



Inspectie van het Onderwijs
*Ministerie van Onderwijs, Cultuur en
Wetenschap*

DE BETROUWBAARHEID VAN INSPECTEURSOORDELEN

EEN VIGNETONDERZOEK

juli 2025

Voorwoord

“Het zou niet moeten uitmaken welke inspecteur mijn school onderzoekt,” zeggen onderwijsbestuurders en schoolleiders regelmatig. Dat is terecht, want te grote verschillen tussen beoordelaars roepen twijfel op over afzonderlijke beoordelingen én over ons toezicht in het algemeen.

Toch zijn verschillen tussen professionele beoordelaars nooit helemaal uit te sluiten. Vooral als zij de vraag krijgen om bij hun oordeel de context in het oog te houden. Dit geldt ook voor ons inspectiewerk, waarin we tot expertoordelen komen door feiten en observaties te wegen in onderlinge samenhang en in samenhang met de context. Daarbij werken we dus niet met afvinklijstjes, zoals wel eens gedacht wordt.

De zogeheten interbeoordelaarsbetrouwbaarheid (IBB) vraagt om voortdurende aandacht. Want als we daar niet op reageren – bijvoorbeeld via scholing en verbeteracties – gaan de oordelen gaandeweg steeds meer uit elkaar lopen.

Daarom is het van belang dat we de mate van IBB periodiek toetsen. Zoals via dit onderzoek. Op basis van de uitkomsten stellen we vast dat de mate van IBB overeenkomt met wat je, op basis van onderzoeken in soortgelijke situaties, mag verwachten. Dat geeft geen reden tot zorg, maar veel belangrijker vind ik dat dit onderzoek aanknopingspunten biedt om de IBB verder te verbeteren. Door dit onderzoek over enkele jaren te herhalen, houden we zicht op de ontwikkeling van de IBB. En kunnen we zo nodig gericht bijsturen.

Dit onderzoek maakt deel uit van het evaluatieprogramma waarmee we de kwaliteit en het effect van ons toezicht doorlopend toetsen. De uitkomsten van de verschillende onderzoeken binnen dit programma gebruiken we daarnaast bij het ontwerp van ons bijgestelde onderzoekskader, dat op 1 augustus 2027 ingaat.

Zo blijven we werken aan toezicht dat het onderwijs verder verbetert.

Alida Oppers,

inspecteur-generaal van het Onderwijs

INHOUD

Voorwoord 2

Samenvatting 5

1	Inleiding 8
1.1	Doel van het onderzoek 9
1.2	Ruis in oordeelsvorming 11
1.3	Resultaten uit vergelijkbaar onderzoek 12
1.4	Leeswijzer 15
2	Methode 16
2.1	Deelnemers 16
2.2	Materiaal 16
2.3	Procedure 17
2.3.1	Fase 1: individuele beoordeling 18
2.3.2	Fase 2: duo-beoordeling 18
2.4	Vorbereiding analyses 18
3	Resultaten primair onderwijs 20
3.1	Hoofdanalyse primair onderwijs 20
3.2	Secundaire analyses primair onderwijs 21
3.3	Exploratieve analyses primair onderwijs 24
4	Resultaten (voortgezet) speciaal onderwijs 27
4.1	Hoofdanalyse (voortgezet) speciaal onderwijs 27
4.2	Secundaire analyses (voortgezet) speciaal onderwijs 28
4.3	Exploratieve analyses (voortgezet) speciaal onderwijs 29
5	Resultaten voortgezet onderwijs 32
5.1	Hoofdanalyse voortgezet onderwijs 32
5.2	Secundaire analyses voortgezet onderwijs 33
5.3	Exploratieve analyses voortgezet onderwijs 35
6	Resultaten middelbaar beroepsonderwijs 37
6.1	Hoofdanalyse middelbaar beroepsonderwijs 37
6.2	Secundaire analyses middelbaar beroepsonderwijs 38
6.3	Exploratieve analyses middelbaar beroepsonderwijs 40
7	Discussie 42
7.1	Bevindingen vergeleken met ander onderzoek 42
7.2	Aanvullende bevindingen 43
7.2.1	Duo's versus individuele inspecteurs 43
7.2.2	Eindoordelen versus oordelen op standaarden 43
7.2.3	Strengheid inspecteurs 44
7.2.4	Groepen inspecteurs 44
7.2.5	Validiteit bevindingen 44
7.2.6	Verschillen in overeenstemming tussen standaarden 45
7.2.7	Voldoende met herstelopdracht als afzonderlijk oordeel 45
7.3	Hoe kunnen we de IBB verbeteren? 45
7.4	Vervolg 46

Bijlage 1 Primair onderwijs 48

Bijlage 2 (Voortgezet) speciaal onderwijs 49

Bijlage 3 Voortgezet onderwijs 50

Bijlage 4 Middelbaar beroepsonderwijs 51

Samenvatting

In dit rapport doen we verslag van een onderzoek naar de betrouwbaarheid van inspecteursoordelen. Dit onderzoek geeft ons inzicht in de vraag in hoeverre inspecteurs dezelfde onderwijssituaties op dezelfde manier beoordelen. Kortom, het geeft inzicht in de mate van interbeoordelaarsbetrouwbaarheid (IBB).

Als inspectie vinden wij het belangrijk om hier zicht op te hebben. De mate van IBB zegt namelijk iets over de kwaliteit van ons toezicht. En daarmee over onze geloofwaardigheid en voorspelbaarheid. Vanuit onze interne kwaliteitszorg besteden we veel aandacht aan het professionaliseren van onze inspecteurs. We evalueren regelmatig of er ongewenste verschillen zijn tussen inspecteurs en beoordelingen. Zo nodig voeren we verbeteracties uit. Toch is volledige overeenstemming tussen professionele beoordelaars niet realistisch. Beoordelen is en blijft mensenwerk.

Met een zogenaamd vignetonderzoek brachten we de mate van IBB binnen de inspectie in beeld. Het onderzoek richtte zich op het primair onderwijs (po), (voortgezet) speciaal onderwijs ([v]so), voortgezet onderwijs (vo) en middelbaar beroepsonderwijs (mbo). Voor elke sector stelden enkele inspecteurs in totaal 16 vignetten op. Elk vignet bevatte een beschrijving van de observaties bij een daadwerkelijk uitgevoerd kwaliteitsonderzoek op een school (po, [v]so), afdeling (vo) of opleiding (mbo).

De vignetten boden informatie over de standaarden die beslissend zijn voor de eindoordelen in de verschillende sectoren. Die standaarden hadden betrekking op de kwaliteitsgebieden Onderwijsproces, Veiligheid en schoolklimaat, Onderwijsresultaten, en Sturen, kwaliteitszorg en ambitie. En daarnaast in het mbo: Borging en afsluiting.

Tijdens het vignetonderzoek beoordeelden inspecteurs de standaarden. Ook gaven zij een eindoordeel over de kwaliteit van de school, afdeling of opleiding. Dit eindoordeel bepaalden zij aan de hand van beslisregels, die uitgaan van de beoordelingen van de standaarden. Voor elk van de sectoren namen we ook de nieuwe standaard OP0 Basisvaardigheden op, omdat we die met ingang van schooljaar 2025-2026 gaan beoordelen. Omdat we OP0 tijdens het onderzoek nog niet in de praktijk beoordeelden, lieten we deze standaard buiten beschouwing in onze analyses. Dat wil zeggen: bij het bepalen van de mate waarin de oordelen van inspecteurs overeenkwamen.

De deelnemende inspecteurs scoorden tijdens het onderzoek individueel enkele vignetten. Daarna gingen zij in duo's in gesprek over de vignetten die zij verschillend hadden beoordeeld. Zij schreven op waarom ze van mening verschilden en kwamen, na overleg, tot gezamenlijke oordelen. Inspecteurs konden tijdens het onderzoek alle informatie raadplegen die zij ook in de dagelijkse praktijk van het toezicht gebruiken. Naast het onderzoekskader¹, waarin onze werkwijze staat en welke standaarden we beoordelen, beschikken inspecteurs over een handleiding kwaliteitsonderzoek. Deze handleiding bevat een zogenaamd

1 <https://www.onderwijsinspectie.nl/onderwerpen/onderzoekskaders>.

afwegingskader/handreiking met richtlijnen die de inspecteurs ondersteunen bij het beoordelen van de standaarden.

De resultaten laten zien dat inspecteurs het in gemiddeld 82% tot 89% van de gevallen met elkaar eens zijn wanneer zij afzonderlijk oordelen. De percentages zijn afhankelijk van de sector. Over de eindoordelen zijn zij het iets minder vaak met elkaar eens: gemiddeld 65% tot 81%. Dat is een verklaarbaar verschil: de inspecteurs kunnen het weliswaar over meerdere standaarden met elkaar eens zijn, maar één afwijkend oordeel op een standaard kan al tot een ander eindoordeel leiden.

Strikt genomen zijn er geen statistisch significante verschillen te zien tussen duo's en individuele inspecteurs. Zij kwamen ongeveer even vaak tot hetzelfde oordeel, behalve bij standaarden in het mbo. Toch zijn er meerdere aanwijzingen dat oordelen in duo's de voorkeur verdient. Dat gebeurt nu al in de huidige toezichtpraktijk. Voor alle sectoren geldt dat de duo's iets vaker overeenstemming bereiken. Namelijk: 89% tot 95% voor de standaarden en 76% tot 82% voor de eindoordelen. Bovendien valt op dat individuele vergissingen van inspecteurs bijna altijd werden gecorrigeerd zodra inspecteurs in duo's gingen werken. Overigens is het aandeel scholen, afdelingen en opleidingen dat eenzelfde oordeel krijgt niet gelijk aan – maar hoger dan – het percentage overeenstemming. Afhankelijk van de sector krijgt 83% tot 90% van de scholen, afdelingen en opleidingen hetzelfde eindoordeel na beoordeling door duo's.

Als verklaring voor verschillende oordelen op de standaarden, gaven inspecteurs het vaakst aan dat zij elementen uit het afwegingskader/de handreiking verschillend wogen. Of deze elementen anders interpreteerden. Het al dan niet meewegen van oordelen op andere standaarden werd ook regelmatig genoemd als verklaring.

We stellen vast dat het onderling vergelijken van resultaten tussen de sectoren niet eenvoudig is. Dus dat raden we af. Verschillen tussen sectoren kunnen duiden op verschillen in de kwaliteit van de oordelen. Ze kunnen ook het resultaat zijn van de manier waarop de vignetten zijn samengesteld in de verschillende sectoren. Voor sommige vignetten sluit de formulering nauw aan op de rapporten van eerdere kwaliteitsonderzoeken waarop ze zijn gebaseerd. Daardoor liggen ze dicht bij de oorspronkelijke oordelen. Dat kan leiden tot een hogere mate van overeenstemming tussen inspecteurs.

Er bestaat geen duidelijke norm voor een voldoende mate van overeenstemming. Het is daarom moeilijk om aan te geven of de vastgestelde overeenstemming hoog of laag is. Wel is het mogelijk om de onderzoeksresultaten te vergelijken met die van studies naar de betrouwbaarheid van oordelen bij andere inspecties. Zo'n vergelijking laat zien dat de vastgestelde mate van overeenstemming in het po, (v)so, vo en mbo in lijn is met die van andere onderzoeken naar de betrouwbaarheid van oordelen.

We beschouwen dit onderzoek als nulmeting. Door dit onderzoek over enkele jaren te herhalen, houden we zicht op de mate van IBB. Ondertussen geven we voortdurend aandacht aan het bevorderen van de IBB via allerlei vormen van professionalisering. Ook voeren inspecteurs onderzoeken nooit alleen uit, maar altijd in wisselende teams. Na afloop van een onderzoek volgt een zogeheten consensusoverleg, waarin de inspecteurs de oordelen bepalen en referenten alle onderzoeksrapporten kritisch meelesen. Daarnaast brengen we met intern

onderzoek in kaart wat er goed gaat en wat er beter kan in het toezicht. De uitkomsten van deze onderzoeken, waaronder die van dit vignetonderzoek, nemen we mee in de professionalisering van onze collega's. En daarnaast in het verbeteren van materialen ter ondersteuning van de oordeelsvorming en de periodieke aanscherping van onze werkwijze.

1 Inleiding

Als Inspectie van het Onderwijs (hierna: inspectie) houden we toezicht op de kwaliteit van het onderwijs. Dat doen we onder meer door onderzoek te doen bij besturen, scholen, instellingen, opleidingen en samenwerkingsverbanden. Aan de hand van onderzoekskaders² beoordelen onze inspecteurs de kwaliteit en geven zij herstelopdrachten, indien nodig. De onderzoekskaders beschrijven de werkwijze van de inspectie en geven aan wat we onderzoeken bij besturen, scholen, afdelingen en opleidingen.

Onze missie luidt: **Effectief toezicht voor beter onderwijs**. Effectief toezicht hangt samen met de geloofwaardigheid van de toezichthouder en met de kwaliteit van het toezichtproces. In de ideale situatie maakt het niet uit welke inspecteur op bezoek komt. Als inspectie willen we daarom weten in hoeverre inspecteurs dezelfde situaties op dezelfde manier beoordelen. In dit rapport doen we verslag van een onderzoek naar deze vraag. Daarbij kijken we naar de oordelen die inspecteurs geven bij kwaliteitsonderzoeken op scholen, afdelingen en opleidingen in het primair onderwijs (po), (voortgezet) speciaal onderwijs ([v]so), voortgezet onderwijs (vo) en middelbaar beroepsonderwijs (mbo). We zijn met het onderzoek in het po gestart in schooljaar 2023-2024. Ervaringen die we bij dat onderzoek opdeden, leidden tot enkele kleine aanpassingen in de onderzoeken naar de inspecteursoordelen in het (v)so, vo en mbo in 2024-2025. Zo kregen de inspecteurs in het (v)so, vo en mbo bijvoorbeeld iets meer tijd om de vignetten te beoordelen.

Professioneel oordelen is **mensenwerk**. Dat betekent dat het oordeel niet automatisch volgt uit het onderzoekskader. De inspecteur weegt feiten en observaties in samenhang met elkaar én met de context. En past daarbij de eigen expertise toe. Hoe inspecteurs alle informatie in de praktijk exact wegen, verschilt uiteraard. Ook kunnen inspecteurs fouten maken. Hierdoor ontstaat '**ruis**' in de beoordeling.

Ruis treedt op bij elk professioneel beoordelingsproces³. Bijvoorbeeld bij rechters, artsen, psychiaters, onderwijzers, asielverleners, patentverstrekkers, arbeidsongeschiktheidsdeskundigen en inspecteurs van kernreactoren, mijnen, milieu en gezondheidszorg⁴. Dat is onvermijdelijk. Inspecteurs van het onderwijs vormen daarop geen uitzondering, laat Engels onderzoek^{5,6,7} zien. Een hoge mate van ruis kan afbreuk doen aan de geloofwaardigheid van de professional of instantie in kwestie. Die is dan minder voorspelbaar en minder betrouwbaar. Ruis binnen de inspectie vertelt ons daarmee iets over de kwaliteit van ons toezicht. Daarom is het wenselijk te streven naar een zo laag mogelijke mate van ruis. En dus een zo hóóg mogelijke **interbeoordelaarsbetrouwbaarheid (IBB)**.

2 <https://www.onderwijsinspectie.nl/onderwerpen/onderzoekskaders>.

3 Kahneman, D., Sibony, O., & Sustain, C. (2021). *Noise – A flaw in human judgement*. HarperCollins Publishers.

4 Ho, D.E. (2016). Does peer review work? An experiment of experimentalism. *Stanford Law Review*, 69, 1.

5 Bokhove, C., Jerrim, J., Sims, S. (2023). *Are some school inspectors more lenient than others?* University of Southampton. Doi:10.5258/SOTON/P1108.

6 Ofsted (2017). *Do two inspectors inspecting the same school make consistent decisions?* Ofsted research report 170004.

7 Ofsted (2019). *Workbook scrutiny*. Ofsted research report 190028.

Het uitgevoerde onderzoek geeft een beeld van de mate van ruis binnen de inspectie. En daarmee van de betrouwbaarheid en kwaliteit van de inspectieoordelen. De uitkomsten van dit onderzoek zijn belangrijk voor onze interne kwaliteitszorg en externe verantwoording. Bovendien bieden zij aanknopingspunten voor verbeteracties. Als inspectie werken we voortdurend aan het verbeteren van de kwaliteit van ons werk. We besteden aandacht aan de professionalisering van onze medewerkers. Ook voeren we allerlei onderzoeken uit om ons werk verder te verbeteren. Het vignetonderzoek is een van deze onderzoeken.

1.1 Doel van het onderzoek

Het doel van dit onderzoek is: de mate van IBB onder inspecteurs bepalen. Tijdens een kwaliteitsonderzoek op een school, afdeling of opleiding beoordelen inspecteurs een aantal standaarden en geven zij een algeheel eindoordeel over de school, afdeling of opleiding. Dit doen zij door lessen te observeren, documenten te bestuderen en verschillende gesprekken te voeren. Bijvoorbeeld met de directie, interne begeleiding, leraren of docenten, leerlingen of studenten, en ouders. Oordelen op de standaarden bepalen zij aan de hand van het onderzoekskader en de handleiding kwaliteitsonderzoek. Deze handleiding bevat voor elke standaard een afwegingskader/handreiking met richtlijnen voor het beoordelen van de standaarden. De inspecteurs bepalen het eindoordeel met de beslisregels uit het onderzoekskader. Leidend daarbij zijn de oordelen op de standaarden.

We beantwoorden onderstaande onderzoeksvragen voor elke sector apart.

Onze **hoofdvraag** luidt:

- Hoe groot is de overeenstemming tussen inspecteurs voor eindoordelen en oordelen op standaarden bij kwaliteitsonderzoeken op scholen, afdelingen of opleidingen?

Bij het beantwoorden van deze vraag vinden we de mate van overeenstemming over het eindoordeel het belangrijkste. Die heeft namelijk de grootste impact. In tweede instantie kijken we naar de mate van overeenstemming over de oordelen op de standaarden. We kunnen de IBB het meest nauwkeurig schatten op basis van de oordelen van individuele inspecteurs. Bovendien kunnen we de resultaten over individuele oordelen het best vergelijken met onderzoek door andere inspecties. De hoofdvraag beantwoorden we dan ook aan de hand van individuele oordelen.

In dit onderzoek legden we aan inspecteurs vignetten met casussen van scholen, afdelingen of opleidingen voor. We vroegen hen om op basis hiervan de kwaliteit van de scholen, afdelingen of opleidingen te beoordelen. Inspecteurs voeren kwaliteitsonderzoeken nooit alleen uit. Daarom lieten we de deelnemende inspecteurs ook in duo's gezamenlijke oordelen geven, na het bepalen van hun individuele oordelen. De schatting van de IBB voor inspecteursduo's is echter minder nauwkeurig. Tijdens het vignetonderzoek kunnen we namelijk veel minder duo's vormen dan in de praktijk.

Bij het bepalen van de mate van overeenstemming van de duo's, namen we de oordelen waarover duo-partners het in de individuele fase eens waren, ook mee. Met als aanname dat inspecteursduo's het over deze oordelen eens zouden blijven.

Om aanknopingspunten voor verbeteracties te vinden, stellen we in dit onderzoek ook enkele **secundaire vragen**:

- Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's?

Een rechtstreekse vergelijking van de overeenstemming tussen individuele oordelen en duo-oordelen is niet mogelijk. Bij de analyse van inspecteursduo's laten we namelijk een (klein) deel van de vignetten waarover de oordelen verschilden, buiten beschouwing. Vanwege de tijdslimiet konden de duo's deze niet bespreken. De analyse richt zich daarom op de besproken vignetten, aangevuld met de vignetten waarover de individuele inspecteurs het al eens waren.

