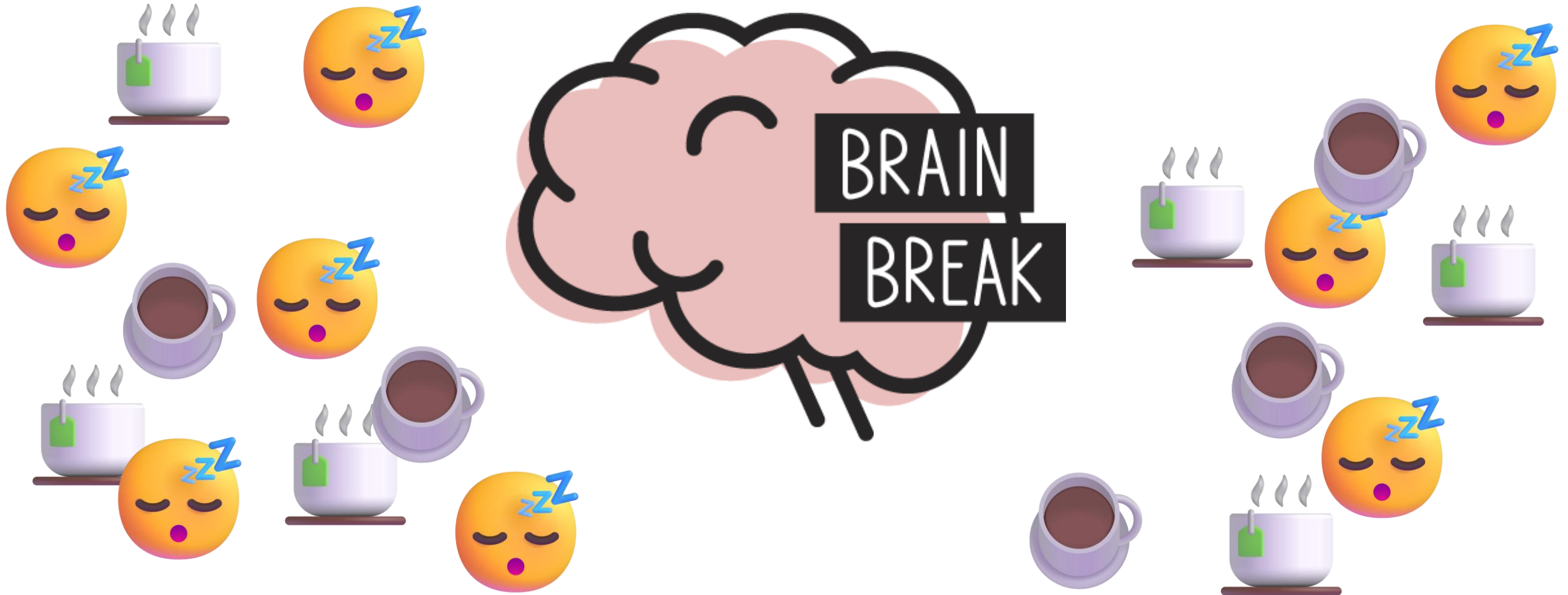
Lostness, L , is then calculated using the following formula:

$$L = \text{sqrt}[(N/S - 1)^2 + (R/N - 1)^2].$$



Voorbeeld 2: succesvolle taken per tijdseenheid





Welke performantiemetrieken bestaan er?

