28/02/2018

Eerlijk en doeltreffend beoordelen met paarsgewijze vergelijking

Een toepassing van D-PAC

Maarten Goossens
FSW Universiteit Antwerpen

Overzicht

- Theoretische achtergrond comparatief beoordelen
- 4 jaar onderzoek: wat hebben we geleerd?
- De D-PAC tool
- Comparatief beoordelen in de praktijk

Waarom comparatief beoordelen? Theoretische achtergrond

Complexe competenties

NIET: enkel kennis, juist of fout,…
WEL: geheel van kennis, vaardigheden en attitudes

Bijvoorbeeld:

Schrijfvaardigheden
Portfolio's
Zelf-reflecties
Visieteksten
Wetenschappelijke rapporten
zelf-reflecties

Spreekvaardigheden
presentaties
voordrachten

Beeldende vaardigheden
moodboards
Grafisch ontwerp

Complexe beoordelingen

Komen we als beoordelaars wel overeen?

➔ betrouwbaarheid

Kijken we naar de belangrijke zaken en hoe waarderen we die?

➔ validiteit

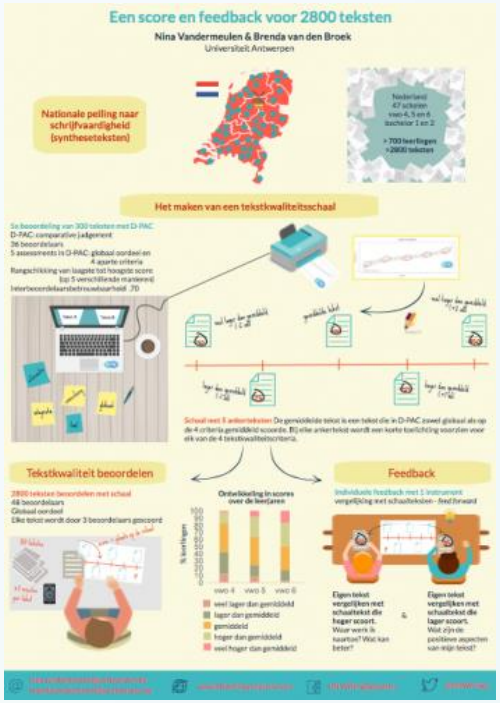
Beoordelingsstrategieën: de families
(Coertjens, 2017)

Holistisch beoordelen

“het geheel is meer dan de som van de delen”

Holistisch beoordelen

- Voorbeeld:



Links ligt een grote stapel ongeordende informatie: rapporten, notities, links naar websites. Rechts staat je tekstverwerker met een maagdelijk leeg scherm. De opdracht: schrijf over die berg informatie een heldere tekst. Een rapport, werkstuk, een scriptie. Hoe pak je dat aan? Waar begin je? Hoe los je de problemen op die zich tijdens het schrijven voordoen? Hoe argumenteer je, hoe gebruik je citaten en voetnoten? En ook: hoe verbeter je je taalgebruik?

Check je tekst helpt je bij de problemen en vragen die je tegenkomt als je een langere tekst moet schrijven, zoals een onderzoeksverslag of een adviesrapport. Het boekje is opgebouwd als een checklist: een lijst met eisen waaraan een goede tekst voldoet. Als je aan de hand van dit boekje de eisen stap voor stap naloopt, krijg je meer greep op je schrijfwerk. Je kunt gemakkelijker beoordelen hoe ver je bent met je tekst, wat er al af is en wat er nog moet gebeuren.

In het eerste deel van dit boek zie je hoe een goede tekst in elkaar zit, in het tweede waar je op moet letten bij het formuleren:

- Deel I, ***De tekst,*** beantwoordt vragen als: 'hoe moet ik mijn tekst beginnen?', 'hoe zit een goed verslag of rapport in elkaar?', 'hoe moet ik een citaat uit een boek verwerken in mijn tekst?'
- Deel II, ***De taal***, gaat in op taalvragen, zoals 'hoe kom ik erachter of mijn d's en t's goed zijn?', 'is het *alles dat* of *alles wat*', 'is het *aan hun* of *aan hen*?'

Achter in het boek vind je een checklist voor de controle achteraf, wanneer je al een eerste versie van je tekst hebt geschreven. Zo kun je snel beoordelen of je concepttekst nog moet worden bijgeschaafd: bij alle controlepunten vind je een verwijzing naar de paragraaf waarin je meer leest. En kun je overal een vinkje zetten? Dan heb je een heel goede kans dat je een prettig leesbare, samenhangende en overtuigende tekst hebt geschreven.

Waarop beoordelen?
Globale omschrijving van de competentie of 5 hoofdcriteria

Probleem:
De weging van de onderdelen binnen de competentie

➔ Betrouwbaarheid tussen beoordelaars

Beoordelingsstrategieën: de families
(Coertjens, 2017)

Analytisch beoordelen

Criterialijsten, rubrics

“opgeruimd staat netjes”

Holistisch beoordelen

“het geheel is meer dan de som van de delen”

Betrouwbaarheid

Opdracht: neem je pennenzak en een blad papier en teken een “mens”
Respondenten: kleuter en lagere school

➔ Scoor de volgende 4 tekeningen

Criteria “Kunstig competent”

- De vakman → Technische uitvoering
- De onderzoeker → Aanwezigheid kenmerken mens
- De kunstenaar → Kleur, Compositie, Fantasie
- (De performer)
- (De samenspeler)

Schaal van 1-5 per criteria

Resultaat?

	Tekening 1	Tekening 2	Tekening 3	Tekening 4
Technische uitvoering				
Kenmerken mens				
Kleur				
Compositie				
Fantasie				

Problemen met analytisch beoordelen

Lastig om te construeren

- Definiëring van het construct

Geen garantie voor betrouwbare en valide oordelen

- Nog steeds ruimte voor interpretatie van criteria
  - ➔ lage interbeoordelaarsbetrouwbaarheid
- Verschil in strengheid/mildheid van de beoordelaar
- Volgorde-effecten
- Zeer taakafhankelijk

Beoordelingsstrategieën: de families
(Coertjens, 2017)

Absoluut (per product)

Analytisch beoordelen

Criterialijsten, rubrics

“opgeruimd staat netjes”

Betrouwbaarheid

Validiteit

Holistisch beoordelen

“het geheel is meer dan de som van de delen”

Betrouwbaarheid

Analytisch

Holistisch

Vergelijkende methoden

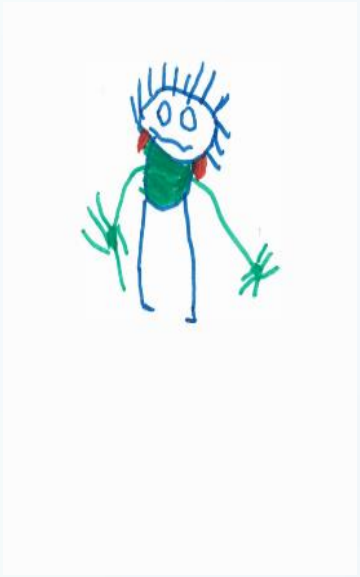
Paarsgewijs vergelijken

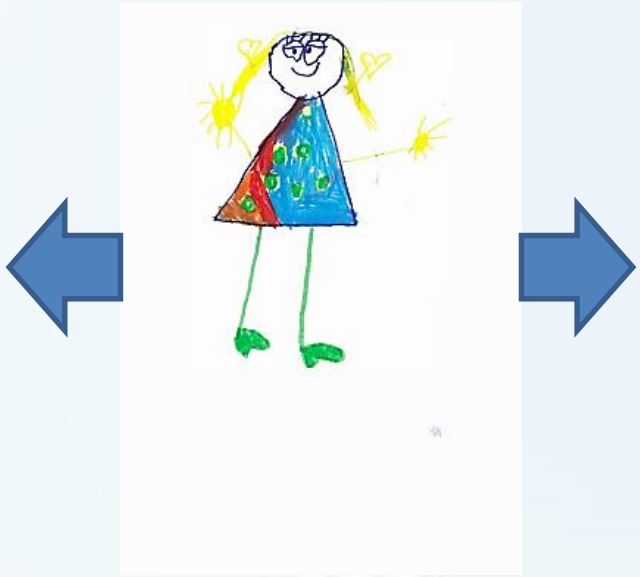
“vergelijken is puur natuur”

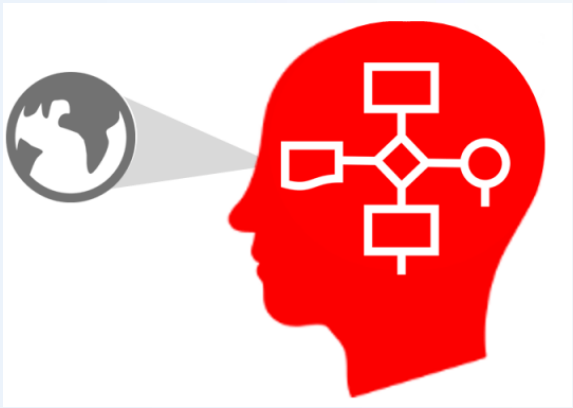
Comparatief

Wat gebeurt er in ons hoofd?

- Intern ga je vergelijken
  - Deze tekening is beter, hoger cijfer
  - Trachten te differentiëren d.m.v. die cijfers
  - Cijfers trachten te matchen met gevoel





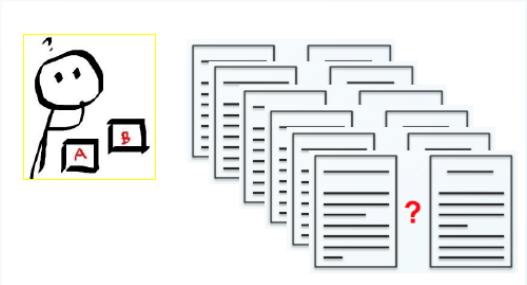


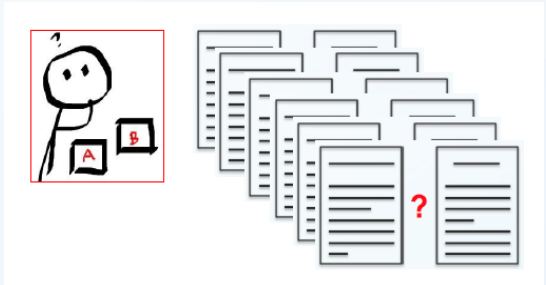
Paarsgewijs vergelijken

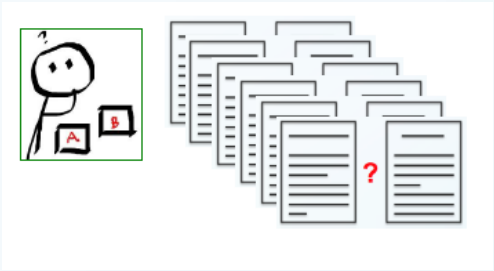
- Welke van de twee tekeningen is beter?

