

Hoorcollege 9

Experimenten



Jeroen Ooge

(Credits: gebaseerd op slides Christof Van Nimwegen)

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

Opbouw van dit hoorcollege

Deel 1: experimenteel onderzoek

Intermezzo: statistische toetsen

Deel 2: experimenteel design

Deel 3: inleiding opdracht 2

Literatuur:

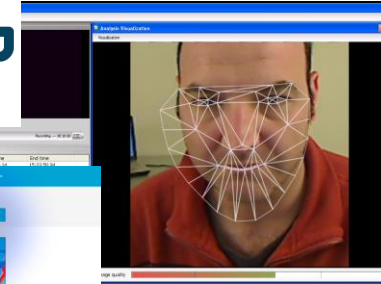
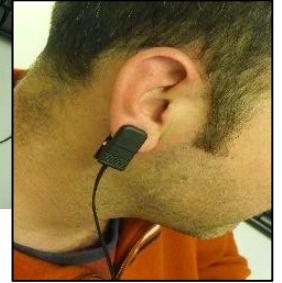
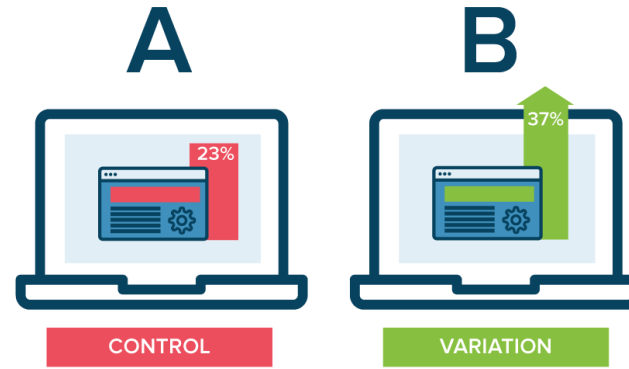
Lazar et al.: H2, H3

Stel gerust vragen!



Experimenten zijn veelzijdig

- Kunnen plaatsvinden in een lab of in de echte wereld
- Veel soorten meetinstrumenten



Type of Research	Focus	General Claims	Typical Methods
Descriptive			
Relational			
Experimental			

Type of Research	Focus	General Claims	Typical Methods
Descriptive	Describe a situation or a set of events	X is happening	Observations, field studies, focus groups, interviews
Relational			
Experimental			

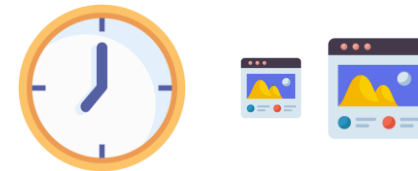


Efficiëntie



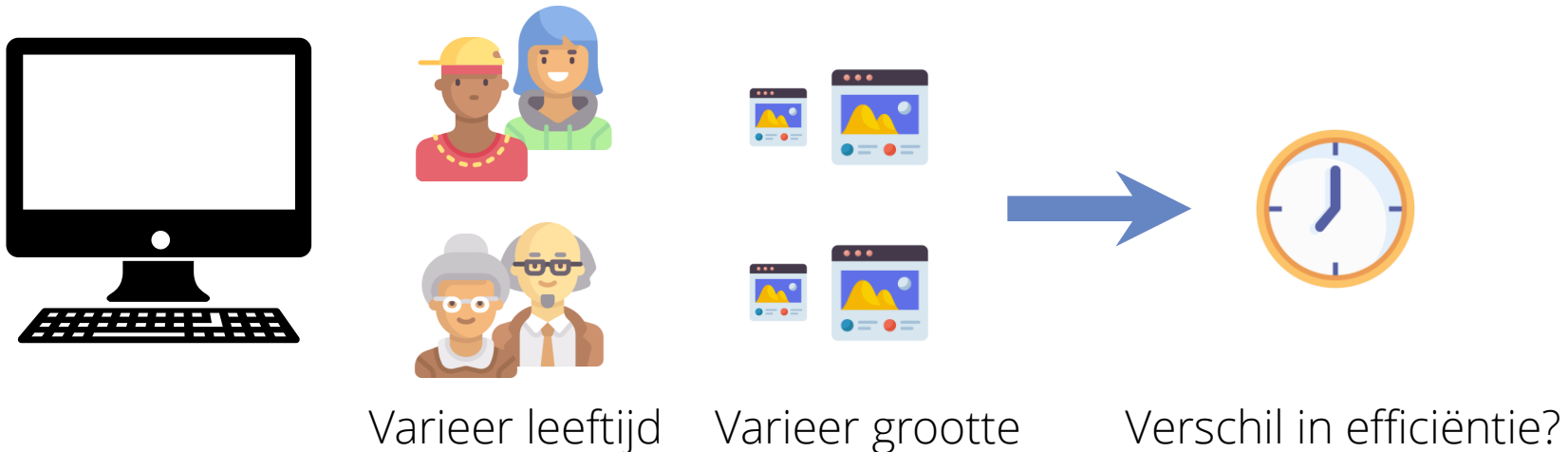
Grootte van afbeeldingen

Type of Research	Focus	General Claims	Typical Methods
Descriptive	Describe a situation or a set of events	X is happening	Observations, field studies, focus groups, interviews
Relational	Identify relations between multiple variables	X is related to Y	Observations, field studies, surveys
Experimental			

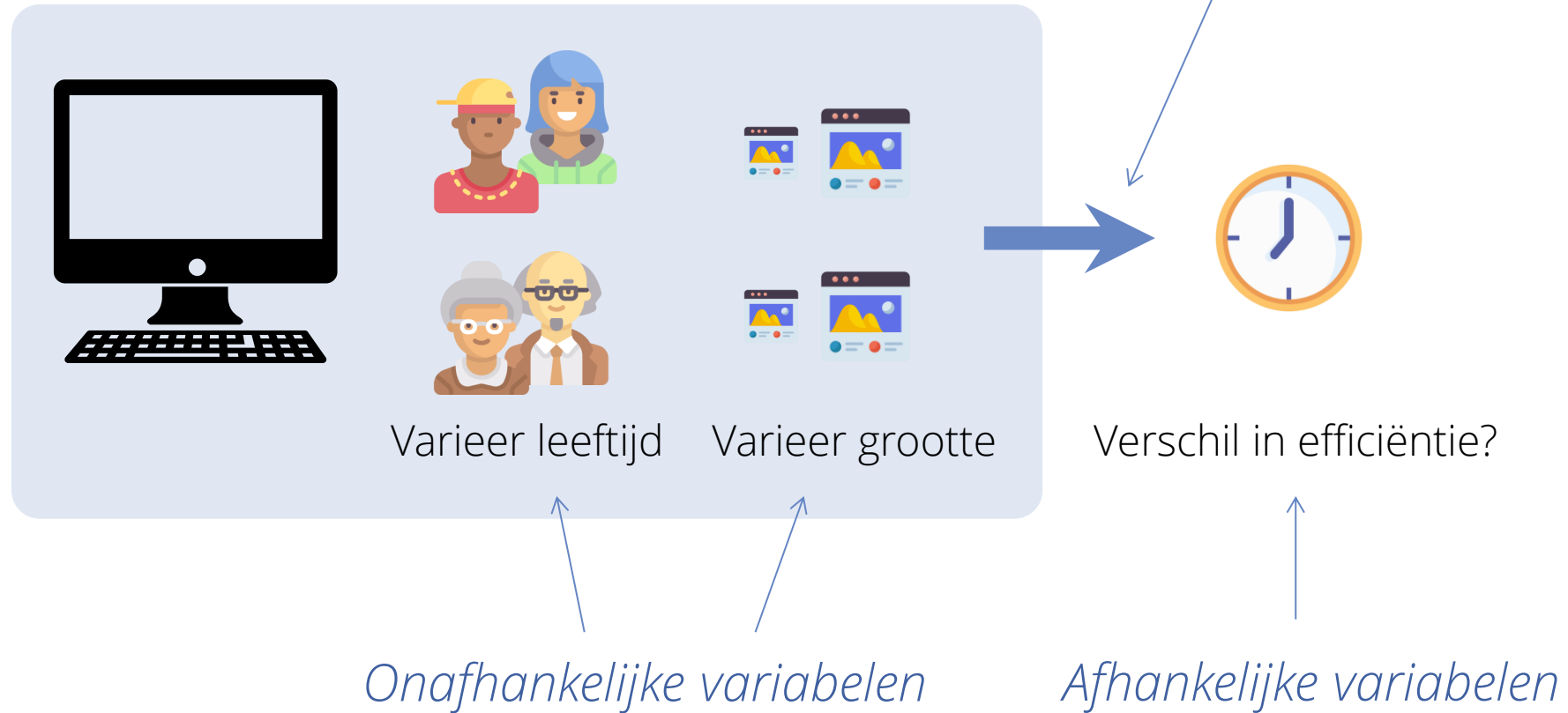


Grootte van afbeeldingen kan efficiëntie beïnvloeden

Type of Research	Focus	General Claims	Typical Methods
Descriptive	Describe a situation or a set of events	X is happening	Observations, field studies, focus groups, interviews
Relational	Identify relations between multiple variables	X is related to Y	Observations, field studies, surveys
Experimental	Identify causes of a situation or a set of events	X is responsible for Y	Controlled experiments

