

Om op basis van die datapunten betrouwbare en rechtvaardige beslissingen te nemen, is het van belang dat examinatoren een gedeeld kwaliteitsbesef ontwikkelen

Kalibratie: van goed gesprek naar zichtbaar effect

■ **Bregje van Dijk**
B. van Dijk is hogeschooldocent bij de opleiding ALO van de Hogeschool van Arnhem en Nijmegen en tweedejaarsstudent van de opleiding Master Toets- en Onderwijskwaliteit. E-mail: bregje.vanDijk@han.nl

■ **Femke Snieders**
F. N. Snieders is hogeschooldocent bij de opleiding Facility Management van de Hogeschool van Arnhem en Nijmegen en tweedejaarsstudent van de opleiding Master Toets- en Onderwijskwaliteit.
E-mail: femke.snieders@han.nl

Dit artikel is een uitwerking van een van de beroepsproducten die zij tijdens de opleiding Master Toets- en Onderwijskwaliteit hebben opgeleverd.

Twee examinatoren beoordelen onafhankelijk hetzelfde portfolio. Beiden hebben deelgenomen aan kalibreersessies, gebruiken dezelfde leeruitkomsten en kennen dezelfde procedure. Toch komen zij niet altijd tot hetzelfde oordeel. Is dat erg? Is het onvermijdelijk? En vooral: hoe weet je of kalibratie voldoende effect heeft gehad?

Aanleiding

Vanuit de examencommissie van Fontys Paramedisch kwam de vraag of, en zo ja hoe, het effect van kalibratie te meten zou zijn. Binnen Fontys Paramedisch wordt gewerkt met programmatisch toetsen. Daarbij worden beslissingen over leeruitkomsten niet gebaseerd op één toetsmoment, maar op een samenhangend geheel van datapunten.

Om op basis van die datapunten betrouwbare en rechtvaardige beslissingen te nemen, is het van belang dat examinatoren een gedeeld kwaliteitsbesef ontwikkelen. Met kwaliteitsbesef bedoelen we dat examinatoren een gedeeld beeld hebben van wat kwaliteit in studentenwerk betekent: wanneer is een prestatie onder, op of boven niveau? Het gaat dus niet alleen om het kennen van de leeruitkomsten, maar vooral om het vergelijkbaar interpreteren en wegen van bewijsmateriaal.

Fontys Paramedisch werkt hieraan door structureel kalibreersessies te organiseren. In deze sessies bespreken examinatoren gezamenlijk studentenwerk en stemmen zij af hoe leeruitkomsten geïnterpreteerd en toegepast worden. Kalibratie is daarmee een belangrijk middel om tot meer vergelijkbare beoordelingen te komen, zeker wanneer beslissingen grote gevolgen hebben voor studenten.

Binnen Fontys Paramedisch is kalibratie stevig ingericht. Kalibreersessies worden zowel aan het begin als gedurende het onderwijs gepland, deelname heeft een verplicht karakter en er is draagvlak binnen de organisatie. Ook zijn inmiddels twee opleidingen succesvol geaccrediteerd. Op papier én in de organisatie lijkt de basis dus op orde.

Toch kwamen bij de examencommissie enkele casussen binnen waarin studenten het niet eens waren met besluiten van examinatoren. In ge-

sprekken bleek dat examinatoren niet altijd op dezelfde manier gekalibreerd leken te handelen. Daarmee ontstond een spanningsveld: als het proces goed is ingericht, hoe weet je dan of kalibratie ook daadwerkelijk doorwerkt in het beoordelaarsgedrag?
Deze vraag vormde het vertrekpunt van het onderzoek. We onderzochten daarom niet alleen hoe het effect van kalibratie zichtbaar gemaakt kan worden, maar ook hoe kalibratie binnen Fontys Paramedisch momenteel is ingericht en welke kwaliteitsaspecten daarbij van belang zijn. Pas wanneer duidelijk is wat goede kalibratie vraagt én of het huidige proces voldoet aan de gestelde kwaliteitseisen kan een passend advies worden gegeven over het meten van het effect ervan.