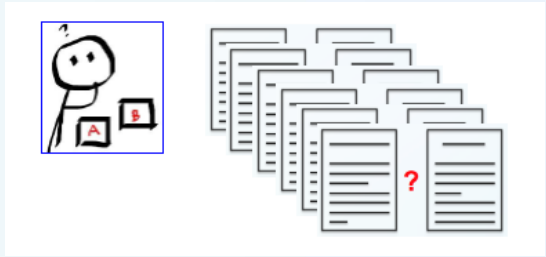
vs

Comparatief beoordelen: paarsgewijs vergelijken





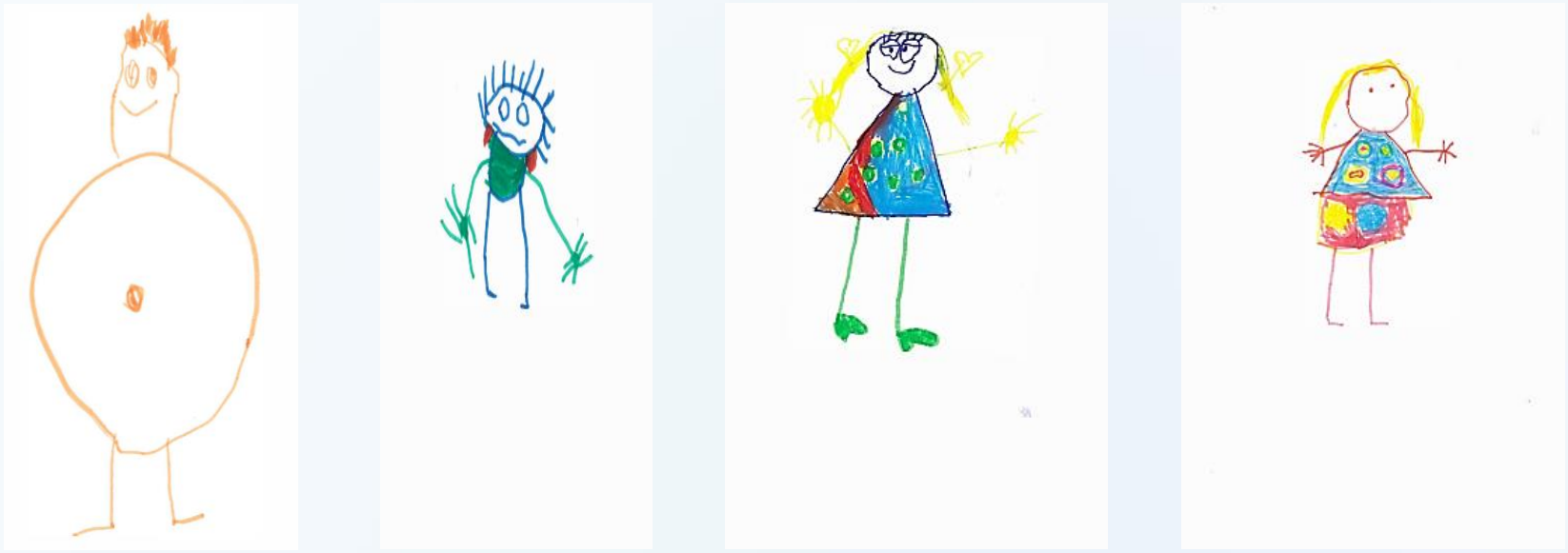




- Betrouwbaarheid = # vergelijkingen per product (voor .70=12, voor >.80= 20)
- Validiteit = minimaal 4 beoordelaars

➔ Groepsconsensus

Comparatief beoordelen: paarsgewijs vergelijken



Waarom comparatief beoordelen?

- Holistisch (Pollitt, 2012)
- Gedeelde consensus ~ groep van beoordelaars (van Daal et al., 2016)
- Gebruik van expertise (Pollitt, 2012; Jones et al., 2015)

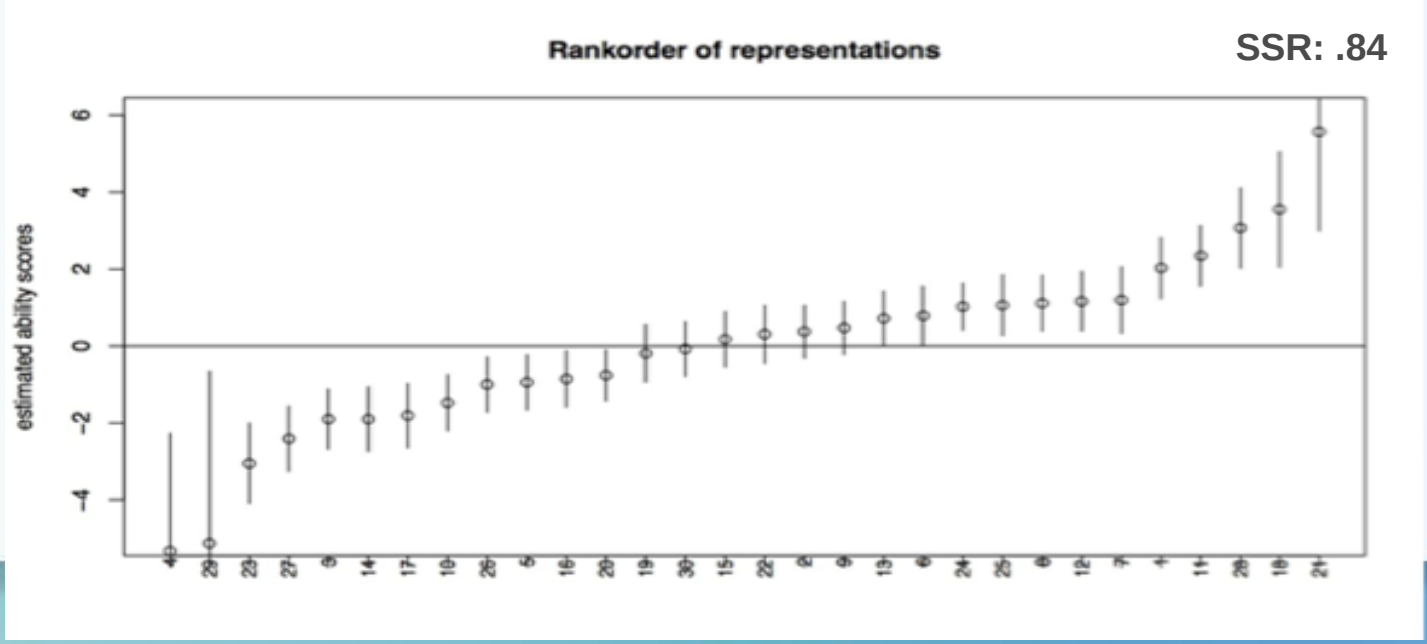
➔ Verhoogt validiteit! (Jones & Inglis, 2015; Pollitt & Crisp, 2004; Pollitt, 2012)

- Breed toepasbaar (e.g., Heldsinger & Humphry, 2010; Jones & Alcock, 2014; Pollitt, 2012):
  - Competenties die moeilijk te vervatten zijn in criteria
  - Verwachte verschillen in antwoorden
  - Performances

Informatie uit comparatief beoordelen

- Bradley-Terry-Luce model:
  - Kwaliteitsscores (in logits) met betrouwbaarheidsinterval
  - Rangorde
- Scale Separation Reliability

  Geeft schatting van interbeoordelaars betrouwbaarheid (Verhavert, 2017)



Informatie uit comparatief beoordelen

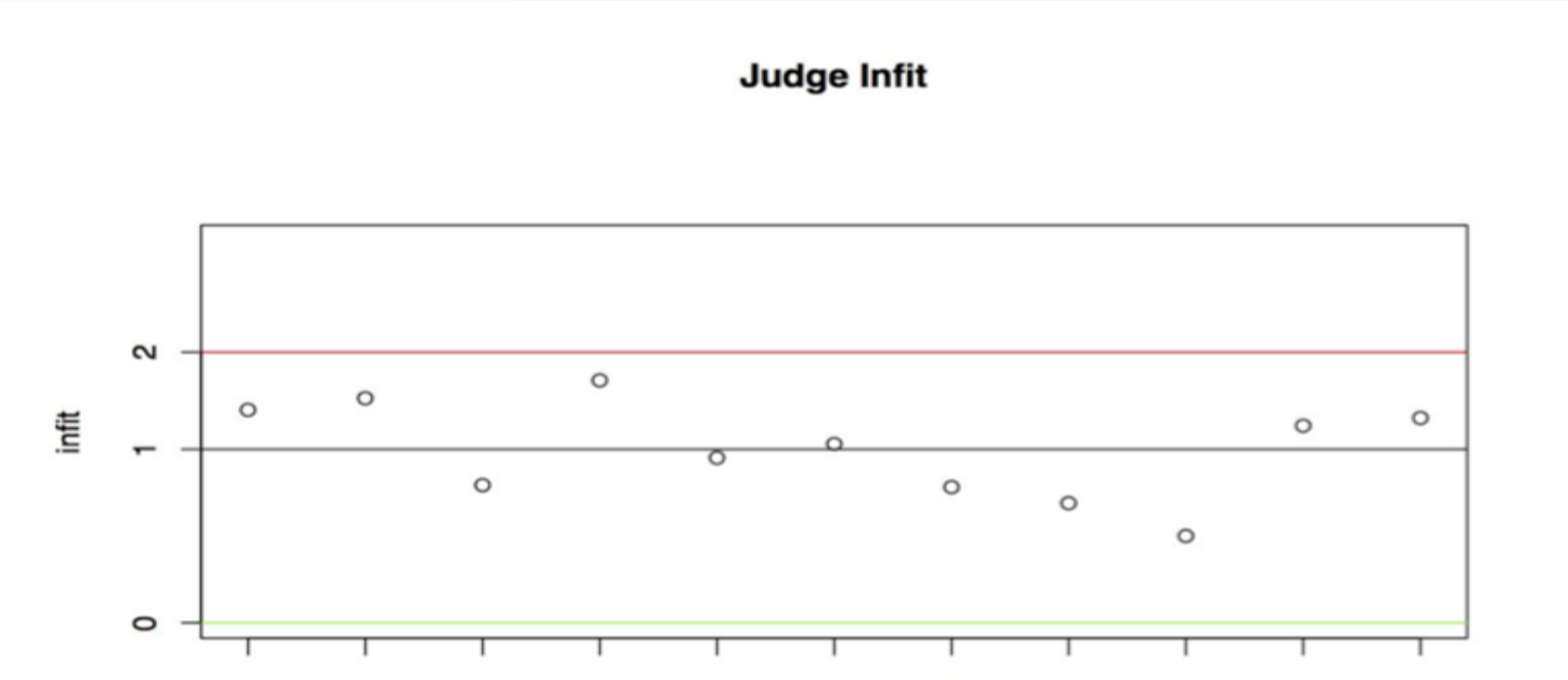
- Bradley-Terry-Luce model:
  - Kwaliteitsscores (in logits) met betrouwbaarheidsinterval
  - Rangorde
- Scale Separation Reliability

  Geeft schatting van interbeoordelaars betrouwbaarheid (Verhavert, 2017)
- Misfit data voor beoordelaars

  Welke beoordelaars hebben een 'andere' kijk (in vergelijking met de gedeelde consensus van beoordelaars?)
- Misfit data voor representaties

  Welke representatie(s) zijn moeilijk te beoordelen?

Misfit van beoordelaars



Taak 1: welke families gebruik je?

- Ga na welke families jij gebruikt in je beoordelingspraktijk + geef ook aan waarom je die familie gebruikt voor die bepaalde test.
- *Een goede keuze?*

4 jaar onderzoek: wat hebben we geleerd?

4 jaar onderzoek: wat hebben we geleerd?

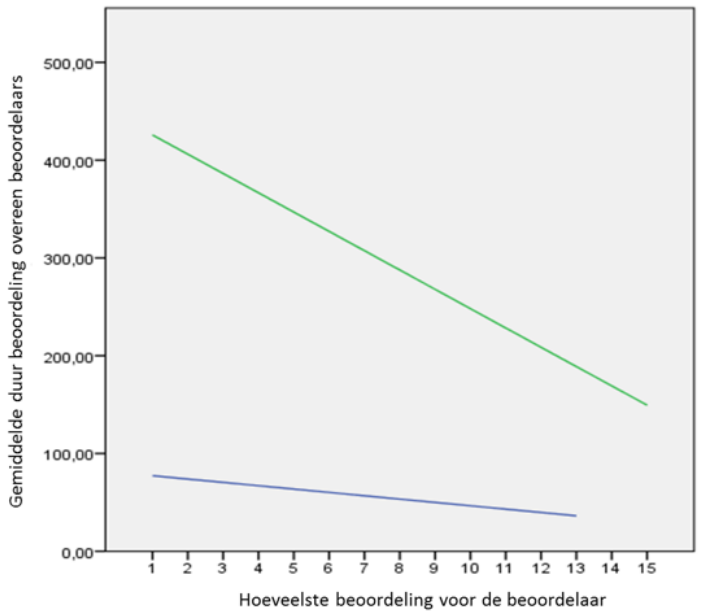
Hoe efficiënt is deze methode eigenlijk?

Hoeveel vergelijkingen & beoordelaars zijn er minimaal nodig om tot betrouwbare oordelen te komen?