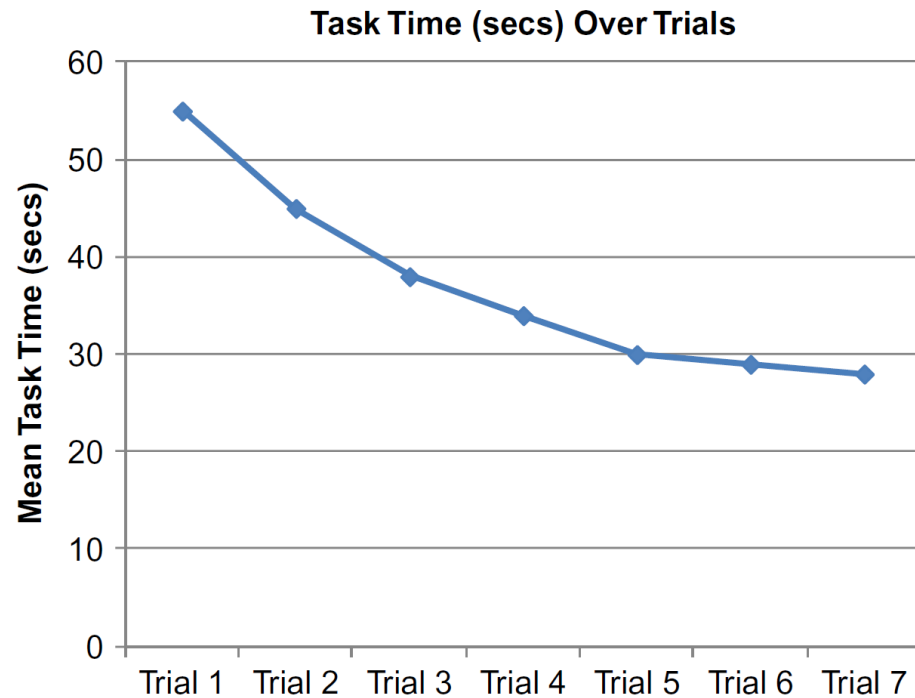
1. Succesvol voltooide taken
2. Tijd om taken te voltooien
3. Fouten
4. Efficiëntie
5. Leerbaarheid



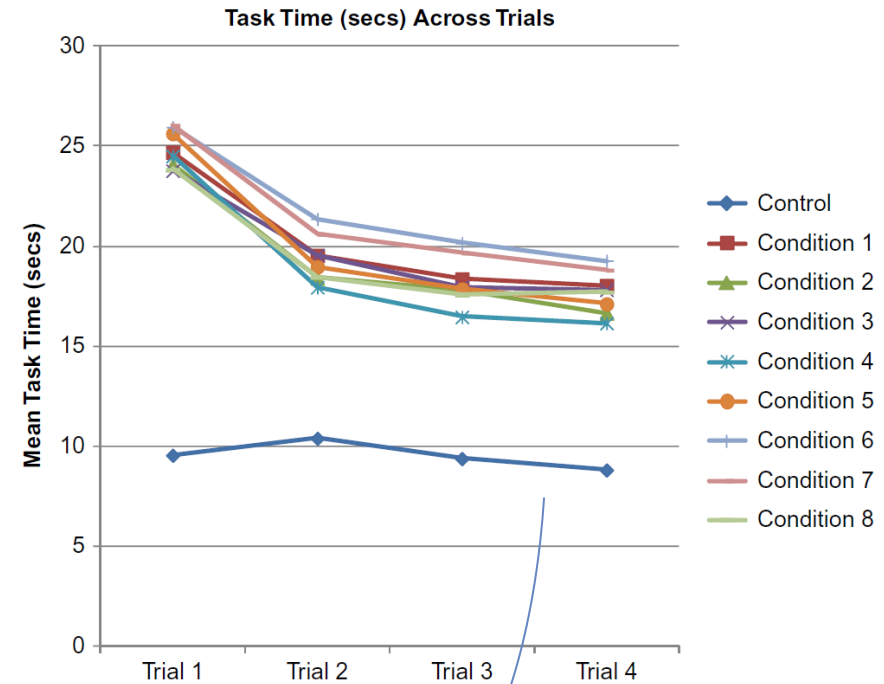
Leerbaarheid

- Maat voor hoe gemakkelijk een applicatie geleerd kan worden: **leren vergt tijd, niet alleen de eerste indruk telt**
- Deelnemers moeten niet letterlijk onthouden hoe taken voltooid moeten worden, wel strategieën optimaliseren
- **Hoe meten?** Eenzelfde performantiemetriek op verschillende momenten (*trials*) meten en evolutie bekijken

