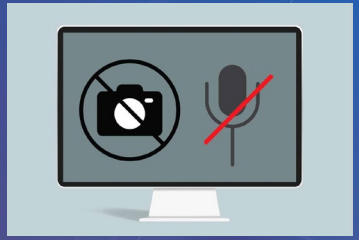
Fontys > FOR SOCIETY



HET EFFECT VAN KALIBRATIE: EEN EERSTE VERKENNING

Vraagstuk examencommissie Fontys Paramedisch

Bregje van Dijk & Femke Snieders
Studenten Master Toets- en Onderwijskwaliteit

17 maart 2026

Even voorstellen…..

Femke Snieders
Docent HAN – Facility Management
Student 1e jaar MTOK

Bregje van Dijk
Docent HAN – ALO
Student 1e jaar MTOK

Programma

- Aanleiding
- Onderzoeksvraag
- Methode
- Resultaten
- Advies

Aanleiding

-Examencommissie heeft een borgende taak

- Accreditatiepanel tevreden, kalibratie is goed ingericht

- Maar zijn examinatoren ook echt gekalibreerd?

Onderzoeksvraag

In welke mate voldoet kalibratie binnen Fontys Paramedisch aan de in de literatuur gestelde kwaliteitseisen en hoe kan het effect van kalibratie gemeten worden door de examencommissie van Fontys Paramedisch?

Methode

Literatuuronderzoek

Documentanalyse

Interviews

- *Wat is kalibratie en wat zijn kwaliteitseisen?*
- *Op welke wijze is het effect van kalibratie te meten?*

- *Hoe is kalibratie ingericht binnen Fontys Paramedisch?*
- *Hoe verloopt kalibratie in de praktijk?*
- *In welke mate komen de kwaliteitseisen van kalibratie terug in het proces?*

Wat verstaan we onder kalibratie?

Kalibratie:

Een activiteit waar beoordelaars door middel van dialoog en het bespreken van authentiek studentenwerk streven naar een gedeeld begrip van leeruitkomsten en de daarbij behorende kwaliteitseisen (Sadler, 2013; Crimmins et al., 2016; Mason & Roberts, 2023).

Kalibratie draagt bij aan feedbackgeletterdheid en maakt duidelijk wat de verwachtingen zijn van studenten (Middleton et al., 2024; O'Donovan et al., 2024).

Kwaliteitseisen van kalibratie

- Vindt plaats vóór en tijdens het leerproces, liefst fysiek
- Duidelijke leeruitkomsten
- Voldoende tijd ter beschikking
- Dialoog, geen discussie → veilig klimaat
- Nieuwe docenten meenemen

(Andriessen, 2015; Mason & Roberts, 2023; Middleton et al., 2024; Prichard et al., 2025)

Kanttekening: 100% kalibratie is niet mogelijk

(Bloxham et al., 2016; O'Donovan et al, 2024)

Resultaten (1)

- Kalibratie Fontys Paramedisch voldoet aan alle kwaliteitseisen
- Op papier én in de praktijk
- Licht aandachtspunt: duur van de kalibreersessies en wordt er ook gehandeld volgens de afspraken?

Resultaten (2)

Mogelijkheden om het effect van kalibratie te meten:

1. Interbetrouwbaarheid berekenen (Stemler, 2004)
   - Consensus: percents agreement / Cohen's kappa
   - Consistentie: pearson correlatie / Spearman's rho
2. Ankerportfolio's – IBB (Sadler,1989; Boud & Falchikov, 2006)
   - Normbesef

Resultaten (3)

Mogelijkheden om het effect van kalibratie te meten:

3. Interbetrouwbaarheid berekenen met drift
   - Sliding windows (Casabianca et al, 2015)
   - Duurzaam kalibreren (Myford & Wolfe, 2009)
4. Herbeoordelen (evt. steekproefsgewijs)

Tabel 1:
Voor- en nadelen van effectmetingen

	Voordelen	Nadelen
IBB consensus (percent agreement of Cohen's kappa)	• Duidelijk of dezelfde normgrenzen worden gehanteerd • Eenvoudig • Transparant	• Controlerend • Zegt niets over waarom er verschillen zijn, schijnzekerheid. • Zegt niets over duurzame effect, alleen op dat moment
IBB consistentie (Pearsoncor-relatie of Spearman's rho)	• Duidelijk of er sprake is van eenzelfde kwaliteitsbeeld	• Kan nog steeds samengaan met een lage consensus. Is er dan sprake van een eerlijke beoordeling?
Ankerportfolio's	• Check of oordeel in lijn is met de vooraf gestelde norm • Minder controlerend • Duurzamer	• Statistisch minder hard meetbaar • Tijdrovend om goede ankerwerken te maken
Drifting (sliding windows)	• Check of kalibratie ook duurzaam effect heeft	• Geen sprake van normwaardes • Met kleine groepen (windows) instabielere cijfers
Herbeoordelen	• Leidt tot verbeterd normbegrip na dialoog	• Tijdrovend • Controlerend