Definitie, doel en kwaliteitsaspecten van kalibratie

In de literatuur wordt onderscheid gemaakt tussen moderatie en kalibratie. Waar moderatie vaak een eenmalige kwaliteitscontrole achteraf is om variatie tussen beoordelaars te beperken, richt kalibratie zich op het borgen van kwaliteit aan de voorkant (Sadler, 2013). Kalibratie is een proces van professionele dialoog waarbij beoordelaars, door het gezamenlijk bespreken van studentenwerk, een gedeeld begrip van kwaliteitseisen ontwikkelen (Mason & Roberts, 2023). Het doel is hierbij tweeledig. Enerzijds gaat het om consensus over de interpretatie van leeruitkomsten: beoordelaars spreken samen af wat bijvoorbeeld ‘onder niveau’, ‘op niveau’ en ‘boven niveau’ betekent. Consensus gaat dus over de mate waarin beoordelaars tot hetzelfde oordeel komen. Anderzijds gaat het om consistentie in het beoordelen van studentenwerk: beoordelaars hanteren een vergelijkbaar beoordelingspatroon (Stemler, 2004). Dat betekent bijvoorbeeld dat zij hetzelfde portfolio als sterker of zwakker herkennen, ook wanneer zij de grens tussen ‘op niveau’ en ‘boven niveau’ net iets anders leggen.

Een effectieve kalibreersessie voldoet aan bepaalde kwaliteitseisen. Dit houdt in dat de kalibreersessies plaatsvinden zowel vóór de onderwijseenheid start en nogmaals vóór de beoordeling van de leeruitkomsten. Daarbij is een veilig klimaat onmisbaar, waarin een dialoog wordt gevoerd onder begeleiding van een moderator en afspraken vastgelegd worden door een notulist. Deze aanpak versterkt de sociale samenhang binnen een team en vergroot de kans dat beoordelaars niet alleen vaker tot hetzelfde oordeel komen, maar ook vanuit een vergelijkbaar kwaliteitsbeeld beoordelen. Het is daarvoor van belang dat er voldoende tijd en middelen beschikbaar worden gesteld om het proces zorgvuldig te doorlopen (Andriessen, 2015; Middleton et al., 2024). Ten slotte speelt inclusiviteit een grote rol, waarbij het expliciet meenemen van nieuwe docenten in de gedeelde standaard essentieel is voor een duurzaam resultaat (Prichard et al., 2025). Hoewel variatie in beoordeling nooit volledig uit te sluiten is zorgt een goed ingericht kalibratieproces voor een stabieler en meer gedragen kwaliteitsbesef dan traditionele moderatie (Bloxham et al., 2016).

Effect meten: verschillende mogelijkheden

Het effect van kalibratie is niet met één enkel meetinstrument vast te stellen. Afhankelijk van de informatiebehoefte van een examencommissie kunnen verschillende methodes worden ingezet, waarbij een zorgvuldige afweging gemaakt dient te worden tussen de werkdruk voor examinatoren versus de mate van controle en de diepgang van de kwaliteitsborging.

Om een eerste indruk te krijgen of examinatoren gekalibreerd zijn is het mogelijk om naar de interbeoordelaarsbetrouwbaarheid (IBB) te kijken. Dit is de mate waarin examinatoren onafhankelijk van elkaar tot vergelijkbare beoordelingen komen. Daarbij is het belangrijk om onderscheid te maken tussen consensus en consistentie.

Consensus gaat over de vraag of beoordelaars tot hetzelfde eindoordeel komen. Geven twee examinatoren bijvoorbeeld allebei het oordeel

‘op niveau’? Dan is er sprake van overeenstemming in de gekozen beoordelingscategorie. Consensus kan eenvoudig worden uitgedrukt als het percentage gelijke beoordelingen; wie wil corrigeren voor toevallige overeenstemming kan gebruikmaken van Cohen’s kappa. In de praktijk kan het uiteraard voorkomen dat twee examinatoren beide het oordeel ‘op niveau’ geven, maar de één vindt het portfolio net voldoende en de ander bijna excellent. Er is dan wel consensus, maar mogelijk geen gedeeld kwaliteitsbeeld. Daarom is het goed om naast consensus ook naar consistentie te kijken.

Consistentie gaat over de vraag of beoordelaars studentenwerk op een vergelijkbare manier ordenen. Vinden beide examinatoren bijvoorbeeld portfolio A sterker dan portfolio B? Dan wijst dat op een vergelijkbaar beoordelingspatroon. Ook wanneer de ene examinator iets strenger is dan de andere, kunnen zij dus toch consistent beoordelen. Consistentie kan bijvoorbeeld worden berekend met een rangordemaat, zoals Spearman’s rho.