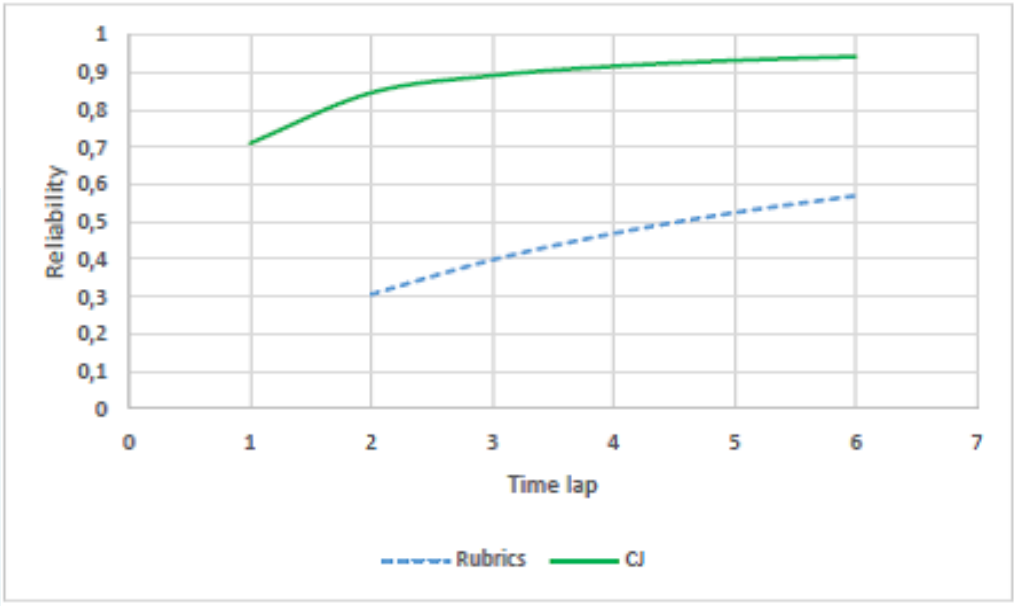
Zijn de uiteindelijke oordelen wel valide?

Kan het ook voor formatieve doeleinden gebruikt worden?

Efficiëntie



Coertjens, Lesterhuis, & De Maeyer (2017)



Goossens & De Maeyer (2017)

Hoeveel vergelijkingen nodig?

Meta-analyse met 49 assessments

Betrouwbaarheid van minimaal .70:

- gemiddeld 12 vergelijkingen per representatie
- minimaal 9, maximaal 20

Betrouwbaarheid van minimaal .80:

- gemiddeld 17 vergelijkingen per representatie
- minimaal 13, maximaal 25

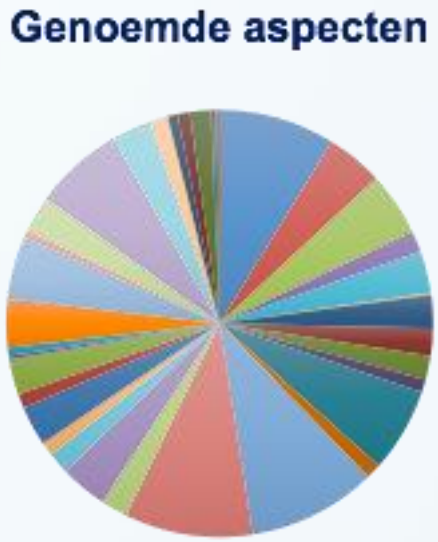
Voorbeeld met 20 studenten:

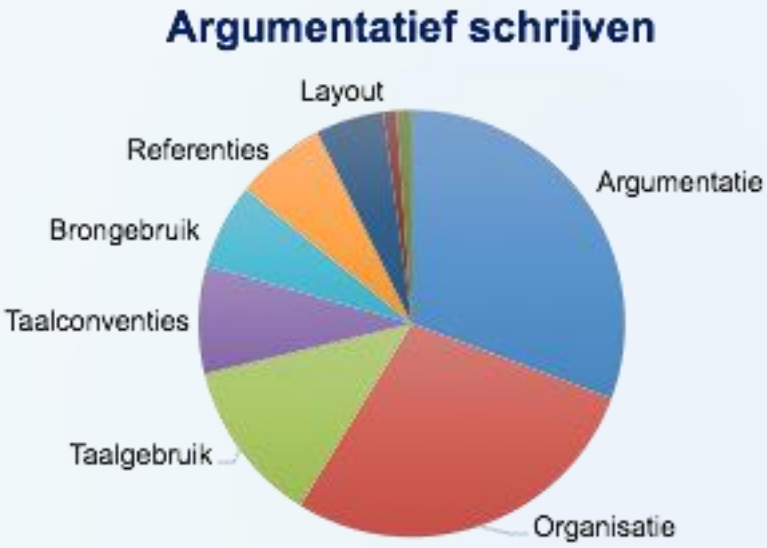
*20*12 vergelijkingen/2 = 120 vergelijkingen*

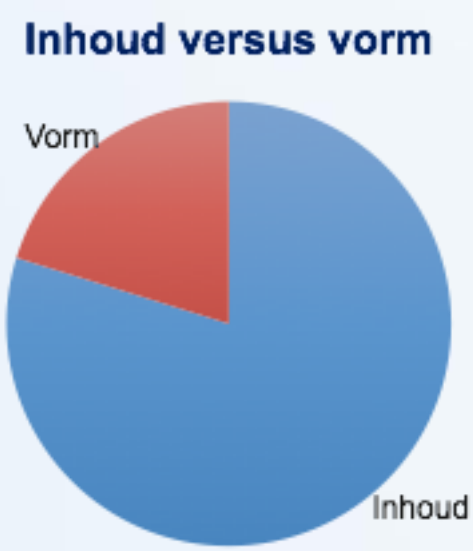
Verhavert, Bouwer, Donche, & De Maeyer (2017)

Validiteit: wat wordt beoordeeld?

Argumentatief schrijven in VO & HO







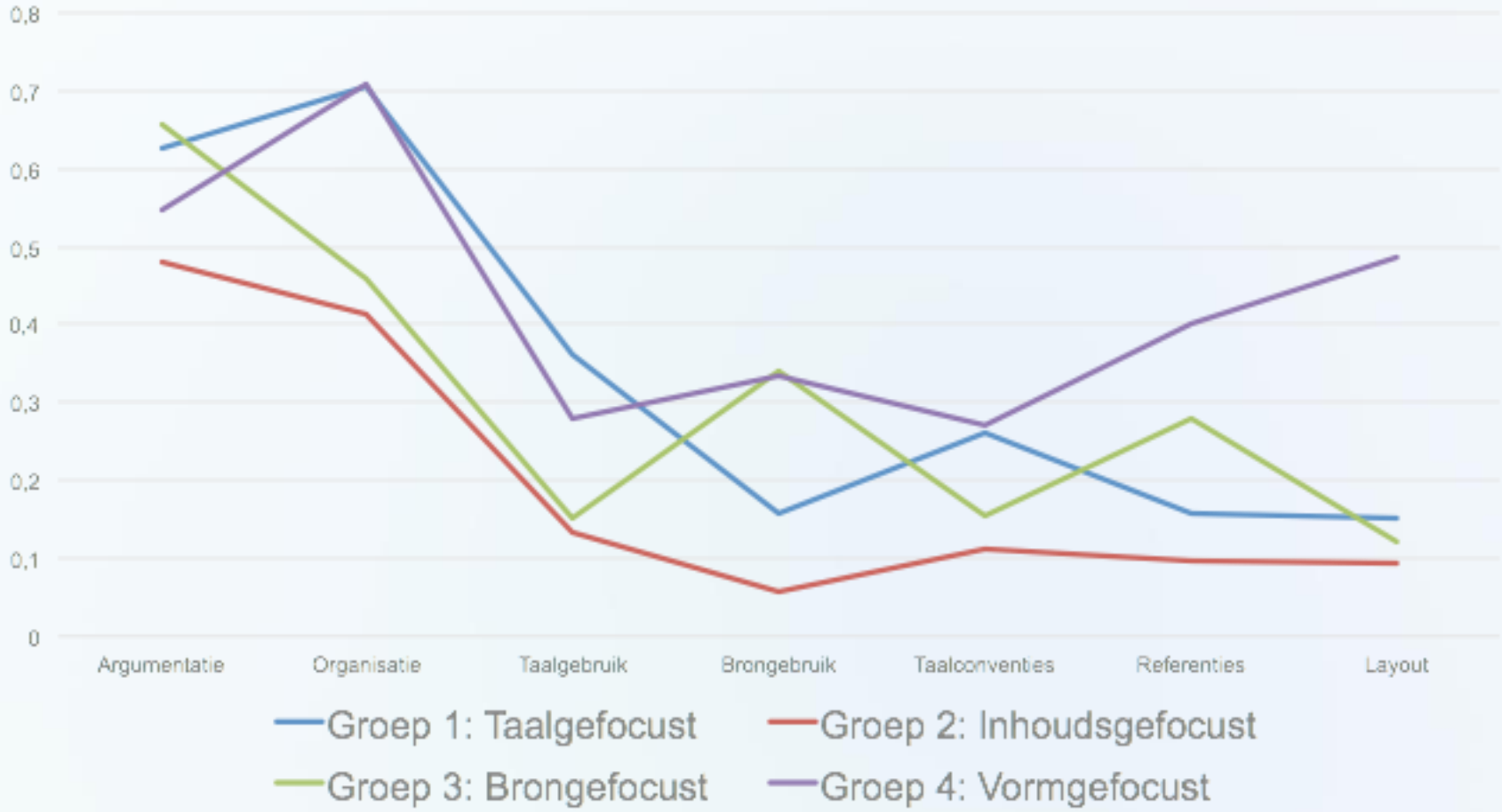
Lesterhuis et al. (2017); Van Daal et al. (2017)

Validiteit: overeenkomst met andere methoden?

Bouwer et al (2017), Coertjens et al (2017), Coertjens et al (2017)

D-PAC

Letten beoordelaars op dezelfde aspecten?



Betrouwbaarheid en validiteit

Meer beoordelaars:

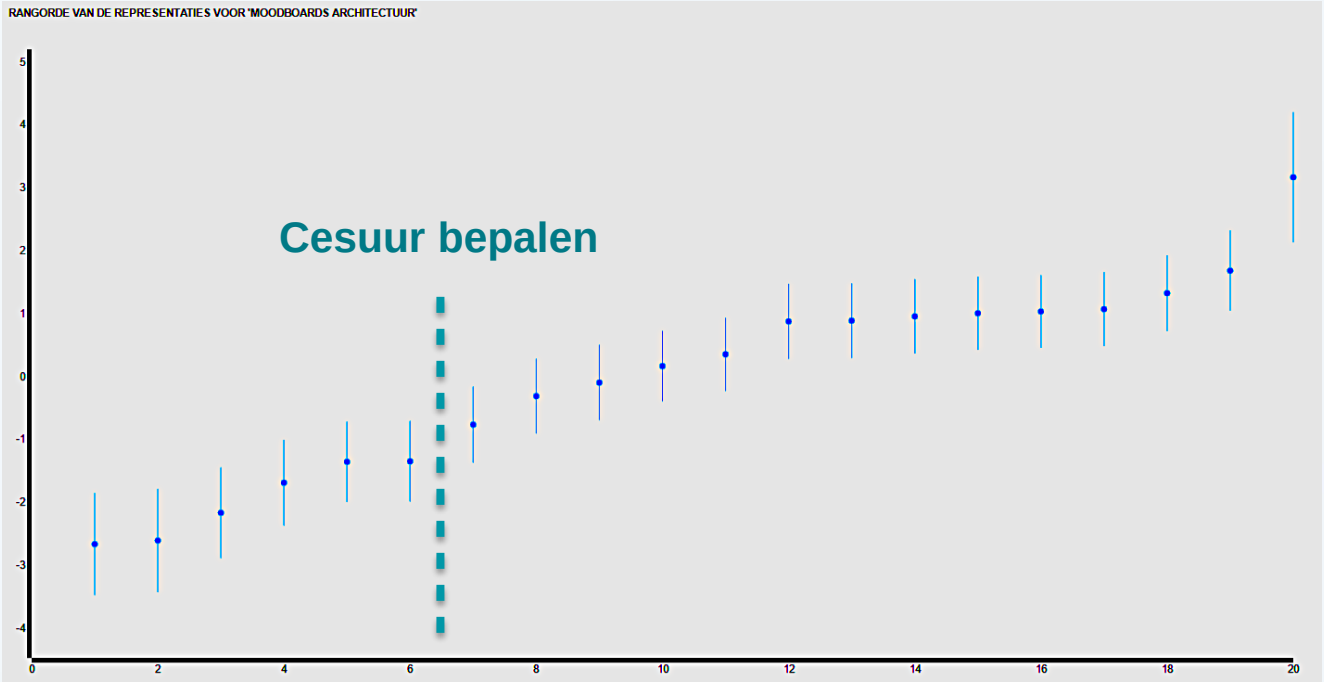
- Verhogen de generaliseerbaarheid (vanaf 4) van de resultaten: groep consensus (Van Daal et al., 2017)
- Verhogen de validiteit van de resultaten: verschillen tussen beoordelaars (Lesterhuis et al., 2017)

Wie dan?

