

# **TECHNISCH RAPPORT**

DE BETROUWBAARHEID VAN INSPECTEURSOORDELEN

EEN VIGNETONDERZOEK

Juli 2025



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



# Inhoudsopgave

<b>TECHNISCH RAPPORT .....</b>	<b>1</b>
<b>1 Definities en databronnen .....</b>	<b>7</b>
1.1 Definities.....	7
1.2 Databronnen.....	8
<b>2 Onderzoeksopzet.....</b>	<b>9</b>
2.1 Methode en analyses .....	9
2.1.1 <i>Methodologisch uitgangspunt</i> .....	9
2.1.2 <i>Afname studie</i> .....	10
2.2 Steekproeftrekking .....	11
2.3 Veranderingen ten opzichte van pre-analyseplan .....	11
<b>3 Sector PO.....</b>	<b>13</b>
3.1 Diagnostische data-analyse .....	13
3.1.1 <i>Respons en representativiteit</i> .....	13
3.1.2 <i>Validiteit</i> .....	15
3.2 Hoofdanalyse .....	18
3.2.1 <i>Primaire uitkomsten</i> .....	18
3.2.2 <i>Secundaire uitkomsten</i> .....	21
3.3 Exploratieve analyse .....	25
3.3.1 <i>Verdiepende analyse van verschillen in eendoordelen</i> .....	25
3.3.2 <i>Pa per standaard</i> .....	27
3.3.3 <i>Strengheid per inspecteur</i> .....	29
3.3.4 <i>Alternatieve specificaties</i> .....	35
<b>4 Sector (v)so.....</b>	<b>40</b>
4.1 Diagnostische data-analyse .....	40
4.1.1 <i>Respons en representativiteit</i> .....	40
4.1.2 <i>Validiteit</i> .....	41
4.2 Hoofdanalyse .....	41
4.2.1 <i>Primaire uitkomsten</i> .....	41
4.2.2 <i>Secundaire uitkomsten</i> .....	42
4.3 Exploratieve analyse .....	44
4.3.1 <i>Verdiepende analyse van verschillen in eendoordelen</i> .....	44
4.3.2 <i>Pa per standaard</i> .....	45
4.3.3 <i>Strengheid per inspecteur</i> .....	47
4.3.4 <i>Alternatieve specificaties</i> .....	51
<b>5 Sector vo .....</b>	<b>54</b>



5.1	Diagnostische data-analyse .....	54
5.1.1	<i>Respons en representativiteit</i> .....	54
5.1.2	<i>Validiteit</i> .....	55
5.2	Hoofdanalyse .....	56
5.2.1	<i>Primaire uitkomsten</i> .....	56
5.2.2	<i>Secundaire uitkomsten</i> .....	56
5.3	Exploratieve analyse .....	59
5.3.1	<i>Verdiepende analyse van verschillen in eendoordelen</i> .....	59
5.3.2	<i>Pa per standaard</i> .....	60
5.3.3	<i>Strengheid per inspecteur</i> .....	62
5.3.4	<i>Alternatieve specificaties</i> .....	66
<b>6</b>	<b>Sector mbo</b> .....	<b>69</b>
6.1	Diagnostische data-analyse .....	69
6.1.1	<i>Respons en representativiteit</i> .....	69
6.1.2	<i>Validiteit</i> .....	70
6.2	Hoofdanalyse .....	71
6.2.1	<i>Primaire uitkomsten</i> .....	71
6.2.2	<i>Secundaire uitkomsten</i> .....	71
6.3	Exploratieve analyse .....	74
6.3.1	<i>Verdiepende analyse van verschillen in eendoordelen</i> .....	74
6.3.2	<i>Pa per standaard</i> .....	75
6.3.3	<i>Strengheid per inspecteur</i> .....	77
6.3.4	<i>Alternatieve specificaties</i> .....	81
<b>7</b>	<b>Referenties</b> .....	<b>84</b>
	<b>Bijlage I</b> .....	<b>85</b>
	<b>Bijlage II</b> .....	<b>86</b>
	<b>Bijlage III</b> .....	<b>91</b>



## Inleiding

Het toezicht op de kwaliteit van het onderwijs door de Inspectie van het Onderwijs richt zich op drie niveaus: op het stelsel, op besturen en op hun scholen, afdelingen of opleidingen. In deze studie richten we ons op de oordeelsvorming van inspecteurs bij kwaliteitsonderzoeken op scholen (po; [v]so), afdelingen (vo) en opleidingen (mbo).

Het belangrijkste doel van het voorliggende onderzoek is om de mate van overeenstemming in de oordelen bij kwaliteitsonderzoeken te schatten. Daarnaast proberen we inzicht te krijgen in de factoren die van invloed zijn op de mate van overeenstemming.