- Is de overeenstemming groter binnen kantoren⁸ en binnen cohorten inspecteurs?

Voor de sectoren (v)so, vo en mbo waren er te weinig inspecteurs per cohort voor een zinvolle vergelijking tussen de overeenstemming per cohort en de gemiddelde overeenstemming van alle inspecteurs. We rapporteren hierover daarom alleen voor de sector po.

- Welke redenen noemen duo's inspecteurs voor verschillen in individuele oordelen?

Na het individueel beoordelen, bespraken inspecteurs in duo's de mogelijke verklaringen voor de verschillen in oordeel. Daarbij hadden zij keus uit één of meerdere opties van een vooraf gestructureerde lijst met mogelijke verklaringen.

De analyses waarmee we de hoofdvraag en de secundaire vragen beantwoorden, legden we voorafgaand aan de uitvoering van het onderzoek vast in een pre-analyseplan⁹. Om extra grip te krijgen op de data stellen we aanvullend enkele **exploratieve vragen**:

- Hoe vaak krijgen scholen, afdelingen of opleidingen hetzelfde eindoordeel?

In dit onderzoek weten we niet wat de juiste oordelen zijn. Wel kunnen we nagaan hoe vaak scholen, afdelingen of opleidingen hetzelfde eindoordeel krijgen.

⁸ Dit geldt alleen voor het onderzoek in po. Alleen po-inspecteurs zijn verbonden aan een specifiek kantoor (Zwolle, Utrecht of Tilburg) en werken meer samen met collega's binnen hetzelfde kantoor dan met collega's verbonden aan een ander kantoor.

⁹ <https://osf.io/b68qe> | Registratie d.d. 19 januari 2024. Uitvoering eerste onderzoek d.d. 24 januari 2024.

- Hoe groot zijn de verschillen in overeenstemming tussen standaarden?

Bij deze onderzoeksvraag beschouwen we de verschillende standaarden apart, inclusief OP0 Basisvaardigheden. Bij de voorgaande onderzoeksvragen beschouwen we de percentages overeenstemming over oordelen op de standaarden samen.

We merken hierbij op dat sommige standaarden zich mogelijk beter lenen voor een vignetstudie dan andere. Een voorbeeld: OP3 Pedagogisch-didactisch handelen – normaal gesproken gebaseerd op lesobservaties – is nu eenmaal moeilijker te vangen in tekst dan de standaarden VS1 Veiligheid en OR1 Resultaten/ Studiesucces. Het kan zijn dat we de overeenstemming over OP3 daardoor sterker overschatten dan de overeenstemming over VS1 en OR1.

- Hoe groot zijn de verschillen in strengheid tussen inspecteurs?

Met strengheid bedoelen we hier de kans dat een inspecteur een willekeurige standaard als Onvoldoende beoordeelt. Naarmate de strengheid van inspecteurs sterker varieert, verklaart dit een groter deel van de totale ruis in de beoordeling. Voor een betrouwbare schatting van strengheid per inspecteur baseerden we ons op de oordelen op de standaarden.

1.2 Ruis in oordeelsvorming

Oordelen kunnen op 2 manieren afwijken: door bias en/of ruis. Bias, eigenlijk 'systeembias', verwijst naar de zogeheten systematische afwijking in een groep oordelende mensen. Als inspecteurs bijvoorbeeld veel Onvoldoendes uitdelen en gemiddeld genomen te streng zijn, dan wordt dat bias genoemd. Om na te gaan of hiervan sprake is, is het nodig om te weten wat het 'juiste' oordeel is. Ruis gaat daarentegen om willekeurige variaties in oordelen. Hiervan is sprake als dezelfde school of opleiding verschillende oordelen krijgt, afhankelijk van de inspecteur die ter plaatse komt kijken. Om te constateren of sprake is van ruis, is het niet nodig om te weten wat het 'juiste' oordeel is. Een hoge mate van ruis betekent simpelweg dat inspecteurs dezelfde situatie verschillend beoordelen.

Er zijn verschillende methodes om de mate van ruis te bepalen. Allereerst is er het zogeheten **veldexperiment**. In 2007 en in 2011 voerden wij 2 van dit soort onderzoeken uit in het po. Inspecteursduo's hielden samen kwaliteitsonderzoeken, waarbij ze de opdracht kregen om aan het eind van het onderzoek onafhankelijk van elkaar oordelen te geven. Achteraf werd de mate van overeenstemming berekend. Het voordeel van deze methode is dat de inspecteurs 'levensechte' situaties beoordelen. Een beperking van deze methode is de mate waarin de oordeelsvorming daadwerkelijk onafhankelijk plaatsvindt. Gesprekken met een deelnemer aan de onderzoeken uit 2007 en 2011 suggereren dat onderlinge beïnvloeding voorkwam. Soms was sprake van tussentijds overleg. Een collega kon zo'n gesprek een bepaalde kant op sturen of bijvoorbeeld via mimiek de individuele oordeelsvorming beïnvloeden. Dit is een groot nadeel van dit type onderzoek.

Een tweede methode, die we toepasten in het huidige onderzoek, is het zogeheten **vignetonderzoek**. Een vignet is een beschrijving van een situatie uit de toezichtpraktijk, die voldoende relevante informatie bevat om oordelen te kunnen geven. Aan inspecteurs wordt gevraagd om de situatie te beoordelen zoals zij dat in de dagelijkse praktijk doen. Daarbij is de externe validiteit lager dan in een veldexperiment. Een vignet geeft namelijk maar een deel van de werkelijkheid weer.

Bovendien zijn er geen consequenties voor de school, afdeling of opleiding verbonden aan de oordelen. In de praktijk is de overeenstemming daarom mogelijk lager. Immers, hoe meer informatie er beoordeeld moet worden, hoe groter de kans op interpretatie¹⁰. Bovendien krijgen inspecteurs in een vignetonderzoek dezelfde informatie aangereikt, terwijl het in de praktijk kan verschillen welke informatie zij meenemen in hun oordeelsvorming. Het voordeel van een vignetonderzoek is daarom de grote experimentele controle. Inspecteurs beoordelen exact dezelfde informatie en doen dit volledig onafhankelijk van elkaar. Een tweede voordeel is de herhaalbaarheid, dus de mogelijkheid om resultaten over tijd te vergelijken. Als we dezelfde vignetten opnieuw aan inspecteurs voorleggen, dan zijn verschillen in overeenstemming niet te herleiden naar veranderingen in de te beoordelen situatie(s). Daarentegen is het niet mogelijk om een werkelijk onderzoek op een school of opleiding op die manier te herhalen.

We gebruiken het **percentage overeenstemming** als maat voor de IBB. Hiervoor delen we het totaal aantal overeenkomstige oordelen over een vignet door het totaal aantal beoordelingen. Zijn 2 beoordelaars het in 8 van de 10 gevallen eens? Dan is het percentage overeenstemming 80%.

1.3 Resultaten uit vergelijkbaar onderzoek

Hoe groot mag de ruis zijn? Welke mate van overeenstemming is acceptabel? Hierop bestaat geen sluitend antwoord. In algemene zin geldt: hoe groter de consequenties van de beoordeling, hoe belangrijker een hoge mate van overeenstemming (IBB)¹¹. In wetenschappelijke literatuur zijn vuistregels voor de interpretatie van IBB-maten te vinden. Deze zijn echter arbitrair en sluiten niet aan op elke specifieke context. Om iets te kunnen zeggen over de resultaten van ons onderzoek, vergelijken we deze met de resultaten van IBB-onderzoek in contexten die het meest lijken op ons inspectiewerk.

Onze inzet is om de eigen resultaten met een groter aantal studies te kunnen vergelijken. Daarvoor beschouwen we niet alleen het percentage overeenstemming als maat van IBB, maar ook **kappa**. Deze maat corrigeert toevallige overeenstemming¹². Kappa kent een waarde van 0 tot 1. Met als een van de vuistregels: van geringe overeenstemming bij $\text{kappa} < 0,21$ tot bijna perfecte overeenstemming bij kappa tussen 0,81 en 1¹³. De Engelse onderwijsinspectie hanteert een percentage overeenstemming van 80% en een kappa-waarde van 0,7 als richtlijn voor een 'hoge IBB onder onderwijsinspecteurs'¹⁴. Een Amerikaans onderzoeksinstituut dat keek naar de beoordeling van lerarenprestaties, ging uit van vergelijkbare cijfers: een acceptabel percentage overeenstemming tussen 75% (ondergrens) en 90% (hoog); een acceptabele kappa-waarde tussen 0,61 (ondergrens) en 0,81 (hoog)¹⁵. Het Amerikaanse instituut voor examinering hanteert een ondergrens van 70% overeenstemming voordat nieuwe examinatoren

10 Kahneman, D., Sibony, O., & Sustain, C. (2021). *Noise – A flaw in human judgement*. HarperCollins Publishers.

11 LeBreton, J. M., & Sentor, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815–852.

12 Beoordelaars kunnen met verschillende referentiekaders, dus met verschillende overwegingen, toch tot dezelfde oordelen komen. Dit wordt toevallige overeenstemming genoemd.

13 Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159-74 .

14 Ofsted (2017). *Do two inspectors inspecting the same school make consistent decisions?* Ofsted research report 170004.

15 Graham, M., Milanowski, A. T., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*. Center for Educator Compensation Reform. Distributed by ERIC Clearinghouse.

in de praktijk examens mogen beoordelen¹⁶. Alle maten die toevalsovereenstemming corrigeren, zijn echter minder betrouwbaar. (Zie voor meer toelichting het Technisch Rapport dat we naast dit publieksrapport publiceren.) Daardoor geldt dit ook voor de betreffende normen.

IBB is in het onderwijsveld geen onbekend terrein. De beoordeling van leerlingenprestaties – door docenten – is geregeld onder de loep genomen. Deze beoordeling blijkt logischerwijs samen te hangen met de complexiteit van het te beoordelen materiaal. Overeenstemming over meerkeuzevragen is bijvoorbeeld hoog: tussen 93% en 100%¹⁷. De becijfering van meer complexe open vragen of examenonderdelen gaat gepaard met een lagere mate van overeenstemming: respectievelijk 42% tot 67%¹⁸ en 54% tot 97,4%¹⁹. Dergelijke cijfers illustreren dat een percentage overeenstemming van 100% niet realistisch is. En dat een realistische norm sterk afhankelijk is van de complexiteit van de casus.

Zoals gezegd zijn er **geen duidelijke richtlijnen** voor IBB in de praktijk en is het beste dat we kunnen doen een vergelijking maken met IBB-onderzoek in vergelijkbare omstandigheden. Zie Tabel 1.1. Een uitgebreidere bespreking van deze onderzoeken is te vinden in Bijlage II van het Technisch Rapport. Het gaat hier om overeenstemming in oordelen van onderwijsinspecteurs, andersoortige inspecteurs en (medisch) professionals waarbij de oordelen serieuze consequenties hebben. Tot slot bespreken we onderzoek naar de beoordeling van leskwaliteit.

Tabel 1.1

Beknopt literatuuroverzicht

Onderzoek	Niveau	Onafhankelijk	Overeenkomst	Kappa
Inspecteurs onderwijs				
Scholen po NL (2007)	indicatoren	twijfelachtig	95%	
Scholen po NL (2011)	indicatoren	twijfelachtig	94%	
Scholen po ENG (2017)	eindoordelen	twijfelachtig	92%	0,80
Curriculum ENG (2019)	indicatoren	in orde		0,38-0,49
Inspecteurs andere inspecties				
Restaurants VS (2016)	indicatoren	laag	96%	
Zorgcentra Uganda (2019)	indicatoren	in orde	71%	
Ziekenhuizen ENG (2016)	domeinen	in orde	61%	
Medische professionals				
Diagnoses (2012)	ziektebeelden	in orde		0,31
Verpleeghuizen VS (2003)	indicatoren	in orde	89%	0,60
Beoordeling leraarprestaties				
Leskwaliteit VS (2012)	onbekend	onbekend	70%	0,54

Bij de interpretatie van de resultaten in de tabel is een aantal variabelen van belang:

- 1) De afstand tot ons eigen werk.
- 2) De onafhankelijkheid van beoordeling.

¹⁶ AlphaPlus Consultancy Ltd (2014). *Standardisation methods, mark schemes, and their impact on marking reliability*, Ofqual/14/5380.

¹⁷ Dhawan, V., & Bramley, T. (2013). Estimation of inter-rater reliability. Cambridge Assessment, Ofqual/13/5260.

¹⁸ Ibidem.

¹⁹ Fowles, D. (2009) How reliable is marking in GCSE English? *English in Education*, 43, 1.

- 3) De samenstelling van de onderzochte populatie in termen van verwachte risico's.
- 4) De complexiteit en het niveau van beoordeling.

Zo verwachten we een lagere overeenstemming bij meer complexe indicatoren dan bij eenduidige, losse indicatoren. Ook zullen (Engelse) inspecteurs van het onderwijs meer verschillen in hun oordeel over 'de kwaliteit van curriculum-implementatie', waarvoor 4 categorieën bestaan, dan (Amerikaanse) inspecteurs van voedselveiligheid bij hun oordeel over de 'temperatuur van de vleeskoeling', die alleen acceptabel of onacceptabel kan zijn.

Drie veldexperimenten waarin Nederlandse en Engelse onderwijsinspecteurs de kwaliteit van scholen beoordeelden (de eerste 3 onderzoeken in Tabel 1.1), laten een overeenstemming zien van 92% tot 95%²⁰. Echter, er bestaan bij deze onderzoeken twijfels over de onafhankelijkheid van beoordeling. Zie Bijlage II van het Technisch Rapport voor een uitgebreide bespreking. Een soortgelijk onderzoek onder inspecteurs van het Amerikaanse equivalent van de Voedsel- en Warenautoriteit, waar onderlinge beïnvloeding sterk voor de hand ligt, toont een vergelijkbare uitkomst: 96%²¹. Vermoedelijk lag de werkelijke mate van overeenstemming lager. In lijn met deze veronderstelling ziet de Engelse onderwijsinspectie een lagere mate van overeenstemming tussen onderwijsinspecteurs die, volledig onafhankelijk van elkaar, de kwaliteit van curriculum-implementatie beoordeelden: $\kappa = 0,38$ tot $0,49$ ²². Het eerdergenoemde veldexperiment van diezelfde toezichthouder resulteerde voor eendoordelen in $\kappa = 0,80$ ²³.

Bovendien werden in de 3 besproken veldexperimenten relatief homogene groepen scholen bezocht. Deze scholen voldeden in de regel (onze eigen onderzoeken uit 2007 en 2011) of als selectiecriteria voor deelname²⁴ aan de basiskwaliteit. In het huidige onderzoek zijn de vignetten voor een aanzienlijk deel gebaseerd op scholen met kwaliteitsrisico's; dit is in lijn met ons huidige risicogerichte toezicht. Dit betekent dat we, in vergelijking met de eerdergenoemde veldexperimenten, meer complexe situaties aan de inspecteurs voorlegden. Het is om bovengenoemde redenen te verwachten dat het percentage overeenkomst in ons onderzoek lager uitvalt.

Onderzoek naar de IBB onder inspecteurs van andere toezichthouders geeft het volgende beeld. De Ugandese geneesmiddelenautoriteit rapporteerde een percentage overeenstemming van 71% over kwaliteitsindicatoren in een veldexperiment bij gezondheidscentra²⁵. Britse inspecteurs van ziekenhuizen voor acute zorg beoordeelden een tiental vignetten en waren het gemiddeld in 61% van de gevallen eens²⁶.

20 Twee onderzoeken waarover de inspectie van het onderwijs in 2007 en 2011 intern verslag deed en een onderzoek van de Engelse onderwijsinspectie: Ofsted (2017). *Do two inspectors inspecting the same school make consistent decisions?* Ofsted research report 170004.

21 Ho, D.E. (2016). Does peer review work? An experiment of experimentalism. *Stanford Law Review*, 69, 1.

22 Ofsted (2019). *Workbook scrutiny*. Ofsted research report 190028.

23 Ofsted (2017). *Do two inspectors inspecting the same school make consistent decisions?* Ofsted research report 170004.

24 Ibidem.

25 Sekayombya, B., Nahamya, D., Garabedian L., Seru, M. and Trap, B. (2019). Inter-rater reliability and validity of good pharmacy practices measures in inspection of public sector health facility pharmacies in Uganda. *Journal of Pharmaceutical Policy and Practice*, 12, 2.

26 Boyd, A., Addicott, R., Robertson, R., Ross, S., and Walshe, K. (2016). Are inspectors' assessments reliable? Ratings of NHS acute hospital trust services in England. *Journal of Health Services Research & Policy*, 22, 1.

Tot slot noemen we 3 onderzoeken naar IBB waarvan de afstand tot ons werk groter is, maar waarvan de bevindingen niettemin interessant zijn. Een meta-analyse uit 2012 toont een 'matige' gemiddelde overeenstemming tussen getrainde medische professionals in de beoordeling van ziektebeelden, $\kappa = 0,31$ ²⁷. Dit onderzoek is relevant omdat het gaat om oordelen (diagnoses) van getrainde professionals (artsen), met bovendien serieuze consequenties (behandeltrajecten). Een grootschalig veldexperiment bij 219 verpleeghuizen in de Verenigde Staten liet 89% overeenstemming ($\kappa = 0,60$) zien tussen speciaal voor het onderzoek getrainde verplegers en praktiserende verplegers bij de beoordeling van 12 kwaliteitsindicatoren²⁸. Tot slot toont een literatuurreview naar de beoordeling van lerarenprestaties een gemiddelde overeenkomst tussen beoordelaars van 70%, beschouwd over 18 studies. En een gemiddelde kappawaarde van 0,54, gebaseerd op 6 studies²⁹.

Samenvattend laten de onderzoeken, waarbij is gekeken naar de beoordeling door inspecteurs, een percentage overeenstemming zien tussen 61% en 95% en kappawaarden tussen 0,38 en 0,80. Waar over de onafhankelijkheid van beoordeling geen twijfel bestaat, betreft dit 61% tot 89% en kappawaarden van 0,38 tot 0,60. Het gaat hier echter om een beperkt aantal studies, met verschillende onderzochte populaties, verschillende onderzoeksdesigns en variërende complexiteit van te beoordelen indicatoren. Daarom is er geen eenduidige 'acceptabele ondergrens' van overeenstemming te bepalen. Wel biedt dit ons een onderbouwd referentiekader om onze resultaten te duiden. In de discussiesectie van dit rapport reflecteren we op onze bevindingen aan de hand van dit referentiekader.

1.4 Leeswijzer

In het vervolg van dit rapport lichten we allereerst toe hoe we de opzet en uitvoering van het vignetonderzoek vormgaven. Daarna beschrijven we in afzonderlijke hoofdstukken de resultaten van het onderzoek in het po, (v)so, vo en mbo. We sluiten het rapport af met een discussie.

In dit rapport beperken we ons tot een weergave van de resultaten die een antwoord geven op de in paragraaf 1.1 beschreven onderzoeksvragen.

In het Technisch Rapport staat een uitgebreide diagnostische analyse. Daarin onderbouwen we de representativiteit en validiteit van de opgehaalde onderzoeksgegevens. Zoals vastgelegd in het pre-analyseplan³⁰ en zoals beargumenteerd in het Technisch Rapport vormt het percentage overeenstemming onze primaire uitkomstmaat. Voor de volledigheid, en om vergelijkingen met ander onderzoek mogelijk te maken, rapporteren we daarnaast 2 maten waarin een toevalscorrectie is toegepast: Fleiss Kappa en AC1³¹.