Een perfecte manier om het effect van kalibratie te meten bestaat niet

Juist het onderscheid tussen consensus en consistentie is belangrijk. Alleen consensus kan schijnzekerheid geven: beoordelaars kunnen hetzelfde oordeel geven, zonder dat zij dit oordeel op hetzelfde kwaliteitsbeeld baseren. Andersom kan lage consensus samengaan met hoge consistentie: beoordelaars herkennen dan wel dezelfde kwaliteitsverschillen, maar leggen de grens tussen bijvoorbeeld ‘op niveau’ en ‘boven niveau’ net anders (Stemler, 2004).

Binnen Fontys Paramedisch sluit dit aan bij de bestaande werkwijze. Twee examinatoren geven onafhankelijk van elkaar een conceptbeoordeling: onder niveau, op niveau of boven niveau. Door deze conceptbeoordelingen vast te leggen vóórdat examinatoren met elkaar in overleg gaan, kan zichtbaar worden waar zij overeenkomen en waar verschillen ontstaan.

Een tweede methode om de IBB te bereken is aan de hand van ankerportfolio’s. Ankerportfolio’s zijn zorgvuldig gekozen voorbeelden van studentenwerk die samen een gedeelde kwaliteitsstandaard zichtbaar maken (Sadler, 1989). Een opleiding kan bijvoorbeeld twee of drie portfolio’s selecteren die jaarlijks gezamenlijk besproken worden tijdens een kalibreersessie. Door de IBB te berekenen over de beoordelingen van examinatoren van de ankerportfolio’s wordt niet alleen gekeken of beoordelaars onderling overeenstemmen, maar ook of hun oordeel aansluit bij de vooraf besproken norm. Deze methode helpt om het gesprek te voeren over wat binnen de opleiding wordt verstaan onder de beoogde leeruitkomsten. Een voordeel is dat dit minder controlerend kan voelen dan het rechtstreeks vergelijken van beoordelingen van collega’s. Tegelijk vraagt deze werkwijze om zorgvuldig gekozen ankerportfolio’s en zijn de uitkomsten statistisch minder hard te duiden, omdat de interpretatie van de gemeten IBB erg afhankelijk is van het niveau van het samengestelde ankerportfolio. Ook is het zeer tijdrovend

om ankerportfolio's op te stellen (of je moet hiervoor studentwerk gebruiken).
Wanneer de metingen met betrekking tot de consensus en/of consistentie reden tot zorg geven kan ervoor gekozen worden om portfolio's steekproefsgewijs te herbeoordelen. Daarbij worden enkele portfolio's opnieuw bekeken door andere examinatoren. Herbeoordeling kan helpen om scherper te krijgen waar verschillen in interpretatie ontstaan en kan bijdragen aan gedeeld normbegrip. Tegelijk is deze werkwijze tijdrovend en kan zij als zeer controlerend worden ervaren. Daarom wordt het aanbevolen om herbeoordeling vooral als gerichte verdiepingsactie in te zetten, niet als standaardinstrument voor alle beoordelingen.

Ook na goede kalibreersessies kan drift optreden: een geleidelijke verschuiving in beoordelaarsgedrag. Examinatoren kunnen na verloop van tijd ongemerkt strenger of milder gaan beoordelen, criteria anders interpreteren of afspraken minder consequent toepassen. Dat is niet vreemd; beoordelen blijft mensenwerk en wordt beïnvloed door routine, ervaring en nakijkbelasting. Daarom is het belangrijk om ook ná kalibratie alert te blijven op verschuivingen in beoordelaarsgedrag.
Een manier om zulke verschuivingen zichtbaar te maken is werken met sliding windows. Daarbij worden beoordelingen in opeenvolgende blokken verdeeld en wordt per blok gekeken naar consensus en consistentie tussen beoordelaars. Als de interbeoordelaarsbetrouwbaarheid in latere blokken afneemt, kan dat een signaal zijn dat herkalibratie nodig is.
Deze aanpak kent wel beperkingen. Verschillen tussen blokken kunnen ook ontstaan doordat de kwaliteit van het studentenwerk verschilt. Bovendien zijn uitkomsten bij kleine aantallen minder stabiel. Sliding windows zijn daarom vooral bruikbaar als er voldoende beoordelingsgegevens zijn en als de resultaten worden gezien als signaal, niet als bewijs (Casabianca et al., 2014).