We maken hieronder een onderscheid tussen de hoofdvraag en secundaire vragen, om te benadrukken dat de schatting van overeenstemming het belangrijkste doel is. De secundaire analyses zijn bedoeld om de primaire uitkomst te duiden en om aanknopingspunten te vinden voor eventuele verbeteracties.

Voor zowel het beantwoorden van de hoofdvraag als secundaire vragen hebben we van tevoren vastgelegd hoe we de analyses uitvoeren in een pre-analyseplan<sup>1</sup>. Daarnaast beschrijven we exploratieve analyses, waarvoor we van tevoren niet hebben vastgelegd hoe we deze uitvoeren, omdat ze pas in een latere fase zijn uitgekristalliseerd.

### *Hoofdvraag:*

1. Hoe groot is de overeenstemming tussen inspecteurs voor eindoordelen en oordelen op standaarden bij kwaliteitsonderzoek op scholen/afdelingen/opleidingen?

### *Secundaire vragen:*

1. Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's?
2. Is de overeenstemming groter binnen kantoren<sup>2</sup> en binnen cohorten inspecteurs?
3. Welke redenen noemen duo's van inspecteurs voor verschillen in individuele oordelen?

### *Exploratieve vragen:*

1. Welke oordelen geven inspecteurs die verschillende eindoordelen geven?<sup>3</sup>
2. Hoe vaak krijgen scholen/afdelingen/opleidingen hetzelfde eindoordeel?
3. Hoe groot zijn verschillen in overeenstemming tussen standaarden?
4. Hoe groot zijn verschillen in strengheid tussen inspecteurs?

De opzet van dit technisch rapport (TR) is als volgt. We beschrijven eerst beknopt de definities, bronnen en onderzoeksozet. Vervolgens presenteren we per sector de analyses in drie stappen: diagnostische analyses gericht de respons, representativiteit en validiteit, de hoofdanalyse en ten slotte

---

<sup>1</sup> <https://osf.io/b68ge/>

<sup>2</sup> 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.

<sup>3</sup> De beantwoording van deze vraag behandelen we alleen in dit Technisch Rapport, ter verdieping.



exploratieve analyses. Bij de beschrijving van de eerste sector die we behandelen, po in paragraaf 3, is de toelichting uitgebreid. Voor de overige drie sectoren, (v)so, vo en mbo, geven we tabellen en grafieken, maar lichten we alleen opvallende zaken eruit.



## 1 Definities en databronnen

### 1.1 Definities

#### *Standaarden*

Het eindoordeel dat een school, afdeling of opleiding krijgt tijdens een kwaliteitsonderzoek, wordt bepaald door de oordelen op zogenaamde standaarden. Deze standaarden betreffen aspecten van de onderwijskwaliteit, zoals: zicht op ontwikkeling, pedagogisch-didactische handelen en veiligheid. Inspecteurs volgen beslisregels om van oordelen op standaarden tot een eindoordeel over een school, afdeling of opleiding te komen.

#### *Vignetten*

In dit onderzoek vatten we vignetten op als beschrijvingen van een aantal aspecten van onderwijskwaliteit op een school die relevant zijn voor inspecteuroordelen. We gebruiken de vignetten alleen als middel om de informatie die inspecteurs gebruiken om tot een oordeel te komen gelijk te houden. De opzet van ons vignetonderzoek wijkt daarmee af van een andere gangbare vorm, gericht op het meten van effecten op keuzeprocessen. Hierbij worden vignetten aangeboden die tussen deelnemers op bepaalde aspecten verschillen.

Elk vignet bevat de volgende informatie over een school/afdeling/opleiding: (1) contextinformatie: 'aanleiding onderzoek', 'school/afdeling/opleidingskenmerken' (i.e., toezichthistorie; leerlingen/studentenpopulatie; prestatie-monitor; overige aspecten) en 'bestuurskenmerken', en (2) zo objectief mogelijk geformuleerde observaties per standaard voor de standaarden die beslissend zijn voor het eindoordeel uit het betreffende onderzoekskader.

#### *Vignetstandaarden*

Met een vignet-standaard bedoelen we de informatie in een vignet over een bepaalde standaard uit het onderzoekskader. In totaal hebben we voor alle sectoren 16 scholen vignetten gemaakt met voor de sectoren po, (v)so en vo 6 standaarden (zie paragraaf 2.2.1). We hebben voor deze sectoren dus  $16 \cdot 6 = 96$  vignet-standaarden gemaakt. Voor mbo zijn er  $16 \cdot 9 = 144$  vignet-standaarden gemaakt.