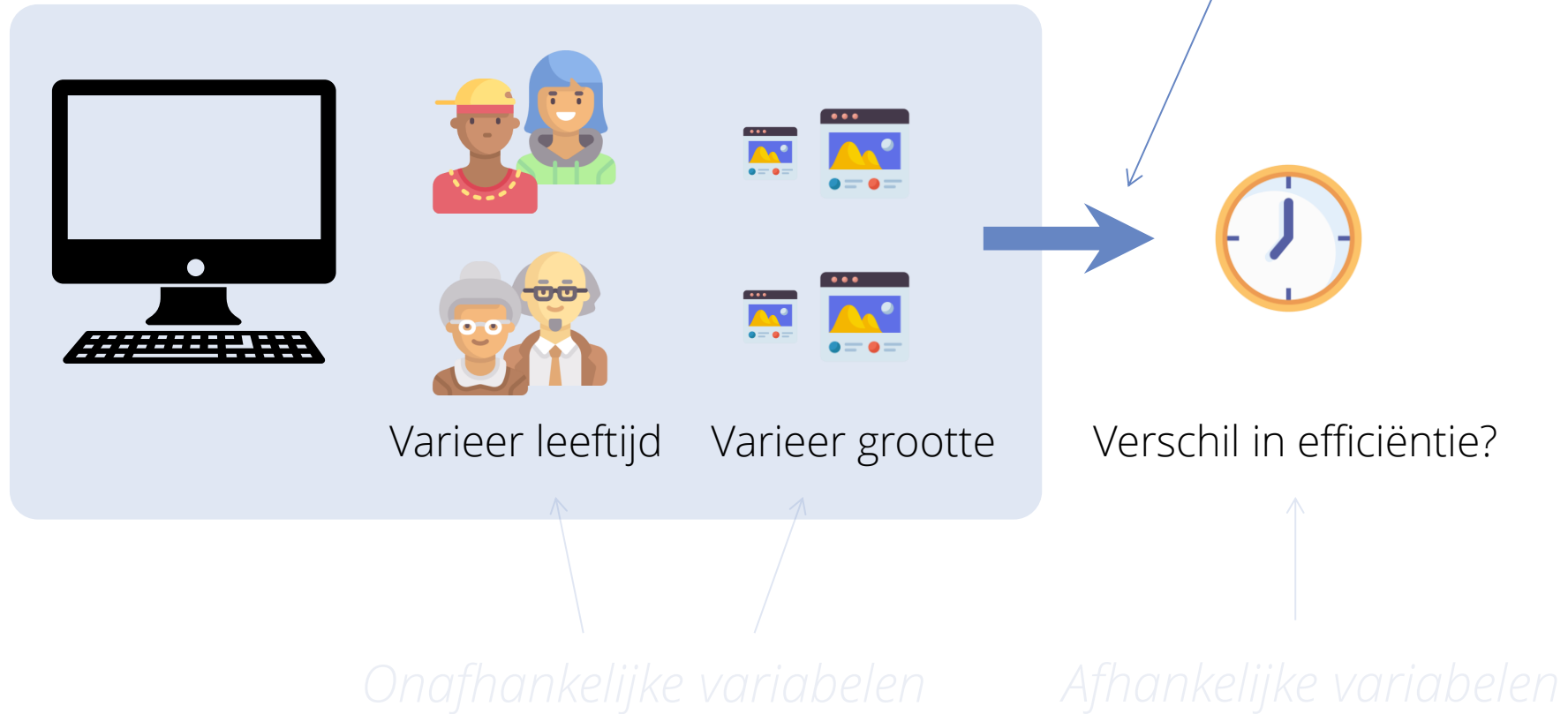


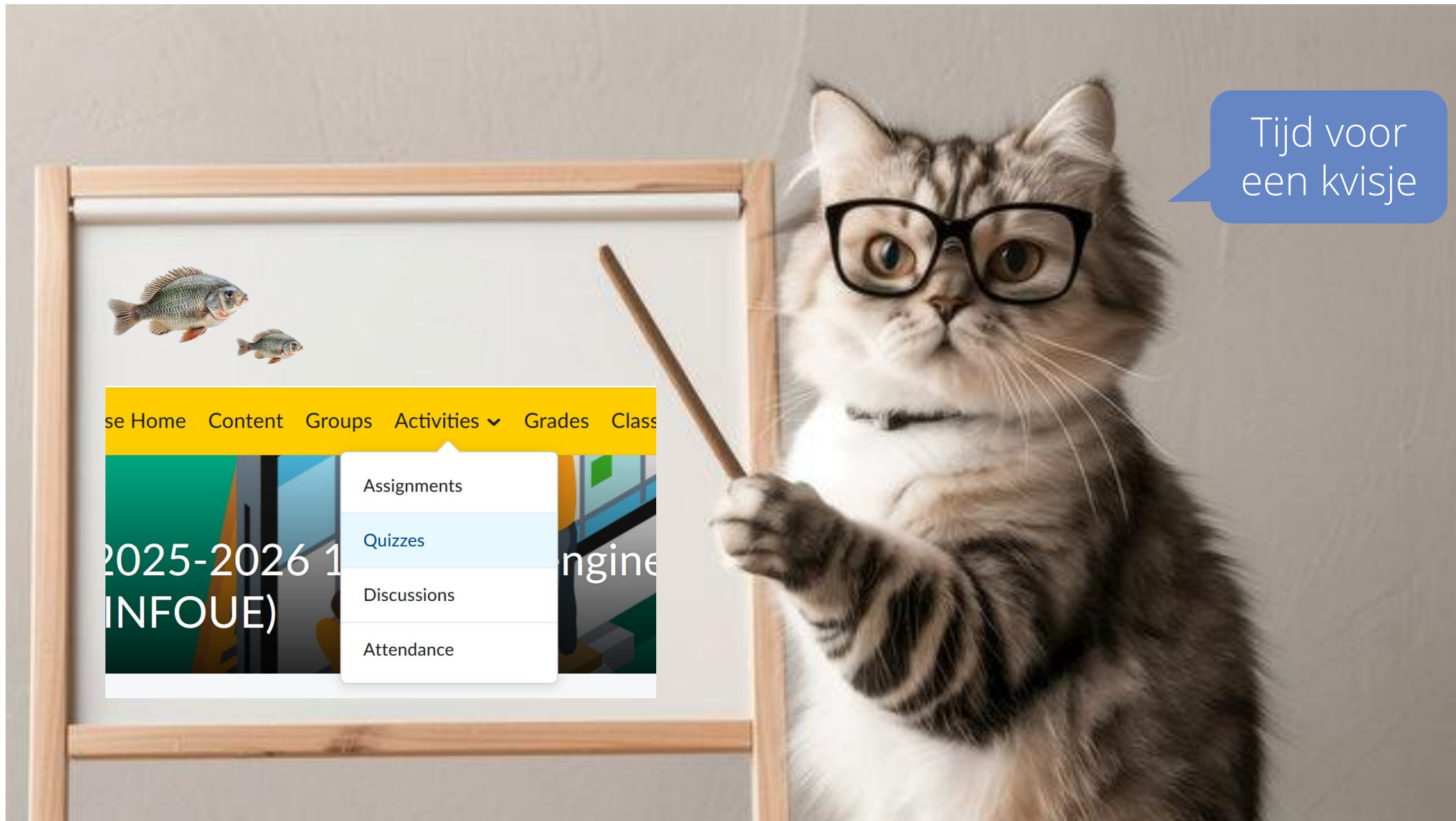
Experimenteel design



Experimenteel design

*Onderzoeksvraag +
onderzoekshypothese*





Tijd voor
een kvisje

Onderzoekshypothese $\neq$ theorie

- Een **hypothese** is een korte, precieze probleem-beschrijving die getest kan worden met 1 experiment
- Een **theorie** is breder en vereist verschillende experimenten om geverifieerd te worden

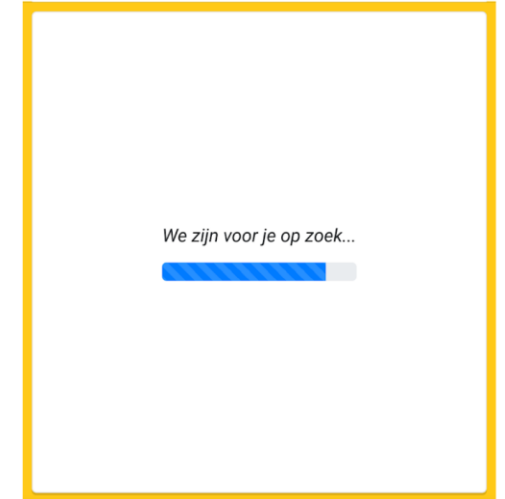
Voorbeeld uit een UEUX-paper

Onderzoeksvraag:

"With longer waits on mobile devices, is there a difference between how waiting is experienced between different waiting strategies (no activity, passive waiting or active waiting)?"

Onderzoekshypothese:

Experiencing waiting can be improved with elaborate feedback and enhanced progress indicators e.g. animations and visuals. Another option is distraction by providing a passive or active task during the wait.



(b) No activity



(c) Passive waiting



(d) Active waiting

Nulhypothese en alternatieve hypothese

- **Nulhypothese:** experimentele condities verschillen niet

vb. H_0 : Er is **geen verschil** in tijd nodig om een pagina te vinden bij een pull-down of een pop-up menu.

- **Alternatieve hypothese:** tegendeel van nulhypothese

vb. H_1 : Er is **een verschil** in tijd nodig om een pagina te vinden bij een pull-down of een pop-up menu.

Waarom is
dit zo?



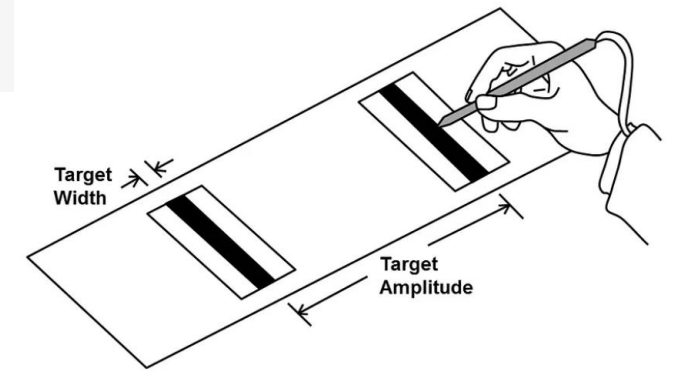
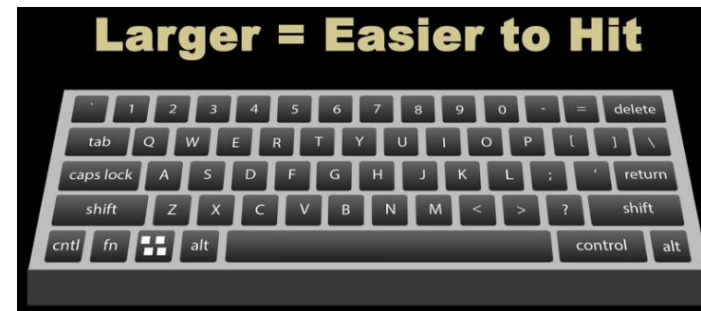
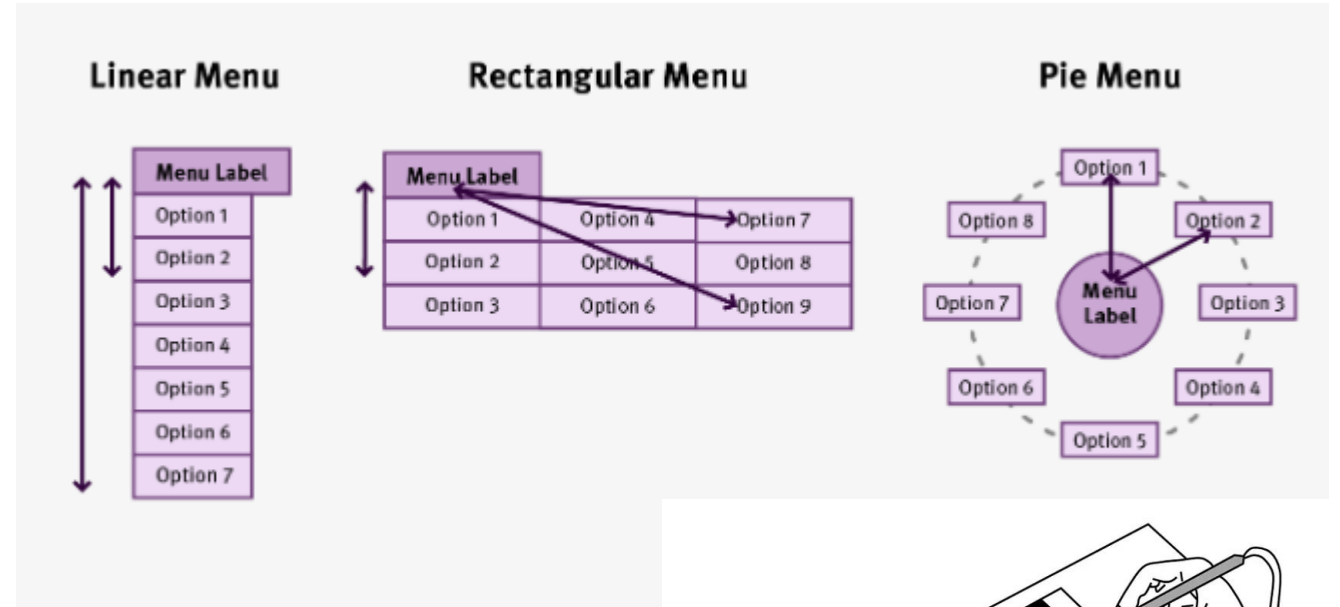
Falsificatie!



Een bekende theorie in UEUX: Fitt's law

Fitt's law beschrijft de relatie tussen beweegtijd, afstand en doelgrootte:

bij beweegtaken neemt de beweegtijd loglineair toe naargelang de beweegafstand toeneemt en de grootte van het doel afneemt