27 Tuijn S.M., Janssens F.J.G., Robben P.B.M., Van den Bergh H. (2012) Reducing interrater variability and improving health care: A meta-analytic review. *Journal of Evaluation in Clinical Practice*, 18, 887-895.

28 Mor, V. et al. (2003). Inter-rater reliability of nursing home quality indicators in the U.S. *BMC Health Services Research*, 3.

29 Graham, M., Milanowski, A. T., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*. Center for Educator Compensation Reform. Distributed by ERIC Clearinghouse.

30 <https://osf.io/b68ge>

31 In het Technisch Rapport, dat we afzonderlijk publiceren, bespreken we de keuze voor deze twee maten. De meest gebruikte maat voor overeenstemming (met toevalscorrectie) bij twee beoordelaars is Kappa. Deze maat is op een aantal manieren uitgebreid naar gevallen waarin er meer dan twee beoordelaars zijn. De bekendste daarvan is waarschijnlijk Fleiss Kappa en de meest geavanceerde is AC1. De maten verschillen alleen in de manier waarop ze toevallige overeenstemming berekenen. Fleiss Kappa houdt alleen rekening met de kans om een antwoordcategorie per toeval te kiezen, terwijl AC1 daarnaast ook rekening houdt met hoe moeilijk het is om een vignet eenduidig te beoordelen. Als AC1 hoger is dan Fleiss Kappa, dan schrijft AC1 dus minder overeenstemming toe aan toeval en zullen er dus relatief veel makkelijk te beoordelen vignetten zijn geweest.

2 Methode

2.1 Deelnemers

Dit vignetonderzoek is voor elk van de 4 sectoren apart ontwikkeld en vond op verschillende dagen plaats: in het po in schooljaar 2023-2024 en in het (v)so, vo en mbo in schooljaar 2024-2025. Om te bepalen welk deel van de inspecteurs deelnam aan het onderzoek, beschouwden we alleen inspecteurs die in de praktijk kwaliteitsonderzoeken uitvoeren als doelpopulatie. We sloten een aantal van hen uit van deelname. Namelijk de inspecteurs die betrokken waren bij het opstellen van de vignetten.

In het po namen 73 van de 89 inspecteurs deel (82%), in het (v)so 21 van de 25 inspecteurs (84%), in het vo 48 van de 55 inspecteurs (87%) en in het mbo 31 van de 38 inspecteurs (82%). Enkelen waren verhinderd op de genoemde onderzoeksdagen en namen later deel aan een apart voor hen georganiseerde sessie.

2.2 Materiaal

Voor elk van de sectoren stelde een ontwikkelteam van enkele inspecteurs uit de betreffende sector **16 vignetten** op, onder begeleiding van het projectteam. Elk vignet betrof een schoolbeschrijving (voor het vo van een afdeling, en voor het mbo van een opleiding) met een lengte van maximaal 2 A4. Deze vignetten waren gebaseerd op een kwaliteitsonderzoek waarbij de auteur van het vignet de betreffende school, afdeling of opleiding had bezocht. Het eerste deel van elk vignet bestond uit de volgende contextinformatie: 'aanleiding onderzoek', 'kenmerken school, afdeling of opleiding' (oftewel: de toezichthistorie, de leerlingen- of studentenpopulatie, prestatieindicator³² en overige aspecten) en 'bestuurskenmerken'. Het tweede deel van elk vignet bestond uit zo objectief mogelijk geformuleerde observaties voor standaarden uit het onderzoekskader 2021³³. Voor vignetten in het po, (v)so en vo betrof het de volgende zes standaarden:

OP0 – Basisvaardigheden
OP2 – Zicht op ontwikkeling en begeleiding
OP3 – Pedagogisch-didactisch handelen
VS1 – Veiligheid
OR1 – Resultaten
SKA1 – Visie, ambitie en doelen

De vignetten in het mbo beschreven eveneens OP0, OP2, OP3, VS1, OR1 (Studiesucces), *niet* SKA1, en aanvullend:

OP5 – Beroepspraktijkvorming
BA1 – Borging diplomering
BA2 – Afsluiting
SKA2 – Uitvoering en kwaliteitscultuur

³² Het risicomodel waarmee de inspectie een risicosortering van onderwijsinstellingen maakt om gericht bureau-onderzoek te kunnen uitvoeren. Dit model hanteren we in het po, so en vo.

³³ <https://www.onderwijsinspectie.nl/onderwerpen/onderzoekskaders>.

We kozen voor deze standaarden omdat ze in de beslisregels voor het eindoordeel over een school, afdeling of opleiding staan. Dit geldt overigens niet voor OPO. Deze standaard is toegevoegd omdat dit een nieuwe standaard is die we vanaf schooljaar 2025-2026 gaan beoordelen.

De auteurs maakten vignetten van hun **meest recente kwaliteitsonderzoeken**³⁴, zodat het om onderzoeken ging die ze zich nog relatief goed konden herinneren. Daarnaast zorgde deze aanpak voor een representatieve spreiding in kwaliteit, vergelijkbaar met die van onze kwaliteitsonderzoeken (KO's). De 16 vignetten per sector zijn gebaseerd op onderzoeken waarvan de eindoordeelen zo goed mogelijk overeenkwamen met de verdeling van eindoordeelen in het jaar vóór het onderzoek. Zie Tabel 2.1.

Tabel 2.1

Verdeling eindoordeelen in de selectie van onderzoeken voor vignetconstructie en in het jaar voorafgaand aan het vignetonderzoek, uitgesplitst naar sector

	Voldoende	Onvoldoende	Zeer zwak
po selectie vignetten	56%	31%	12%
po kalenderjaar 2023	55%	26%	18%
(v)so selectie vignetten	69%	25%	6%
(v)so schooljaar 2023-2024	78%	18%	4%
vo selectie vignetten	50%	38%	6%
vo schooljaar 2023-2024	64%	34%	2%
mbo selectie vignetten	50%	50%	0%
mbo kalenderjaar 2024	50%	48%	2%

Let op: Door afronding tellen sommige rijen op tot 99% in plaats van 100%. De selectie vignetten in het vo telt op tot 94%, omdat in één geval geen eindoordeel bekend was.

In een pilot zagen we dat inspecteurs ongeveer 8 vignetten in de toegewezen tijd (1 uur) konden beoordelen. Daarom maakten we 4 deels overlappende sets van elk 10 vignetten, voorzien van unieke kleuren: blauw, geel, groen en rood. De toewijzing van de vignetten aan de sets beschrijven we in detail in het Technisch Rapport³⁵.

2.3 Procedure

Het onderzoek duurde tussen 2 uur (voor het po) en 2,5 uur (voor het [v]so, vo en mbo)³⁶ en bestond uit 2 fases: 1) individuele beoordeling en 2) duo-beoordeling. De werkwijze in deze fases lichten we hieronder toe. In de week voorafgaand aan het onderzoek informeerden we deelnemers mondeling en schriftelijk over de doelstelling en het verloop van het onderzoek. Zij kregen onder meer de aanmoediging om, indien van toepassing, materiaal mee te nemen dat zij in de dagelijkse praktijk ook gebruiken bij kwaliteitsonderzoeken. Zoals het onderzoekskader en de handleiding kwaliteitsonderzoek met daarin het afwegingskader/de handreiking per standaard. Tijdens het onderzoek was dit materiaal ook beschikbaar voor inspecteurs die het waren vergeten.

³⁴ Dit waren zowel kwaliteitsonderzoeken (onderzoeken vanwege geconstateerde risico's op de scholen) als steekproefkwaliteitsonderzoeken (onderzoeken op scholen uit een landelijk representatief steekproef).

³⁵ Het Technisch Rapport betreft een uitgebreide weergave van de methode en resultaten van dit onderzoek en vormt een op zichzelf staande publicatie naast het huidige publieksrapport.

³⁶ We startten het onderzoek in het po en optimaliseerde de procedure op enkele vlakken. Zo namen we meer tijd in het (v)so, vo en mbo om tot een groter aantal gescoorde vignetten te komen.

2.3.1 *Fase 1: individuele beoordeling*

Na een mondelinge instructie (die identiek was voor alle deelnemers en alle sectoren) deelden de proefleiders aan elk van de deelnemers 1 van de 4 geprinte sets met **10 vignetten** uit. Deelnemers kregen de opdracht om, zonder overleg, de standaarden te beoordelen en het eindoordeel te bepalen voor zoveel mogelijk vignetten. Op een antwoordformulier vulden zij per vignet deze oordelen in. Net als in de praktijk kon een standaard de volgende oordelen krijgen: Goed, Voldoende, Voldoende met herstelopdracht, Onvoldoende of Niet te beoordelen (alleen bij OR1³⁷). Bij het eindoordeel waren er in het po, (v)so en vo 3 opties: Voldoende, Onvoldoende of Zeer Zwak. In het mbo waren er 2 aanvullende opties voor het eindoordeel: Goed en Onvoldoende met risico op bekostigings sanctie. Op het antwoordformulier was bovendien ruimte voor het achterlaten van enkele persoonsgegevens voor nadere analyse. Op vrijwillige basis en na schriftelijk informeren over het doel van gegevensverzameling noteerden deelnemers hun ervaring als inspecteur (bijvoorbeeld in het po: indiensttreding vóór 2017, tussen januari 2017 en augustus 2023, of na augustus 2023)³⁸. In het po vroegen we ook naar de standplaats (kantoor)³⁹. Deze persoonsgegevens zijn volledig geanonimiseerd in de analyses. Na ongeveer 60 minuten (po) dan wel 70 minuten ([v]so, vo en mbo) liep Fase 1 af en volgde een korte instructie voor Fase 2.

2.3.2 *Fase 2: duo-beoordeling*

Inspecteurs vormden duo's op basis van de kleur van de set vignetten. In het po kregen inspecteurs in dienst ná augustus 2023 de opdracht om onderling duo's te vormen. Naderhand gaven zij de feedback dat ze liever toehoorder waren geweest bij meer ervaren duo's, om daarvan te leren. In de andere 3 sectoren lieten we, vanwege deze ervaring, nieuwe inspecteurs direct als toehoorder aansluiten bij andere duo's.

Deelnemers kregen de opdracht om in deze fase te komen tot gezamenlijke oordelen. Zo moesten zij allereerst tot **consensus** (overeenstemming) zien te komen over de vignetten waarop hun oordelen verschilden – zowel ten aanzien van de onderliggende standaard(en) als over de eindoordeelen. De resterende tijd konden zij besteden aan de vignetten waarvoor het eindoordeel overeenkwam, maar waarbij zij van mening verschilden over oordelen op onderliggende standaard(en). Deelnemers noteerden de gezamenlijke oordelen op een tweede antwoordformulier. Zij kregen daarnaast de opdracht om voor standaarden waar hun oordeel niet overeenkwam, samen na te gaan waarom dit het geval was. Dit registreerden zij aan de hand van een lijst met voorgestructureerde **verklaringen**. Zij hadden in deze fase ongeveer 45 minuten (voor het po) dan wel 75 minuten (voor het [v]so, vo of mbo).

2.4 **Vorbereiding analyses**

Bij enkele deelnemers ontbraken oordelen op 1 of meer standaarden, of was het genoteerde oordeel niet eenduidig. Bijvoorbeeld 'Onvoldoende/Voldoende'. In die gevallen was geen sprake van een samenhangend eindoordeel en werd voor die

37 Alleen bij de standaard OR1 komt het voor dat geen oordeel kan worden gegeven, zie bv. de Regeling leerresultaten po: <https://wetten.overheid.nl/BWBR0043066/2023-10-05> (geldig tot 31-7-2024).

38 De grens 2017 correspondeert grofweg met het van kracht zijn van het vigerende toezichtkader (2021) en het daar sterk op lijkende voorgaande toezichtkader (2017). Deelnemers in dienst na augustus 2023 waren nog in opleiding en/of hadden beperkte ervaring met kwaliteitsonderzoeken.

39 Dit geldt alleen voor het onderzoek in po. Alleen po-inspecteurs zijn verbonden aan een specifiek kantoor (Zwolle, Utrecht of Tilburg) en werken meer samen met collega's binnen hetzelfde kantoor dan met collega's verbonden aan een ander kantoor.

deelnemers het eindoordeel voor het betreffende vignet uitgesloten van verdere analyses, zoals vooraf vastgelegd⁴⁰.

In alle analyses, behalve enkele exploratieve analyses, pasten we 4 **aanpassingen** op de data toe:

1. De data van de deelnemers die pas kort in dienst waren (bijvoorbeeld in het po: in dienst na augustus 2023) namen we niet mee in de hoofdanalyses. Deze groep was ten tijde van het onderzoek nog in opleiding en/of had beperkte ervaring met kwaliteitsonderzoeken.
2. De optie 'Voldoende met herstelopdracht' voegden we voor de hoofdanalyses samen met de optie 'Voldoende'. In de beslisregels voor het eindoordeel bestaat dit onderscheid namelijk niet.
3. Oordelen op de standaard OPO namen we niet mee in de hoofdanalyses, omdat dit een nieuwe standaard⁴¹ is, die we tijdens het vignetonderzoek nog niet in de praktijk beoordeelden. Bovendien had dit oordeel geen invloed op het eindoordeel.
4. Tot slot namen we de oordelen op het negende en tiende vignet in elke set niet mee in de hoofdanalyses. Hiermee voorkomen we een oververtegenwoordiging van oordelen van 'snelle inspecteurs', die mogelijk meer of minder overeenkomen met die van ander inspecteurs. Juist deze oordelen kunnen de resultaten kleuren.

40 Zie het pre-analyseplan: <https://osf.io/b68qe>

41 OPO werd tijdens het vignetonderzoek nog niet beoordeeld; er werden al wel herstelopdrachten gegeven.

3 Resultaten primair onderwijs

In dit hoofdstuk bespreken we de resultaten van het vignetonderzoek in het primair onderwijs (po) en geven we antwoord op de onderzoeksvragen zoals geformuleerd in paragraaf 1.1. In het po namen 73 van de 89 inspecteurs (82%) deel aan het onderzoek.

3.1 Hoofdanalyse primair onderwijs

- Hoe groot is de overeenstemming tussen inspecteurs in het po voor eindoordelen en oordelen op standaarden bij kwaliteitsonderzoeken op scholen?

Voor de beantwoording van deze vraag kijken we in de eerste plaats naar de overeenstemming tussen individuele inspecteurs over de eindoordelen.

Tabel 3.1

Percentage overeenstemming tussen individuele inspecteurs

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	65%	56% tot 73%
Standaardoordelen	82%	78% tot 85%

Tabel 3.1 laat zien dat de mate van overeenstemming tussen individuele inspecteurs lager uitvalt voor de eindoordelen dan voor oordelen op de standaarden. Daarnaast is, zoals verwacht, voor oordelen op de standaarden de schatting het nauwkeurigst: het betrouwbaarheidsinterval is hier het kleinst⁴².

Tabel 3.2

Percentage overeenstemming tussen inspecteursduo's

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	76%	63% tot 88%
Standaardoordelen	89%	84% tot 93%

Tabel 3.2 geeft de overeenstemming tussen inspecteursduo's weer. In Fase 2 kregen deelnemers de opdracht om tot gezamenlijke oordelen te komen voor vignetten en standaarden waarover zij het in de individuele fase oneens waren.

We konden een klein deel van de oordelen niet meenemen in de berekening. Vanwege de tijdslimiet werd namelijk een deel van de vignetten waarover de duo-partners het in de individuele fase oneens waren, niet besproken. Dit betrof respectievelijk 6% en 7% van het totaal aantal oordelen op standaarden en eindoordelen. Het is waarschijnlijk dat wanneer deze situaties wel waren besproken

⁴² De vignetten weerspiegelen een willekeurige steekproef van scholen. De mate van overeenstemming over de kwaliteit van die steekproef is daarmee een *schatting* van de overeenstemming over de volledige groep scholen die we bezoeken. Die schatting gaat gepaard met een onzekerheid. Het 95%-betrouwbaarheidsinterval betekent dat er 95% kans is dat het werkelijke populatiegemiddelde binnen de getoonde grenzen valt.

de totale overeenstemming tussen duo's zou afnemen. Het feit dat individuele inspecteurs het over deze situaties oneens waren, maakt het waarschijnlijk dat dit om relatief complexe situaties ging. Omdat dit deel van de gezamenlijke oordelen ontbreekt, schatten we in Tabel 3.2 dus een bovengrens aan overeenstemming tussen de inspecteursduo's.

Net als bij de individuele oordelen zien we dat de overeenstemming tussen inspecteursduo's voor de eindoordelen lager is dan voor de oordelen op de standaarden. Ook hier zien we, zoals verwacht, een meer nauwkeurige schatting van de overeenstemming over standaardoordelen. Verder suggereert Tabel 3.2 in vergelijking met Tabel 3.1 dat verschillende inspecteursduo's het vaker met elkaar eens zijn dan individuele inspecteurs onderling. Die statistische vergelijking behandelen we bij de volgende onderzoeksvraag.

Tabel 3.3 laat zien in hoeverre individuele inspecteurs tot hetzelfde oordeel komen, met een correctie voor de kans op toeval. Dit gebeurt aan de hand van 2 aanvullende maten.

Tabel 3.3

Mate van overeenstemming tussen inspecteurs, na correctie voor toeval

Type oordelen	Maat	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	Fleiss Kappa	0,44	0,31–0,57
	AC1	0,48	0,35–0,62
Standaardoordelen	Fleiss Kappa	0,60	0,52–0,68
	AC1	0,78	0,73–0,83

3.2

Secundaire analyses primair onderwijs

- Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's?

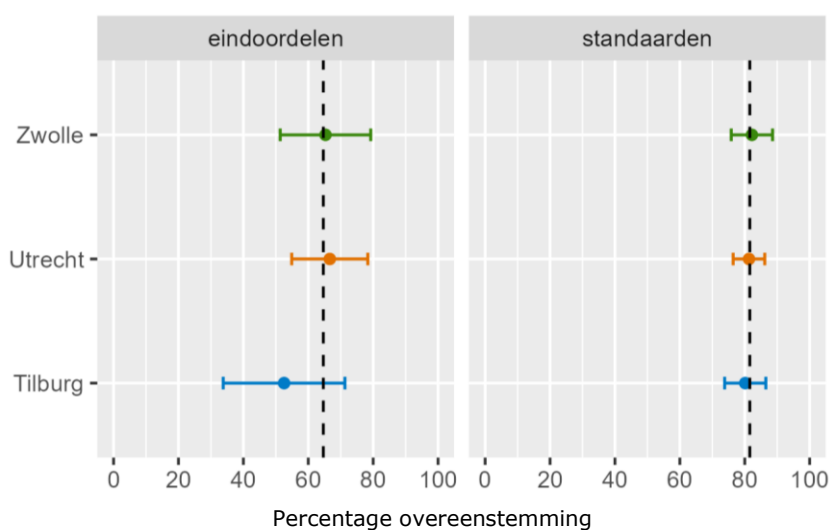
Een vergelijking tussen individuele inspecteurs en duo's vereist dat we hetzelfde deel van de oordelen weglaten in de berekening van het percentage overeenstemming tussen individuele inspecteurs. Namelijk: het deel dat de duo's niet bespraken. Als we dit doen, zien we een hogere mate van overeenstemming tussen individuele inspecteurs voor zowel eindoordelen (67%) als oordelen op standaarden (85%). De waarden voor inspecteursduo's zoals weergegeven in Tabel 3.2 liggen respectievelijk (afgerond) 8 procentpunt en 4 procentpunt hoger.