- Zij die verwacht worden instaat te zijn de competentie te beoordelen
- Training niet noodzakelijk
- Peers?
  - Even betrouwbaar
  - Cor .65 met experts-> valide?

En dan? Van rangorde naar cijfers

Omzetten naar cijfers



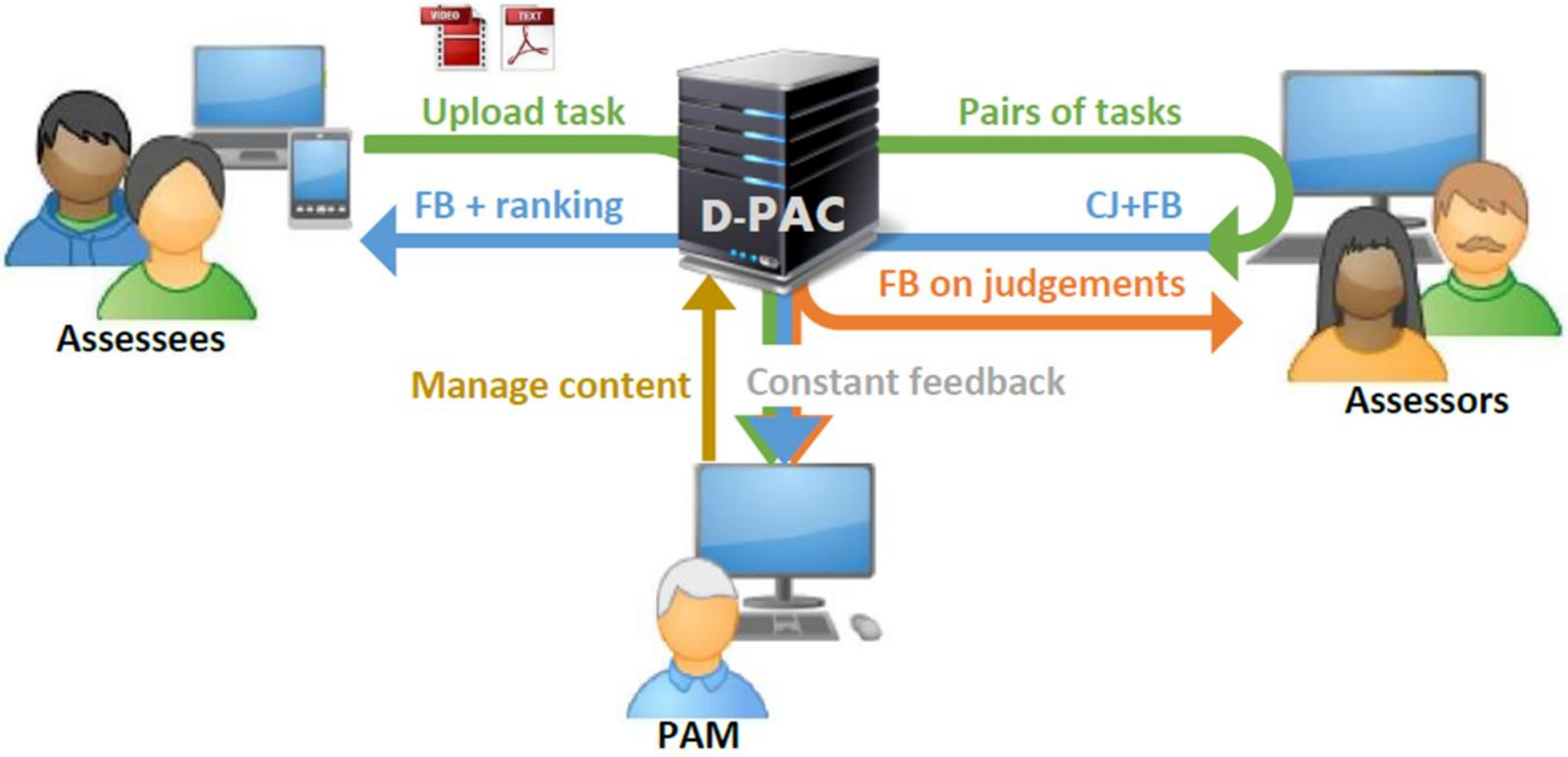
Taak 2: bedenk een case waar comparatief beoordelen zou voor passen.

- Hoe zou jij comparatief beoordelen kunnen gebruiken in jouw (onderzoeks)context?
  Welke doelen & vragen heb je daarbij?

- Waarmee dien je rekening te houden bij de implementatie van comparatief beoordelen in jouw (onderzoeks)context?
  Denk aan verschillende actoren, doelen, taken, etc.

De D-PAC tool

D-PAC tool


VIDEO
TEXT
Upload task
Pairs of tasks
FB + ranking
D-PAC
CJ+FB
FB on judgements
Assessees
Assessors
Manage content
Constant feedback
PAM

Een live assessment

- Welke competentie + welke opdracht (nu uit te voeren)?
- Creëer jouw representatie
- Volg de link in je mail ‘Uitnodiging D-PAC’
- Laad je werk op via ‘upload’
- Maak al je vergelijkingen via ‘beoordeel’
- Bekijk resultaten

Comparatief beoordelen in de praktijk ...

Peer assessment

Eenvoudig studenten beoordelen even betrouwbaar als docenten (SSR beide groepen = .77; correlatie = .65)

Veilig want alles gaat anoniem

Leren door te vergelijken

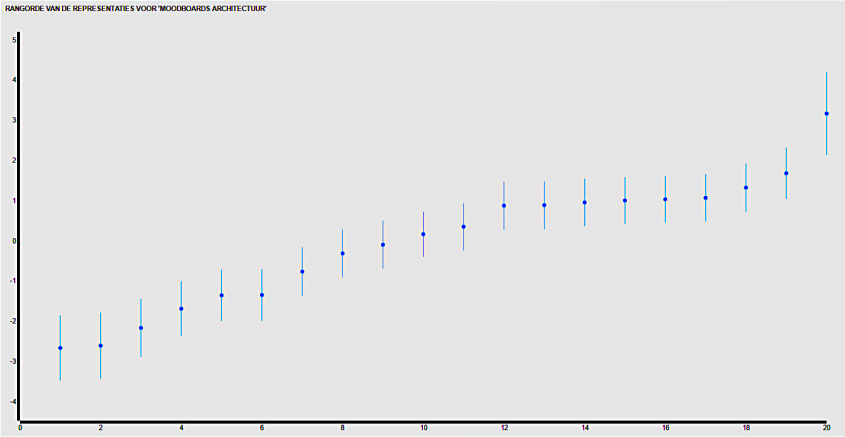
- Voorbeelden van uiteenlopende kwaliteit
- Kwaliteitscriteria vaststellen (bottom-up)
- Veel en on time feedback
- Bereidheid om feedback te gebruiken
- Focus op hogere orde aspecten van de taak

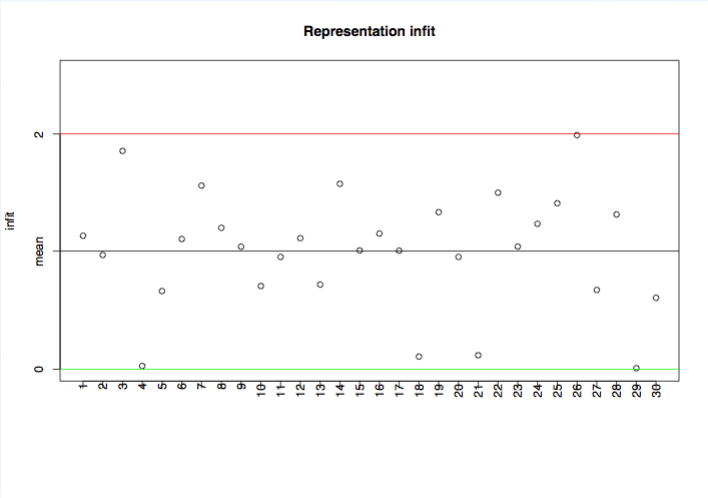
Professionaliseren van docententeam

Beoordeelt iedereen op dezelfde manier?

> waar zou dat aan kunnen liggen?

> extra richtlijnen/professionalisering?





Professionaliseren van beoordelaars

AHOVOKS
AGENTSCHAP VOOR HOGER ONDERWIJS, VOLWASSENENONDERWIJS, KWALIFICATIES & STUDIETOELAGEN

Vlaanderen
is onderwijs & vorming

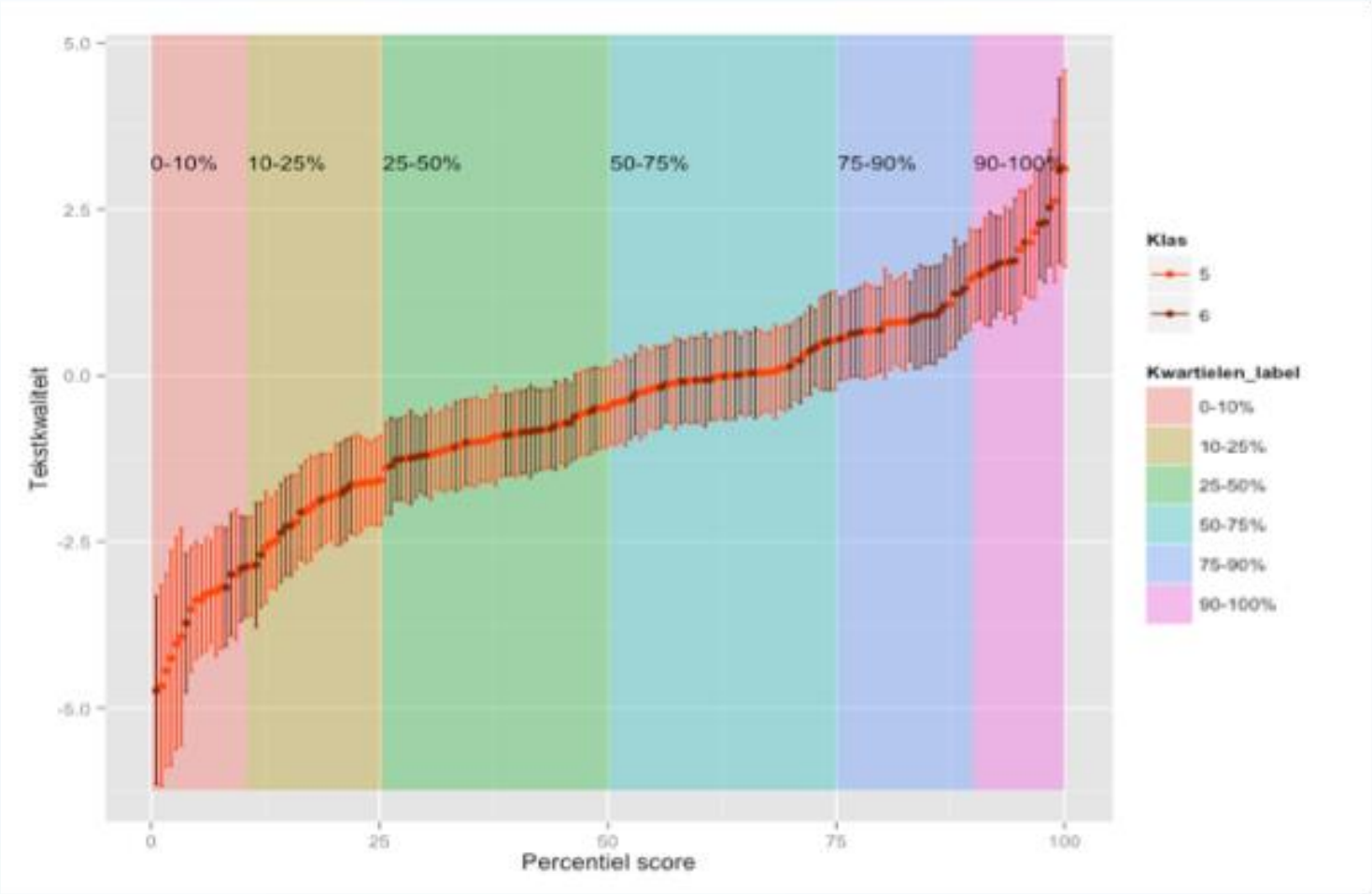
Interne beoordelaars (9 experts) & externen (16 freelancers) beoordelen 37 teksten in D-PAC

In hoeverre beoordelen ze op dezelfde wijze?