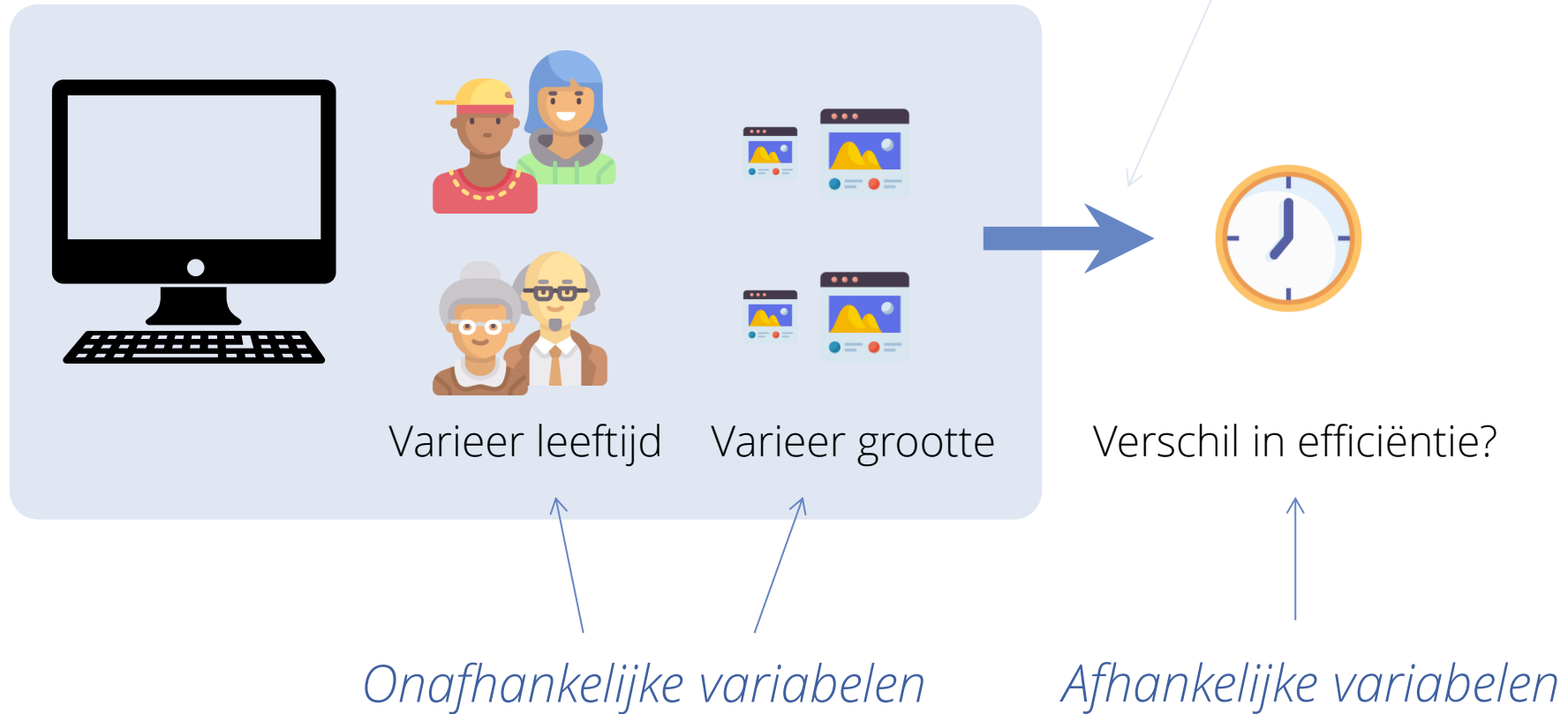


Source: <https://www.yorku.ca/mack/hci1992.html>



Experimenteel design

*Onderzoeksvraag +
onderzoekshypothese*

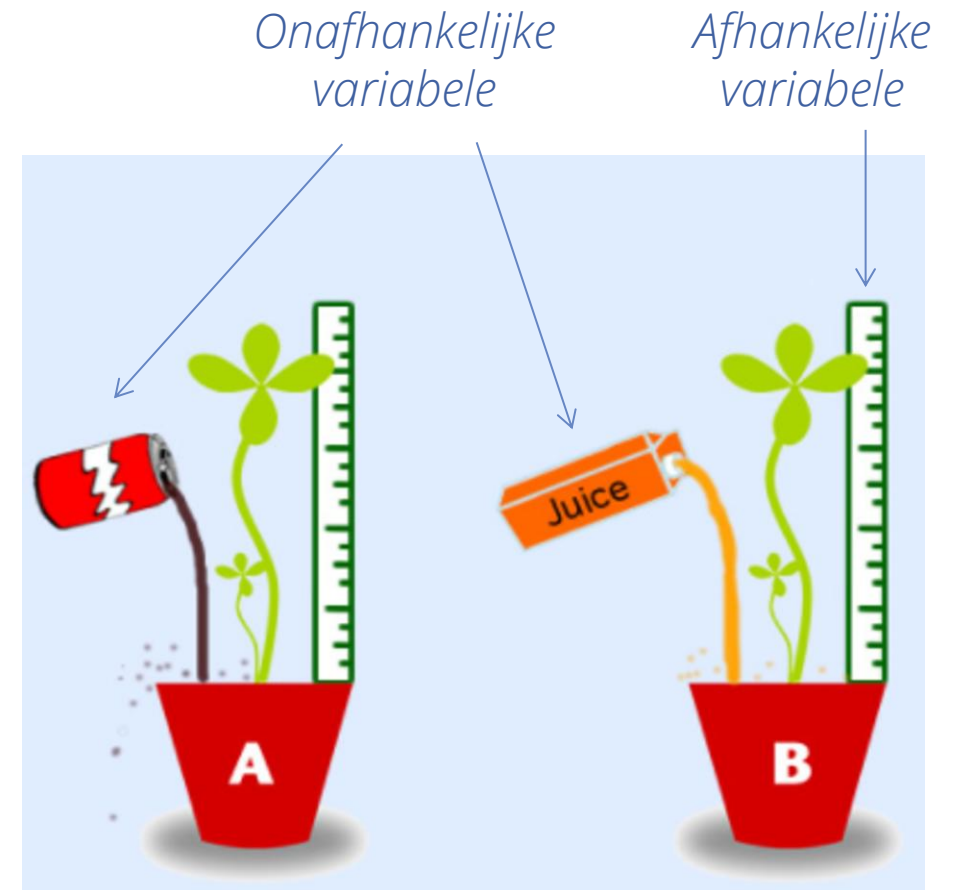


Afhankelijke en onafhankelijke variabelen

“Afhankelijk” = de variabele hangt af van het gedrag van deelnemers

“Onafhankelijk” = de variabele waarin onderzoekers geïnteresseerd zijn

Algemeen idee: leiden veranderingen in de onafh. variabelen tot veranderingen in de afh. variabelen en zo ja, welke?

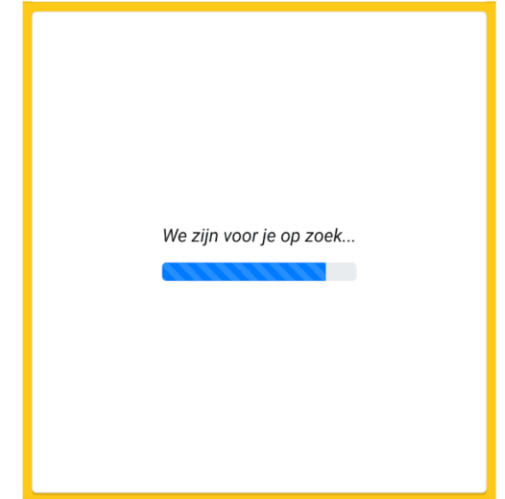


Voorbeeld uit een UEUX-paper

*Onafhankelijke variabele
= verschillende wachtervaringen:*

Afhankelijke variabelen:

- **Perceived waiting time** (feeling) - *What do you think of the duration of the waiting?* A 7-point Likert-scale ranging from 'very long' to 'very short' is used.
- **Estimated waiting time** (seconds) - *How many seconds do you think you have waited?* Participants saw a grid with values one to twenty where they had to select a number .
- **Enjoyment** (feeling) - *What do you think of the waiting in general?* A 7-point Likert-scale ranging from 'very boring' to 'very entertaining' was used to determine the user's affective response to the waiting experience.



(b) No activity



(c) Passive waiting



(d) Active waiting

Typische onafhankelijke variabelen in UEUX



Over technologie: type (typen vs dicteren, muis vs joystick vs touch pad ...) of design (pull-down vs pop-up menu, tekstgrootte, kleurgebruik ...)

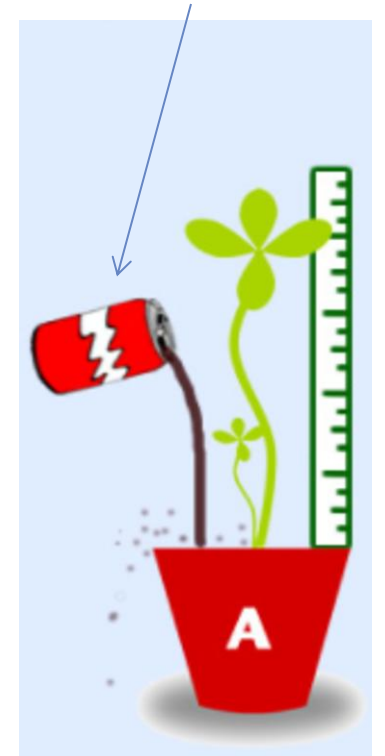


Over mensen: leeftijd, gender, ervaring, beroep, opleiding, cultuur, beperkingen ...



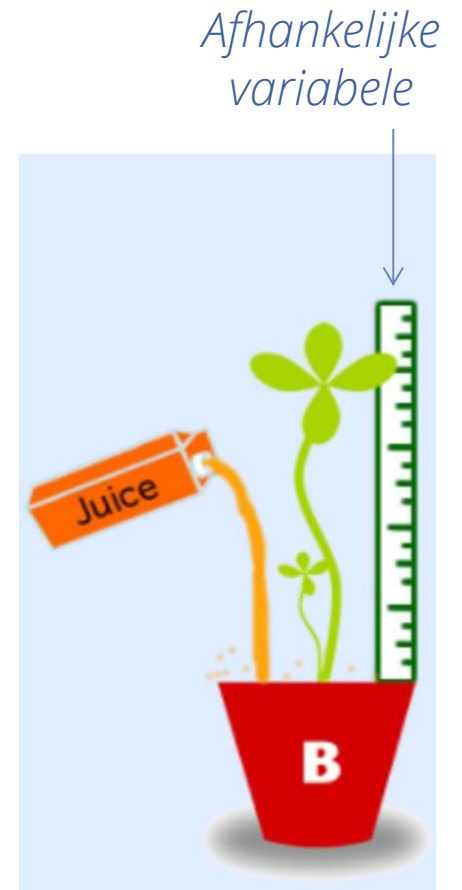
Over gebruikscontext: omgevingsfactoren (geluid, licht, temperatuur ...) en sociale factoren (groepsdruk ...)

*Onafhankelijke
variabele*



Typische afhankelijke variabelen in UEUX

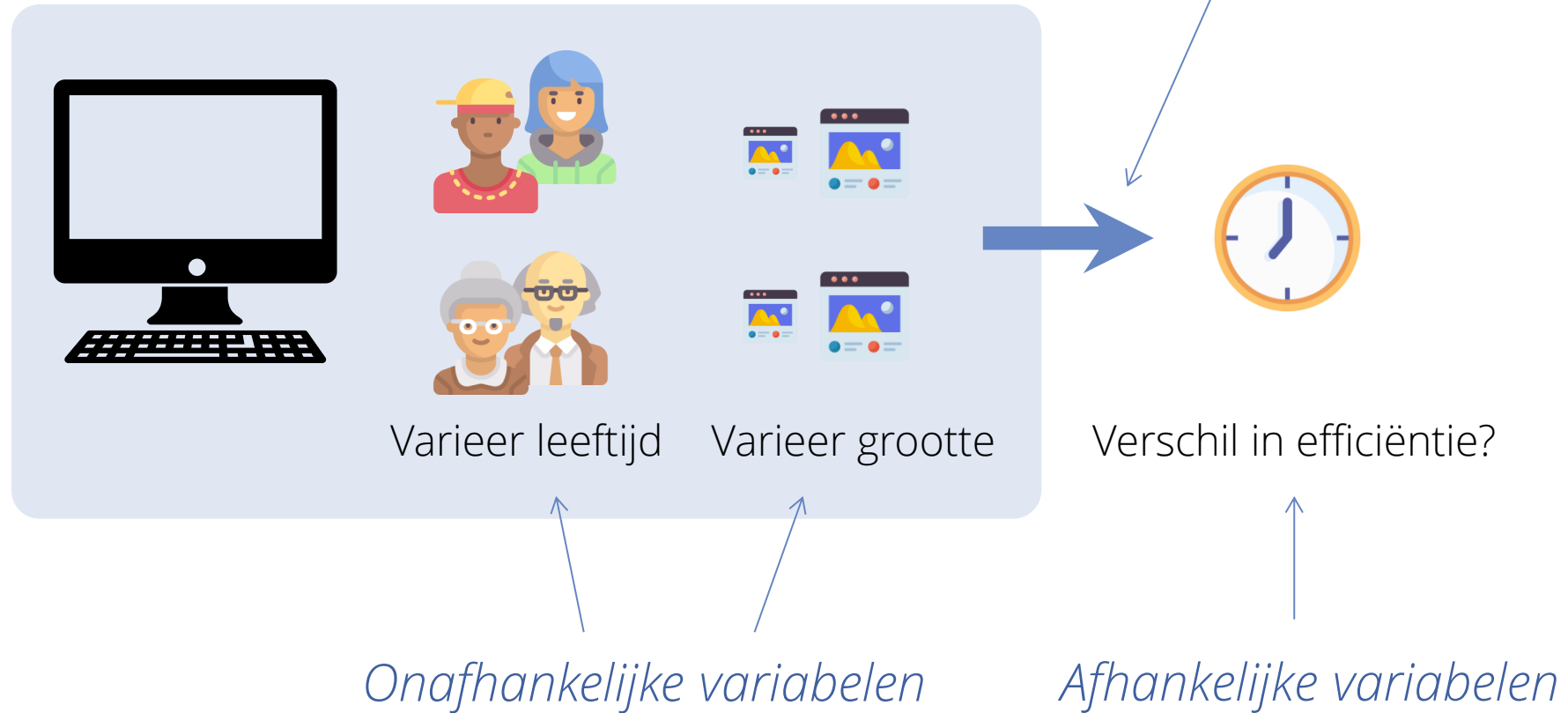
- **Efficiëntie:** snelheid, tijd (bv. woorden/minuut) ...
- **Accuraatheid:** aantal fouten
- **Tevredenheid:** zelfgerapporteerd met Likert-schalen
- **Leergemak en retentie:** hoe vlot kunnen taken geleerd worden en hoe lang zijn ze te onthouden
- **Fysieke en mentale belasting:** stress, vermoeidheid, biosignalen ...



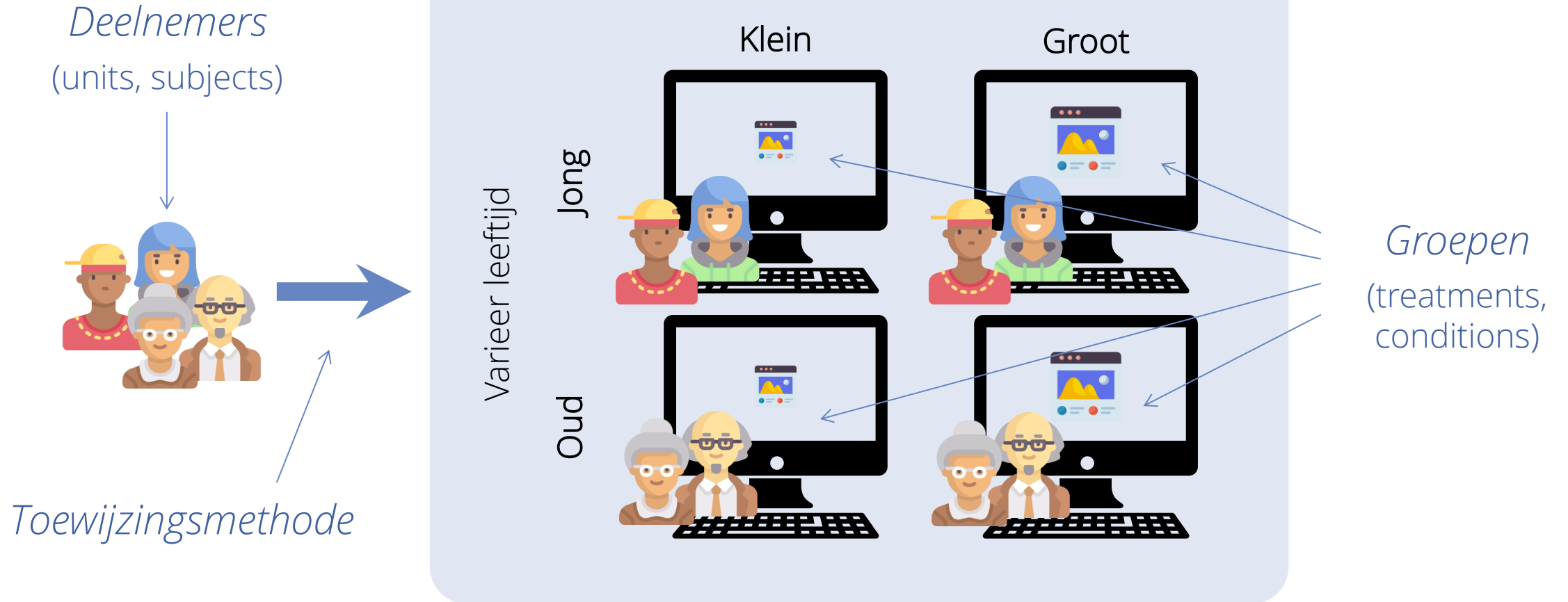
En hoe helpt dit
allemaal om causale
verbanden te testen?