Het verschil dat we zien tussen individuele inspecteurs en duo's voor de eindoordelen is echter niet statistisch significant: $t(15) = 1,08, p = 0,30$. Dat geldt ook voor oordelen op de standaarden, $t(79) = 1,49, p = 0,14$. Met andere woorden: we kunnen niet uitsluiten dat de verschillen op toeval berusten. Dit onderzoek geeft geen bewijs voor de stelling dat overleg in duo's in het po leidt tot meer betrouwbare oordelen. Een kanttekening hierbij is dat het negeren van een (klein) deel van de vignetten – waarbij oordelen verschilden – waarschijnlijk zorgt voor een lichte onderschatting van de toename in overeenstemming. Bovendien werden vergissingen van individuele inspecteurs bijna altijd gecorrigeerd in de duo-fase. Dat ging bijvoorbeeld om het beoordelen van OR1 en het toepassen van de beslisregels voor het eindoordeel.

- Is de overeenstemming groter binnen kantoren en binnen cohorten inspecteurs?

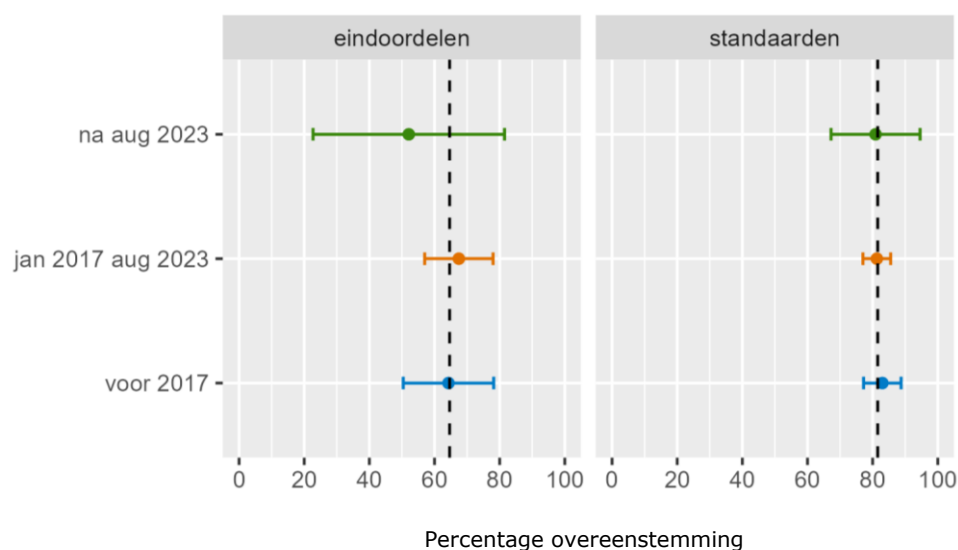
Hier gaan we na of verschillen in oordelen samenhangen met het kantoor⁴³ waaraan inspecteurs verbonden zijn. Of met de mate van ervaring van inspecteurs (indiensttreding vóór 2017, tussen 2017 en augustus 2023, of na augustus 2023). Het jaartal 2017 is gekozen, omdat het onderzoekskader in dat jaar sterk is gewijzigd ten opzichte van eerdere jaren. Voor inspecteurs die na augustus 2023 in dienst zijn gekomen, geldt dat zij meestal nog in opleiding waren. En daarom minder ervaring hebben met het uitvoeren van kwaliteitsonderzoeken. Als de overeenstemming binnen bepaalde groepen hoger is in vergelijking met de totale groep deelnemers, dan is het aannemelijk dat die groepen onderling verschillen in de wijze van beoordeling. We vergeleken dus telkens de subgroep (het kantoor of het cohort) met de totale groep inspecteurs. Hierbij keken we alleen naar de individuele oordelen, omdat er te weinig duo-oordelen zijn om noemenswaardige afwijkingen per groep vast te stellen.

Figuur 3.1 toont het percentage overeenstemming voor elk van de 3 kantoren en, als referentie, de gemiddelde overeenstemming tussen alle inspecteurs aan de hand van de stippellijn. Figuur 3.2 toont een uitsplitsing naar de 3 ervaringscohorten, opnieuw in vergelijking tot de totale groep inspecteurs – uitgebeeld met de stippellijn. Ter informatie nemen we hier ook de groep inspecteurs mee die na augustus 2023 zijn begonnen. We zien geen sterke afwijkingen in de mate van overeenstemming binnen de kantoren of binnen de ervaringscohorten, ten opzichte van de totale groep inspecteurs. Bij kantoor Tilburg en bij de groep inspecteurs in dienst na augustus 2023, zien we meer onzekerheid in de schatting van het percentage overeenstemming. Namelijk, grotere betrouwbaarheidsintervallen. Dit komt doordat in deze 2 groepen relatief weinig inspecteurs zitten. Desondanks overlappen de foutenbalken met het gemiddelde van de totale groep inspecteurs. Het is dus niet aannemelijk dat de standplaats of ervaring van inspecteurs een verklaring vormen voor de variatie in oordelen.



Figuur 3.1. Percentage overeenstemming uitgesplitst naar de 3 kantoren en type oordelen, inclusief 95%-betrouwbaarheidsintervallen.

⁴³ Po-inspecteurs zijn verbonden aan een specifiek kantoor (Zwolle, Utrecht of Tilburg) en werken meer samen met collega's binnen hetzelfde kantoor dan met collega's verbonden aan een ander kantoor.



Figuur 3.2. Percentage overeenstemming uitgesplitst naar de 3 ervaringscohorten en type oordelen, inclusief 95%-betrouwbaarheidsintervallen.

- Welke redenen noemen inspecteursduo's voor verschillen in individuele oordelen?

Tabel 3.4 laat zien hoe vaak zij elke verklaring noemden. OPO is hierbij buiten beschouwing gelaten. In Bijlage 1 bij dit rapport splitsen we deze gegevens uit naar de afzonderlijke standaarden. Verreweg het vaakst noemden deelnemers dat zij elementen uit het afwegingskader verschillend hadden gewogen. We zien dat dit de meest genoemde reden was voor het verschillend beoordelen van de standaarden OP3, OP2, SKA1 en VS1. Bij OR1 was de meest genoemde reden het anders toepassen van de beslisregel⁴⁴. Inspecteurs noemen het verschillend interpreteren van het afwegingskader vooral bij OR1 en SKA1. Verder valt op dat deelnemers de contextinformatie van de school relatief vaak verschillend meewogen in het oordeel op SKA1. En dat zij met name bij het beoordelen van OP3 en SKA1 het oordeel op een andere standaard verschillend meewogen (het zogeheten 'doortikeffect'). Het al dan niet beredeneerd oordelen met het oog op het effect van toezicht (de gevolgen voor het eindoordeel) noemden zij vooral bij SKA1. Tot slot, de open antwoorden bij de categorie 'anders' varieerden van vergissingen tot het verschillend interpreteren van de vignettekst ($n = 6$). Bij vergissingen ging het bijvoorbeeld om het verkeerd toepassen van de beslisregel voor het eindoordeel, of het verkeerd noteren van een oordeel. Een enkeling merkte op dat belangrijke informatie ontbrak in het vignet, waardoor oordelen niet overeenkwamen ($n = 3$).

In Bijlage 1 zijn ook de verklaringen opgenomen die inspecteurs gaven voor verschillen in oordelen bij de standaard OPO. Zij noemden vooral het wegen van de elementen uit het afwegingskader en de interpretatie van het afwegingskader.

⁴⁴ In het po werd relatief veel vergissingen gemaakt als OR1 Resultaten niet kon worden beoordeeld. Tijdens de duo-fase werden deze vergissingen grotendeels gecorrigeerd. In het schooljaar 2023-2024 waarin het vignetonderzoek is uitgevoerd, werden naast de signaleringswaarden ook correctiewaarden gehanteerd bij de beoordeling van OR1. In het schooljaar 2024-2025 zijn de correctiewaarden vervallen en is de beslisregel vereenvoudigd, zodat dit nu minder vaak zal voorkomen.

Tabel 3.4

Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OPO Basisvaardigheden)

Reden voor afwijkend oordeel	Aantal keer (%) ^a
Elementen uit het afwegingskader verschillend gewogen	83 (37%)
Afwegingskader verschillend geïnterpreteerd	34 (15%)
Contextinformatie school anders gewogen	18 (8%)
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	15 (7%)
Beslisregel OR1 anders toegepast	13 (6%)
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school')	12 (5%)
Informatie in het vignet over het hoofd gezien	11 (5%)
Handleiding met afwegingskader wel/niet gebruikt	9 (4%)
Kenmerken leerlingenpopulatie anders gewogen	4 (2%)
Contextinformatie bestuur anders gewogen	1 (<1%)
Toezichthistorie anders gewogen	1 (<1%)
Anders	25 (11%)

^a) Let op: Niet altijd werd een reden opgegeven. De percentages reflecteren dus het aandeel van het totaal aantal opgegeven redenen, niet van het totaal aantal beoordeelde standaarden in de duo-fase.

3.3

Exploratieve analyses primair onderwijs

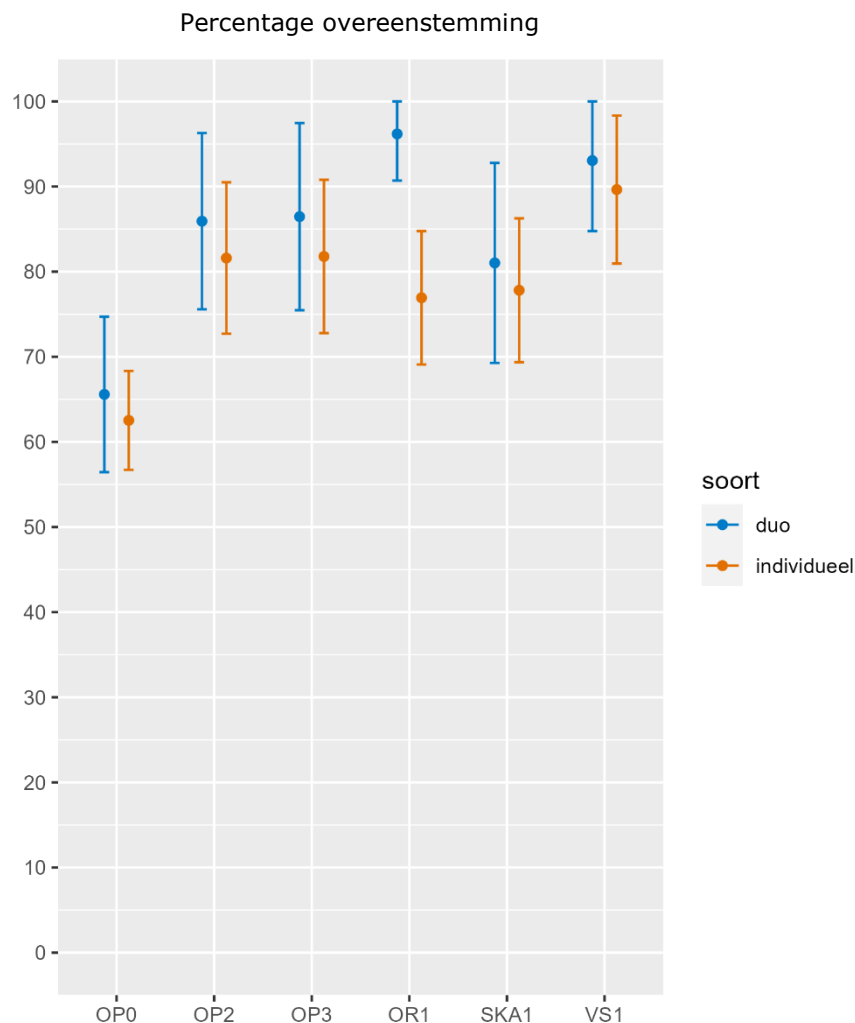
- Hoe vaak krijgen scholen hetzelfde eindoordeel?

Voor individuele inspecteurs blijkt dat gemiddeld 76% het meerderheidsoordeel geeft (95%-betrouwbaarheidsinterval: 72% tot 80%). Voor duo's blijkt dat gemiddeld 83% het meest gekozen oordeel geeft (95%-betrouwbaarheidsinterval: 78% tot 88%).

- Hoe groot zijn de verschillen in overeenstemming tussen standaarden?⁴⁵

Figuur 3.3 geeft de verschillen in overeenstemming per standaard weer. Opvallend is de lagere overeenstemming over oordelen voor OPO. Zoals gezegd is dit een nieuwe standaard die inspecteurs in de praktijk nog niet beoordeelden en waarvoor de scholing tijdens dit onderzoek nog volop in gang was. Verder valt op dat de oordelen van individuele inspecteurs over de standaard OR1 Resultaten weinig overeenkomen, in vergelijking met de andere standaarden. Terwijl de beslisregels heel eenduidig zijn. Relatief veel inspecteurs vergisten zich bij het correct toepassen van deze beslisregels. In de duo-fase corrigeerden zij deze vergissingen grotendeels en nam de overeenstemming sterk toe.

⁴⁵ In de hier gerapporteerde analyse laten we zoals besproken de optie 'Voldoende met herstelopdracht' buiten beschouwing. In het Technisch Rapport rapporteren we de analyse ook met deze vierde optie.



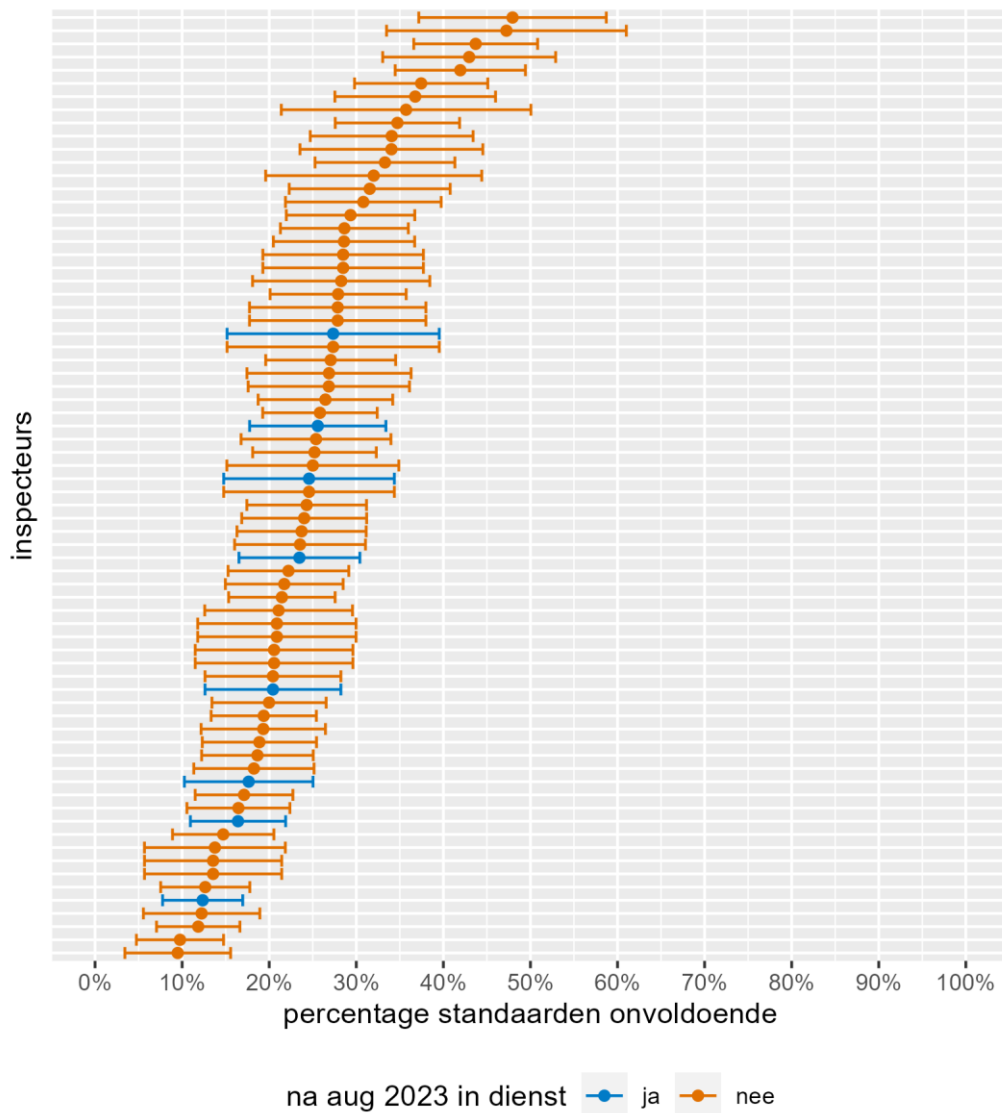
Figuur 3.3. Percentage overeenstemming over individuele oordelen en duo-oordelen, uitgesplitst naar standaard, inclusief 95%-betrouwbaarheidsintervallen.

- Hoe groot zijn de verschillen in strengheid tussen inspecteurs?

We betrokken bij deze analyse ook de oordelen op OP0 en de oordelen op de eventueel afgeronde negende en tiende vignetten. Dat deden we om de hoeveelheid gegevens te maximaliseren. Ook de oordelen van inspecteurs in dienst na augustus 2023 betrokken we in deze analyse.

Omdat niet elke inspecteur dezelfde vignetten beoordeelde, en de kans op een oordeel Onvoldoende afhangt van het vignet, pasten we een regressiemodel toe. Met dit model schatten we het verwachte oordeel voor elke standaard per inspecteur. En daarmee de verwachte strengheid voor alle 96 standaarden in de vignetten. Uiteindelijk schatten we hiermee per inspecteur de gemiddelde kans op een Onvoldoende oordeel op een willekeurige standaard.

Figuur 3.4 laat zien dat de kans op een oordeel Onvoldoende bij de meest strenge inspecteur flink verschilt van de kans hierop bij de minst strenge inspecteur. Ook wanneer we rekening houden met de betrouwbaarheidsintervallen. Verder laat de figuur zien dat de meeste inspecteurs qua strengheid niet sterk van elkaar verschillen. Met name aan de bovenkant van de figuur zien we een kleine groep relatief strenge inspecteurs.



Figuur 3.4. Gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard, uitgebeeld voor elke deelnemer van hoog naar laag. De 95%-betrouwbaarheidsintervallen variëren in grootte, afhankelijk van het aantal gescoorde vignetten.

4 Resultaten (voortgezet) speciaal onderwijs

In dit hoofdstuk bespreken we de resultaten van het vignetonderzoek in het (voortgezet) speciaal onderwijs (v)so en geven we antwoord op de onderzoeksvragen zoals geformuleerd in paragraaf 1.1. In het (v)so namen 21 van de 25 inspecteurs (84%) deel aan het onderzoek.

4.1 Hoofdanalyse (voortgezet) speciaal onderwijs

- Hoe groot is de overeenstemming tussen inspecteurs in het (v)so voor eindoordelen en oordelen op standaarden bij kwaliteitsonderzoeken op scholen?

Voor de beantwoording van deze vraag kijken we in de eerste plaats naar de overeenstemming tussen individuele inspecteurs over de eindoordelen.

Tabel 4.1

Percentage overeenstemming tussen individuele inspecteurs

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	72%	59% tot 85%
Standaardoordelen	85%	80% tot 89%

Tabel 4.1 laat zien dat de mate van overeenstemming tussen individuele inspecteurs lager uitvalt voor de eindoordelen dan voor oordelen op de standaarden. Daarnaast is, zoals verwacht, voor oordelen op de standaarden de schatting het nauwkeurigst: het betrouwbaarheidsinterval is hier het kleinst⁴⁶.