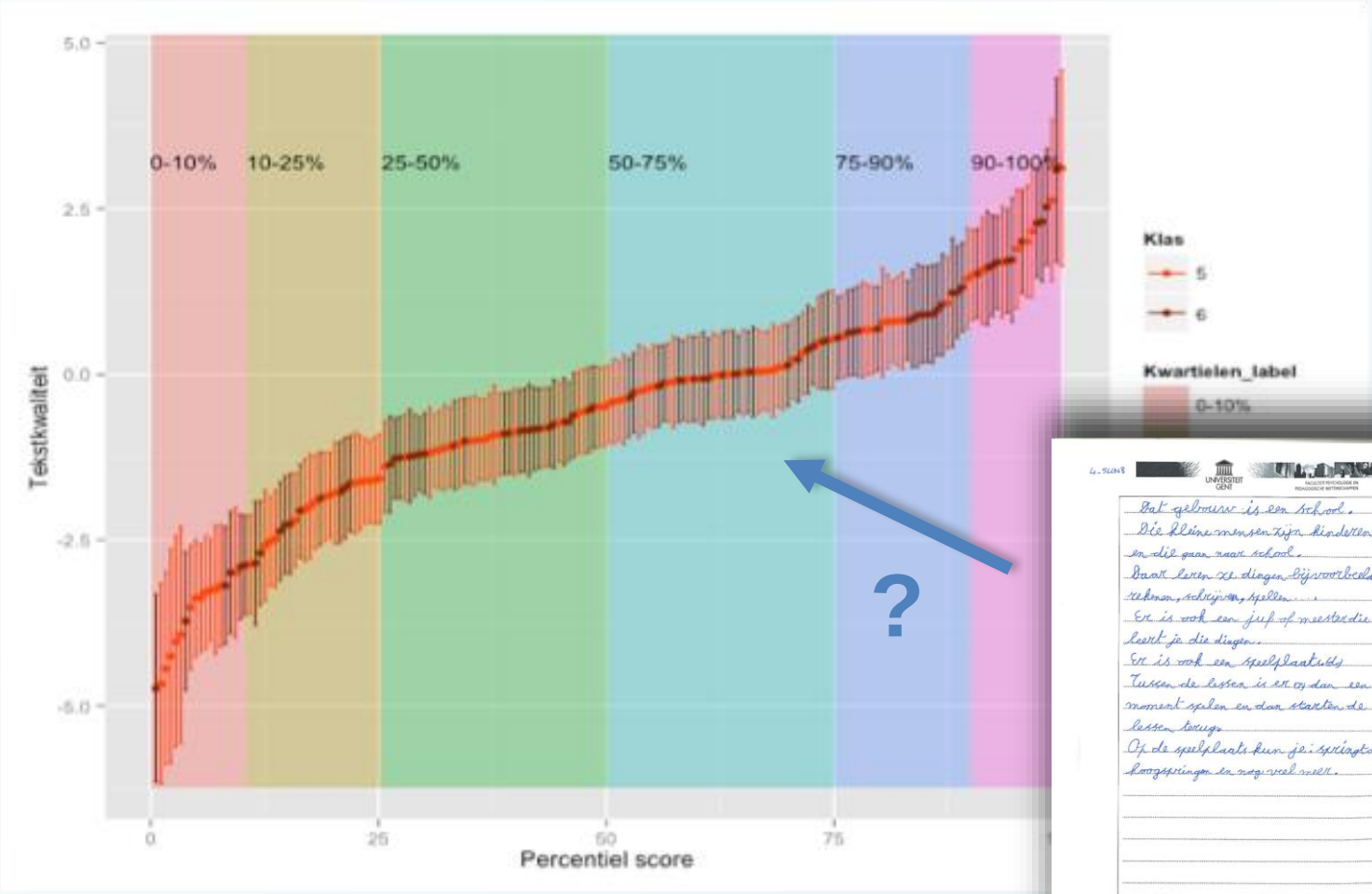
- Pretest: 5 externen wijken significant af van de internen
- Professionaliseringsdag
- Posttest: externen op één lijn met experts (nog maar 1 misfit)

D-PAC

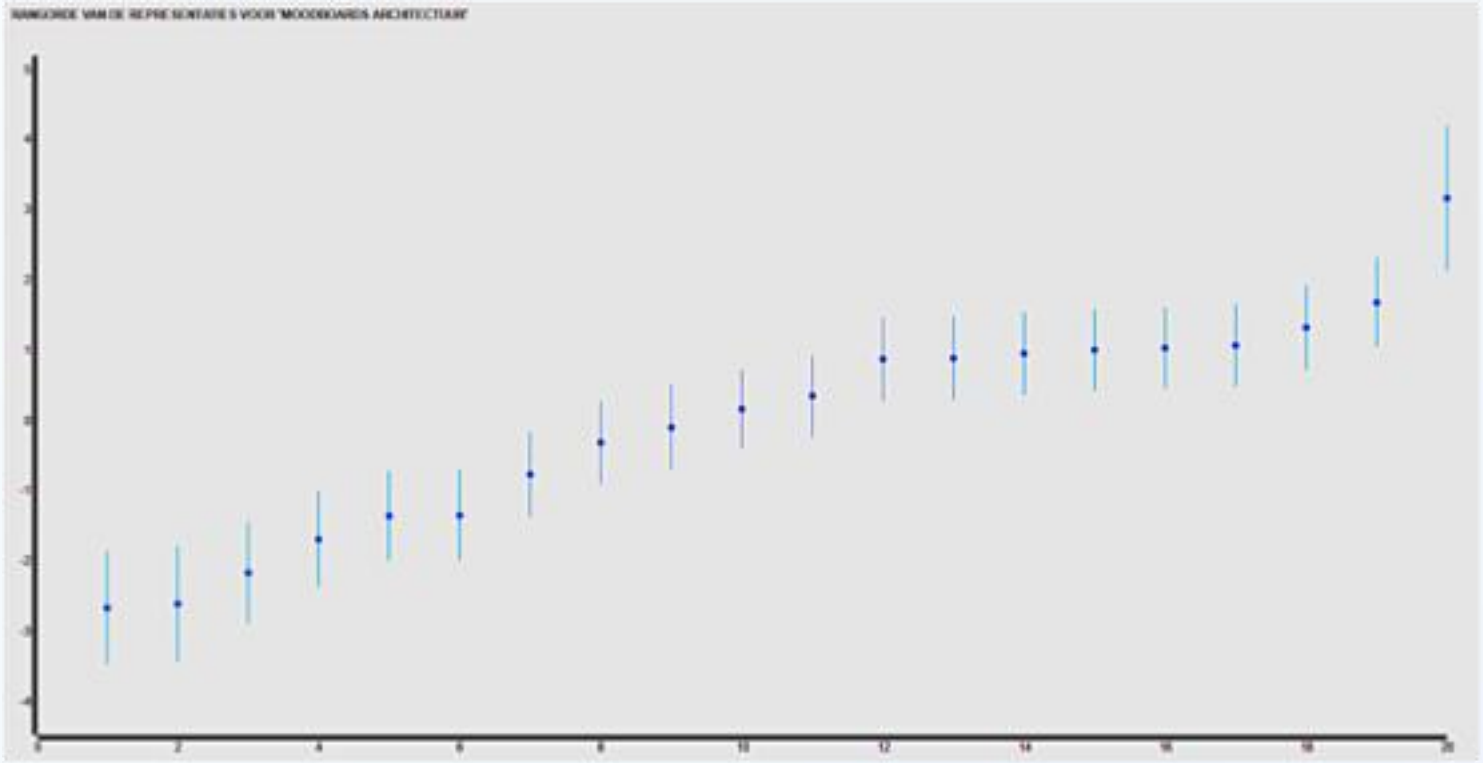
Doorlopende leerlijn: niveau per leerjaar



Plaatsingsalgoritme: leerwinst monitoren



Selecteren van benchmarks



Taak 3: finaliseer jouw casus

- Concretiseer je casus die je tot hiertoe ontwikkelde
  Wees zo specifiek mogelijk

- Stel gerust vragen!

www.d-pac.be | d-pac@uantwerpen.be

Edubron Universiteit Antwerpen

UNIVERSITEIT GENT

iMinds ILAB.O

D-PAC

D-PAC tool

- Maak enkele vergelijkingen:

  URL: https://sandbox.d-pac.be

  Ga naar ‘Beoordeel’

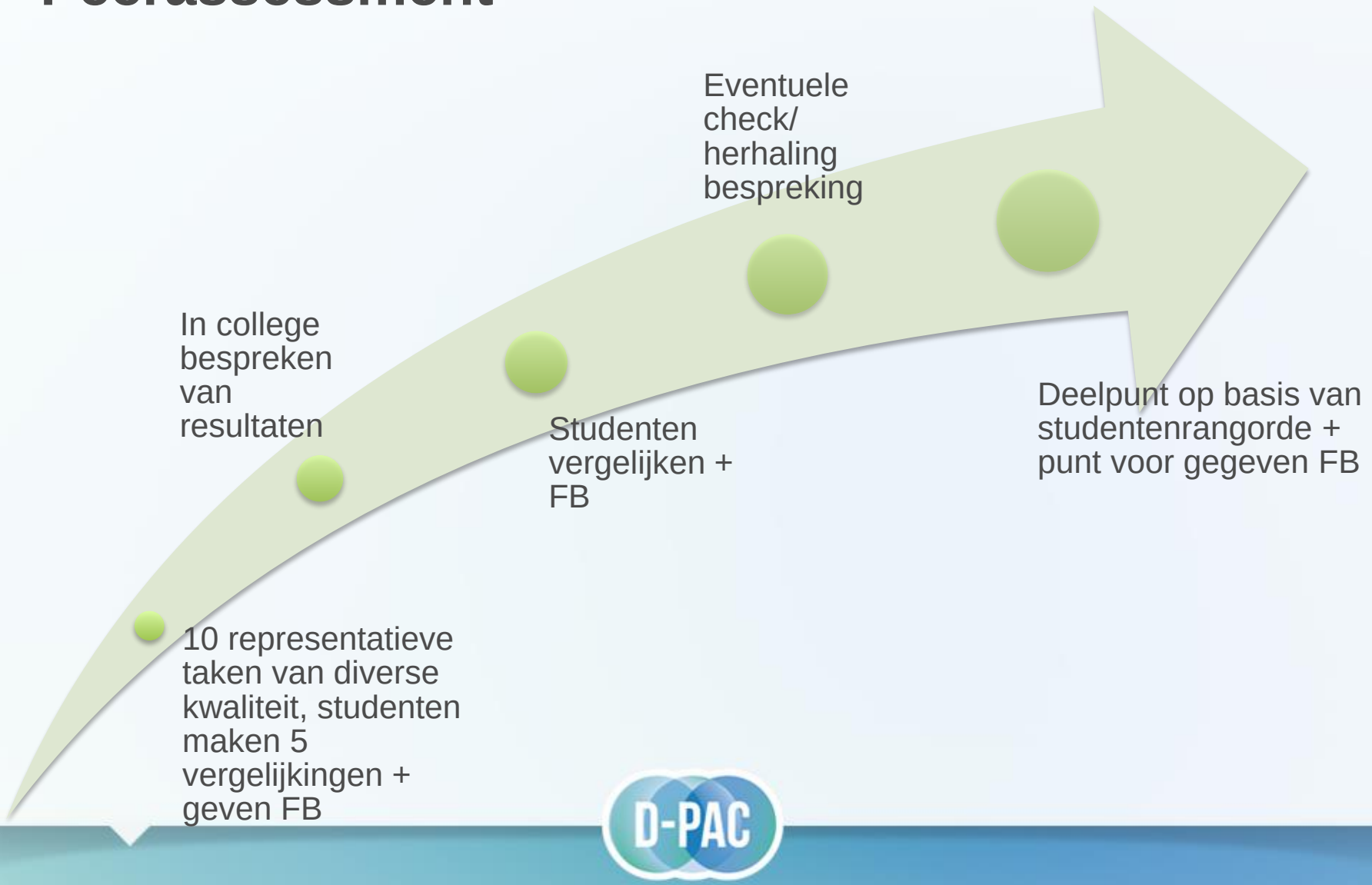
- Bekijk de resultaten:

  Ga naar ‘Resultaten’ > ‘Schrijfvaardigheid’

Implementatie in de praktijk: Peerassessment

- Eenvoudig om te doen: studenten beoordelen even betrouwbaar als docenten (SSR = .75-.79)
- Veilig: alles gaat anoniem
- Leren door te vergelijken
  - Voorbeelden van uiteenlopende kwaliteit
  - Kwaliteitscriteria vaststellen (bottom-up)
  - Self-efficacy
  - Bereidheid om feedback te gebruiken

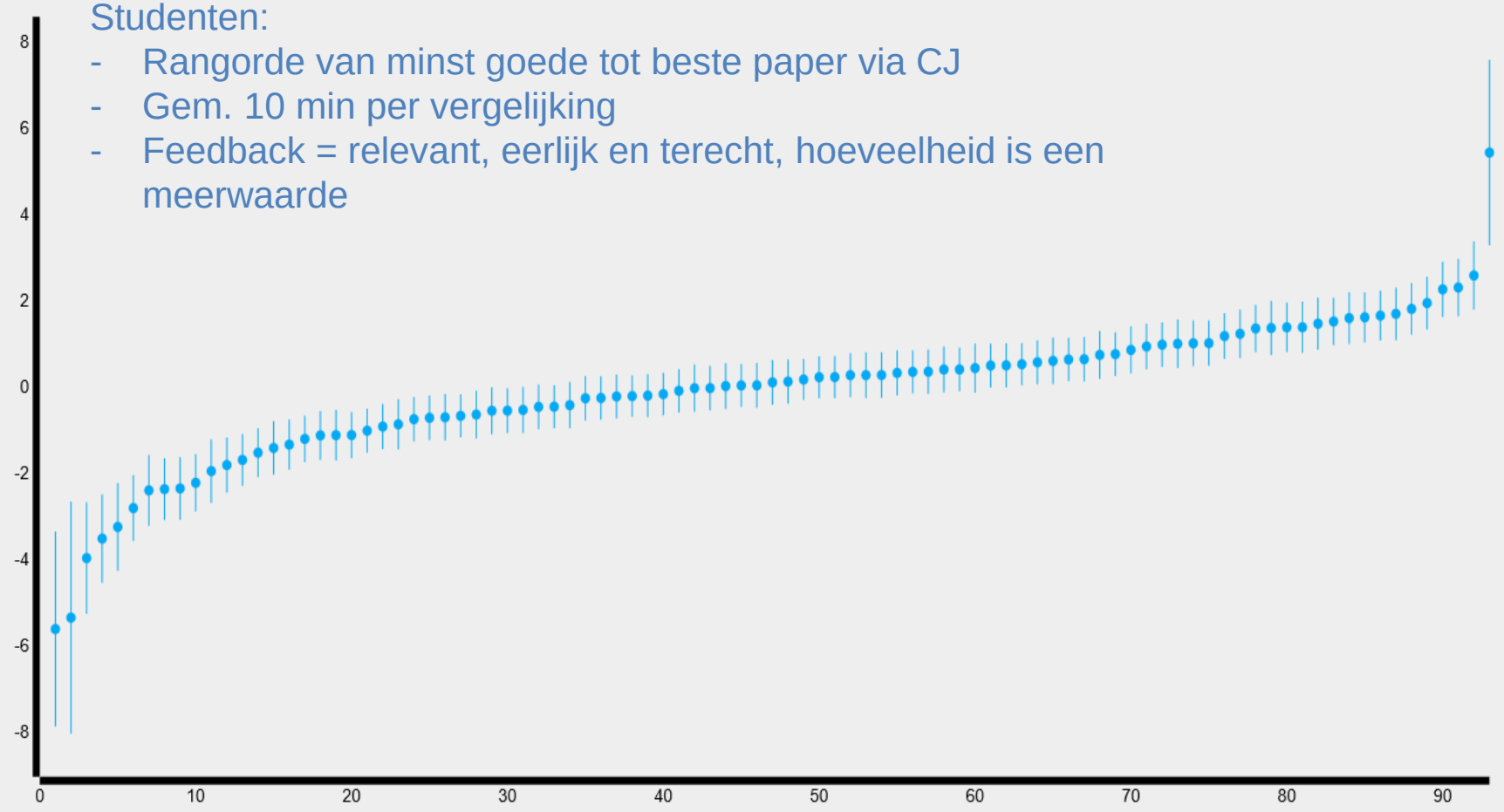
Implementatie in de praktijk: Peerassessment



Implementatie in de praktijk: Peerassessment -> tijdswinst

Studenten:

- Rangorde van minst goede tot beste paper via CJ
- Gem. 10 min per vergelijking
- Feedback = relevant, eerlijk en terecht, hoeveelheid is een meerwaarde



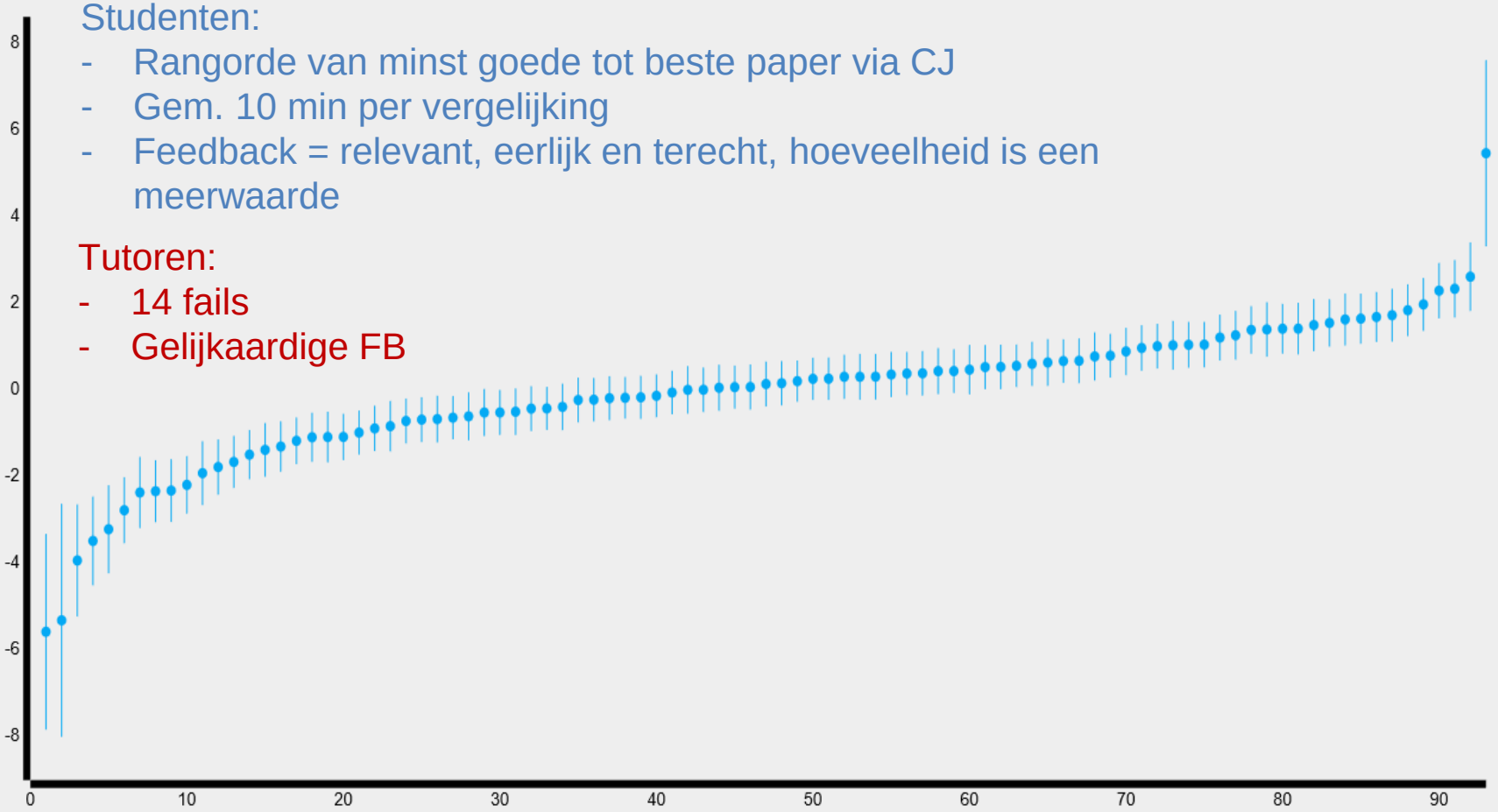
Implementatie in de praktijk: Peerassessment -> tijdswinst

Studenten:

- Rangorde van minst goede tot beste paper via CJ
- Gem. 10 min per vergelijking
- Feedback = relevant, eerlijk en terecht, hoeveelheid is een meerwaarde

Tutoren:

- 14 fails
- Gelijkaardige FB



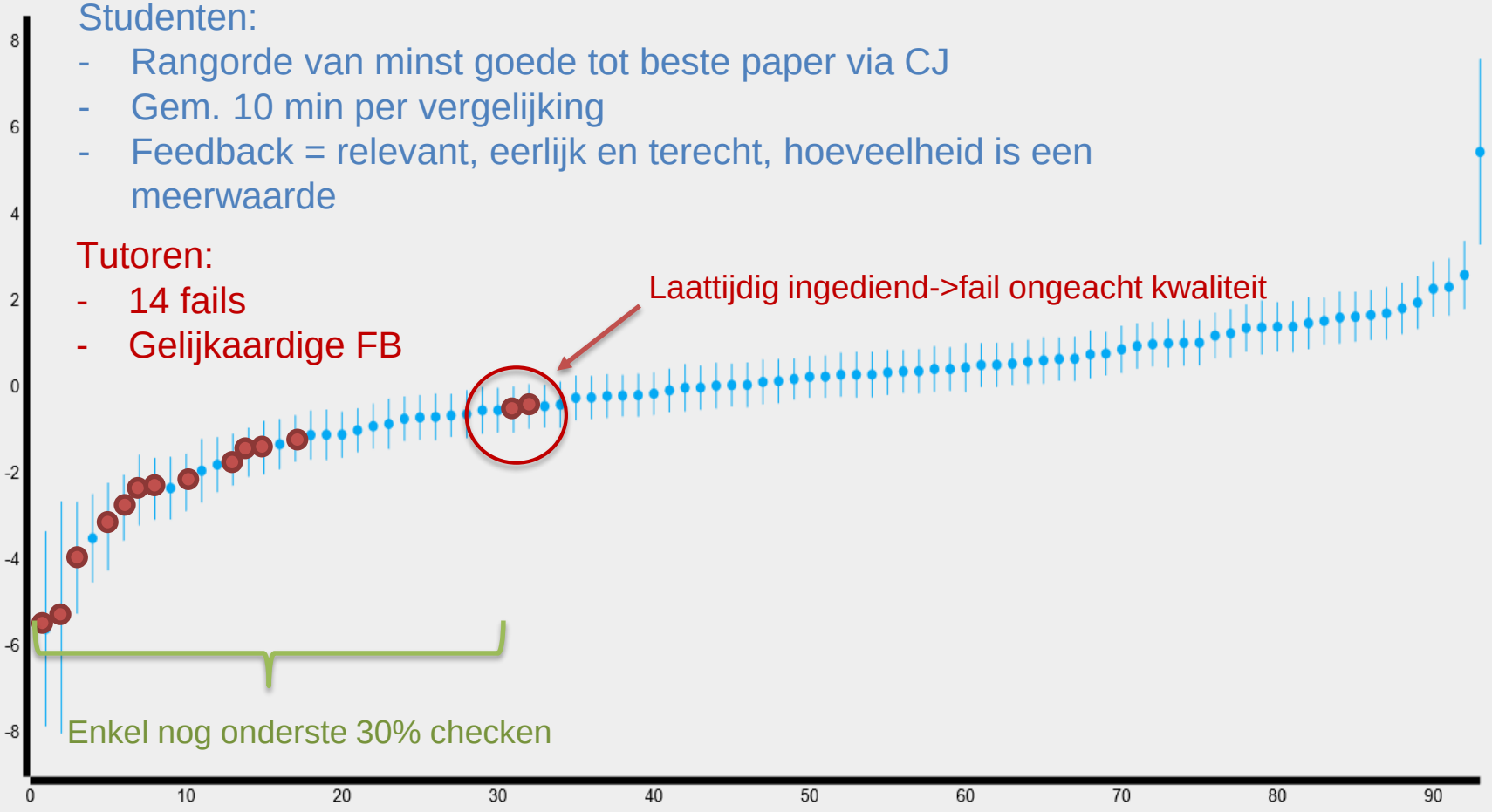
Implementatie in de praktijk: Peerassessment -> tijdswinst