#### *Individuele fase*

De fase van het onderzoek waarin inspecteurs individueel de vignetten beoordelen.

#### *Duo-fase*

De fase van het onderzoek waarin inspecteurs in duo's (of in een enkel geval in trio's) overeenstemming proberen te bereiken op basis van de vignet-standaarden uit de individuele fase. Ze hebben een vignet-standaard alleen dan besproken wanneer hun oordeel verschilde én wanneer de tijd dit toeliet: niet alle oordelen die verschilden werden daadwerkelijk besproken.

#### *Beoordelaar*

Met deze term verwijzen we naar de entiteit die een oordeel geeft in de studie. In de individuele fase is dit de inspecteur, in de duo-fase is dit het duo. In de praktijk van het toezicht zijn beoordelaars duo's, die samen een oordeel geven.



#### *Percentage overeenstemming (Pa)*

Dit is onze belangrijkste uitkomstmaat. We hebben de keuze daarvoor beargumenteerd in het pre-analyseplan<sup>4</sup> en doen dit hier opnieuw in paragraaf 2.2.1. We berekenen deze maat als volgt. We maken alle mogelijke koppels van beoordelaars die hebben deelgenomen aan de studie. Per standaard en per eindoordeel noteren we voor elk mogelijk koppel of ze al dan niet hetzelfde oordeel hebben gegeven. Het percentage overeenstemming voor een beschreven vignet-standaard is gelijk aan het percentage van de mogelijke koppels dat hetzelfde oordeel geeft. De Pa voor alle inspecteurs over alle vignetten is het gemiddelde van de Pa van de vignet-standaarden.

#### *Oordelen*

Deelnemers kregen de opdracht om de standaarden te beoordelen en het eindoordeel te bepalen voor zoveel mogelijk vignetten. Net als in de praktijk kon een standaard de volgende oordelen krijgen: Goed, Voldoende, Voldoende met herstelopdracht, Onvoldoende of Niet te beoordelen (alleen bij OR1<sup>5</sup>). Bij het eindoordeel waren in het po, (v)so en vo drie opties: Voldoende, Onvoldoende of Zeer Zwak. In het mbo waren er twee aanvullende opties voor het eindoordeel: Goed en Onvoldoende met risico op bekostigings sanctie.

## **1.2 Databronnen**

In deze studie vormen de oordelen van inspecteurs op de vignetten de primaire databron. Deze databron is aangevuld met de historische oordelen voor de 16 beschreven scholen/afdelingen/opleidingen uit een intern inspectiebestand met recente oordelen per sector. De aantallen inspecteurs die per sector de doelpopulatie vertegenwoordigen, zijn verkregen door het personeelsbestand te raadplegen en navraag te doen welke inspecteurs ingezet worden bij kwaliteitsonderzoeken op scholen/afdelingen/opleidingen.

---

<sup>4</sup> <https://osf.io/b68ge/>

<sup>5</sup> Alleen bij de standaard OR1 komt het voor dat geen oordeel kan worden gegeven, zie bv. <https://wetten.overheid.nl/BWBR0043066/2023-10-05> (geldig tot 31-7-2024).



## 2 Onderzoeksopzet

### 2.1 Methode en analyses

#### 2.1.1 *Methodologisch uitgangspunt*

Dit onderzoek is geïnspireerd door de zogenaamde *noise audits* van Kahneman, Sibony en Sunstein (2021). In dergelijk onderzoek lezen professionele beoordelaars in een organisatie dezelfde tekst over een casus (hierna een 'vignet'), waarna ze een oordeel moeten geven. Voor rechters is dit oordeel bijvoorbeeld de duur van een gevangenisstraf. Het uitgangspunt bij deze studies is dat de overeenstemming over teksten altijd groter zal zijn dan die over casussen in de praktijk. In de praktijk zal er immers veel meer informatie afkomen op een beoordelaar dan in een vignet. En hoe meer informatie, hoe groter de kans op verschillende interpretaties van die informatie, en hoe minder overeenstemming.

Dit uitgangspunt is alleen dan aannemelijk als inspecteurs zich bij het oordelen beperken tot de informatie die wordt aangeboden in de vignetten, en er geen informatie bij bedenken. We hebben daarom van tevoren door een aantal inspecteurs laten verifiëren of de vignetten voldoende relevant geachte informatie bevatten om een oordeel te kunnen geven. Bovendien hebben inspecteurs voor afname van de studie de instructie ontvangen om bij gebrek aan informatie over een aspect van onderwijskwaliteit aan te nemen dat er geen zwaarwegende contra-indicaties zijn.