Tabel 4.2

Percentage overeenstemming tussen inspecteursduo's

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	77%	53% tot 100%
Standaardoordelen	89%	80% tot 98%

Tabel 4.2 laat de overeenstemming zien tussen inspecteursduo's. We konden een klein deel van de oordelen niet meenemen in de berekening. Een deel van de vignetten waarover de duo-partners het in de individuele fase oneens waren, werd namelijk niet besproken. Dit betrof respectievelijk 8% en 6% van het totaal aantal oordelen op standaarden en eindoordelen. Het is waarschijnlijk dat, wanneer deze situaties wel waren besproken, de totale overeenstemming tussen duo's zou afnemen. Het feit dat individuele inspecteurs het over deze situaties oneens waren, maakt het waarschijnlijk dat dit om relatief complexe situaties ging. Omdat dit deel van de gezamenlijke oordelen ontbreekt, schatten we in Tabel 4.2 dus een hogere

⁴⁶ De vignetten weerspiegelen een willekeurige steekproef van scholen. De mate van overeenstemming over de kwaliteit van die steekproef is daarmee een *schatting* van de overeenstemming over de volledige groep scholen die we bezoeken. Die schatting gaat gepaard met een onzekerheid. Het 95%-betrouwbaarheidsinterval betekent dat er 95% kans is dat het werkelijke populatiegemiddelde binnen de getoonde grenzen valt.

bovengrens aan overeenstemming tussen de inspecteursduo's. We zien opnieuw dat de overeenstemming tussen inspecteursduo's voor de eindoordelen lager is dan voor de oordelen op de standaarden. Ook hier zien we een meer nauwkeurige schatting van de overeenstemming over standaardoordelen. Verder suggereert Tabel 4.2 in vergelijking met Tabel 4.1 dat inspecteursduo's het onderling vaker eens zijn dan individuele inspecteurs onderling. Die statistische vergelijking behandelen we bij de volgende onderzoeksvraag.

Tabel 4.3 laat zien in hoeverre individuele inspecteurs tot hetzelfde oordeel komen, met een correctie voor de kans op toeval. Dit gebeurt aan de hand van 2 aanvullende maten.

Tabel 4.3

Mate van overeenstemming tussen inspecteurs, na correctie voor toeval

Type oordelen	Maat	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	Fleiss Kappa	0,49	0,30–0,68
	AC1	0,61	0,41–0,82
Standaardoordelen	Fleiss Kappa	0,52	0,38–0,66
	AC1	0,82	0,76–0,88

4.2

Secundaire analyses (voortgezet) speciaal onderwijs

- Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's?

Een vergelijking tussen individuele inspecteurs (Tabel 4.1) en duo's (Tabel 4.2) vereist dat we hetzelfde deel van de oordelen weglaten in de berekening van het percentage overeenstemming tussen individuele inspecteurs. Namelijk: het deel dat de duo's niet bespraken. Als we dit doen, zien we een hogere mate van overeenstemming tussen individuele inspecteurs voor zowel eindoordelen (75%) als oordelen op standaarden (89%). De waarde voor inspecteursduo's ligt (afgerond) 2 procentpunt hoger voor eindoordelen, zoals te zien is in Tabel 4.2. Voor standaardoordelen is de waarde gelijk.

Het verschil dat te zien is tussen individuele inspecteurs en duo's voor de eindoordelen, is echter niet statistisch significant: $t < 1$ (idem voor standaardoordelen, $t < 1$). Met andere woorden: we kunnen niet uitsluiten dat de verschillen op toeval berusten. Dit onderzoek geeft geen bewijs voor de stelling dat overleg in duo's in het (v)so leidt tot meer betrouwbare oordelen. Een kanttekening hierbij is dat het negeren van een (klein) deel van de vignetten – waarbij oordelen verschilden – waarschijnlijk zorgt voor een lichte onderschatting van de toename in overeenstemming. Bovendien werden vergissingen van individuele inspecteurs, bijvoorbeeld bij het toepassen van de beslisregels voor het eindoordeel, bijna altijd gecorrigeerd in de duo-fase.

- Welke redenen noemen inspecteursduo's voor verschillen in individuele oordelen?

Tabel 4.4 laat zien hoe vaak zij elke verklaring noemden. OP0 is hierbij buiten beschouwing gelaten. In Bijlage 2 bij dit rapport splitsen we deze gegevens uit naar de afzonderlijke standaarden. Verreweg het vaakst noemden deelnemers dat zij elementen uit het afwegingskader verschillend hadden gewogen. We zien dat dit de meest genoemde reden was voor het verschillend beoordelen van de standaarden OP2, SKA1 en VS1. Bij OR1 noemden inspecteurs relatief vaak dat zij informatie in het vignet over het hoofd hadden gezien. Tot slot bleek uit de open antwoorden bij de categorie 'anders' dat 2 deelnemers niet hadden gezien dat het oordeel Goed een optie was.

In Bijlage 2 zijn ook de verklaringen opgenomen die inspecteurs gaven voor verschillen in oordelen bij de standaard OP0. Daarbij noemden zij vooral de interpretatie van het afwegingskader en het wege van de elementen uit het afwegingskader.

Tabel 4.4

Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OP0 Basisvaardigheden)

Reden voor afwijkend oordeel	Aantal keer (%) ^a
Elementen uit het afwegingskader verschillend gewogen	18 (37%)
Afwegingskader verschillend geïnterpreteerd	8 (16%)
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	5 (10%)
Informatie in het vignet over het hoofd gezien	4 (8%)
Handleiding met afwegingskader wel/niet gebruikt	3 (6%)
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school')	2 (4%)
Contextinformatie school anders gewogen	2 (4%)
Kenmerken leerlingenpopulatie anders gewogen	1 (2%)
Contextinformatie bestuur anders gewogen	0
Toezichthistorie anders gewogen	0
Anders	6 (12%)

^a) Let op: Niet altijd werd een reden opgegeven. De percentages reflecteren dus het aandeel van het totaal aantal opgegeven redenen, niet van het totaal aantal beoordeelde standaarden in de duo-fase. In het (v)so werd de optie 'Beslisregel OR1 anders toegepast' niet voorgelegd, omdat richtlijnen voor de beoordeling van OR1 niet in een beslisregel zijn vastgelegd. In de andere sectoren legden we deze optie wel voor.

4.3

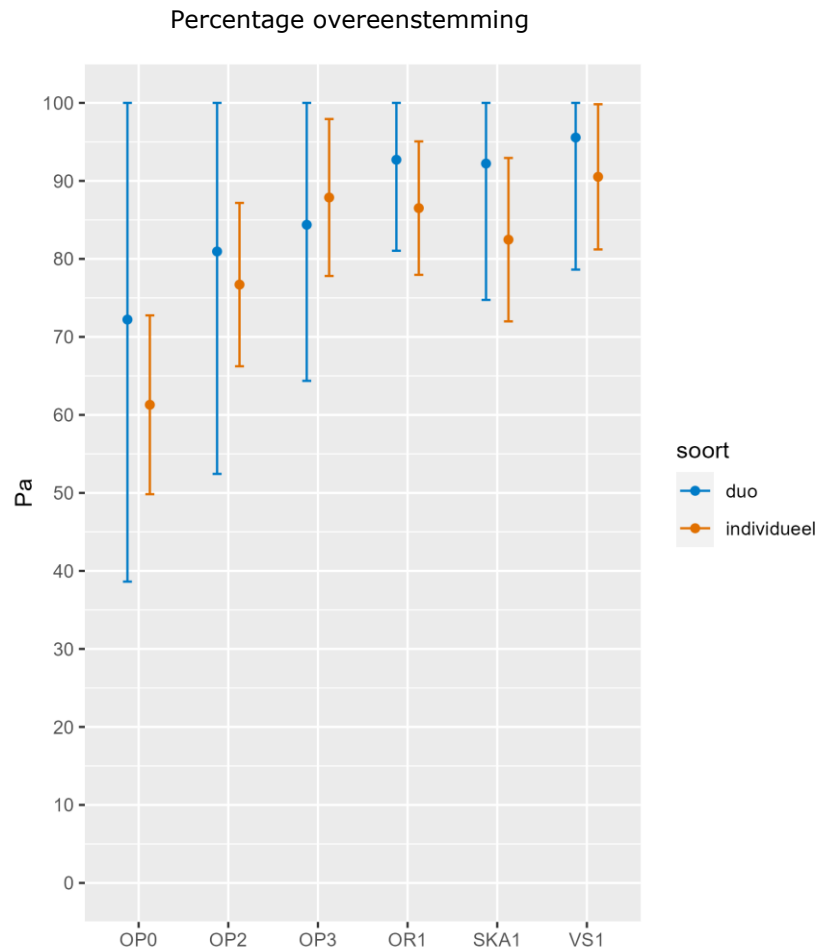
Exploratieve analyses (voortgezet) speciaal onderwijs

- Hoe vaak krijgen scholen hetzelfde eindoordeel?

Voor individuele inspecteurs blijkt dat gemiddeld 80% het meerderheidsoordeel geeft (95%-betrouwbaarheidsinterval: 75% tot 85%). Voor duo's blijkt dat gemiddeld 89% het meest gekozen oordeel geeft (95%-betrouwbaarheidsinterval: 84% tot 94%).

- Hoe groot zijn de verschillen in overeenstemming tussen standaarden?⁴⁷

Figuur 4.2 geeft de verschillen in overeenstemming per standaard weer. De lagere mate van overeenstemming over oordelen voor OP0 valt op. Zoals gezegd is dit een nieuwe standaard die inspecteurs in de praktijk nog niet beoordeelden en waarvoor de scholing tijdens dit onderzoek nog volop in gang was.



Figuur 4.2. Percentage overeenstemming over individuele oordelen en duo-oordelen, uitgesplitst naar standaard, inclusief 95%-betrouwbaarheidsintervallen.

- Hoe groot zijn de verschillen in strengheid tussen inspecteurs?

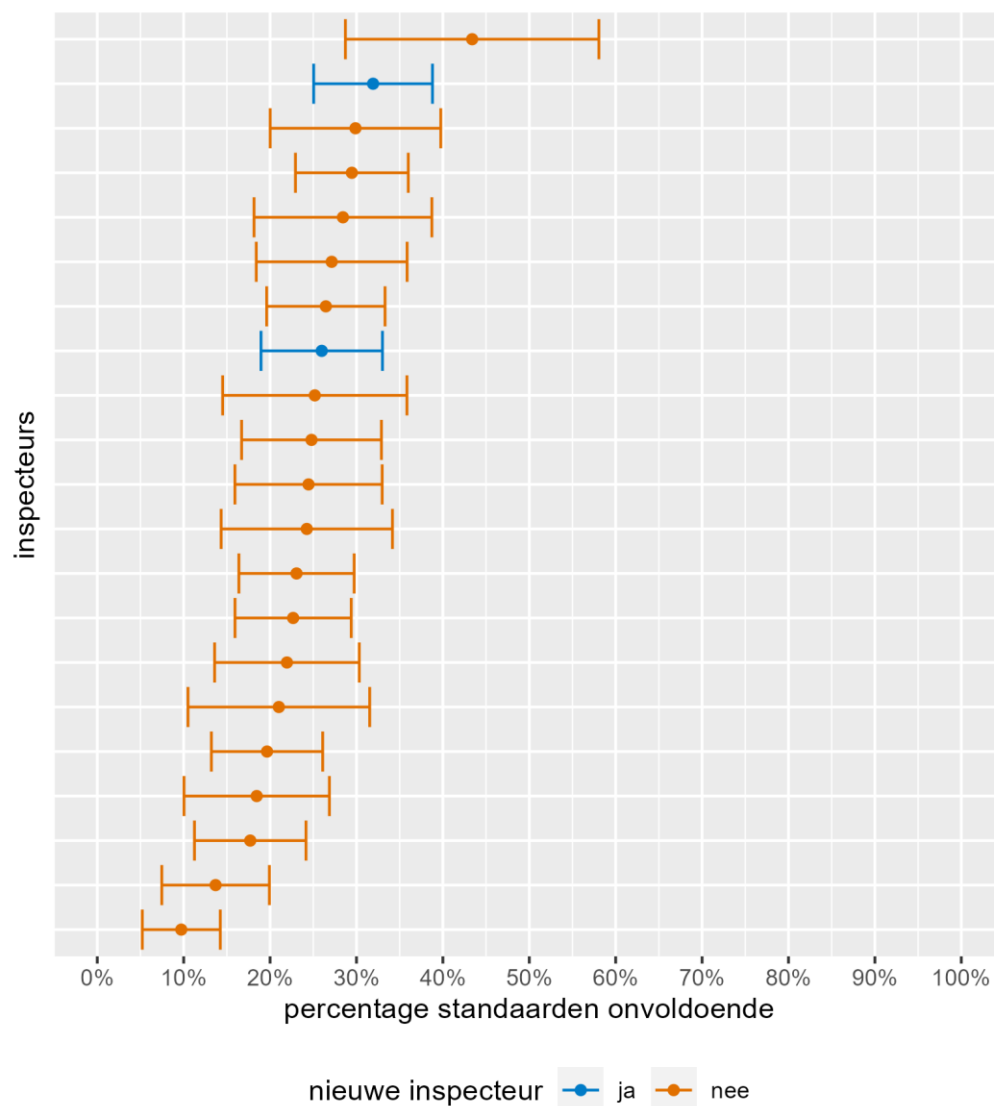
We betrokken bij deze analyse ook de oordelen op OP0 en de oordelen op de eventueel afgeronde negende en tiende vignetten. Dit deden we om de hoeveelheid gegevens te maximaliseren. Ook de oordelen van inspecteurs in dienst na augustus 2024 betrokken we in deze analyse.

Omdat niet elke inspecteur dezelfde vignetten beoordeelde, en de kans op een oordeel Onvoldoende afhangt van het vignet, pasten we een regressiemodel toe.

⁴⁷ In de hier gerapporteerde analyse laten we zoals besproken de optie 'Voldoende met herstelopdracht' buiten beschouwing. In het Technisch Rapport rapporteren we de analyse ook met deze vierde optie.

Met dit model schatten we het verwachte oordeel voor elke standaard per inspecteur. En daarmee de verwachte strengheid voor alle 96 standaarden in de vignetten. Uiteindelijk schatten we hiermee per inspecteur de gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard.

Figuur 4.3 laat zien dat de kans op een oordeel Onvoldoende bij de meest strenge inspecteur flink verschilt van de kans hierop bij de minst strenge inspecteur. Ook wanneer we rekening houden met de betrouwbaarheidsintervallen. Verder laat de figuur zien dat de meeste inspecteurs min of meer even streng oordelen. Aan de bovenkant van de figuur zien we een relatief strenge inspecteur. Maar met een grote mate van onzekerheid in de geschatte waarde, door het beperkt aantal gescoorde vignetten.



Figuur 4.3. Gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard, uitgebeeld voor elke deelnemer van hoog naar laag. De 95%-betrouwbaarheidsintervallen variëren in grootte, afhankelijk van het aantal gescoorde vignetten.

5 Resultaten voortgezet onderwijs

In dit hoofdstuk bespreken we de resultaten van het vignetonderzoek in het voortgezet onderwijs (vo) en geven we antwoord op de onderzoeksvragen zoals geformuleerd in paragraaf 1.1. In het vo namen 48 van de 55 inspecteurs (87%) deel aan het onderzoek.

5.1 Hoofdanalyse voortgezet onderwijs

- Hoe groot is de overeenstemming tussen inspecteurs in het vo voor eindoordelen en oordelen op standaarden bij kwaliteitsonderzoeken op afdelingen?

Voor de beantwoording van deze vraag kijken we allereerst naar de overeenstemming tussen individuele inspecteurs over de eindoordelen.

Tabel 5.1

Percentage overeenstemming tussen individuele inspecteurs

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	81%	71% tot 90%
Standaardoordelen	89%	85% tot 92%

Tabel 5.1 laat zien dat de mate van overeenstemming tussen individuele inspecteurs lager uitvalt voor de eindoordelen dan voor oordelen op de standaarden. Daarnaast is, zoals verwacht, voor oordelen op de standaarden de schatting het nauwkeurigst: het betrouwbaarheidsinterval is hier het kleinst⁴⁸.

Tabel 5.2

Percentage overeenstemming tussen inspecteursduo's

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	82%	66% tot 97%
Standaardoordelen	94%	89% tot 99%

Tabel 5.2 laat de overeenstemming zien tussen inspecteursduo's. Een klein deel van de oordelen konden we niet meenemen in de berekening. Een deel van de vignetten waarover de duo-partners het in de individuele fase oneens waren, werd namelijk niet besproken. Dit ging om respectievelijk 5% en 4% van het totale aantal oordelen op standaarden en eindoordelen. Het is waarschijnlijk dat, wanneer deze situaties wel waren besproken, de totale overeenstemming tussen duo's zou afnemen. Het feit dat individuele inspecteurs het over deze situaties oneens waren, maakt het waarschijnlijk dat dit om relatief complexe situaties ging. Omdat dit deel van de

⁴⁸ De vignetten weerspiegelen een willekeurige steekproef van scholen. De mate van overeenstemming over de kwaliteit van die steekproef is daarmee een *schatting* van de overeenstemming over de volledige groep scholen die we bezoeken. Die schatting gaat gepaard met een onzekerheid. Het 95%-betrouwbaarheidsinterval betekent dat er 95% kans is dat het werkelijke populatiegemiddelde binnen de getoonde grenzen valt.

gezamenlijke oordelen ontbreekt, schatten we in Tabel 5.2 dus een bovengrens aan overeenstemming tussen de inspecteursduo's.

Net als bij de individuele oordelen zien we dat de overeenstemming tussen inspecteursduo's voor de eindoordelen lager is dan voor de oordelen op de standaarden. Ook hier zien we, zoals verwacht, een meer nauwkeurige schatting van de overeenstemming over standaardoordelen. Verder suggereert Tabel 5.2 in vergelijking met Tabel 5.1 dat verschillende inspecteursduo's het vaker met elkaar eens zijn dan individuele inspecteurs onderling. Die statistische vergelijking behandelen we bij de volgende onderzoeksvraag.

Tabel 5.3 laat zien in hoeverre individuele inspecteurs tot hetzelfde oordeel komen, met een correctie voor de kans op toeval. Dit gebeurt aan de hand van 2 aanvullende maten.

Tabel 5.3

Mate van overeenstemming tussen inspecteurs, na correctie voor toeval

Type oordelen	Maat	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	Fleiss Kappa	0,66	0,50–0,82
	AC1	0,73	0,59–0,87
Standaardoordelen	Fleiss Kappa	0,69	0,60–0,79
	AC1	0,87	0,83–0,91

5.2

Secundaire analyses voortgezet onderwijs

- Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's?

Een vergelijking tussen individuele inspecteurs (Tabel 5.1) en duo's (Tabel 5.2) vereist dat we hetzelfde deel van de oordelen weglaten in de berekening van het percentage overeenstemming tussen individuele inspecteurs. Namelijk: het deel dat de duo's niet bespraken. Als we dit doen, zien we een hogere mate van overeenstemming tussen individuele inspecteurs voor zowel eindoordelen (82%) als oordelen op standaarden (92%). De waarde voor inspecteursduo's ligt 2 procentpunt hoger voor standaardoordelen, zoals te zien is in Tabel 5.2. Voor eindoordelen is de waarde gelijk.