Leerbaarheid analyseren en rapporteren

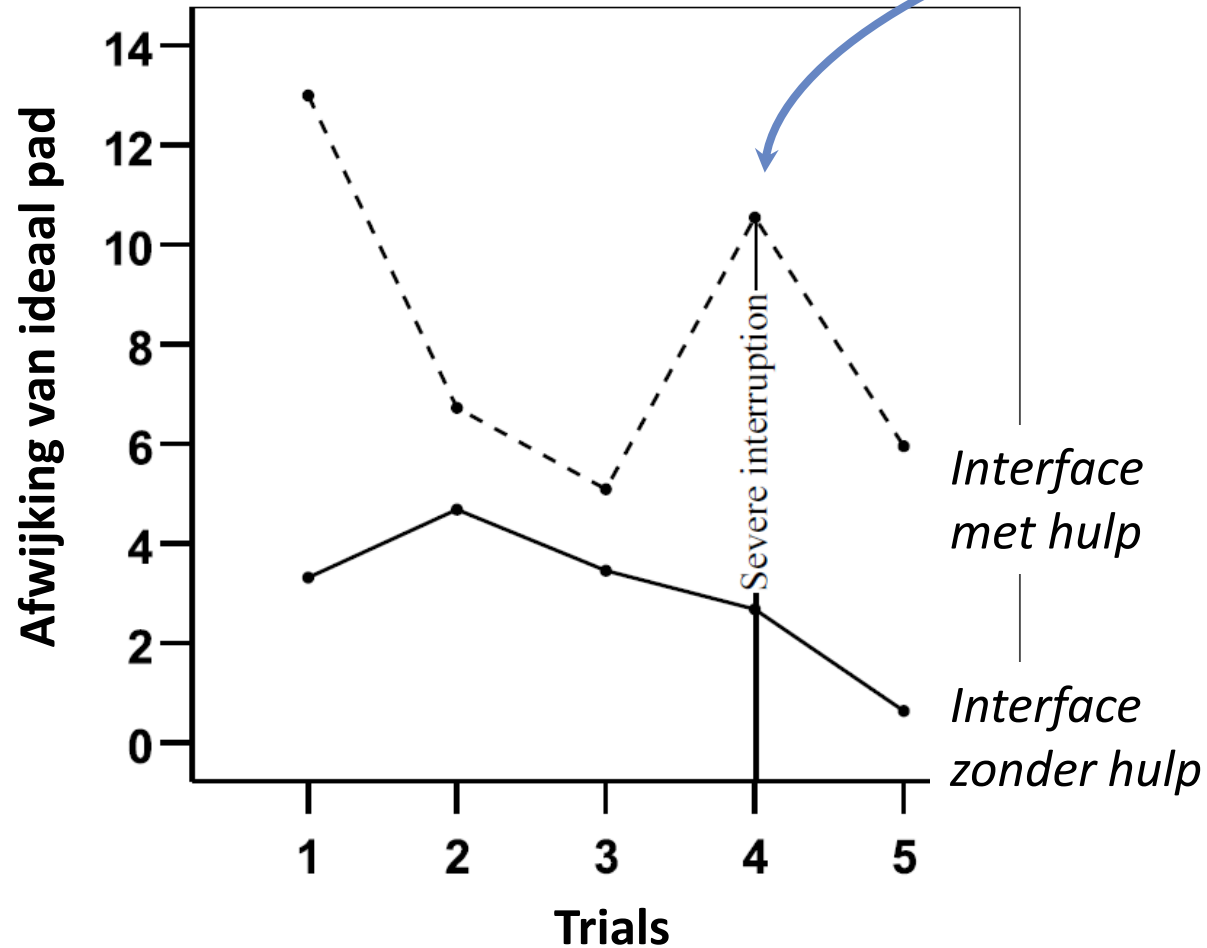


Wanneer metingen doen?



Wanneer convergeert een metriek?

Voorbeeld: impact van een onverwachte onderbreking



Bij de interface met hulp weken deelnemers vaker af van het ideale pad en was de impact van de onderbreking groot (= lage leerbaarheid)

Welke performantiemetrieken bestaan er?

1. Succesvol voltooide taken
2. Tijd om taken te voltooien
3. Fouten
4. Efficiëntie
5. Leerbaarheid



Welke limitaties zouden
performantiemetrieken
toch kunnen hebben?