#### *Meting overeenstemming*

Ons doel is om te bepalen in hoeverre inspecteurs het met elkaar eens zijn wanneer ze scholen/afdelingen/opleidingen beoordelen. Alle maten van overeenstemming nemen als beginpunt voor de berekening het percentage overeenstemming (Pa), dat we hebben toegelicht in hoofdstuk 1. Het is gebruikelijk om in studies, die overeenstemming tussen beoordelaars meten, vervolgens te corrigeren voor 'toevalsovereenstemming'. Deze zogenaamde toevallige overeenstemming wordt als niet 'echt' gezien, en daarom afgetrokken van de 'werkelijke' overeenstemming. Uit de oordelen is echter niet op te maken welk deel van de overeenstemming op toeval berust. Daarom moet dit deel worden geschat op basis van het aantal antwoordcategorieën of de verdeling van de antwoorden over deze categorieën. Er zijn vele manieren om dit aan te vliegen, die leiden tot verschillende schattingen van de 'werkelijke' overeenstemming. Al deze correctiemethodes, zoals Fleiss Kappa of AC1, leiden echter tot paradoxen (McHugh 2012, Tong et al 2020, Rau & Shih 2021, Zhao et al 2023). Het is daarom niet duidelijk of ze niet meer vertekeningen toevoegen dan opheffen. Bovendien zijn de resultaten van deze correctiematen niet stabiel. Ze zijn namelijk gevoelig voor veranderingen in de verdelingen van de antwoorden, die niet hoeven samen te hangen met de mate van toevalsovereenstemming.

We kiezen daarom als primaire uitkomstmaat het ruwe percentage overeenstemming (Pa), waarin we dus niet corrigeren voor toevalsovereenstemming. We rapporteren in de exploratieve analyse wel de waarden van de maten AC1 en Fleiss Kappa, voor onderzoekers die daarin geïnteresseerd zijn en om vergelijkbaarheid met ander onderzoek (voor zover dat mogelijk wordt geacht) mogelijk te maken. We doen dit alleen voor overeenstemming in de individuele fase en niet voor overeenstemming in de duo-fase, vanwege de grotere onzekerheid van die schatting.



### *Context*

Tijdens kwaliteitsonderzoeken bij scholen/afdelingen/opleidingen beoordelen inspecteurs een aantal standaarden en geven zij een eindoordeel over de kwaliteit van het onderwijs. Het eindoordeel bepalen inspecteurs aan de hand van beslisregels, waarin de oordelen op de standaarden zijn verwerkt. In de studie zijn in elke sector 16 vignetten (met omvang van maximaal twee A4) gebruikt. In de vignetten voor de sectoren po, (v)so en vo zijn – naast contextinformatie over de school/afdeling en het bestuur – observaties opgenomen over de onderstaande 6 standaarden.

OP0 – Basisvaardigheden (Nederlandse taal, rekenen, burgerschap)  
OP2 – Zicht op ontwikkeling en begeleiding  
OP3 – Pedagogisch-didactisch handelen  
VS1 – Veiligheid  
OR1 – Resultaten  
SKA1 – Visie, ambitie en doelen

Gekozen is voor deze standaarden (uitgezonderd OP0), omdat zij een rol spelen in de beslisregels voor het eindoordeel (Voldoende, Onvoldoende of Zeer zwak). OP0 Basisvaardigheden is toegevoegd aan de vignetten, omdat dit een nieuwe standaard is. Deze standaard onderzoeken we sinds september 2023, maar werd ten tijde van deze studie nog niet beoordeeld en niet betrokken bij het bepalen van het eindoordeel. Daarom sluiten we deze standaard uit van alle analyses behalve de exploratieve analyses.

Voor de sector mbo verschillen de beslisregels om van het oordeel voor een standaard tot het eindoordeel te komen en zijn de volgende standaarden beschreven in de vignetten. Naast de bovengenoemde eindoordelen kunnen in het mbo ook de eindoordelen Goed en Onvoldoende met risico op bekostigings sanctie worden gegeven.

OP0 – Basisvaardigheden  
OP2 – Ontwikkeling en begeleiding  
OP3 – Pedagogisch-didactisch handelen  
OP5 – Beroepspraktijkvorming  
VS1 – Veiligheid  
BA1 – Borging diplomering  
BA2 – Afsluiting  
OR1 – Studiesucces  
SKA2 – Uitvoering en kwaliteitscultuur

### *2.1.2 Afname studie*

De studie is in de sector po begin 2025 tegelijkertijd uitgevoerd op de drie kantoren in Zwolle, Utrecht en Tilburg. Voor de andere sectoren is de studie uitgevoerd op de locatie Utrecht. Voor (v)so en vo was het afnamemoment eind 2024 en voor mbo begin 2025. Voor de afname in alle sectoren geldt dat er een extra moment is ingelast om inspecteurs, die niet aanwezig konden zijn bij de centrale afname, de kans te geven om de vignetten alsnog te beoordelen en zo deel te nemen aan de studie.