Het verschil dat te zien is tussen individuele inspecteurs en duo's voor de standaardoordelen, is echter niet statistisch significant: $t < 1$ (idem voor eindoordelen, $t < 1$). Met andere woorden: we kunnen niet uitsluiten dat de verschillen op toeval berusten. Dit onderzoek geeft geen bewijs voor de stelling dat overleg in duo's in het vo leidt tot meer betrouwbare oordelen. Een kanttekening hierbij is dat het negeren van een (klein) deel van de vignetten – waarbij oordelen verschilden – waarschijnlijk zorgt voor een lichte onderschatting van de toename in overeenstemming. Bovendien werden vergissingen van individuele inspecteurs, bijvoorbeeld bij het toepassen van de beslisregels voor het eindoordeel, bijna altijd gecorrigeerd in de duo-fase.

- Welke redenen noemen inspecteursduo's voor verschillen in individuele oordelen?

Tabel 5.4 laat zien hoe vaak zij elke verklaring noemden. OP0 is hierbij buiten beschouwing gelaten. In Bijlage 3 bij dit rapport splitsen we deze gegevens uit naar de afzonderlijke standaarden. Verreweg het vaakst noemden deelnemers dat zij elementen uit het afwegingskader verschillend hadden gewogen. We zien dat dit de meest genoemde reden was voor het verschillend beoordelen van de standaarden OP2, SKA1 en VS1. Bij OR1, VS1 OP2 en SKA1 gaven inspecteurs als reden dat zij informatie in het vignet over het hoofd hadden gezien. Dit was ook het geval bij enkele andere standaarden. Het verschillend interpreteren van het afwegingskader noemden zij vooral bij VS1 en OP2. Bij OP2 noemden inspecteurs het verschillend meewegen van informatie over de leerlingenpopulatie meerdere keren. Tot slot, de open antwoorden bij de categorie 'anders' varieerden van vergissingen ($n = 1$) tot het verschillend interpreteren van de vignettekst ($n = 2$). Een enkeling merkte op dat belangrijke informatie ontbrak in het vignet, waardoor oordelen niet overeenkwamen ($n = 1$).

In Bijlage 3 zijn ook de verklaringen opgenomen die inspecteurs gaven voor verschillen in oordelen bij de standaard OP0. Het valt op dat vooral het wegen van de elementen uit het afwegingskader wordt genoemd.

Tabel 5.4

Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OP0 Basisvaardigheden)

Reden voor afwijkend oordeel	Aantal keer (%) ^a
Elementen uit het afwegingskader verschillend gewogen	29 (41%)
Informatie in het vignet over het hoofd gezien	13 (18%)
Afwegingskader verschillend geïnterpreteerd	8 (11%)
Kenmerken leerlingenpopulatie anders gewogen	6 (9%)
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	3 (4%)
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke afdeling')	2 (3%)
Beslisregel OR1 anders toegepast	2 (3%)
Contextinformatie afdeling anders gewogen	2 (3%)
Contextinformatie bestuur anders gewogen	0
Toezichthistorie anders gewogen	0
Handleiding met afwegingskader wel/niet gebruikt	0
Anders	6 (9%)

^a) Let op: Niet altijd werd een reden opgegeven. De percentages reflecteren dus het aandeel van het totaal aantal opgegeven redenen, niet van het totaal aantal beoordeelde standaarden in de duo-fase.

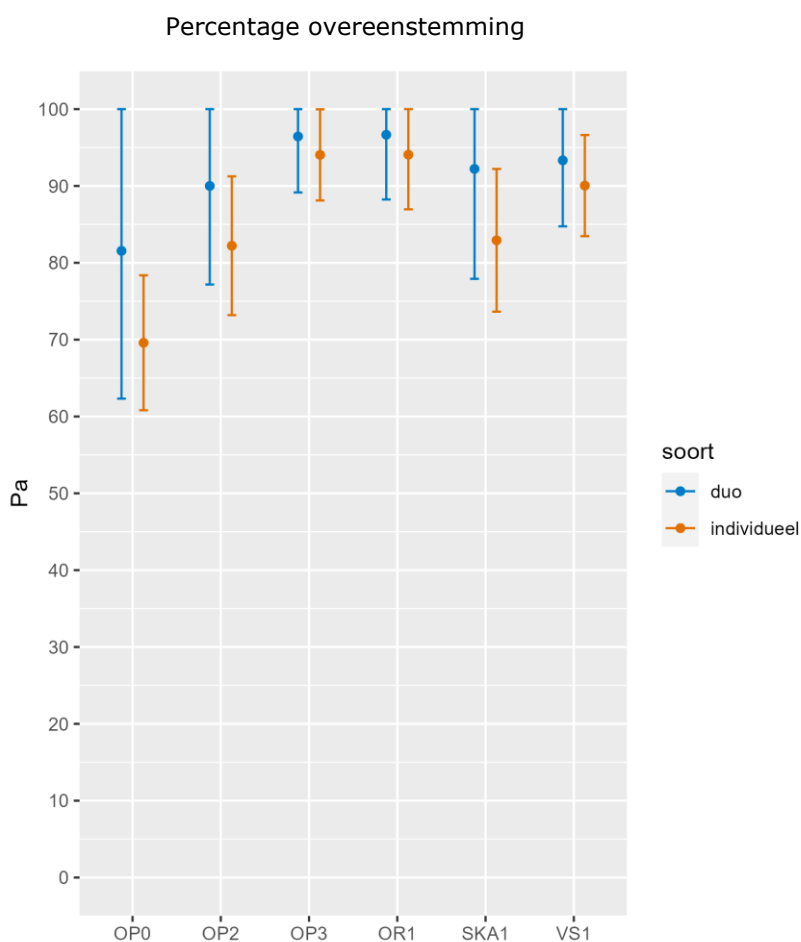
5.3 Exploratieve analyses voortgezet onderwijs

- Hoe vaak krijgen afdelingen hetzelfde eindoordeel?

Voor individuele inspecteurs blijkt dat gemiddeld 87% het meerderheidsoordeel geeft (95%-betrouwbaarheidsinterval: 84% tot 91%). Voor duo's blijkt dat gemiddeld 90% het meest gekozen oordeel geeft (95%-betrouwbaarheidsinterval: 86% tot 93%).

- Hoe groot zijn de verschillen in overeenstemming tussen standaarden?⁴⁹

In Figuur 5.2 valt op dat er een lagere mate van overeenstemming is over oordelen voor OP0. Zoals gezegd is dit een nieuwe standaard, die inspecteurs in de praktijk nog niet beoordeelden en waarvoor de scholing tijdens dit onderzoek nog volop in gang was.



Figuur 5.2. Percentage overeenstemming over individuele oordelen en duo-oordelen, uitgesplitst naar standaard, inclusief 95%-betrouwbaarheidsintervallen.

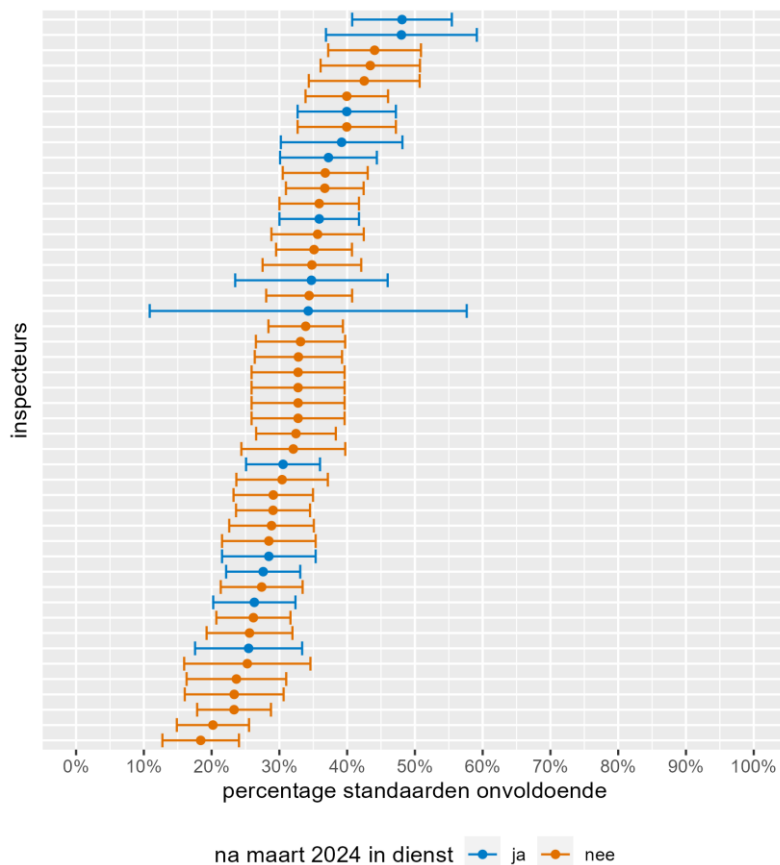
⁴⁹ In de hier gerapporteerde analyse laten we zoals besproken de optie 'Voldoende met herstelopdracht' buiten beschouwing. In het Technisch Rapport rapporteren we de analyse ook met deze vierde optie.

- Hoe groot zijn de verschillen in strengheid tussen inspecteurs?

We betrokken bij deze analyse ook de oordelen op OP0 en de oordelen op de eventueel afgeronde negende en tiende vignetten. Dit deden we om de hoeveelheid gegevens te maximaliseren. Ook de oordelen van inspecteurs in dienst na maart 2024 betrokken we in deze analyse.

Omdat niet elke inspecteur dezelfde vignetten beoordeelde, en de kans op een oordeel Onvoldoende afhangt van het vignet, pasten we een regressiemodel toe. Met dit model schatten we het verwachte oordeel voor elke standaard per inspecteur. En daarmee de verwachte strengheid voor alle 96 standaarden in de vignetten. Uiteindelijk schatten we hiermee per inspecteur de gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard.

Figuur 5.3 laat zien dat de kans op een oordeel Onvoldoende bij de meest strenge inspecteur flink verschilt van de kans op dit oordeel bij de minst strenge inspecteur. Ook wanneer we rekening houden met de betrouwbaarheidsintervallen. Verder laat de figuur zien dat de meeste inspecteurs min of meer even streng oordelen. Aan de bovenkant van de figuur zien we enkele relatief strenge inspecteurs.



Figuur 5.3. Gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard, uitgebeeld voor elke deelnemer van hoog naar laag. De 95%-betrouwbaarheidsintervallen variëren in grootte, afhankelijk van het aantal gescoorde vignetten.

6 Resultaten middelbaar beroepsonderwijs

In dit hoofdstuk bespreken we de resultaten van het vignetonderzoek in het middelbaar beroepsonderwijs en geven we antwoord op de onderzoeksvragen zoals geformuleerd in paragraaf 1.1. In het mbo namen 31 van de 38 inspecteurs (82%) deel aan het onderzoek.

6.1 Hoofdanalyse middelbaar beroepsonderwijs

- Hoe groot is de overeenstemming tussen inspecteurs in het mbo voor eindoordelen en oordelen op standaarden bij kwaliteitsonderzoeken op opleidingen?

Voor de beantwoording van deze vraag kijken we in de eerste plaats naar de overeenstemming tussen individuele inspecteurs over de eindoordelen.

Tabel 6.1

Percentage overeenstemming tussen individuele inspecteurs

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	76%	65 tot 87%
Standaardoordelen	89%	85 tot 92%

Tabel 6.1 laat zien dat de mate van overeenstemming tussen individuele inspecteurs lager uitvalt voor de eindoordelen dan voor oordelen op de standaarden. Daarnaast is, zoals verwacht, voor oordelen op de standaarden de schatting het nauwkeurigst: het betrouwbaarheidsinterval is hier het kleinst⁵⁰.

Tabel 6.2

Percentage overeenstemming tussen inspecteursduo's

Type oordelen	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	80%	64% tot 95%
Standaardoordelen	95%	91% tot 99%

In Tabel 6.2 geven we de overeenstemming tussen inspecteursduo's weer. We konden een klein deel van de oordelen niet meenemen in de berekening. Een deel van de vignetten waarover de duo-partners het in de individuele fase oneens waren, werd namelijk niet besproken. Dit ging respectievelijk om 1% en 5% van het totaal aantal oordelen op standaarden en eindoordelen. Het is waarschijnlijk dat, wanneer deze situaties wel waren besproken, de totale overeenstemming tussen duo's zou afnemen. Het feit dat individuele inspecteurs het over deze situaties oneens waren, maakt het waarschijnlijk dat dit om relatief complexe situaties ging. Omdat dit deel

⁵⁰ De vignetten weerspiegelen een willekeurige steekproef van scholen. De mate van overeenstemming over de kwaliteit van die steekproef is daarmee een *schatting* van de overeenstemming over de volledige groep scholen die we bezoeken. Die schatting gaat gepaard met een onzekerheid. Het 95%-betrouwbaarheidsinterval betekent dat er 95% kans is dat het werkelijke populatiegemiddelde binnen de getoonde grenzen valt.

van de gezamenlijke oordelen ontbreekt, schatten we in Tabel 6.2 dus een bovengrens aan overeenstemming tussen de inspecteursduo's.

Net als bij de individuele oordelen zien we dat de overeenstemming tussen inspecteursduo's voor de eindoordelen lager is dan voor de oordelen op de standaarden. Ook hier zien we, zoals verwacht, een meer nauwkeurige schatting van de overeenstemming over standaardoordelen. Verder wekt Tabel 6.2 de indruk, vergeleken met Tabel 6.1, dat verschillende inspecteursduo's het vaker met elkaar eens zijn dan individuele inspecteurs onderling. Die statistische vergelijking behandelen we bij de volgende onderzoeksvraag.

Tabel 6.3 laat zien in hoeverre individuele inspecteurs tot hetzelfde oordeel komen, met een correctie voor de kans op toeval. Dit gebeurt aan de hand van 2 aanvullende maten.

Tabel 6.3

Mate van overeenstemming tussen inspecteurs, na correctie voor toeval

Type oordelen	Maat	Overeenstemming	95%-betrouwbaarheidsinterval
Eindoordelen	Fleiss Kappa	0,53	0,32–0,73
	AC1	0,71	0,57–0,85
Standaardoordelen	Fleiss Kappa	0,56	0,43–0,70
	AC1	0,87	0,83–0,91

6.2

Secundaire analyses middelbaar beroepsonderwijs

- Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's?

Een vergelijking tussen individuele inspecteurs (Tabel 6.1) en duo's (Tabel 6.2) vereist dat we hetzelfde deel van de oordelen weglaten in de berekening van het percentage overeenstemming tussen individuele inspecteurs. Namelijk: het deel dat de duo's niet bespraken. Als we dit doen zien we een hogere mate van overeenstemming tussen individuele inspecteurs voor zowel eindoordelen (79%) als oordelen op standaarden (90%). De waarden voor inspecteursduo's zoals getoond in Tabel 6.2 liggen respectievelijk 1 procentpunt en 5 procentpunt hoger.

Het verschil dat te zien is tussen de individuele inspecteurs en duo's voor de eindoordelen, is niet statistisch significant: $t < 1$. Wel laten duo's meer overeenstemming zien over oordelen over de standaarden, $t(121) = 2,54, p < 0,01$. Dit onderzoek toont dus aan dat overleg in duo's leidt tot meer betrouwbare standaardoordelen. Dat er geen bewijs is dat dit zich vertaalt in een hogere overeenstemming over eindoordelen, kan komen doordat inspecteurs het in eerste instantie vaak al eens waren over een doorslaggevende (Onvoldoende) standaard. Een kanttekening bij deze bevindingen is, dat het negeren van een (klein) deel van de vignetten – waarbij oordelen verschilden – waarschijnlijk zorgt voor een lichte onderschatting van de toename in overeenstemming. Bovendien werden vergissingen van individuele inspecteurs, bijvoorbeeld bij het toepassen van de beslisregels voor het eindoordeel, bijna altijd gecorrigeerd in de duo-fase.

- Welke redenen noemen inspecteursduo's voor verschillen in individuele oordelen?

Tabel 6.4 laat zien hoe vaak inspecteurs elke verklaring noemden. OP0 is hierbij buiten beschouwing gelaten. In Bijlage 4 bij dit rapport splitsen we deze gegevens uit naar de afzonderlijke standaarden. Verreweg het vaakst noemden deelnemers dat zij elementen uit de handreiking verschillend hadden gewogen. We zien dat dit de meest genoemde reden was voor het verschillend beoordelen van de standaarden OP3, SKA2, OP5, BA1, OP2, VS1 en BA2. Verder valt op dat deelnemers de contextinformatie van de opleiding relatief vaak verschillend meewogen in het oordeel op OP3, SKA2 en BA1. Bij OR1 werd het anders toepassen van de beslisregel het meest genoemd. Het verschillend interpreteren van de handreiking vond vooral plaats bij SKA2. Bij OP3 – maar af en toe ook bij BA1, BA2, VS1 en SKA1 – gaven inspecteurs aan dat ze het oordeel op een andere standaard verschillend meewogen (het zogeheten 'doortikeffect'). Met name bij OP5 zagen inspecteurs een aantal keer informatie in het vignet over het hoofd. Tot slot, de open antwoorden bij de categorie 'anders' varieerden van vergissingen ($n = 4$), bijvoorbeeld het verkeerd noteren van een oordeel, tot het verschillend interpreteren van de vignettekst ($n = 2$). Een enkeling merkte op dat belangrijke informatie ontbrak in het vignet, waardoor oordelen niet overeenkwamen ($n = 1$).

Bijlage 4 laat ook de verklaringen van inspecteurs voor verschillen in oordelen bij de standaard OP0 zien. Inspecteurs noemden vooral het wege van de elementen uit de handreiking en de interpretatie van de handreiking. Maar ook dat ze informatie in het vignet over het hoofd hadden gezien.

Tabel 6.4

Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OP0 Basisvaardigheden)

Reden voor afwijkend oordeel	Aantal keer (%) ^a
Elementen uit de handreiking verschillend gewogen	33 (40%)
Contextinformatie opleiding anders gewogen	9 (11%)
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	7 (9%)
Handreiking verschillend geïnterpreteerd	6 (7%)
Informatie in het vignet over het hoofd gezien	6 (7%)
Beslisregel OR1 anders toegepast	3 (4%)
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke opleiding')	2 (2%)
Contextinformatie bestuur anders gewogen	2 (2%)
Kenmerken studentenpopulatie anders gewogen	2 (2%)
Handleiding met handreiking wel/niet gebruikt	0
Toezichthistorie anders gewogen	0
Anders	12 (15%)

^a) Let op: Niet altijd werd een reden opgegeven. De percentages reflecteren dus het aandeel van het totaal aantal opgegeven redenen, niet van het totaal aantal beoordeelde standaarden in de duo-fase.

6.3

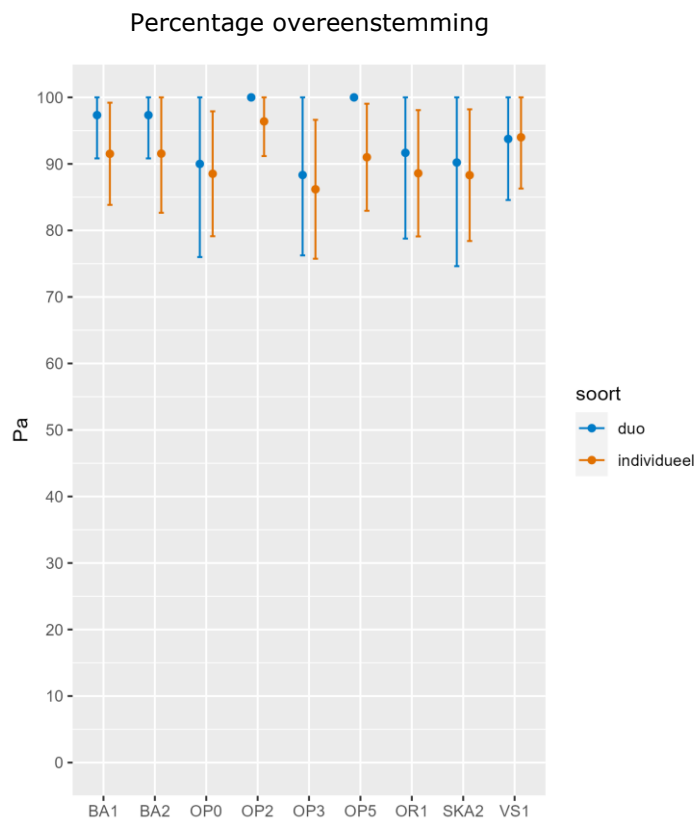
Exploratieve analyses middelbaar beroepsonderwijs

- Hoe vaak krijgen opleidingen hetzelfde eindoordeel?