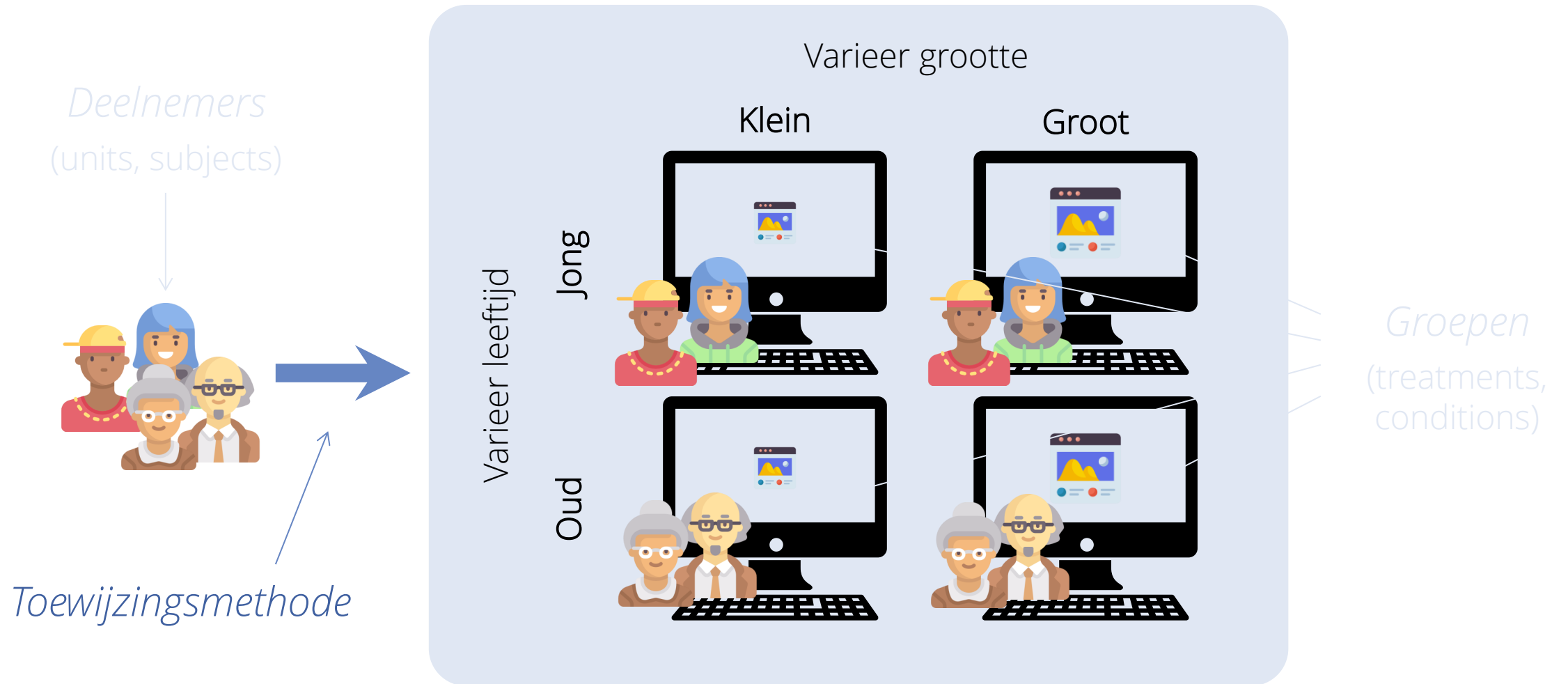
Experimenteel design



Experimenteel design



Experimenteel design



Toewijzingsmethodes

Line	Random Digits				
000	10097	32533	76520	13586	34673
001	37542	04805	64894	74296	24805
002	08422	68953	19645	09303	23209
003	99019	02529	09376	70715	38311
004	12807	99970	80157	36147	64032
005	66065	74717	34072	76850	36697

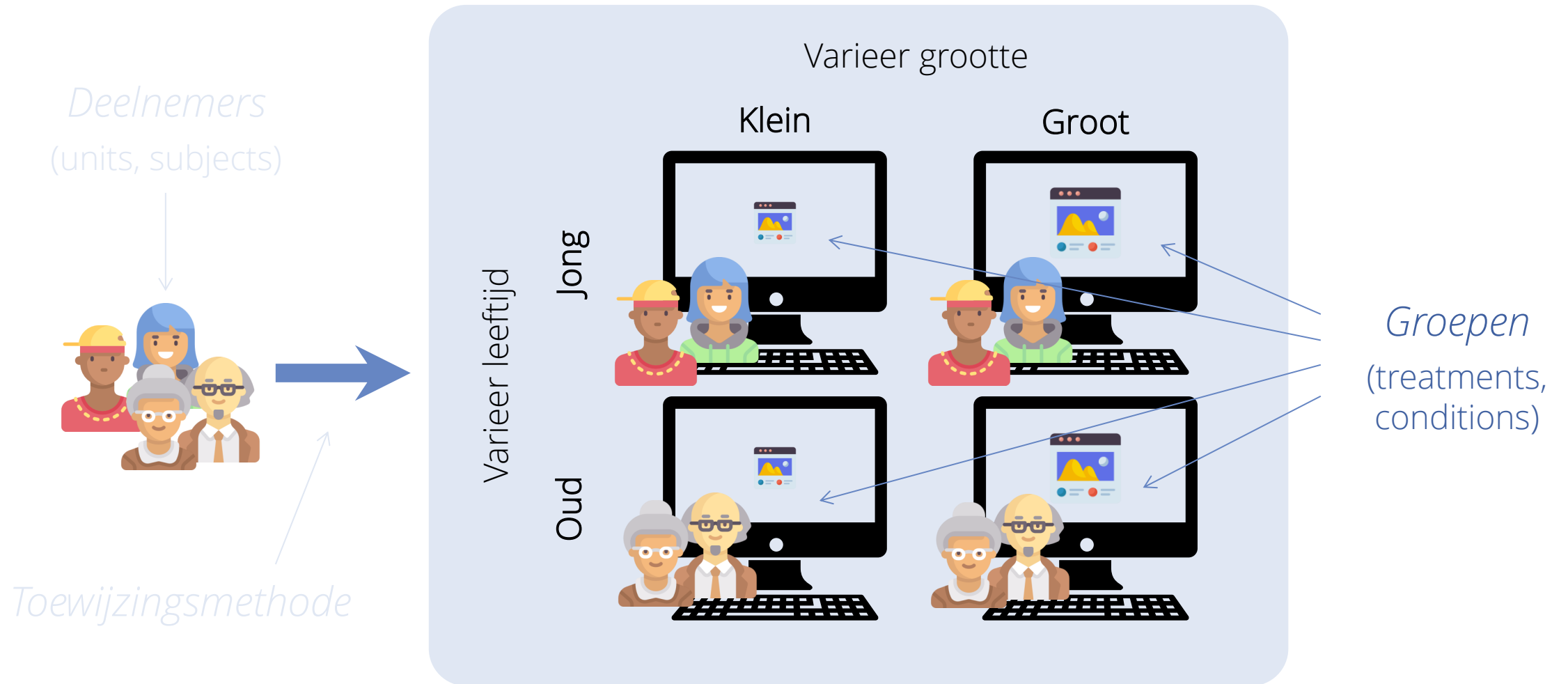


Volledig gerandomiseerd: niemand kan voorspellen in welke groep deelnemers zullen belanden, inclusief de onderzoekers (vb. random digit table, tossen ...)

Balancing (counterbalancing): gebruik een Latijns vierkant wanneer testpersonen meerdere condities doorlopen om leereffecten in rekening te brengen

1	2	3	4	5
2	1	5	3	4
3	4	2	5	1
4	5	1	2	3
5	3	4	1	2

Experimenteel design



Opbouw van dit hoorcollege

Deel 1: experimenteel onderzoek

Intermezzo: statistische toetsen

Deel 2: experimenteel design

Deel 3: inleiding opdracht 2

Literatuur:

Lazar et al.: H2, H3

Om relaties tussen groepen te testen heb je statistiek nodig...





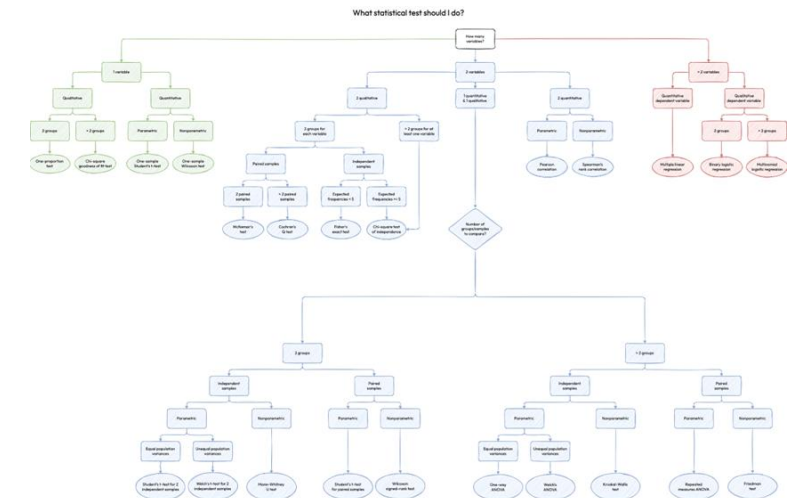
Het doel van statistische toetsen

Hoe zeker kunnen we zijn dat geobserveerde effecten in een steekproef ook gelden in de volledige populatie?



Tip: doe niet zomaar wat... 🤪

- Begrijp hoe de toets werkt
- Controleer aannames voordat je een toets uitvoert (vb. verdeling van de data)
- De toets hangt af van de hypothese, de steekproef en het experimenteel design



Statistische toetsen zijn feilbaar!

		Populatie	
Steekproef	H_0 waar	H_0 onwaar	
	H_0 waar 	Type II-fout β <i>vals negatief</i>	
	H_0 onwaar Type I-fout α <i>vals positief</i>		



StatQuest with Josh Starmer 

@statquest · 1.1M subscribers · 267 videos

Statistics, Machine Learning and Data Science can sometimes seem like very scary topics,...

patreon.com/statquest and 4 more links

Subscribe

Join

<https://youtube.com/shorts/qioNUZGVH1A>
<https://youtube.com/shorts/P4aYC1kS4d0>
<https://www.youtube.com/@statquest>

Type 1 Errors:

False Positive!

Type 2 Errors:

False Negative!



Universiteit
Utrecht

Populatie

Steekproef

	H_0 waar	H_0 onwaar
H_0 waar		Type II-fout β <i>vals negatief</i>
H_0 onwaar	Type I-fout α <i>vals positief</i>	

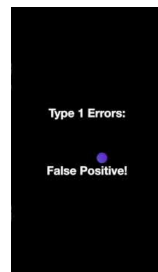
“When you say something's false but it really is true”



Hangt af van α , n , σ en de effectgrootte

$1 - \beta$ heet de “power” van de toets en in HCI willen we vaak $1 - \beta = 0,80$

“When you say something's true but it's really not true”



In HCI kiezen we vaak $\alpha = 0,05$

Opbouw van dit hoorcollege

Deel 1: experimenteel onderzoek

Intermezzo: statistische toetsen

Deel 2: experimenteel design

Deel 3: inleiding opdracht 2

Literatuur:

Lazar et al.: H2, H3

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

zie hoorcollege 3

Wacht even, wat is het verschil met usability engineering?