Limitaties van performantiemetrieken

Performantiemetrieken zijn geschikt om *wat*-vragen te beantwoorden, maar niet *waarom*-vragen

vb. Performantiemetrieken kunnen uitwijzen welke delen van een interface ongebruiksvriendelijk zijn, maar niet wat de oorzaak daarvan is of hoe mensen zich daarbij voelen

Oplossing: observaties en **zelfgerapporteerde metrieken**

Opbouw van dit hoorcollege

Deel 1: usabilityproblemen

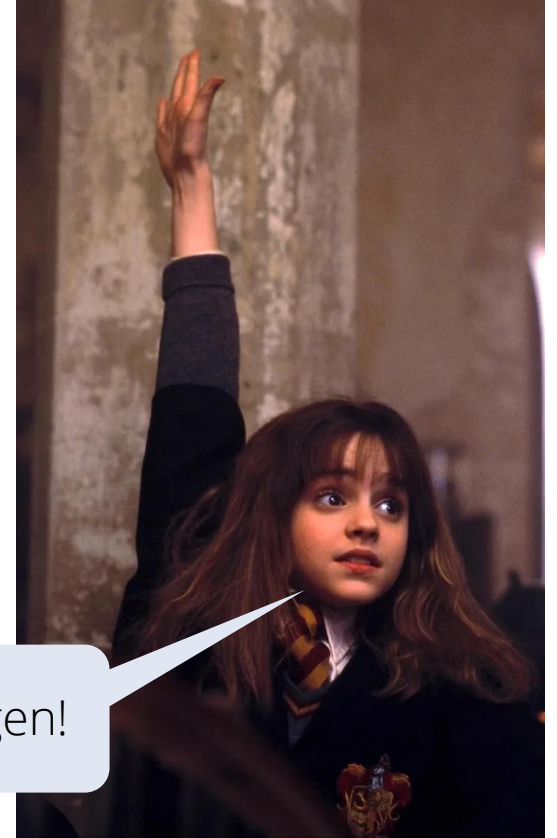
Deel 2: performantiemetrieken

Deel 3: zelfgerapporteerde metrieken

Literatuur:

Tullis & Albert: H4, H5, H6

Stel gerust vragen!



Voordelen van zelfgerapporteerde metrieken

- Vullen “objectieve” performantiemetrieken aan met “subjectieve” informatie over **hoe mensen zich voelen**, bv. blijheid, woede, frustratie, uitdaging, vertrouwen, algemeen ervaren gebruiksgemak ...
- Vergeleken met performantiemetrieken kan informatie over hoe mensen zich voelen bij producten beter voorspellen of ze die blijven gebruiken

Zelfgerapporteerde metrieken verzamelen

Vaak in de vorm van *vragenlijsten* met

- open tekstvelden
- “reaction cards”
- Likert-schalen of Likert-type vragen (zie hoorcollege 5)

Kan gecombineerd worden met vervolginterviews (zie hoorcollege 5), bv. met een deel van de deelnemers

Open tekstvelden analyseren en rapporteren

Wordclouds zijn visueel aantrekkelijk, maar niet inzichtelijk:

- Kleuren zijn redundant
- Moeilijk om groottes van woorden te vergelijken
- Verbatim reacties bevatten veel irrelevante woorden

Beter: kwalitatieve analyse



Zelfgerapporteerde metrieken verzamelen

Vaak in de vorm van *vragenlijsten* met

- open tekstvelden
- “reaction cards”
- Likert-schalen of Likert-type vragen (zie hoorcollege 5)

Kan gecombineerd worden met vervolginterviews (zie hoorcollege 5), bv. met een deel van de deelnemers

“Reaction cards”