Advies (1)

- Geen enkele meting is perfect
- Metingen toepassen in bestaande organisatie, passend bij doel
- Niet te controlerend en werkdrukverhogend
- Tweetrapsraket: IBB als signaal berekenen bij conceptbeoordelingen, niet als bewijs
  - Conceptbeoordelingen administreren
  - Woordbeoordelingen coderen
- Consensus: Cohen's kappa > 0.60 (Landis & Koch, 1977)
- Consistentie: Spearman's rho > 0.70 (Schober, Boer & Schwarte, 2018)
- Signalering zorgwekkend: overige methoden overwegen

Advies (2)

	Consensus hoog	Consensus laag
Consistentie hoog	*Gekalibreerd*	*Verschillen bespreken of 3e beoordelaar. Eigen protocol*
Consistentie laag	*Herkalibratie, kwaliteitsbeeld verschilt*	*Ongekalibreerd*

Advies (3)

- Accepteer drift. Sliding windows lastig toepasbaar. Blijf regelmatig kalibreren om drift te voorkomen en duurzaamheid van kalibratie te behouden
- Ankerportfolio's bij kalibratie om kwaliteitsbesef te ontwikkelen, niet ook tijdens beslismoment. Hiervan geen IBB berekenen
- Herbeoordelen / steekproeven alleen indien scores zeer zorgwekkend zijn, in hoge nood (zeer controlerend en tijdrovend)

Advies (4)

- 100% gekalibreerd zijn lukt niet
- Meet nooit zonder dialoog: belangrijk dat het niet gaat om individuele controle maar gezamenlijke normafstemming
- Spagaat tussen controle vs vertrouwen zal altijd aanwezig blijven

Met dank aan: examencommissie Fontys Paramedisch

Fontys › FOR SOCIETY

Femke Snieders (f.snieders@student.fontys.nl)
Bregje van Dijk (b.vaneijck@student.fontys.nl)

Bronnenlijst (1)

Andriessen, D. (2015). Handreiking kalibreersessies (Versie 1.1). Hogeschool Utrecht. Geraadpleegd op 20 november 2025, van https://www.hu.nl/onderzoek/publicaties/handreiking-kalibreersessies

Bloxham, S., den-Outer, B., Hudson, J. & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466-481. https://doi.org/10.1080/02602938.2015.1024607

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399–413. https://doi.org/10.1080/02602930600679050

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in Classroom Observation Scores. *Educational and psychological measurement*, *75*(2), 311–337. https://doi.org/10.1177/0013164414539163

Crimmins, G., Nash, G., Oprescu, F., Alla, K., Brock, G., Hickson-Jamieson, B. & Noakes, C. (2016). Can a systematic assessment moderation process assure the quality and integrity of assessment practice while supporting the professional development of casual academics? *Assessment & Evaluation in Higher Educatio*n, *41*(3), 427–441. https://doi.org/10.1080/02602938.2015.1017754

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. https://doi.org/10.2307/2529310

Mason, J. & Roberts, L.D. (2023). Consensus moderation: the voices of expert academics. *Assessment & Evaluation in Higher Education, 48*(7), 926-937. https://doi.org/10.1080/02602938.2022.2161999

Middleton, R., Lewer, K., Antoniou, C., Pratt, H., Bowdler, S., Jans, C. & Rolls, K. (2024). Understanding the processes, practices and influence of calibration on feedback literacy in higher education: a qualitative study. *Nurse Education Today*, 135. https://doi.org/10.1016/j.nedt.2024.106106

Bronnenlijst (2)

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and drift. *Journal of Educational Measurement, 46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

O'Donovan, B., Sadler, I. & Reimann, N. (2024). Social moderation and calibration versus codification: a way forward for academic standards in higher education? *Studies in Higher Education, 49*(12), 2693-2706. https://doi.org/10.1080/03075079.2024.2321504

Prichard, R., Peet, J., El Haddad, M., Chen, Y. & Lin, F. (2025). Assessment moderation in higher education: Guiding practice with evidence-an integrative review. *Nurse Education Today, 146.* https://doi.org/10.1016/j.nedt.2024.106512

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy and Practice, 20*, 5-19.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4), 1–11. https://doi.org/10.7275/2vrc-rx66