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

Usability engineering

- Heeft producten verbeteren als doel
- Experimenten lijken nogal sterk op summatief testen
- Heeft gecontroleerde condities voor de test
- Is vaak niet helemaal replicabel
- Resultaten leiden tot inzichten
- Wordt vaak gebruikt met bv. ontwikkelaars

Experimenteel onderzoek

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

*Laat in het
ontwerpproces*



Summatief testen

High-fidelity prototypes eerder formeel testen om de effectiviteit van specifieke ontwerpkeuzes te testen

Voordeel: op taakniveau kwantitatieve vergelijkingen doen

Nadeel: moeilijk om het ontwerp nog sterk te wijzigen

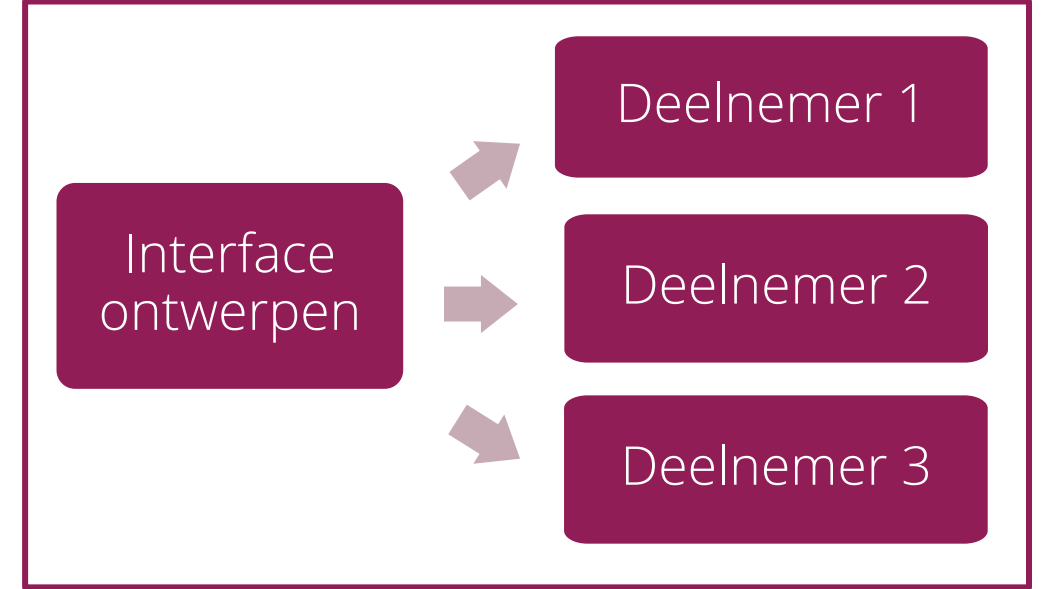
Inderdaad: alleen het doel verschilt als er groepen zijn, er een streng protocol is en er statistiek wordt toegepast



Usability testing vs experimenteel onderzoek

Algemeen doel is kennis vergaren met een ontwerp

Experimenteel onderzoek



Algemeen doel is een ontwerp iteratief verbeteren

Usability testing





Experimenteel onderzoek
is duidelijk perfect

Euh, er zijn
limitaties...

🧠 ?!

Welke limitaties
heeft experimenteel
onderzoek?

BRAIN
BREAK



Limitaties van experimenteel onderzoek

Ambigüiteit: hypothesen opstellen is moeilijk, want veel HCI-problemen zijn niet eenduidig gedefinieerd of hangen van veel factoren af

Verstorende variabelen: niet elke factor die de afhankelijke variabelen beïnvloedt kan worden gecontroleerd

Hawthorne-effect: mensen gedragen zich soms anders tijdens experimenten (<https://www.youtube.com/watch?v=W7RHjwmVGhs>)

Opbouw van dit hoorcollege

Deel 1: experimenteel onderzoek

Intermezzo: statistische toetsen

Deel 2: experimenteel design

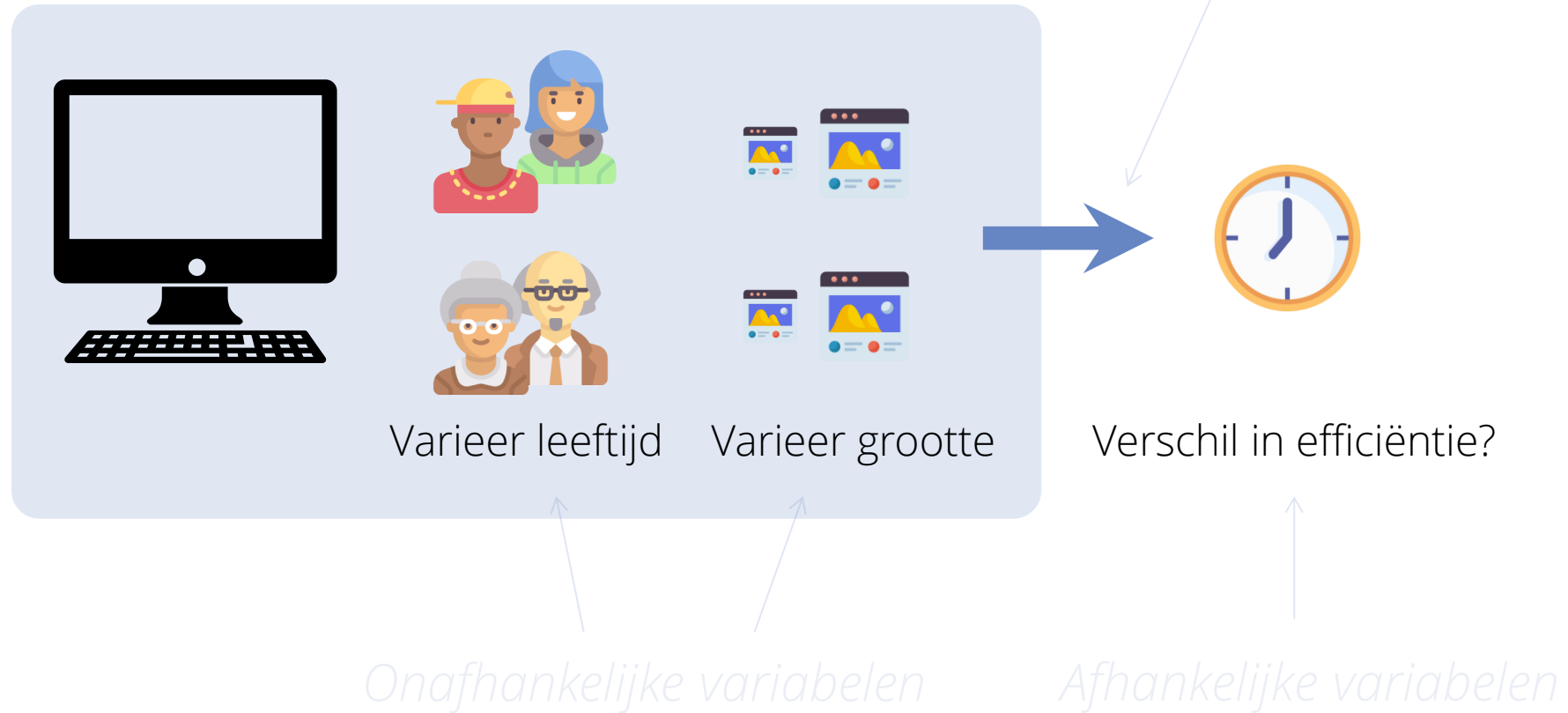
Deel 3: inleiding opdracht 2

Literatuur:

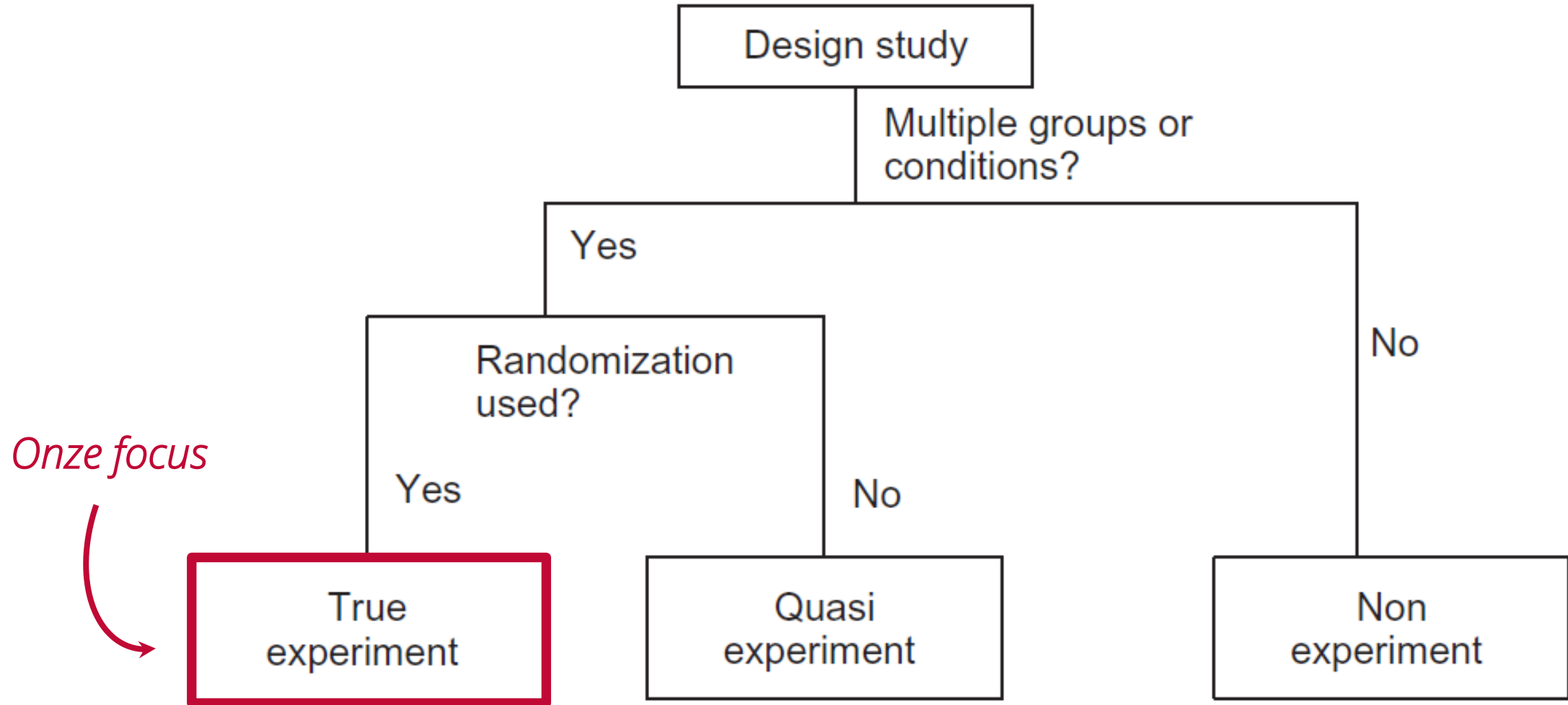
Lazar et al.: H2, H3



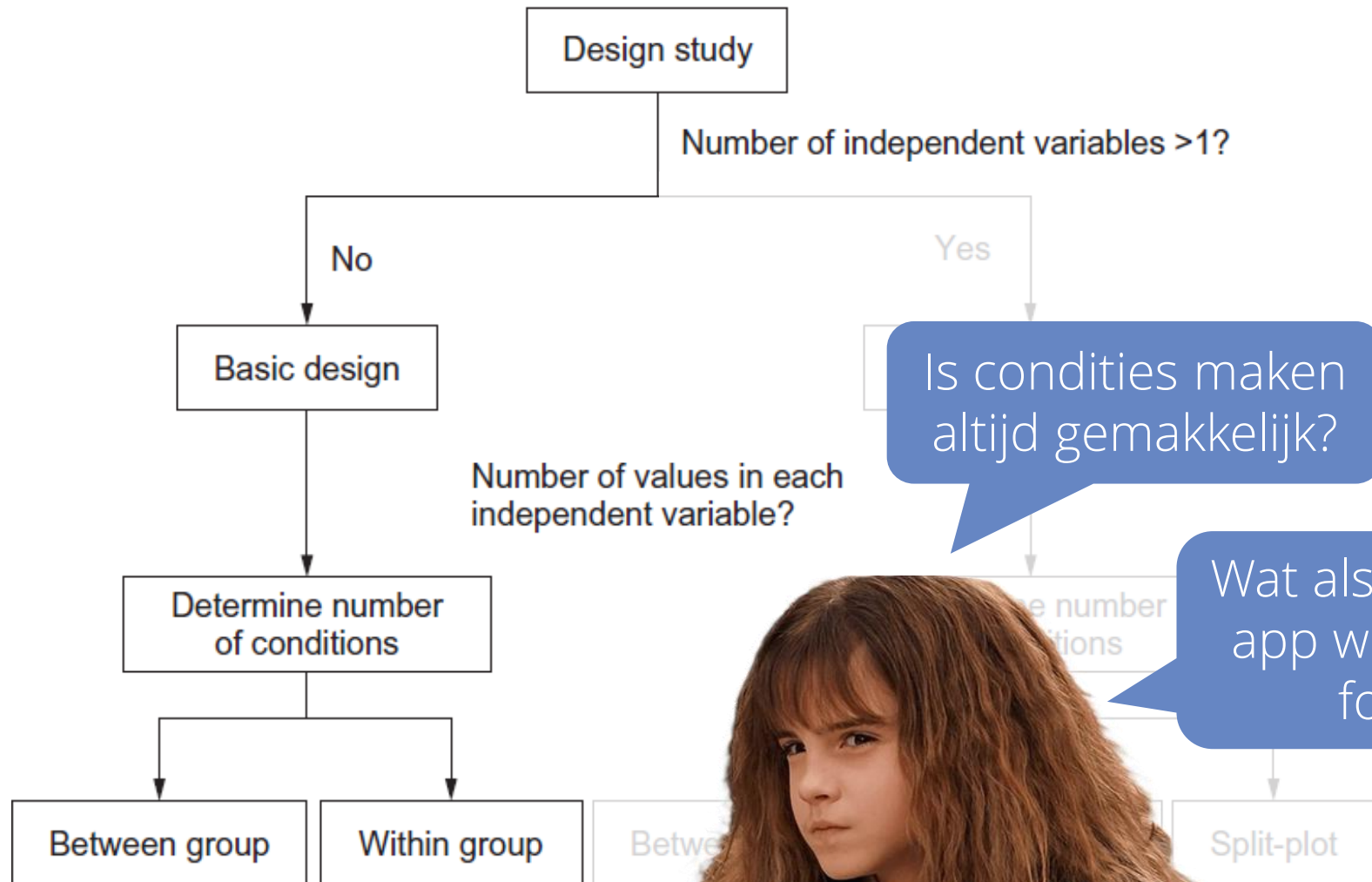
Experimenteel design



Het design hangt af van het type experiment



Verschillende experimentele designs



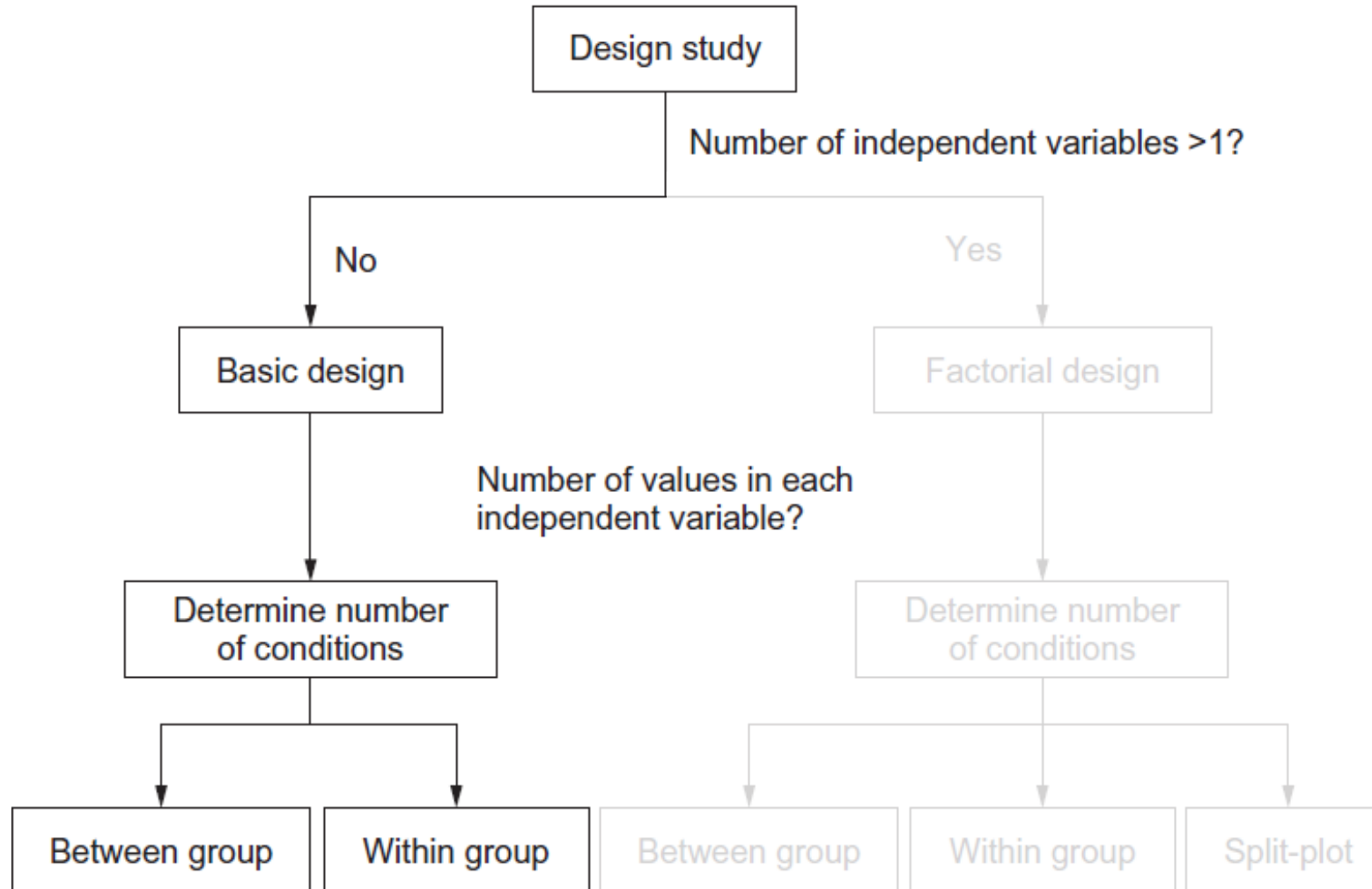
Wizard-of-Oz experimenten

- Verborgen persoon simuleert geavanceerde functionaliteiten
- Nuttig om prototypes voor toekomstige functionaliteiten te testen
- Extra achtergrond:
<https://www.smashingmagazine.com/2025/07/unmasking-magic-wizard-oz-method-ux-research/>



<https://youtu.be/hKrWSXVx1bE>

Verskillende experimentele designs



Verschillende basisdesigns

Deelnemers behoren tot precies 1 groep

Determine number of conditions

Alle deelnemers doorlopen elke conditie

Between group

Within group

Groep 1

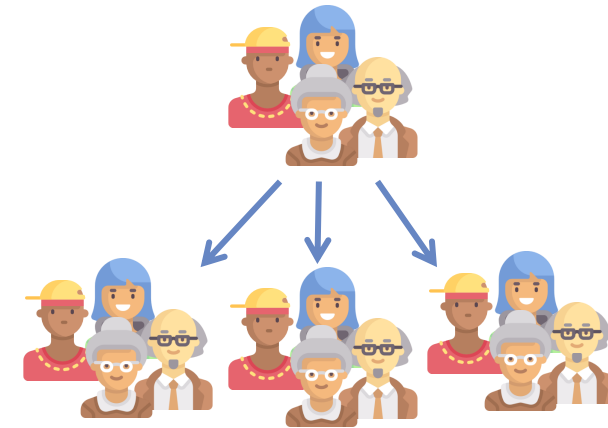
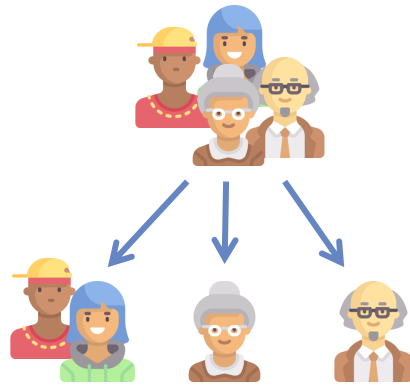
Groep 2

Groep 3

Groep 1

Groep 2

Groep 3



Between-Group Design

Advantages

Cleaner
Avoids learning effect
Better control of confounding factors, such as fatigue

Limitations

Larger sample size
Large impact of individual differences
Harder to get statistically significant results

Within-Group Design

Smaller sample size
Effective isolation of individual differences
More powerful tests

Hard to control learning effect
Large impact of fatigue

Counterbalancing!

Wanneer kies je voor een *between-group* design?

- Het experiment omvat eenvoudige taken waarbij de impact van menselijke verschillen klein is
- Het experiment omvat taken waarbij een groot leereffect kan ontstaan
- Deelnemers kunnen geen deel uitmaken van meerdere groepen, bv. groepen op basis van leeftijd

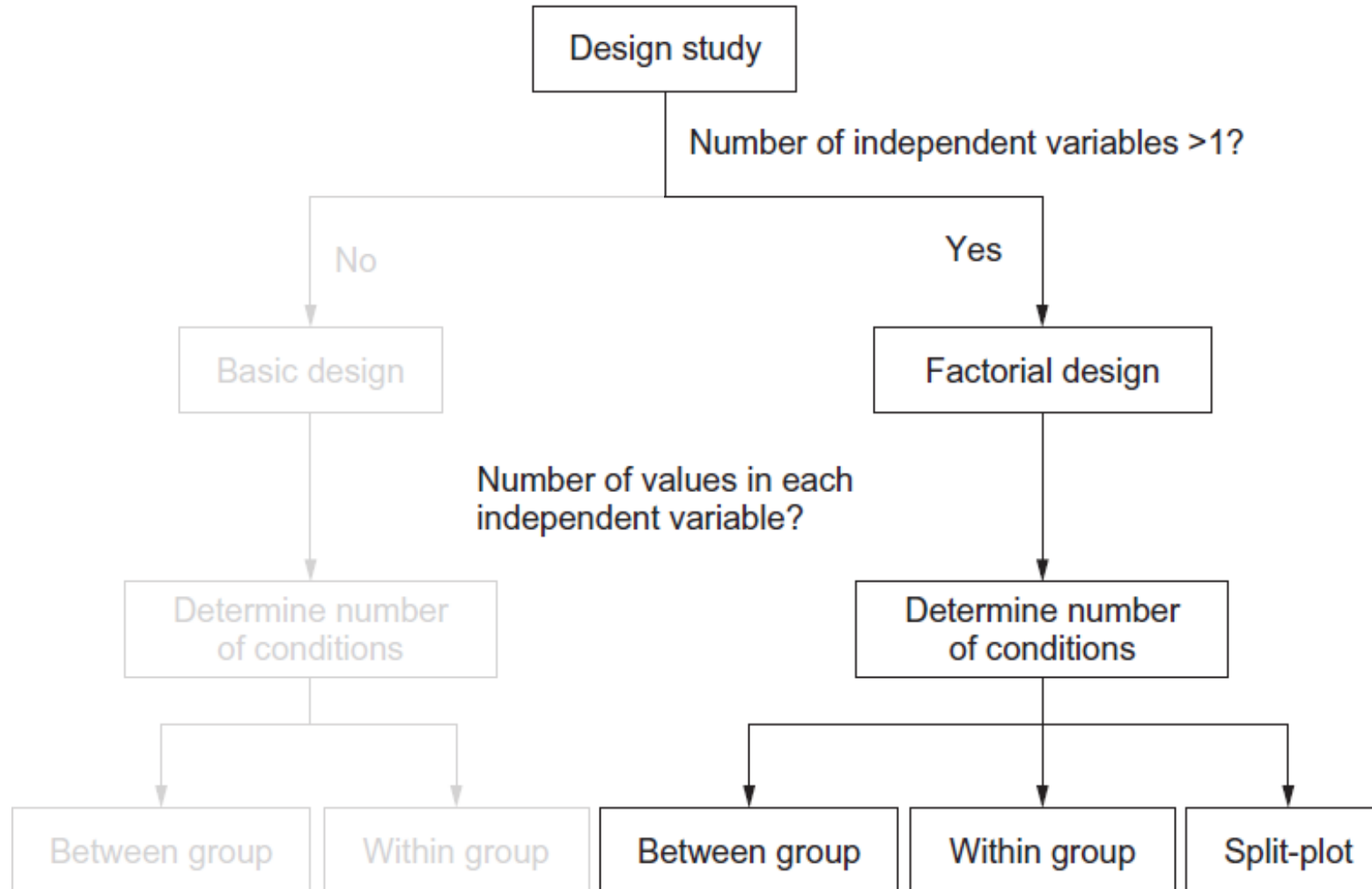
Wanneer kies je voor een *within-group* design?

- Het experiment omvat moeilijke taken waarbij de impact van menselijke verschillen groot is, vb. begrijpend lezen, problemen oplossen ...
- Het experiment omvat taken waarbij weinig leereffecten kunnen ontstaan
- Het aantal beschikbare deelnemers is klein, bv. de populatie is erg specifiek

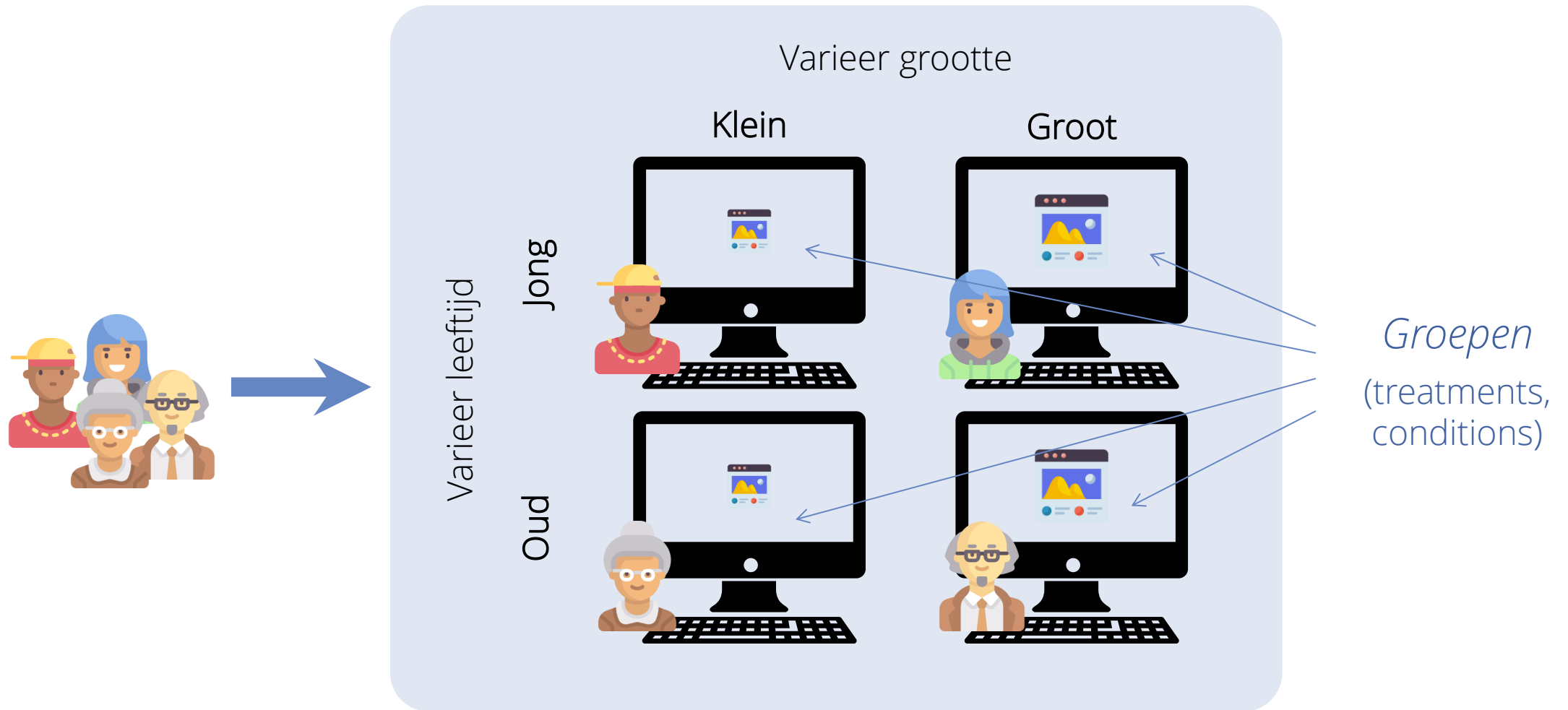
Voorzie rustpauzes en laat experimenten max. 60-90 min duren