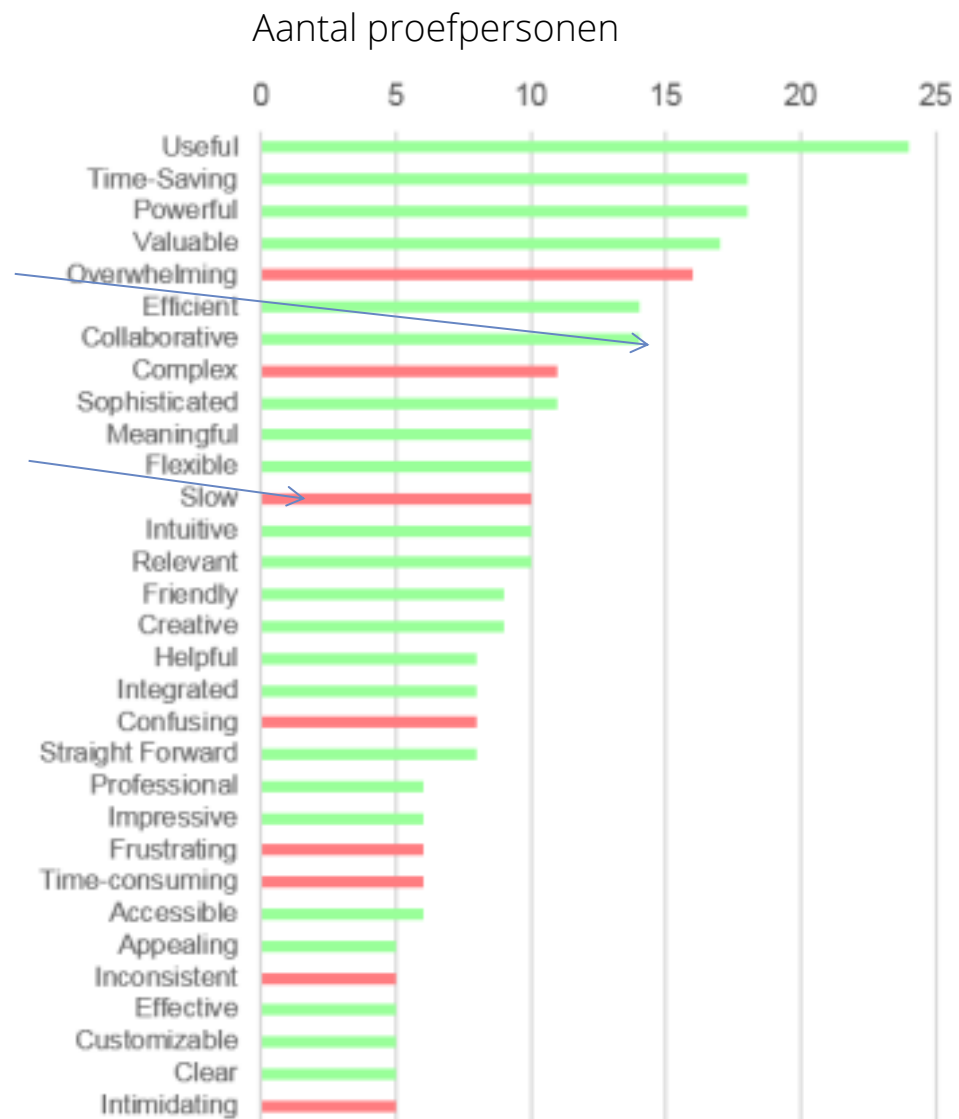
Deelnemers kiezen
adjectieven die volgens hen
het best passen bij de
applicatie en motiveren hun
keuze in een open tekstveld

The complete set of 118 Product Reaction Cards				
Accessible	Creative	Fast	Meaningful	Slow
Advanced	Customizable	Flexible	Motivating	Sophisticated
Annoying	Cutting edge	Fragile	Not Secure	Stable
Appealing	Dated	Fresh	Not Valuable	Sterile
Approachable	Desirable	Friendly	Novel	Stimulating
Attractive	Difficult	Frustrating	Old	Straight Forward
Boring	Disconnected	Fun	Optimistic	Stressful
Business-like	Disruptive	Gets in the way	Ordinary	Time-consuming
Busy	Distracting	Hard to Use	Organized	Time-Saving
Calm	Dull	Helpful	Overbearing	Too Technical
Clean	Easy to use	High quality	Overwhelming	Trustworthy
Clear	Effective	Impersonal	Patronizing	Unapproachable
Collaborative	Efficient	Impressive	Personal	Unattractive
Comfortable	Effortless	Incomprehensible	Poor quality	Uncontrollable
Compatible	Empowering	Inconsistent	Powerful	Unconventional
Compelling	Energetic	Ineffective	Predictable	Understandable
Complex	Engaging	Innovative	Professional	Undesirable
Comprehensive	Entertaining	Inspiring	Relevant	Unpredictable
Confident	Enthusiastic	Integrated	Reliable	Unrefined
Confusing	Essential	Intimidating	Responsive	Usable
Connected	Exceptional	Intuitive	Rigid	Useful
Consistent	Exciting	Inviting	Satisfying	Valuable
Controllable	Expected	Irrelevant	Secure	
Convenient	Familiar	Low Maintenance	Simplistic	

“Reaction cards” analyseren en rapporteren

staaflengtes kun
je snel vergelijken

kleuren hebben betekenis:
positief vs negatief



Zelfgerapporteerde metrieken verzamelen

Vaak in de vorm van *vragenlijsten* met

- open tekstvelden
- “reaction cards”
- Likert-schalen of Likert-type vragen ([zie hoorcollege 5](#))

Kan gecombineerd worden met vervolginterviews ([zie hoorcollege 5](#)), bv. met een deel van de deelnemers

Wanneer zelfgerapporteerde metrieken verzamelen?