De studie bestond uit twee fases. In de eerste fase, hierna de individuele fase genoemd, beoordeelden inspecteurs de vignetten alleen. In de tweede fase, hierna de duo-fase genoemd, gingen duo's van inspecteurs in gesprek om overeenstemming te bereiken over standaard- en eindoordelen waar ze het op basis van hun oordelen uit de individuele fase niet eens waren. De duo's begonnen hun gesprekken met vignetten waarvan hun eindoordelen verschilden.

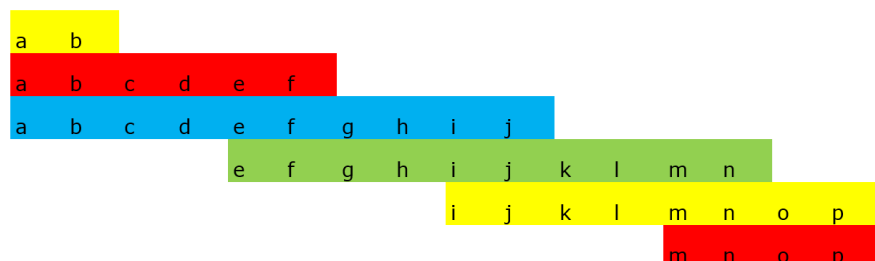


In de praktijk voeren inspecteurs kwaliteitsonderzoeken uit in duo's (of met meer inspecteurs als het een zeer grote school/afdeling/opleiding betreft). Voor het po geldt dat deze duo's in de praktijk worden samengesteld binnen hetzelfde kantoor en veelal ook binnen dezelfde afdeling. Tijdens de studie hebben we inspecteurs duo's laten vormen per kantoor en niet per afdeling, om het makkelijker te maken om duo's met dezelfde kleur boekjes te vormen. Elke inspecteur kreeg maximaal 10 van de 16 vignetten. We hebben vier boekjes met elk 10 vignetten gemaakt, die we hierna aanduiden met de kleuren rood, blauw, geel en groen. Hieronder gaan we in meer detail in op deze boekjes.

## 2.2 Steekproeftrekking

Met de resultaten doen we uitspraken over twee populaties. Ten eerste over inspecteurs, en ten tweede over scholen die in het jaar voorafgaand aan het onderzoek zijn bezocht. Wat betreft de eerste populatie hebben we geen steekproef getrokken, maar hebben we alle inspecteurs die kwaliteitsonderzoeken uitvoeren in de genoemde periode gevraagd om deel te nemen aan het onderzoek. Wat betreft inspecteurs uit de populatie die niet hebben deelgenomen aan de studie nemen we aan dat er sprake is van *missingness completely at random* omdat ze door bijvoorbeeld ziekte ontbraken. Bovendien sluiten we, zoals vastgelegd in het pre-analyseplan, inspecteurs die kortgeleden als inspecteur in dienst zijn gekomen uit van de primaire en secundaire analyses. Dat doen wij omdat zij veelal nog niet zelfstandig kwaliteitsonderzoeken uitvoeren, omdat hun opleiding mogelijk nog niet was afgerond op het moment van de studie.

Wat betreft de tweede populatie zijn per sector vignetten gemaakt voor 16 recent bezochte scholen/afdelingen/opleidingen. We hebben per sector inspecteurs bereid gevonden om deze vignetten te maken op basis van hun meest recente kwaliteitsonderzoeken. We hebben vervolgens getest hoeveel tijd nodig is om vignetten te beoordelen met vijf van de auteurs van de vignetten en drie onderzoekers. Op basis van deze test schatten we voor aanvang van de studie in dat de meeste inspecteurs in staat zouden moeten zijn om ten minste acht vignetten te beoordelen in de toegewezen tijd (1 uur). Daarom hebben we vier overlappende blokken (ook wel: boekjes) van elk tien vignetten gemaakt. Per blok zijn er dus twee vignetten extra opgenomen voor snelle werkers, die we kunnen gebruiken in exploratieve analyses. Zie de tabel hieronder. Elke letter staat voor een uniek vignet. Elke kleur staat voor een boekje met tien vignetten.



Inspecteurs zijn willekeurig toegewezen aan één van de vier blokken. De vignetten zijn ten slotte toegewezen aan deze blokken door middel van een gestratificeerd steekproefschema, dat we beschrijven in bijlage I.