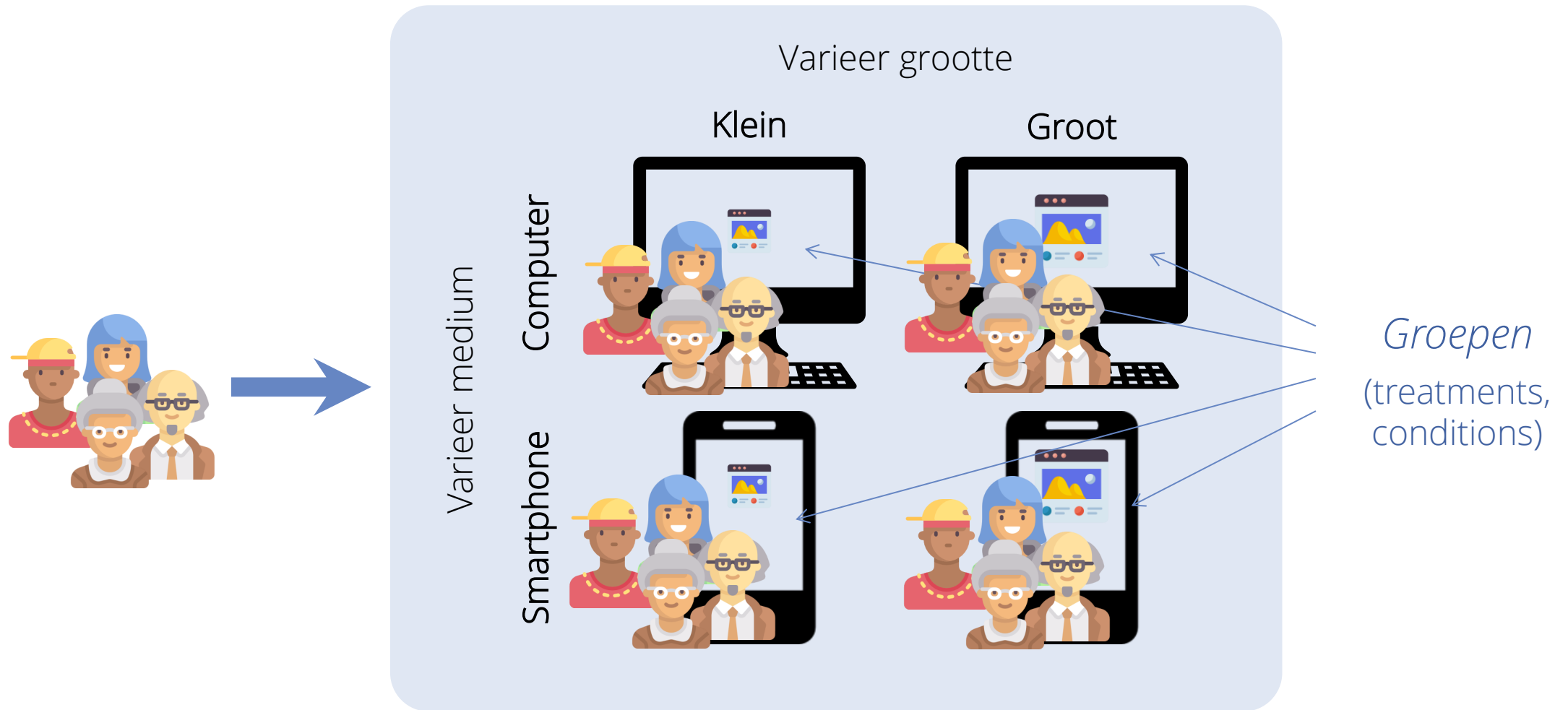
Verskillende experimentele designs



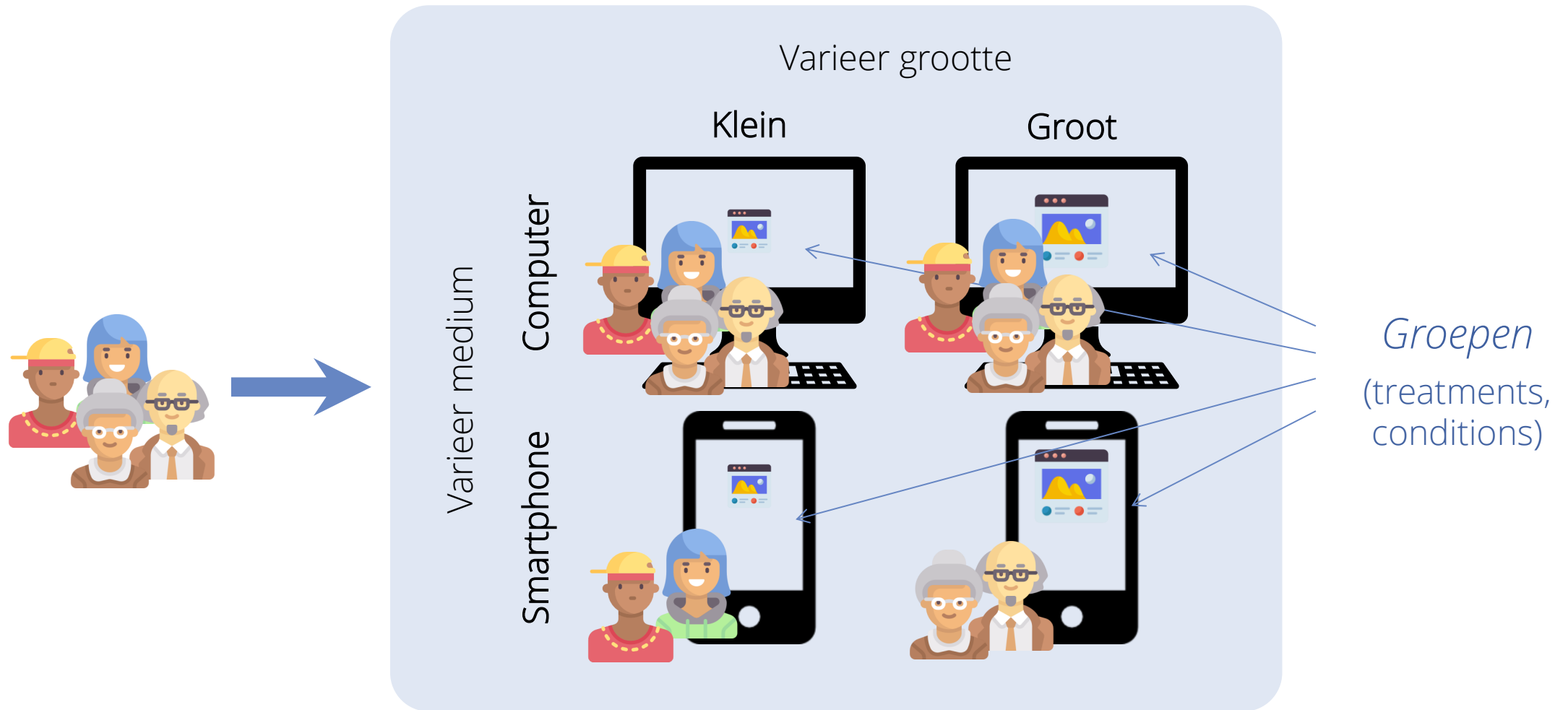
Factorieel *between-group* design



Factorieel *within-group* design



Factorieel *split-plot* design

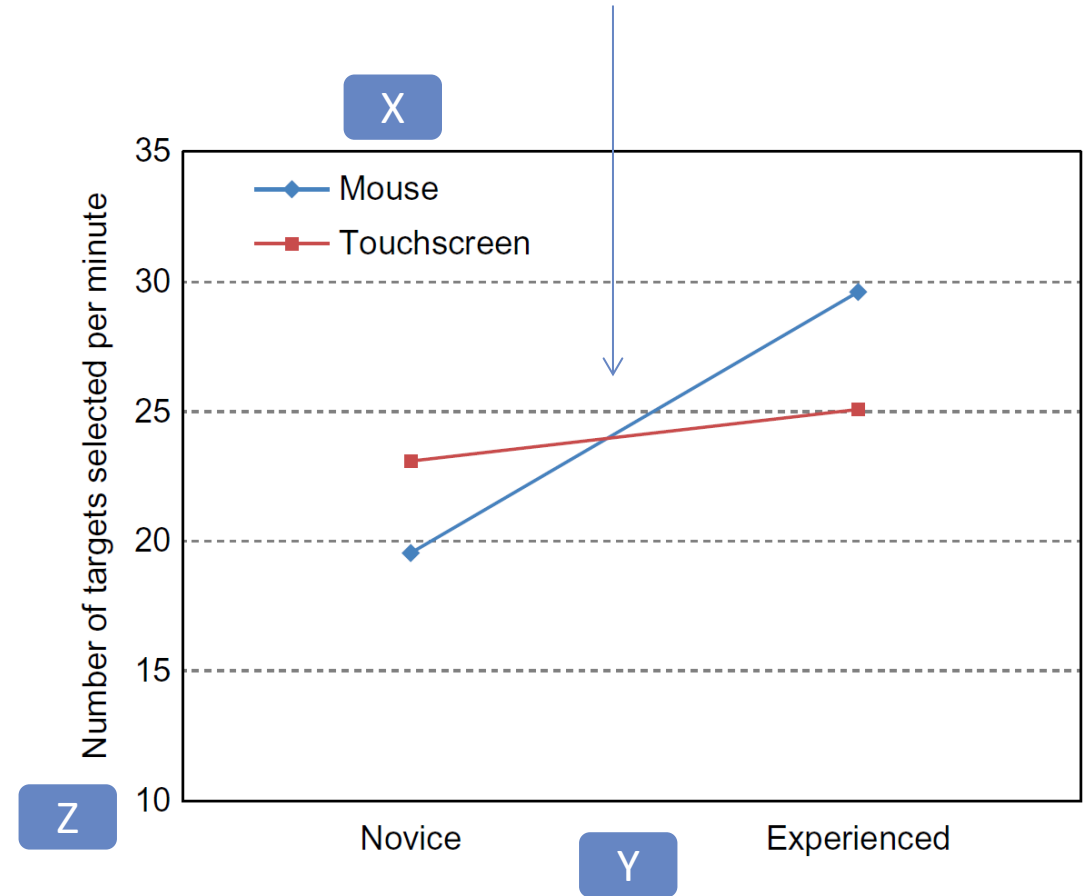


Interactie- vs hoofdeffecten

Er is een **interactie-effect** tussen onafhankelijke variabelen X en Y als ze samen een afhankelijke variabele Z bepalen

Als een onafhankelijke variabele X een afhankelijke variabele Z bepaalt zonder te interageren, dan is dat een **hoofdeffect**

Interactieplot: als verschillende lijnstukken snijden, interageren de onafhankelijke variabelen



Experimenten analyseren
kan dus moeilijk worden ...