Na elke taak (“post-task”)

bv. korte vragenlijsten laten toe precies te bepalen welke delen van een interface problematisch zijn

Overall, this task was?

Very Difficult ☐ ☐ ☐ ☐ ☐ ☐ ☐ Very Easy

Na alle taken (“post-study”)

bv. uitgebreidere vragenlijsten (met open vragen) kunnen gedetailleerd de volledige ervaring met applicaties toetsen

SUS, QUIS, USE, zelf opgestelde vragenlijsten ...



Oké, zelfgerapporteerde
metrieken zijn nuttig, maar er
zijn zeker weer limitaties?



Jep

Limitaties van zelfgerapporteerde metrieken

(zie slides *Persuasive Technologies*)

- *Sociaal wenselijke antwoorden*: deelnemers antwoorden zoals ze verwachten dat ze moeten antwoorden
- *Persoonlijke verschillen*: sommige mensen zijn geneigd vooral extreme (vs neutraal) waarden te kiezen of het voornamelijk eens (vs oneens) te zijn met stellingen
- *Gebrekkige zelfreflectie*: mensen kunnen vaak niet goed inschatten wat hun gevoelens of acties zijn
- *Ontbrekend referentiepunt*: hoe zijn antwoorden geijkt?



Rond usability
engineering nou maar af!



Usability engineering

- Heeft producten verbeteren als doel
- Betrekt weinig deelnemers
- Heeft geen sterk gecontroleerde condities voor de tests
- Is vaak niet helemaal repliceerbaar
- Resultaten leiden tot ontwerpkeuzes
- Wordt vaak alleen intern gedeeld met bv. ontwikkelaars en bestuursleden

Experimenteel onderzoek

- Heeft kennis creëren als doel
- Betrekt veel deelnemers
- Heeft experimenteel design met sterk gecontroleerde condities
- Is repliceerbaar
- Resultaten zijn statistisch gevalideerd
- Wordt gedeeld met wetenschappers wereldwijd

zie hoorcollege 9

Tijd voor
een kvisje



se Home Content Groups Activities ▾ Grades Class

2025-2026 1
(INFOUE)

Assignments

Quizzes

Discussions

Attendance

ngine



Methodes voor (vooral) usability engineering

HC 1: Introductie usability engineering & UX
HC 2: Usability testing & inspection
HC 3: Performance & self-reported metrics

Methodes voor (vooral) user experience

HC 4: UX & working with human participants
HC 5: Surveys & interviews
HC 6: Focus groups & diary studies
HC 7: Ethnography & case studies
HC 9: Experiments

Methodes voor dataverzameling in UX-onderzoek

HC 10: Automated data collection & online research
HC 11: Ubiquitous research & measuring the human

Volgende keer...

*Werkcollege (verplicht) -
Opdracht 1: think-aloud pilot
afnemen en peer feedback*



Lees de paper
van Hassenzahl
& Tractinsky (zie
Brightspace)!



Methodes voor (vooral) usability engineering

HC 1: Introductie usability engineering & UX

HC 2: Usability testing & inspection

HC 3: Performance & self-reported metrics

Methodes voor (vooral) user experience

HC 4: UX & working with human participants

HC 5: Surveys & interviews

HC 6: Focus groups & diary studies

HC 7: Ethnography & case studies

HC 9: Experiments

Methodes voor dataverzameling in UX-onderzoek

HC 10: Automated data collection & online research

HC 11: Ubiquitous research & measuring the human

Zo meteen...

*Werkcollege - Opdracht 1:
think-aloud pilot afwerken
en afnemen*

Niet verdwalen op
weg naar Dalton, hè





**Universiteit
Utrecht**

Sharing science,
shaping tomorrow