Voor individuele inspecteurs blijkt dat gemiddeld 85% het meerderheidsoordeel geeft (95%-betrouwbaarheidsinterval: 81% tot 89%). Voor duo's blijkt dat gemiddeld 87% het meest gekozen oordeel geeft (95%-betrouwbaarheidsinterval: 83% tot 92%).

- Hoe groot zijn de verschillen in overeenstemming tussen standaarden?⁵¹

Figuur 6.2 geeft de verschillen in overeenstemming per standaard weer. Opvallend is dat de overeenstemming over oordelen per standaard weinig verschilt. Dit geldt ook voor OP0, terwijl dit om een nieuwe standaard gaat, die in de praktijk nog niet werd beoordeeld. Bovendien valt op dat alle duo's het onderling eens waren over OP2 en OP5.



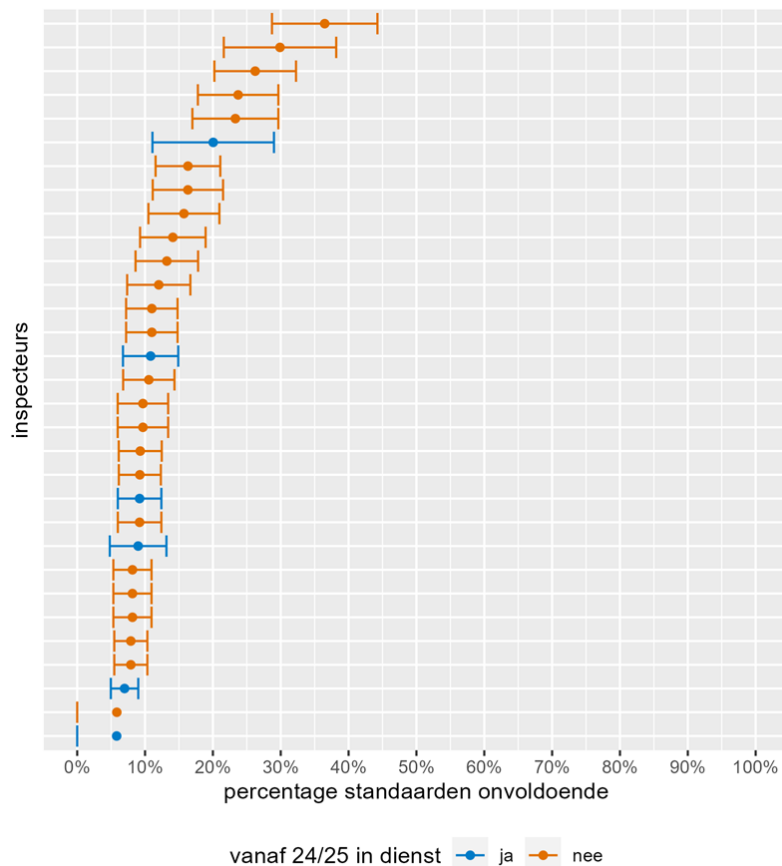
Figuur 6.2. Percentage overeenstemming over individuele oordelen en duo-oordelen, uitgesplitst naar standaard, inclusief 95%-betrouwbaarheidsintervallen.

⁵¹ In de hier gerapporteerde analyse laten we zoals besproken de optie 'Voldoende met herstelopdracht' buiten beschouwing. In het Technisch Rapport rapporteren we de analyse ook met deze vierde optie.

- Hoe groot zijn de verschillen in strengheid tussen inspecteurs?

We betrokken bij deze analyse ook de oordelen op OP0 en de oordelen op de eventueel afgeronde negende en tiende vignetten. Dit deden we om de hoeveelheid gegevens te maximaliseren. Ook de oordelen van inspecteurs in dienst na augustus 2024 betrokken we in deze analyse.

Omdat niet elke inspecteur dezelfde vignetten beoordeelde, en de kans op een oordeel Onvoldoende afhangt van het vignet, pasten we een regressiemodel toe. Met dit model schatten we het verwachte oordeel voor elke standaard per inspecteur. En daarmee de verwachte strengheid voor alle 122 standaarden in de vignetten. Uiteindelijk schatten we hiermee per inspecteur de gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard. Figuur 6.3 laat zien dat de kans op een oordeel Onvoldoende bij de meest strenge inspecteur flink verschilt van de kans hierop bij de minst strenge inspecteur. Ook wanneer we rekening houden met de betrouwbaarheidsintervallen. Verder laat de figuur zien dat de meeste inspecteurs min of meer even streng oordelen. Men name aan de bovenkant van de figuur zien we een kleine groep relatief strenge inspecteurs. Standaardfouten voor de twee inspecteurs met de laagste percentages Onvoldoende konden we niet berekenen. Er was namelijk te weinig variatie in hun oordelen Voldoende/Onvoldoende.



Figuur 6.3. Gemiddelde kans op een oordeel Onvoldoende op een willekeurige standaard, uitgebeeld voor elke deelnemer van hoog naar laag. De 95%-betrouwbaarheidsintervallen variëren in grootte, afhankelijk van het aantal gescoorde vignetten.

7 Discussie

In de ideale situatie maakt het voor de oordeelsvorming niet uit welke inspecteur een school bezoekt. Uit onderzoek weten we echter dat deze ideale situatie niet realistisch is.

Dit onderzoek geeft inzicht in de mate waarin inspecteurs dezelfde situaties hetzelfde beoordelen. Daarmee biedt het aanknopingspunten voor verbeteracties binnen het toezicht. We zien dit onderzoek als **nulmeting**. Door het onderzoek over een paar jaar te herhalen, houden we zicht op de mate van IBB.

In dit rapport geven we de resultaten weer van uitgevoerd onderzoek in het po, (v)so, vo en mbo. We benadrukken hierbij nogmaals dat het niet mogelijk is om een onderlinge vergelijking te maken tussen de percentages overeenstemming bij de verschillende sectoren. De verschillen tussen de sectoren kunnen duiden op een verschil in kwaliteit van de oordelen, maar kunnen ook het resultaat zijn van de manier waarop de vignetten werden samengesteld. Voor sommige vignetten sluit de formulering nauw aan op de rapporten van eerdere kwaliteitsonderzoeken waarop ze zijn gebaseerd. Daardoor liggen deze vignetten dicht bij de oorspronkelijke oordelen. Dat kan leiden tot een hogere mate van overeenstemming tussen inspecteurs.

Het onderzoek laat zien dat individueel oordelende inspecteurs in het po, (v)so, vo en mbo het in gemiddeld 82% tot 89% van de gevallen eens zijn over de oordelen op standaarden. En in gemiddeld 65% tot 81% van de gevallen over de eindoordelen. De percentages zijn iets hoger als inspecteurs de oordelen vervolgens in duo's afstemmen: respectievelijk 89% tot 95% (voor de standaarden) en 76% tot 82% (voor de eindoordelen). In de praktijk is de overeenstemming mogelijk lager, omdat inspecteurs dan over meer informatie beschikken. Waarbij de aanname is dat meer informatie een grotere kans geeft dat die informatie verschillend geïnterpreteerd kan worden⁵².

7.1 Bevindingen vergeleken met ander onderzoek

We stelden eerder vast dat er geen consensus is over wat een lage of hoge mate van overeenstemming is. Het beste dat we kunnen doen, is onze resultaten vergelijken met de resultaten van eerder onderzoek uit de wetenschappelijke literatuur (zie Paragraaf 1.3). We beperken ons hierbij tot de 3 onderzoeken die het meest vergelijkbaar zijn.

In 2 onderzoeken beoordeelden inspecteurs individueel de kwaliteit van ziekenhuizen⁵³ en gezondheidscentra⁵⁴. In het derde onderzoek beoordeelden individuele verpleegkundigen los van elkaar de kwaliteitscriteria van verpleeghuizen⁵⁵. In deze onderzoeken was het percentage overeenstemming over de beoordeelde indicatoren respectievelijk 61%, 71% en 89%. Deze percentages

52 Kahneman, D., Sibony, O., & Sustain, C. (2021). *Noise – a flaw in human judgement*. HarperCollins Publishers.

53 Boyd, A., Addicott, R., Robertson, R., Ross, S., and Walshe, K. (2016). Are inspectors' assessments reliable? Ratings of NHS acute hospital trust services in England. *Journal of Health Services Research & Policy*, 22, 1.

54 Sekayombya, B., Nahamya, D., Garabedian L., Seru, M. and Trap, B. (2019). Inter-rater reliability and validity of good pharmacy practices measures in inspection of public sector health facility pharmacies in Uganda. *Journal of Pharmaceutical Policy and Practice*, 12, 2.

55 Mor, V. et al. (2003). Inter-rater reliability of nursing home quality indicators in the U.S. *BMC Health Services Research*, 3.

kunnen we vergelijken met de percentages overeenstemming in individuele oordelen op de standaarden in ons onderzoek, namelijk: 82% in po, 85% in (v)so, 89% in vo en 89% in mbo. Vanuit dit perspectief zijn de gevonden percentages niet opmerkelijk hoog of laag. We benadrukken echter dat we onze resultaten met slechts 3 studies vergelijken. Het gaat om studies uit andere landen en van andere soorten inspecties, die bovendien sterk uiteenlopende resultaten laten zien.

Naast het percentage overeenstemming gebruikten we ook kappa: een maat die overeenstemming op basis van toeval tussen inspecteurs corrigeert. De Engelse onderwijsinspectie rapporteerde kappa-waarden variërend van 0,38 tot 0,49 in een onderzoek waarin inspecteurs individueel de kwaliteit van curriculum-implementatie beoordeelden⁵⁶. In het onderzoek waarbij verplegers⁵⁷ kwaliteitscriteria van verpleeghuizen beoordeelden, correspondeerde een overeenstemming van 89% met een kappa-waarde van 0,60. In ons onderzoek vinden we kappa-waarden van 0,60 in het po, 0,52 in het (v)so, 0,69 in het vo en 0,64 in het mbo voor de individuele oordelen op standaarden. Uit deze vergelijking concluderen we opnieuw dat de verschillen niet opmerkelijk hoog of laag zijn. De vergelijking met deze studies, met een correctie op toevallige overeenstemming, is echter extra onzeker. (Meer details zijn te vinden in het Technisch Rapport.)

7.2 Aanvullende bevindingen

7.2.1 Duo's versus individuele inspecteurs

Hoewel de percentages overeenstemming meestal hoger zijn voor duo's dan voor individuele inspecteurs, kunnen we niet concluderen dat de IBB in alle sectoren hoger is als inspecteurs in duo's beoordelen. De verschillen in percentages bleken namelijk niet statistisch significant, de standaardoordelen in het mbo uitgezonderd (5% meer overeenstemming onder duo's). Met als kanttekening dat het aantal beoordeelde standaarden in het mbo ($n = 122$) groter was dan in de andere sectoren ($n = 96$). Het is mogelijk dat een groter aantal duo-oordelen zou resulteren in statistisch significante verschillen. We merken hierbij op dat we de toename in overeenstemming tussen individuele beoordelingen en duo-beoordelingen mogelijk te laag hebben ingeschat. Een (klein) deel van de vignetten waarover inspecteurs het niet met elkaar eens waren, is namelijk niet besproken.

Hoewel de overeenstemming bij duo-oordelen over alle standaarden meestal niet significant hoger was, heeft afstemming in duo's wel degelijk voordelen. Zo zagen we dat vergissingen, bijvoorbeeld bij het toepassen van de beslisregels voor het eindoordeel, bijna altijd werden gecorrigeerd in de duo-fase (zie paragraaf 7.2.5). Ook vergissingen in het po bij het beoordelen van OR1 werden bijna allemaal hersteld in de duo-fase.

7.2.2 Eindoordelen versus oordelen op standaarden

Wat opvalt is dat de overeenstemming over de eindoordelen lager is dan de overeenstemming over de oordelen op standaarden. Dit verbaasde de inspecteurs, die het omgekeerde hadden verwacht. Dit verschil is als volgt te verklaren. Als er voor elke standaard een kleine kans bestaat dat inspecteurs verschillend oordelen, dan zijn er veel manieren om tot een ander eindoordeel te komen. De optelsom van al die kleine kansen op verschillende oordelen per standaard leidt daarmee tot een

⁵⁶ Ofsted (2019). *Workbook scrutiny*. Ofsted research report 190028

⁵⁷ Mor, V. et al. (2003). Inter-rater reliability of nursing home quality indicators in the U.S. *BMC Health Services Research*, 3.

grote kans op afwijkende eindoordeelen. Een rekenvoorbeeld (uitgaand van 6 standaarden)⁵⁸: wanneer inspecteurs in 90% van de gevallen het juiste oordeel op een standaard geven, levert dit in slechts 68% van de gevallen hetzelfde juiste eindoordeel op.

Sommige inspecteurs gaven aan dat zij, op basis van het brede beeld van een school, afdeling of opleiding, een passend eindoordeel voor ogen hebben. Daarbij kan het voorkomen dat standaarden die zij normaal gesproken als Onvoldoende zouden beoordelen, nu een Voldoende krijgen, of andersom – redenerend vanuit het eindoordeel. Bovendien zou volgens sommige inspecteurs deze holistische benadering onderling in de regel niet verschillen. Dit zou betekenen dat inspecteurs vaak tot hetzelfde eindoordeel komen, maar dat de onderbouwing ervan – in de vorm van oordelen op standaarden – soms verschilt. Toch ondersteunen de verzamelde data dit scenario niet. Bovendien gebruikten inspecteurs in de duo-fase niet vaak de omschrijving 'beredeneerd geoordeeld met het oog op effect toezicht' als verklaring voor hun verschillende individuele oordelen op standaarden. Het is niet uit te sluiten dat beredeneerd oordelen in de praktijk een grotere rol speelt dan tijdens het onderzoek, omdat de consequenties van de beoordeling groter zijn in de praktijk.

7.2.3 *Strengheid inspecteurs*

Het onderzoek laat zien dat de meeste inspecteurs ongeveer even streng oordelen. En dat een kleine groep inspecteurs relatief strenger oordeelt. Voor de beoordeling van een school, afdeling of opleiding kan het dus uitmaken of 2 relatief strenge of 2 relatief milde inspecteurs op bezoek komen. Al is de kans klein dat dit ook echt gebeurt.

7.2.4 *Groepen inspecteurs*

De resultaten wijzen er niet op dat werkervaring van inspecteurs een belangrijke en aannemelijke verklaring is voor de variatie in oordelen. Tegelijkertijd zijn de vergeleken groepen vaak erg klein, vooral de groepen van nieuwe inspecteurs. Voor het po zijn er daarnaast geen aanwijzingen dat de standplaats (kantoor) van inspecteurs invloed heeft op de oordelen.

7.2.5 *Validiteit bevindingen*

Hoe valide zijn de bevindingen⁵⁹? Zijn er aanwijzingen dat inspecteurs anders oordelen tijdens het vignetonderzoek dan in de praktijk? Om dit te bepalen, keken we eerst hoe vaak inspecteurs afweken van de beslisregels voor het eindoordeel. In het po gebeurde dit in de individuele fase 31 keer (5,5%) en in de duo-fase 6 keer (3,5%)⁶⁰. In de andere sectoren waren de aantallen als volgt. In het (v)so: 9 keer (6,3%) in de individuele fase en 1 keer (2%) in de duo-fase. In het vo: 30 keer (7,3%) in de individuele fase en 1 keer (1,2%) in de duo-fase. En in het mbo: 14 keer (5,4%)⁶¹ in de individuele fase en 1 keer (1,4%) in de duo-fase. Dit beeld komt

58 In het Technisch Rapport schrijven we dit rekenvoorbeeld uit.

59 Ook deze kwestie bespreken we uitgebreider in het Technisch Rapport.

60 In het po werden relatief veel vergissingen gemaakt als OR1 Resultaten niet kon worden beoordeeld. Tijdens de duo-fase werden deze vergissingen grotendeels gecorrigeerd. In het schooljaar 2023-2024 waarin het vignetonderzoek is uitgevoerd, werden naast de signaleringswaarden ook correctiewaarden gehanteerd bij de beoordeling van OR1. In het schooljaar 2024-2025 zijn de correctiewaarden vervallen en is de beslisregel vereenvoudigd, zodat dit nu minder vaak zal voorkomen.

61 Bij mbo is het toegestaan om van de beslisregels af te wijken als alleen OR1 Onvoldoende is; 4 van de 14 keer dat de beslisregels niet werden gevolgd in de individuele fase was dit het geval. Als we dit als juiste oordelen beschouwen gaven inspecteurs in 4% in plaats van 5,4% van de gevallen een verkeerd eindoordeel.

overeen met de dagelijkse praktijk van het toezicht, waarin dit ook nauwelijks voorkomt.

Daarnaast zetten we op een rij hoe vaak deelnemers aangaven dat de informatie in het vignet ontoereikend was om tot een gedegen oordeel te komen. Hiervoor keken we naar de open antwoorden op de vraag waarom inspecteurs in de individuele fase verschillend oordeelden. Dit waren kleine aantallen op het totaal van beoordeelde standaarden. In een beperkt aantal gevallen noemden inspecteurs de vignettekst bij een standaard 'onduidelijk' of 'multi-interpretabel'. Enkele keren noemden inspecteurs dat 'cruciale informatie' bij een standaard ontbrak. Veruit de meeste deelnemers gaven vooral aan dat verschillen in oordelen het resultaat waren van het verschillend wegen van informatie. Of van het verschillend interpreteren van het afwegingskader/de handreiking. Dit soort factoren spelen in de praktijk ook een rol in het oordeelsproces.

In combinatie met de aanvullende analyses in het Technisch Rapport concluderen we dat de bevindingen voldoende valide zijn.

7.2.6 *Verschillen in overeenstemming tussen standaarden*

Het is mogelijk dat sommige standaarden geschikter zijn voor een vignetonderzoek dan andere. De standaard OP3 Pedagogisch-didactisch handelen – die normaal gesproken gebaseerd is op lesobservaties – is bijvoorbeeld lastiger te vangen in tekst dan de standaarden VS1 Veiligheid en OR1 Resultaten. Daardoor bestaat de kans dat we in dit onderzoek de overeenstemming over OP3 sterker overschatten dan de overeenstemming over VS1 en OR1.

7.2.7 *Voldoende met herstelopdracht als afzonderlijk oordeel*

Tijdens het onderzoek vroegen we de inspecteurs de standaarden die beslissend zijn voor het eindoordeel te beoordelen. Naast het oordeel Goed, Voldoende, Onvoldoende of Niet te beoordelen (alleen bij OR1) kregen ze ook de optie om een 'Voldoende met herstelopdracht' te geven. Het geven van een Voldoende met herstelopdracht heeft geen consequenties voor het eindoordeel van een school, afdeling of opleiding. Daarom lieten we deze oordelen buiten beschouwing in de analyses die in dit rapport zijn opgenomen. In het Technisch Rapport doen we hier wel verslag van.