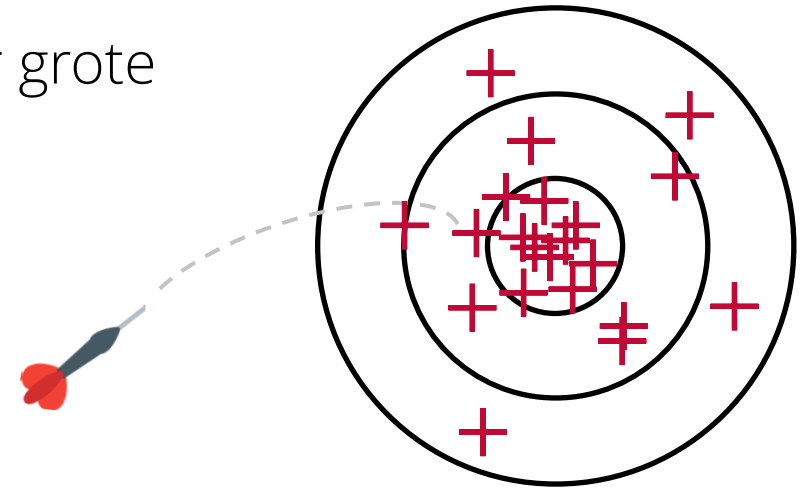
Jep

Experimentele metingen zijn feilbaar

Observaties = echte waarden + willekeurige fouten

Willekeurige fouten ("noise") correleren niet met de echte waarden, maar zijn ruis

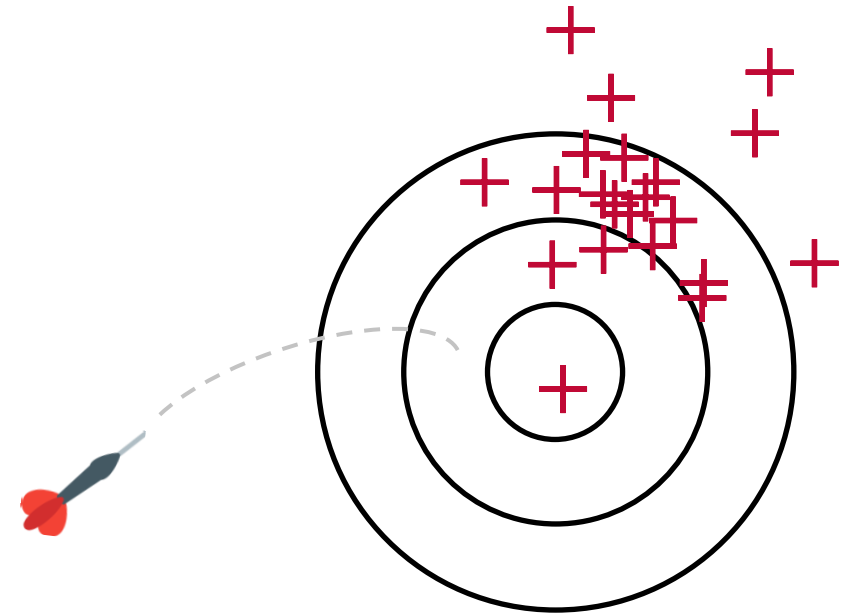
Willekeurige fouten zijn onvermijdelijk, maar grote steekproeven onderdrukken hun impact



Experimentele metingen zijn feilbaar

Systematische fouten ("bias") vertekenen alle metingen in dezelfde richting. Mogelijke oorzaken:

- meetinstrument
- experimentele procedure
- deelnemers
- gedrag van onderzoekers
- omgeving tijdens experiment
- ...



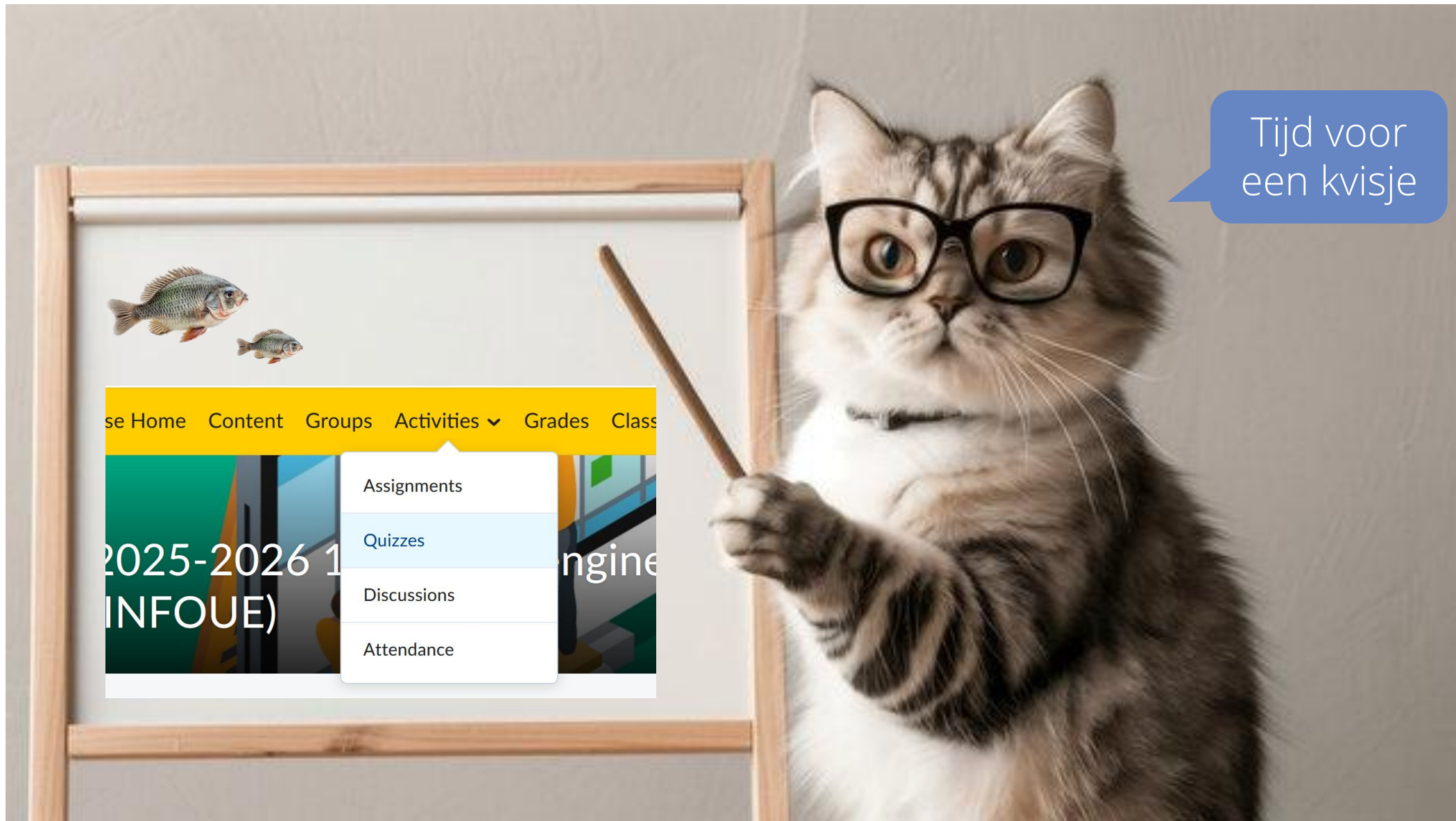


Sociale druk tijdens experimenten: <https://youtu.be/qA-gbpt7Ts8>

Take-aways voor dit hoorcollege



- Experimenten testen causale verbanden tussen onafhankelijke variabelen enerzijds en afhankelijke variabelen anderzijds
- Experimenten kunnen een *between-group*, *within-group* of *split-plot* design hebben
- Experimenten zijn erg krachtig, maar hebben limitaties



Opbouw van dit hoorcollege

Deel 1: experimenteel onderzoek

Intermezzo: statistische toetsen

Deel 2: experimenteel design

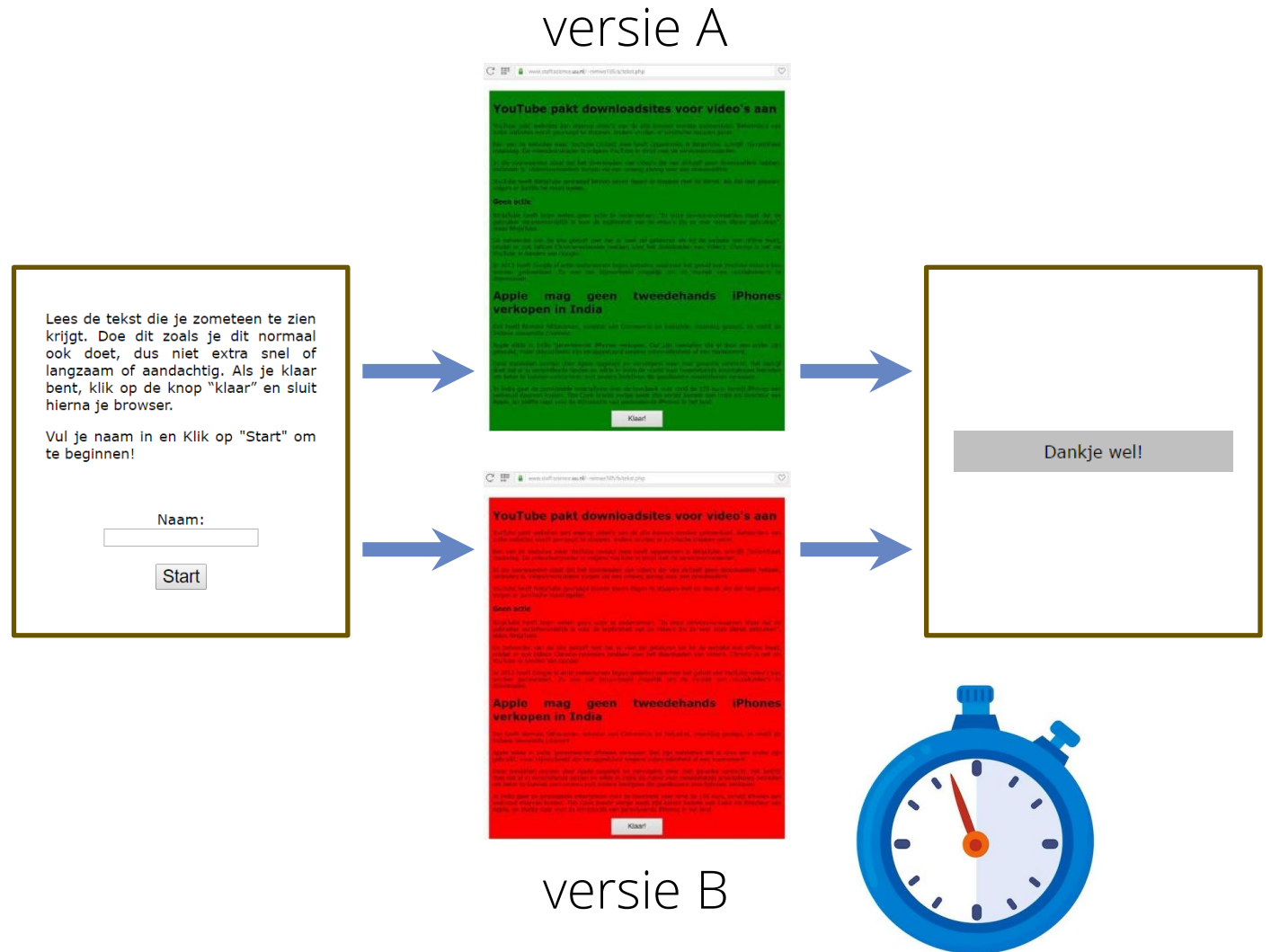
Deel 3: inleiding opdracht 2

Literatuur:

Lazar et al.: H2, H3

Opdracht 2: doe een A/B-test

- Gerandomiseerd *between-group* experiment met 2 condities
= 1 onafhankelijke variabele met 2 waarden (bv. achtergrondkleur)
- Min. 1 afhankelijke variabele, waaronder tijd (bv. leestijd)
- Gebruik een zelfgebouwde timer en **smartlook**



Opdracht 2: planning

Fase 1: A/B-test voorbereiden

- Literatuur verkennen & onderzoeksvraag formuleren
- Website met logging opzetten
- Rekrutering voorbereiden

Fase 2: A/B-test uitvoeren & verzamelde data analyseren

Fase 3: artikel indienen en poster presenteren



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 - Opdracht 2: onderzoeksvraag formuleren en literatuurstudie



COLUMNS | BITS TO BITES
Jie Li, EPAM

How Far Can We Go with Synthetic User Experience Research?

User experience research (UXR) plays an important role in uncovering user personas, their contexts of use, and their detailed needs. UXR employs a range of research methodologies, such as surveys, interviews, diary studies, behavior, I spent days at festivals, theme parks, concert halls, and train stations, observing the construction of sites and architecture, as well as the placement of facilities such as fences and one-way gates. A few years ago, I spent a week conducting observational studies on an outside broadcasting truck at the MotoGP races at Silverstone. To understand live broadcasts and live commentary, I recently, along with a team of researchers, had to spend a week at Silverstone, with its extremely fast-paced live broadcasting sessions. It was often challenging to capture the essence of the experience.

such as observing how 50,000 people were evacuated from a music festival when a thunderstorm was forecast. Getting quickly adapted to the user environment can also be challenging, as evidenced by my experience with the MotoGP study. As I had to remain silent during observation and refrain from asking the broadcasting team any questions to avoid distractions, I struggled to understand each person's responsibility within the team and capture their jargon-heavy conversations and workflows during the extremely fast-paced live broadcasting sessions. It was often challenging to capture the essence of the experience.

with real human participants, to test a new product.

CONDUCTING FIELD USER RESEARCH: COMPLAINTS AND JOY
Since the start of my Ph.D. in 2012, I have had only a few chances to conduct field research. To study crowd management strategies and crowd behavior, I spent days at festivals, theme parks, concert halls, and train stations, observing the construction of sites and architecture, as well as the placement of facilities such as fences and one-way gates. A few years ago, I spent a week conducting observational studies on an outside broadcasting truck at the MotoGP races at Silverstone. To understand live broadcasts and live commentary, I recently, along with a team of researchers, had to spend a week at Silverstone, with its extremely fast-paced live broadcasting sessions. It was often challenging to capture the essence of the experience.

The reliability of facilitated research is now a topic often debated among researchers.

26 INTERACTIONS MAY-JUNE 2024

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 (verplicht) -
Research design exercise:
experimenten*

Niet verdwalen op
weg naar Dalton, hè





**Universiteit
Utrecht**

Sharing science,
shaping tomorrow