Studenten:

- Rangorde van minst goede tot beste paper via CJ
- Gem. 10 min per vergelijking
- Feedback = relevant, eerlijk en terecht, hoeveelheid is een meerwaarde

Tutoren:

- 14 fails
- Gelijkaardige FB



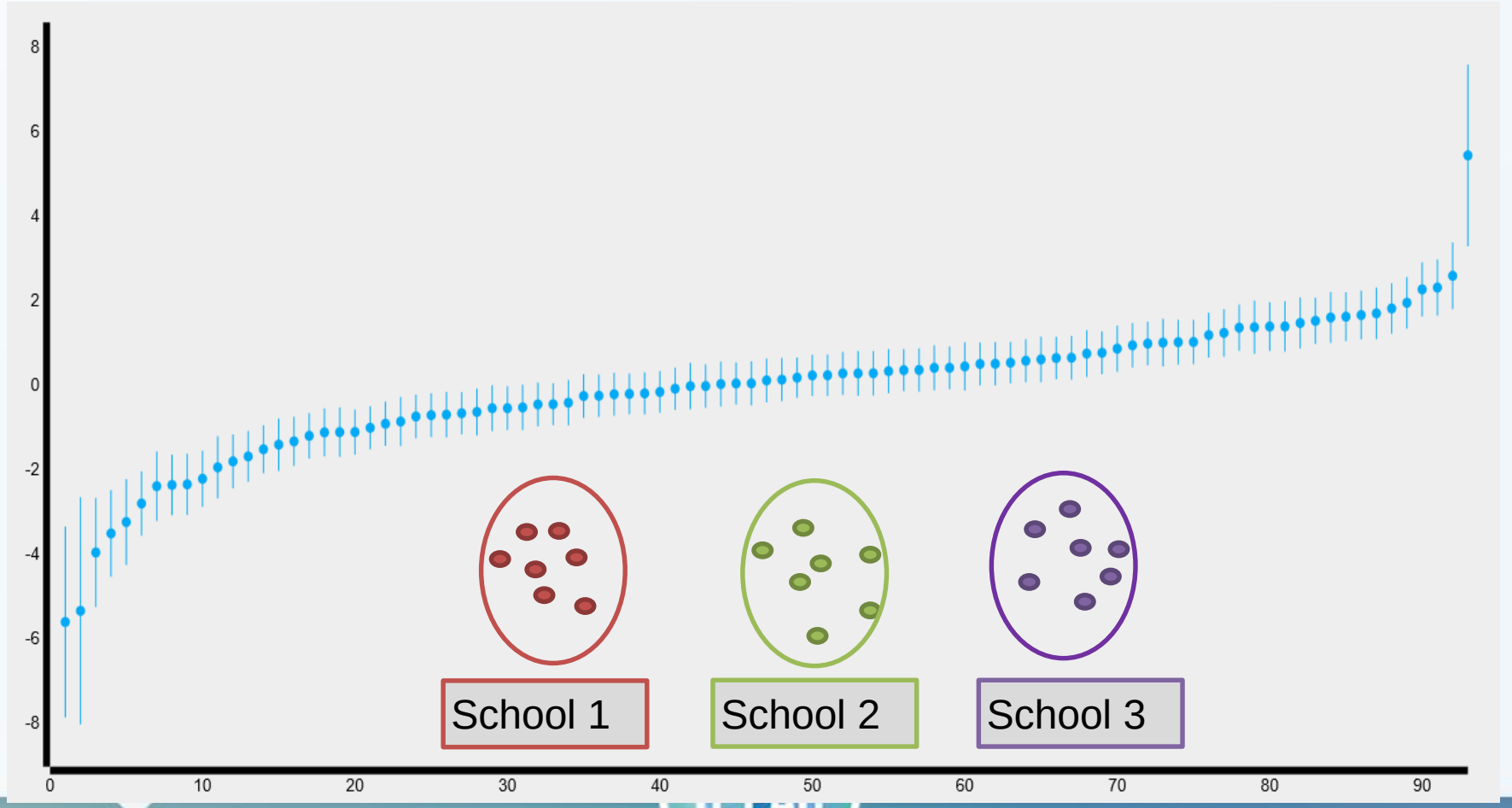
Implementatie in de praktijk: Summatief evalueren

- 50 studenten multimedia -> 50 animatics
- 9 beoordelaars
  - Docenten
  - Professionals: geluidshuis, software bedrijf, grafische vormgever/illustrator
- Werkbelasting:
  - 347 vergelijkingen
    - 14 vergelijkingen per animatic
    - 38 vergelijkingen per beoordelaar
  - 80 sec. per vergelijking = 50 min.
- Resultaat:
  - Betrouwbaarheid rangorde .72

plaats	name	ability	Score	Afgeronde score
1	05-05	4,184700961	18,00	18
2	22-22	3,945798732	17,61	18
3	01-01	2,966230597	15,99	16
4	02-02	2,48836859	15,21	15,5
5	36-36	1,969513954	14,35	14,5
6	20-20	1,329586485	13,30	13,5
7	46-46-2	1,310413379	13,27	13,5
8	06-06	1,306068961	13,26	13,5
9	09-09	1,098741407	12,92	13
10	07-07	1,0518964	12,84	13
11	44-44	0,949689169	12,67	13
12	42-42	0,933822255	12,65	13
…	…	…	…	…
…	…	…	…	…
45	37-37	-1,491379922	8,65	9
46	47-47-1	-1,670968947	8,36	8,5
47	48-48	-1,794772189	8,15	8,5
48	14-14	-3,134576938	5,95	6
49	02-51	-3,885973008	4,71	5
50	49-49	-4,31594817	4,00	4

Implementatie in de praktijk: Leerwinst monitoring

1) Gekalibreerde schaal

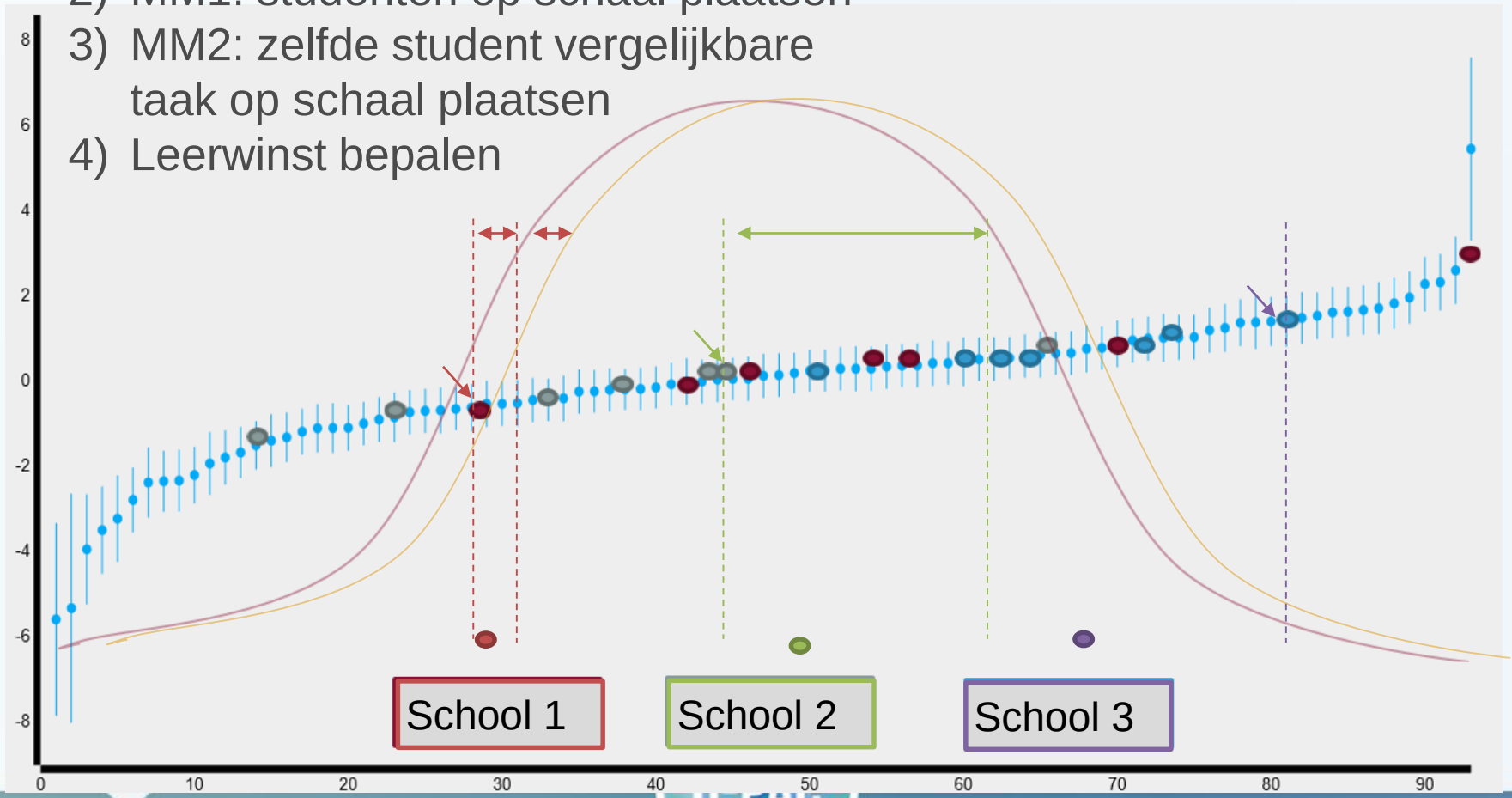


Implementatie in de praktijk: Leerwinst monitoring

1) Gekalibreerde schaal
2) MM1: studenten op schaal plaatsen

Implementatie in de praktijk: Leerwinst monitoring

1) Gekalibreerde schaal
2) MM1: studenten op schaal plaatsen
3) MM2: zelfde student vergelijkbare taak op schaal plaatsen
4) Leerwinst bepalen



Implementatie in de praktijk: professionaliseren

AHOVOKS
AGENTSCHAP VOOR HOGER ONDERWIJS, VOLWASSENENONDERWIJS, KWALIFICATIES & STUDIETOELAGEN

Vlaanderen
is onderwijs & vorming

Interne & externe medewerkers beoordelen duizenden teksten per jaar

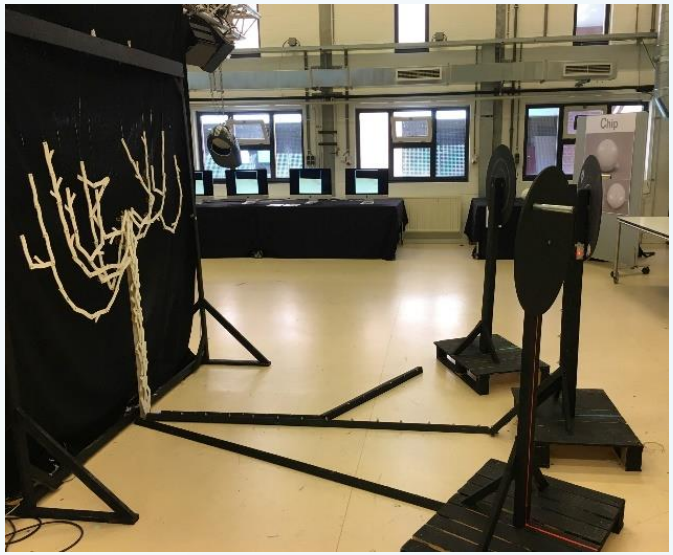
Welke van de externe medewerkers hebben training nodig?