## 2.3 Veranderingen ten opzichte van pre-analyseplan

We zijn bij de hoofd- en secundaire analyse op een aantal punten afgeweken van het pre-analyseplan. We lichten dit hieronder nader toe.



1. In het pre-analyseplan namen we voor om het Pa van groepen inspecteurs, cohorten en kantoren, met elkaar te vergelijken om inzicht te krijgen in de invloed van deze groepen op de oordeelsvorming. Bij nader inzien vinden we de vergelijking van Pa tussen groepen niet erg relevant. Het kan immers zijn dat de Pa 40% is tussen alle inspecteurs en voor elke groep 80%. Er zijn dan geen verschillen tussen groepen in Pa, maar het volgt ook dat er tussen groepen heel verschillend wordt geoordeeld! Voor de vraag of groepsaspecten de mate van Pa verklaren, is dus het verschil tussen Pa van de hele populatie inspecteurs en die binnen een groep relevant. We onderzoeken dit door de betrouwbaarheidsintervallen van het Pa per groep te plotten.
2. We hebben de beoordeling 'Voldoende met herstelopdracht' voor de hoofdanalyses omgecodeerd naar 'Voldoende'. De reden hiervoor is dat deze herstelopdrachten volgens de beslisregels niet tot een ander eindoordeel leiden. Uiteraard zou het Pa op standaarden lager zijn als Voldoende met herstel een aparte categorie was, maar zou de relatie met het Pa voor eindoordelen moeilijker te interpreteren zijn. We rapporteren de resultaten van de hoofdanalyse als Pa met 'Voldoende met herstelopdracht' als aparte optie in de exploratieve analyse voor de volledigheid.
3. Voor de vergelijking van Pa tussen de individuele en duo-fase vullen we de duo-oordelen aan met oordelen waarover inspecteurs het in de individuele fase eens waren. We nemen hierbij aan dat inspecteurs die het eens waren vóór de duo-fase het eens zouden zijn gebleven bij het duo-overleg. Omdat niet alle vignet-standaarden waarover inspecteurs het oneens waren zijn herbeoordeeld, zijn er missende waarden. We nemen aan dat de Pa in deze analyse hoger ligt dan als er geen waarden ontbraken. Dit is in lijn met het uitgangspunt om een bovengrens aan de overeenstemming te schatten. In het pre-analyseplan en de bijbehorende poweranalyse hebben we duo-oordelen niet op deze wijze aangevuld.



## 3 Sector PO

### 3.1 Diagnostische data-analyse

#### 3.1.1 Respons en representativiteit

We bespreken hieronder eerst de representativiteit van onze twee steekproeven van respectievelijk scholen en inspecteurs. De vraag is hierbij of ze representatief zijn voor de doelpopulaties van scholen die worden bezocht voor kwaliteitsonderzoeken en alle inspecteurs in de sector po. Daarna bespreken we de representativiteit van de *beoordeelde* vignet-standaarden ten opzichte van alle vignet-standaarden in een boekje.

Een beperking van de checks op de representativiteit is dat we alleen de verdelingen van geobserveerde kenmerken kunnen controleren.

Representativiteit op niet-geobserveerde kenmerken kan bereikt worden door randomisatie. We checken daarom ook hoe goed de randomisatie van boekjes per kantoor is gelukt. Merk op dat er bij het uitdelen van de boekjes rekening moest worden gehouden met het feit dat er duo's gevormd moesten worden die dezelfde kleur boekje moesten hebben. Per acht inspecteurs werden alle vier verschillende kleuren boekjes uitgedeeld (twee inspecteurs per kleur voor het duo-overleg). Bij een restgetal bij deling door acht werden boekjes van dezelfde kleur per tweetal inspecteurs uitgedeeld. Elk kantoor begon met uitdelen bij een andere kleur boekjes in het geval van een restgetal. In Utrecht vond de randomisatie plaats in twee verschillende zalen, hetgeen de uitschieter voor geel verklaart.