Als we een Voldoende met herstelopdracht als een apart oordeel beschouwen, dan zien we, zoals te verwachten, dat de overeenstemming over oordelen op de standaarden daalt. De overeenstemming over de individuele oordelen op alle standaarden varieert in de sectoren tussen 65% en 84% en bij duo-oordelen tussen 69% en 91%. Ter vergelijking: als we de Voldoende met herstelopdracht niet als apart oordeel beschouwen, maar als Voldoende, dan variëren deze percentages respectievelijk tussen 82% en 89% en tussen 89% en 95%. De genoemde afname in overeenstemming tussen individueel oordelende inspecteurs is het meest prominent in het po, (v)so en vo. Hierbij valt de relatief grote daling op bij de standaard OP3. In het (v)so valt daarnaast de daling op bij OR1.

7.3 **Hoe kunnen we de IBB verbeteren?**

Kahneman en collega's⁶² bespreken richtlijnen om ruis te verminderen. Twee daarvan zijn relevant voor ons als inspectie.

62 Kahneman, D., Sibony, O., & Sustain, C. (2021). *Noise – A flaw in human judgement*. HarperCollins Publishers.

De eerste betreft de **volgorde van informatieverwerking**. Contextinformatie, maar ook het oordeel van een collega, beïnvloedt het oordeelproces. Om die reden stellen onderwijsinspecties in sommige omringende landen het oordelen uit tot na de eerste fase van onderzoek waarin observaties en dataverzameling centraal staan⁶³. In de toezichtspraktijk doen we dit al: het consensusoverleg vindt niet op de dag van het onderzoek zelf plaats en scholen, afdelingen of opleidingen krijgen pas daarna informatie over de oordelen.

De tweede richtlijn gaat over het **gebruik van beslisregels**. Die leiden in de medische wereld aantoonbaar tot minder ruis in de beoordeling (pp. 273⁶⁴). Uit onderzoek naar het verbeteren van de IBB van gezondheidszorgprofessionals⁶⁵ blijkt dat het aanscherpen van het onderzoeksinstrument meer effect heeft dan training in het gebruik van bestaande criteria. Bovendien kan aanscherping van het onderzoekskader ertoe leiden dat scholen eerder het door de inspectie gewenste gedrag laten zien. Tegelijkertijd wordt daardoor duidelijker op welke terreinen zij de ruimte hebben om zelf keuzes te maken, zo stelt de Onderwijsraad⁶⁶. Als inspectie hanteren we eenduidige beslisregels voor het eindoordeel. Zulke regels zijn er niet voor de beoordeling van de standaarden, met uitzondering van de standaard OR1 Resultaten in het po en vo. In een wettelijke regeling is vastgelegd hoe we OR1 beoordelen⁶⁷. Voor de beoordeling van de standaarden beschikken inspecteurs over een afwegingskader/handreiking. Daarin zijn de elementen opgenomen die de inspecteurs moeten meewegen bij het bepalen van hun oordeel. Al schrijft het afwegingskader niet voor hoe zij die verschillende elementen exact moeten wegen.

Overigens is aanscherping van een onderzoeksinstrument niet altijd een goed idee. Het aanscherpen of uitbreiden van criteria kan het lastiger maken om complexe situaties zo te beoordelen dat dit recht doet aan de specifieke situatie. Een verhoging van de IBB door aanscherping van het onderzoeksinstrument kan dan ten koste gaan van de validiteit van de oordelen. Het huidige onderzoekskader en de afwegingskaders/handreiking zijn vormgegeven met het oog op die balans tussen validiteit en betrouwbaarheid. Met de resultaten van dit onderzoek – dat een beeld geeft van de mate waarin we betrouwbaar oordelen – kunnen we die balans opnieuw bepalen.

7.4

Vervolg

De inspectie heeft als missie: 'Effectief toezicht voor beter onderwijs'. Oordelen hebben impact op scholen, afdelingen en opleidingen. Als zij een onjuist eindoordeel krijgen, doet dat af aan de effectiviteit van ons toezicht. Daarnaast zijn de oordelen bepalend voor onze betrouwbaarheid als toezichthouder en vormen ze een informatiebron voor bijvoorbeeld ouders.

In de dagelijkse praktijk van ons toezicht voeren inspecteurs kwaliteitsonderzoeken nooit alleen uit, maar in wisselende teams. Voordat zij hun oordelen aan scholen, afdelingen of opleidingen terugkoppelen, stemmen zij de oordelen eerst af in een zogeheten **consensusoverleg**. Bij dit consensusoverleg sluiten vaak collega's aan

63 Inspectie van het Onderwijs (2022). *Internationale Toezichtscan 2022 Interbeoordelaarsbetrouwbaarheid*. (intern document).

64 Kahneman, D., Sibony, O., & Sustain, C. (2021). *Noise – A flaw in human judgement*. HarperCollins Publishers.

65 Tuijn S.M., Janssens F.J.G., Robben P.B.M., Van den Bergh H. (2012) Reducing interrater variability and improving health care: A meta-analytic review. *Journal of Evaluation in Clinical Practice*, 18, 887-895.

66 Essentie van extern toezicht: <https://www.onderwijsraad.nl/publicaties/adviezen/2022/03/23/essentie-van-extern-toezicht>

67 Regeling leerresultaten po: <https://wetten.overheid.nl/BWBR0043066/2023-10-05> (geldig tot 31-7-2024) en Regeling leerresultaten vo: <https://wetten.overheid.nl/BWBR0038374/2023-03-16>.

die niet bij het onderzoek betrokken waren. Zij stellen kritische vragen en stimuleren de inspecteurs om de oordelen zo goed mogelijk te onderbouwen. Tot slot leest een zogeheten referent het onderzoeksrapport kritisch mee.

Tegelijkertijd besteden we voortdurend aandacht aan het bevorderen van de IBB. In het **programma Evaluatie van het toezicht**⁶⁸ brengen we met diverse onderzoeken in kaart wat er goed gaat in het toezicht, en wat er beter kan. Het vignetonderzoek is onderdeel van dit programma. De resultaten van deze onderzoeken gebruiken we om ons te verantwoorden over onze werkwijze en maatschappelijke meerwaarde. Ook gebruiken we de uitkomsten bij de verdere ontwikkeling van ons toezicht en de periodieke bijstelling van onze huidige **onderzoekskaders**. In 2027 worden de herziene onderzoekskaders van kracht.

De uitkomsten van dit vignetonderzoek dienen daarnaast als input voor de activiteiten waarmee we werken aan verdere verbetering van de IBB. Zo geven we doorlopende aandacht aan de **professionalisering** van onze collega's. Het hele schooljaar door volgen zij trainingen en scholing waarin de beoordeling van de verschillende standaarden centraal staat. Specialisten binnen de inspectie verzorgen dit scholingsaanbod. Het vignetonderzoek laat bovendien zien dat we deze activiteiten voor de verschillende sectoren beter op elkaar af kunnen stemmen.

Daarnaast werken we doorlopend aan het **verbeteren van materialen** die ondersteunen bij de oordeelsvorming, zoals de afwegingskaders/handreiking. Ook analyseren we sinds een paar jaar de rapporten van de kwaliteitsonderzoeken in het po, (v)so, vo en mbo.

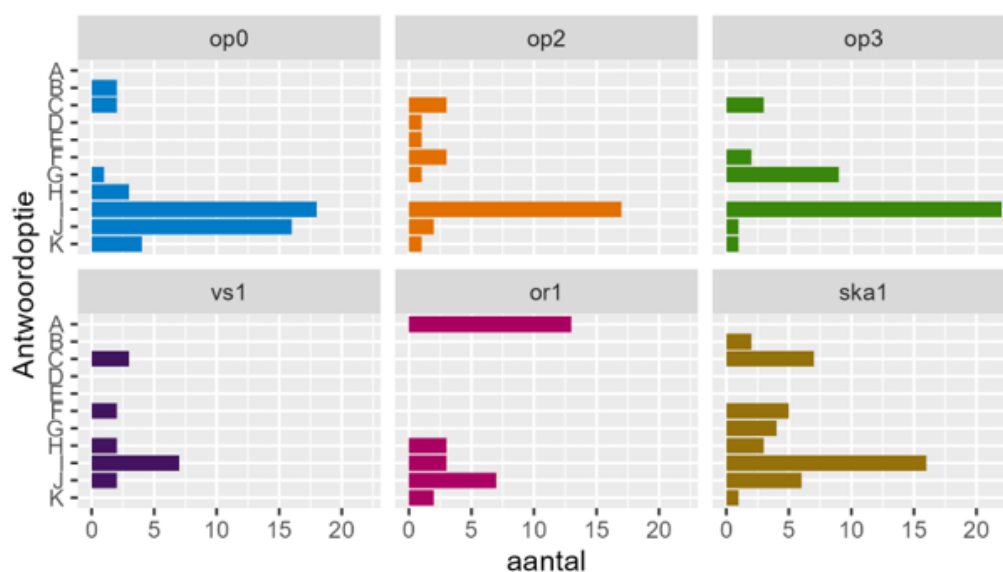
Met ingang van schooljaar 2025-2026 beoordelen we de standaard **OP0 Basisvaardigheden** voor het eerst tijdens kwaliteitsonderzoeken. De resultaten van het vignetonderzoek laten zien dat inspecteurs minder vaak tot gelijke oordelen komen op OP0, vergeleken met de andere standaarden. Tijdens het vignetonderzoek waren inspecteurs al gestart met training en scholing om deze standaard eenduidig te beoordelen. Ook nu blijft die eenduidige beoordeling een belangrijk aandachtspunt. Omdat we veel belang hechten aan de basisvaardigheden, heeft de inspectie recent taal- en rekenspecialisten aangesteld. Zij vormen de nieuw ingerichte expertisegroepen voor taal en rekenen, naast de al bestaande expertisegroep voor burgerschap. Deze expertisegroepen voorzien de inspecteurs in alle sectoren van relevante kennis.

Door dit onderzoek over enkele jaren te herhalen, kunnen we zicht houden op de betrouwbaarheid van onze oordelen en ons toezicht zo nodig gericht bijsturen.

68 <https://www.onderwijsinspectie.nl/onderwerpen/onderzoekskaders/evaluatie-van-het-toezicht>

Bijlage 1 Primair onderwijs

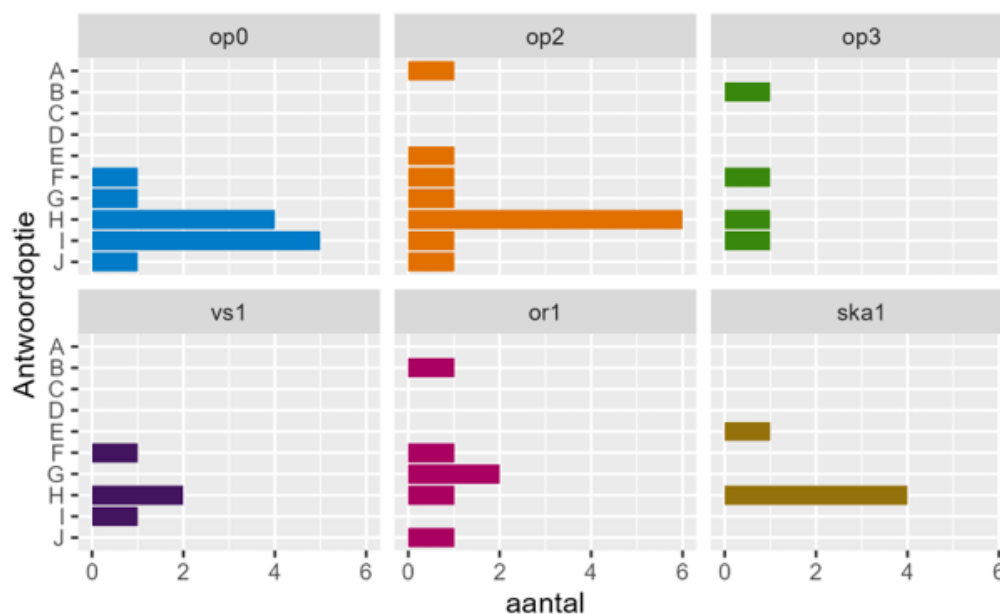
Deze bijlage toont de zelfgerapporteerde verklaringen van inspecteursduo's in het primair onderwijs voor hun verschillende oordelen in de individuele fase, uitgesplitst naar de standaarden.



- A. Beslisregel OR1 anders toegepast.
- B. Kenmerken leerlingenpopulatie anders gewogen.
- C. Contextinformatie school anders gewogen.
- D. Contextinformatie bestuur anders gewogen.
- E. Toezichthistorie anders gewogen.
- F. Beredeneerd geoordeeld met oog op het effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school').
- G. Oordeel op andere standaard meegewogen (voorkomen doortikeffect).
- H. Informatie in het vignet over het hoofd gezien.
- I. Elementen uit het afwegingskader verschillend gewogen.
- J. Afwegingskader verschillend geïnterpreteerd.
- K. Handleiding met afwegingskader wel/niet gebruikt.

Bijlage 2 (Voortgezet) speciaal onderwijs

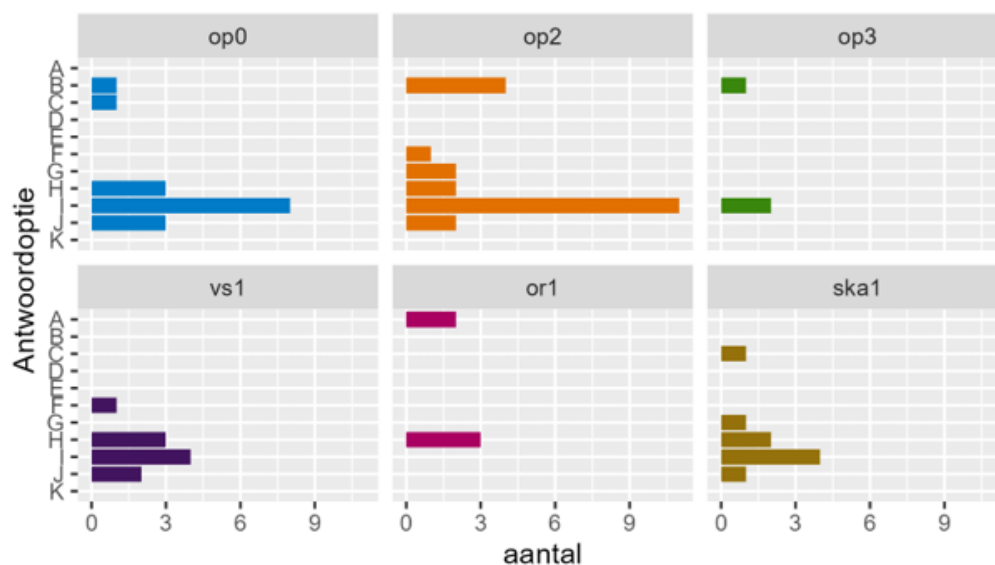
Deze bijlage toont de zelfgerapporteerde verklaringen van inspecteursduo's in het (voortgezet) speciaal onderwijs voor hun verschillende oordelen in de individuele fase, uitgesplitst naar de standaarden.



- A. Kenmerken leerlingenpopulatie anders gewogen.
- B. Contextinformatie school anders gewogen.
- C. Contextinformatie bestuur anders gewogen.
- D. Toezichthistorie anders gewogen.
- E. Beredeneerd geoordeeld met oog op het effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school').
- F. Oordeel op andere standaard meegewogen (voorkomen doortikeffect).
- G. Informatie in het vignet over het hoofd gezien.
- H. Elementen uit het afwegingskader verschillend gewogen.
- I. Afwegingskader verschillend geïnterpreteerd.
- J. Handleiding met afwegingskader wel/niet gebruikt.

Bijlage 3 Voortgezet onderwijs

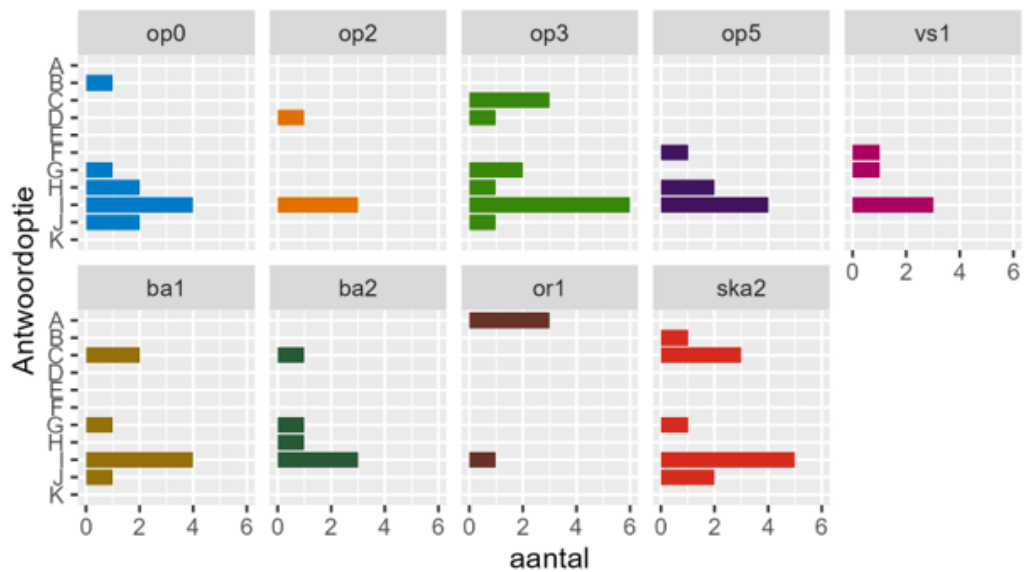
Deze bijlage toont de zelfgerapporteerde verklaringen van inspecteursduo's in het voortgezet onderwijs voor hun verschillende oordelen in de individuele fase, uitgesplitst naar de standaarden.



- A. Beslisregel OR1 anders toegepast.
- B. Kenmerken leerlingenpopulatie anders gewogen.
- C. Contextinformatie afdeling anders gewogen.
- D. Contextinformatie bestuur anders gewogen.
- E. Toezichthistorie anders gewogen.
- F. Beredeneerd geoordeeld met oog op het effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke afdeling').
- G. Oordeel op andere standaard meegewogen (voorkomen doortikeffect).
- H. Informatie in het vignet over het hoofd gezien.
- I. Elementen uit het afwegingskader verschillend gewogen.
- J. Afwegingskader verschillend geïnterpreteerd.
- K. Handleiding met afwegingskader wel/niet gebruikt.

Bijlage 4 Middelbaar beroepsonderwijs

Deze bijlage toont de zelfgerapporteerde verklaringen van inspecteursduo's in het middelbaar beroepsonderwijs voor hun verschillende oordelen in de individuele fase, uitgesplitst naar de standaarden.



- A. Beslisregel OR1 anders toegepast.
- B. Kenmerken studentenpopulatie anders gewogen.
- C. Contextinformatie opleiding anders gewogen.
- D. Contextinformatie bestuur anders gewogen.
- E. Toezichthistorie anders gewogen.
- F. Beredeneerd geoordeeld met oog op het effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke opleiding').
- G. Oordeel op andere standaard meegewogen (voorkomen doortikeffect).
- H. Informatie in het vignet over het hoofd gezien.
- I. Elementen uit de handreiking verschillend gewogen.
- J. Handreiking verschillend geïnterpreteerd.
- K. Handleiding met handreiking wel/niet gebruikt.