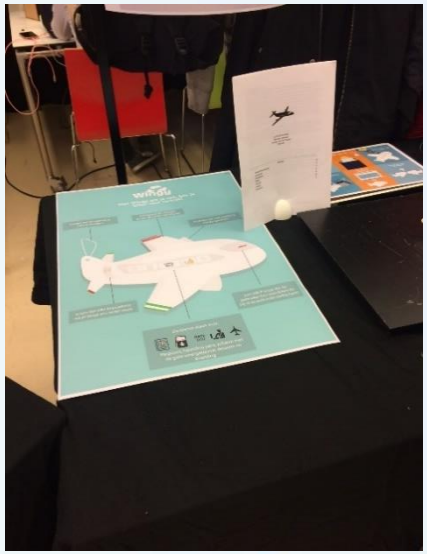
- Rangorde van 37 teksten: experten (*N* = 9) en externen (*N* = 16)
  - ✓ 5 externe medewerkers die significant afweken van de rangorde van experten (misfit analyse)
  - ✓ Misfits letten op andere dimensies van schrijven dan experten
- Professionaliseringsdag: wat is een goede tekst?
- Achteraf nog een ronde met vergelijkingen
  - ✓ Externe medewerkers op één lijn met experts (nog 1 misfit)

D-PAC

Comparatieve beoordeling bij avans hogeschool

- Live beoordelen interactieve installaties (groepswerk)
- Studenten 2e jaars Communication & Multimedia Design





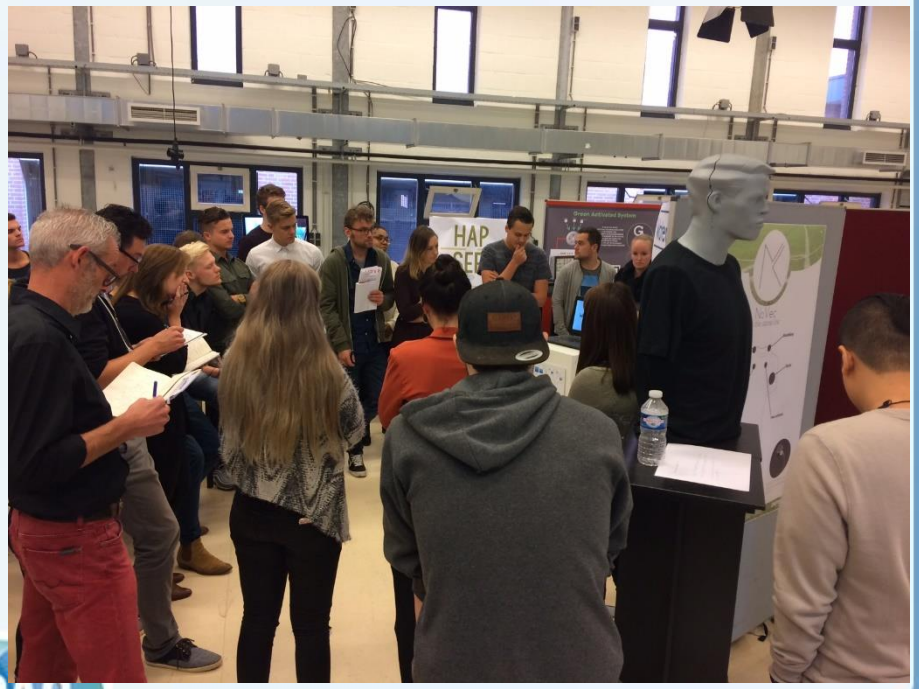
D-PAC

Comparatieve beoordeling bij avans hogeschool

Procedure

A. Demo/presentatie (2u.)

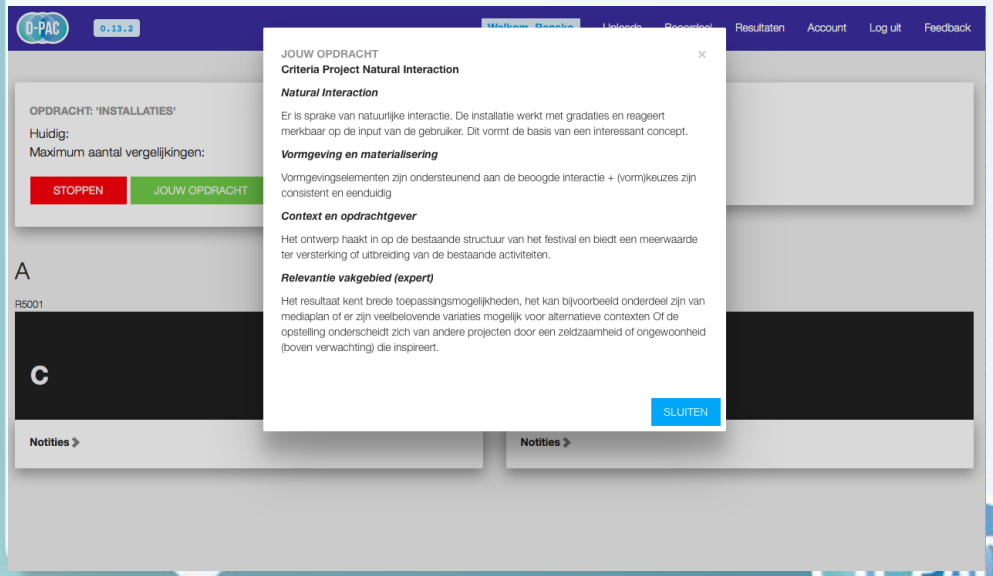
- Door studenten
- 19 groepswerken
- beoordelaars kunnen vragen stellen

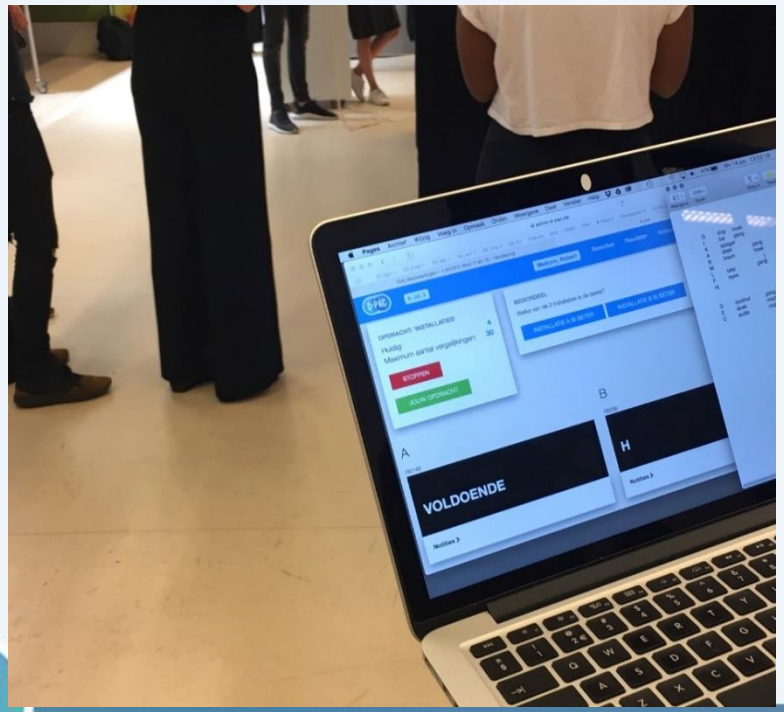


Comparatieve beoordeling bij avans hogeschool

B. Beoordelen met D-PAC (90 minuten)

- 6 beoordelaars, onafhankelijk van elkaar
- 32 vergelijkingen per beoordelaar
- vergelijking holistisch met inachtneming van de 4 criteria
- ‘voldoende’ komt terug in de vergelijkingen.

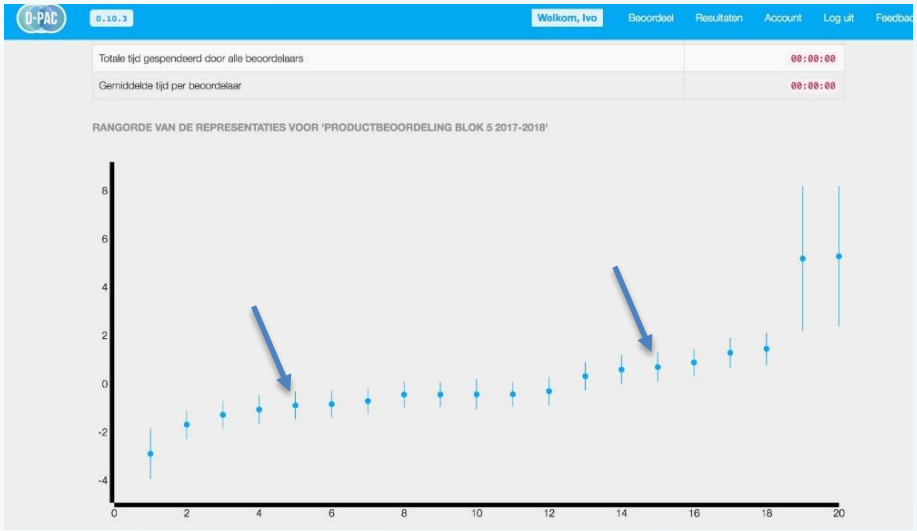




Comparatieve beoordeling bij avans hogeschool

C. Rangorde komt terug, daarin voldoende opgenomen

- Beoordelaars kennen gezamenlijk een tweede cijfer toe (naast voldoende)
- D-PAC (medewerker) berekent ander cijfers.



Comparatieve beoordeling bij avans hogeschool

D. Projectcoördinator modereert feedback

- consistentie en toon;
- communiceert cijfers én feedback met studenten (liefst face-to-face).

DETAILGEGEVENS VOOR 'R5008'

J

Rangorde	2
Waarde	2.94
Aantal beoordelingen van deze representatie	17

FEEDBACK VOOR 'NOTEER HIER STERK- EN WERKPUNTEN PER INSTALLATIE. WERK HIERBIJ VANUIT JOUW EXPERTISE.'

Beoordelaar	Sterke punten	Werkpunten
beoordelaar-98f4	Heel toegankelijk, iedereen interacteert er automatisch mee en het geeft een extra dimensie aan je gesprek. Heel tof als je de woorden kan manipuleren, zeker wat doen met bepaalde woorden die je daar graag wilt horen of die extra betekenis hebben in die context. Misschien leuk als je de plattegrond van het festival ziet als je 'plattegrond' zegt, maar ook een paar gekke dingen erin houden die zijn erg leuk	Ik zou kijken naar vormgeving, dat kan denk ik nog sterker, en hoe je het in de binnentuin goed kunt integreren.
beoordelaar-98f7	Lokt direct uit tot gesprek	Het trekt wel erg de aandacht, en kan ook verstoren.
beoordelaar-98f8	IJzerstekerke subtiele en intuitieve interactie. verrassend en op een heel natuurlijke plek geplaatst.	vormgeving zou nog wel een keer naar gekeken mogen worden.
beoordelaar-98f5	Heldere interactie	Hoe weet je welk woord 'scoort'
beoordelaar-98f6	Goed werkend prototype en je brengt echt een nieuwe dimensie aan in een gesprej	Het kan een gesprek ook afleiden dat het alleen nog maar om de tafel gaat

D-PAC

Comparatieve beoordeling bij

Bevindingen

1. ‘Doet recht’ aan werken
2. Beoordelen snel en simpel
3. Elk jury-lid kan zich vinden in punten
4. Tijd voor feedback