Er kan geconcludeerd worden dat alle methoden voor- en nadelen kennen. Deze zijn afgewogen voordat een expliciet advies aan de examencommissie geformuleerd kon worden. Een overzicht van de voor- en nadelen zijn te vinden in tabel 1.

Van meten naar adviseren: wat kan de examencommissie hiermee?

Een perfecte manier om het effect van kalibratie te meten bestaat niet. In het geval van Fontys Paramedisch is een aanpak geadviseerd die

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 duurzaam effect, alleen op dat moment
IBB consistentie (Pearson-correlatie 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
Herbeoordelen	• Leidt tot verbeterd normbegrip na dialoog	• Tijdrovend • Controlerend
Drifting (sliding windows)	• Check of kalibratie ook duurzaam effect heeft	• Geen sprake van normwaardes • Met kleine groepen (windows) instabielere cijfers

paste bij de bestaande werkwijze. In besliscommissies geven twee examinatoren onafhankelijk van elkaar een conceptbeoordeling: onder niveau, op niveau of boven niveau. Door twijfelgevallen ook dusdanig te registreren worden beoordelingen op een vijf-puntsschaal vastgelegd. Door deze conceptbeoordelingen vast te leggen vóórdat examinatoren met elkaar in overleg gaan, kan zichtbaar worden in hoeverre zij tot vergelijkbare oordelen komen. Daarbij is het belangrijk om niet alleen te kijken naar consensus, maar ook naar consistentie. De uitkomsten moeten niet worden gebruikt als hard bewijs dat examinatoren wel of niet gekalibreerd zijn. Ze zijn vooral een aanleiding voor gesprek. Wanneer de uitkomsten vragen oproepen, is het advies om eerst te onderzoeken waar verschillen ontstaan. Daarna kan worden besloten of verdere meetmethoden ingezet moeten worden.



Figuur 1. Advies naar aanleiding van het berekenen van de interbeoordelaarsbetrouwbaarheid
Noot. Afbeelding gegenereerd met behulp van ChatGPT (Open AI, 2026)

Tussen vertrouwen en controle

Het advies is om met name de IBB in kaart te brengen (zowel consensus als consistentie) en die uitkomsten als signaalfunctie te gebruiken, maar niet als harde meting. Dit is schematisch weergegeven in Figuur 1. Het nadeel van dit voorstel is echter dat examinatoren het gevoel kunnen hebben gecontroleerd te worden en dit defensief gedrag kan opwekken. De belangrijkste voorwaarde van het meten van het effect van kalibratie is dat vooraf helder wordt gecommuniceerd waarom er gemeten wordt. Het doel is niet controle van individuele examinatoren, maar het versterken van gezamenlijke normontwikkeling. Meten krijgt pas waarde wanneer de uitkomsten leiden tot professionele dialoog.

Kalibratie als kompas, niet als keurslijf

Beoordelen blijft mensenwerk. Examinatoren kunnen goed voorbereid, zorgvuldig geschoold en regelmatig gekalibreerd zijn, maar volledig identieke oordelen zijn onrealistisch. Ook 'drift' is een natuurlijk proces: na verloop van tijd kunnen interpretaties ongemerkt verschuiven. De opdracht is daarom niet om alle verschillen uit te bannen, maar om ze zichtbaar en bespreekbaar te houden. Dáár ligt de duurzame waarde





Duurzame waarde van kalibratie ligt een blijvende aandacht voor in gedeeld kwaliteitsbesef en de professionele dialoog

van kalibratie: niet in het bereiken van honderd procent consensus, maar in het behouden van een blijvende aandacht voor een gedeeld kwaliteitsbesef en de professionele dialoog. ■

Referenties

- 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
- Casabianca, J. M., Junker, B. W., & Patz, R. J. (2014). Detecting changes in rater behavior over time. Applied Measurement in Education, 27(4), 303-320. https://doi.org/10.1080/08957347.2014.944093
- 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
- 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. https://doi.org/10.1080/0969594x.2012.714742
- 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/10.7275/96jp-xz07


onderdeel van
Examenjaar 2026
Door de ogen van de examenmakers
Docent en AI
“Waar het om gaat is de optimale samenwerking tussen docent en AI.”
Collega Wouter Jansen vertelt in ons examenverslag hoe we kijken naar de inzet van AI bij het nakijken van centrale examens.
Ontdek het zelf - veel leesplezier!
Naar het artikel
CITO CTE