**Tabel 3.1: Aantal respondenten per kleur**

	kleur	n
Tilburg	blauw	2
	geel	4
	groen	4
	rood	5
Utrecht	blauw	9
	geel	12
	groen	8
	rood	7
Zwolle	blauw	6
	geel	4
	groen	6
	rood	6

##### 3.1.1.1 Representativiteit van scholen

Vanwege de tijd die het kost om een vignet te schrijven hebben we ons moeten beperken tot 16 vignetten. Wat betreft de selectie van beschreven scholen hebben we een strategie van quasi-randomisatie gevolgd: we hebben de inspecteurs gevraagd om hun vignetten te baseren op hun meest recente uitgevoerde (steekproef)kwaliteitsonderzoeken (KO/SKO). Als deze bezoeken gelijkmatig over het jaar verspreid zijn, en dat is het geval, dan leidt deze selectie tot dezelfde verwachte eenduidigheid. Een balanscheck van de selectie van scholen verkrijgen we door de verdeling van de historische eindoordelen van de 16 scholen waarover vignetten zijn gemaakt te vergelijken met de verdeling in alle (steekproef)kwaliteitsonderzoeken van de inspectie op scholen in de



sector po met een eindoordeel in 2023. We zien dat de verhoudingen behoorlijk goed overeenkomen.

**Tabel 3.2: Historische- en vignetoordelen in percentages**

	<b>Zeer zwak</b>	<b>Onvoldoende</b>	<b>Voldoende</b>
KO/SKO 23	18	26	55
vignetten	12	31	56
beoordeelde vignetten	13	32	56

Aangezien we vignetten maken bij situaties op een school kan er ook bias ontstaan in de schatting van Pa door de *beschrijvingen* van de schoolkwaliteit. Het kan bijvoorbeeld zo zijn dat de situatie op twee scholen even duidelijk was, maar dat het vignet van de ene school toch vager is geschreven dan het vignet van de andere school.. Een te vage omschrijving van een schoolsituatie kan dan onbedoeld leiden tot meer ruis in oordeelsvorming. Wat betreft deze zorg is onze strategie geweest om de door inspecteurs geschreven vignetten door het hele onderzoeksteam intensief te laten redigeren. Bovendien heeft een klankbordgroep, bestaande uit een strategisch inspecteur en een afdelingshoofd, zich daarna nog eens gebogen over de kwaliteit van de vignetten.

#### 3.1.1.2 Representativiteit van inspecteurs

Er zijn 89 inspecteurs die behoren tot de doelpopulatie: inspecteurs die op het moment van de studie en de periode daarvoor kwaliteitsonderzoeken hebben uitgevoerd en nog werkzaam zijn als inspecteur. Aangezien er 73 deelnamen aan de studie is de response 82%. Hieronder staat de verdeling per kantoor.

**Tabel 3.3: Aantal respondenten per kantoor**

	<b>steekproef</b>
Tilburg	15
Utrecht	36
Zwolle	22

Inspecteurs die na 1 augustus 2023 in dienst zijn gekomen, zijn zoals we eerder hebben opgemerkt uitgesloten voor de hoofdanalyse en secundaire analyse. Er hebben 8 van dergelijke inspecteurs deelgenomen aan de studie.

#### 3.1.1.3 Representativiteit van gegeven oordelen

Het aantal inspecteurs dat alle vignetten volledig heeft beoordeeld, is lager dan we hadden verwacht op basis van onze try-out. Slechts 51% van de inspecteurs heeft 8 vignetten of meer volledig beoordeeld. We houden voor de hoofdanalyses en secundaire analyses vast aan het voornemen (uit het pre-analyseplan) om per inspecteur het negende en tiende vignet weg te filteren. Hiertoe hadden we besloten teneinde geen oververtegenwoordiging van snelwerkende inspecteurs te krijgen.

Er is één inspecteur die geen enkel vignet volledig heeft ingevuld. Deze persoon heeft in geen enkel vignet OP0 gescoord. Alweer volgens het pre-analyseplan worden al deze oordelen weggefilterd. De reden om alleen volledig ingevulde vignetten mee te nemen is het oordeel op een vignet-standaard in samenhang met de andere vignet-standaarden tot stand komt. Als een inspecteur maar een deel van de vignet-standaarden heeft beschouwd zou een gebrek aan overeenstemming door dat gegeven kunnen komen in plaats van een daadwerkelijk verschil van inzicht.



**Tabel 3.4: aantal volledig ingevulde vignetten**

ingevulde_vignetten	n	cumulatieve_n	perc
10	21	21	29
9	7	28	38
8	9	37	51
7	13	50	68
6	13	63	86
5	6	69	95
4	1	70	96
3	2	72	99
0	1	73	100

Inspecteurs hebben gemiddeld 78% van de vignet-standaarden ingevuld. Hieronder geven we patronen in (non-)response van vignet-standaarden per boekje, per kantoor en per periode van indiensttreding.

**Tabel 3.5: Respons vignetten per kleur boekje**

	response (%)
blauw	81
geel	75
groen	75
rood	83

**Tabel 3.6: Respons vignetten per kantoor**

	response (%)
Tilburg	73
Utrecht	76
Zwolle	85

**Tabel 3.7: Respons vignetten per indiensttreding**

	response (%)
voor 2017	84
tussen jan 2017 en aug 2023	78
na aug 2023	82
geen antwoord	53

### 3.1.2

#### *Validiteit*

Het methodologische uitgangspunt van deze studie is dat de overeenstemming in het veld kleiner zal zijn dan op papier (dat wil zeggen: tijdens het vignetonderzoek). De onderbouwing hiervoor is theoretisch: hoe meer informatie, hoe minder overeenstemming. Deze onderbouwing veronderstelt dat tussen veld en papier de *wijze van beoordelen* gelijk is, en dat alleen de *hoeveelheid informatie* verschilt. We kunnen deze aanname iets afzwakken, omdat het niet erg is als de wijze van beoordelen verschilt, zolang deze in de vignetconditie maar niet tot minder overeenstemming leidt dan in het veld. Desalniettemin is onze onderbouwing sterker naarmate er minder aanwijzingen zijn dat de wijze van oordelen verschilt. Deze aanwijzingen onderzoeken we hieronder.

De eerste manier waarop de wijze van beoordelen kan verschillen, is doordat inspecteurs de opdracht niet serieus hebben genomen. Dit was bij de afname niet de perceptie van het onderzoeksteam. Ook het eerder genoemde feit dat slechts 51% van de deelnemers acht vignetten of meer heeft ingevuld steunt de lezing dat de vignetten met aandacht zijn ingevuld. Een andere test die we kunnen uitvoeren is om te controleren hoe vaak inspecteurs zich hebben vergist bij het toepassen van de beslisregel voor het eindoordeel. Van de 584



eindoordelen die in totaal zijn gegeven is dit 31 keer gebeurd in de individuele fase van de studie. Dat is dus in 5,5% van de gevallen. Er was hierbij één inspecteur die zich meer dan twee keer heeft vergist (5 keer). Deze inspecteur is na augustus 2023 begonnen en is dus mogelijk nog niet goed bekend met de beslisregels. Nadere analyse leert dat er vooral vergissingen werden gemaakt wanneer de standaard onderwijsresultaten OR1 als 'niet te beoordelen' werd gescoord. In dat geval gelden andere beslisregels dan gebruikelijk. Van de eindoordelen die besproken zijn in de duo-fase is in 6 gevallen afgeweken van de beslisregel. Dit is gelijk aan 3,6% van de gevallen. Deze gegevens ondersteunen naar onze mening de lezing dat de vignetten serieus zijn beoordeeld: zelfs als we nieuwe inspecteurs meenemen zijn de beslisregels over het algemeen correct toegepast.

In tabel hieronder verwijst EOS naar Eindoordeel School, is zz = Zeer zwak (alleen voor eindoordelen), o = Onvoldoende, vh = Voldoende met herstelopdracht, v = Voldoende en ntb = Niet te beoordelen.

**Tabel 3.8: Inspecteurs die afweken van de beslisregels in de individuele fase**

	OP0	OP2	OP3	VS1	OR1	SKA1	EOS	bereken d eindoor deel
o	vh	o	o	v	ntb	vh	o	ZZ
b	vh	o	o	vh	ntb	vh	o	ZZ
f	o	v	o	o	ntb	v	v	ZZ
	o	v	o	o	ntb	vh	o	ZZ
p	o	o	o	v	v	v	ZZ	o
o	vh	o	o	v	ntb	v	o	ZZ
f	vh	v	o	o	ntb	v	o	ZZ
m	vh	o	v	v	v	o	v	o
f	v	v	o	o	ntb	v	o	ZZ
p	o	o	o	v	v	v	ZZ	o
o	vh	o	o	v	ntb	v	o	ZZ
g	o	vh	vh	v	ntb	v	o	v
f	vh	vh	o	o	ntb	vh	o	ZZ
h	vh	v	vh	vh	o	v	o	ZZ
l	o	o	v	v	v	v	v	o
m	v	vh	vh	v	ntb	o	v	o
e	o	v	v	v	v	o	o	v
m	vh	o	v	v	ntb	o	o	ZZ
l	o	o	v	v	v	v	v	o
m	vh	o	vh	v	v	vh	v	o
n	v	o	v	v	v	vh	v	o
o	vh	v	o	v	ntb	vh	v	o
p	vh	o	vh	v	v	v	v	o
i	o	v	v	v	ntb	v	o	v
m	vh	o	o	v	ntb	vh	o	ZZ
o	o	vh	o	v	ntb	o	o	ZZ
b	vh	vh	o	vh	ntb	o	o	ZZ
l	o	o	vh	v	v	v	v	o
o	vh	o	o	v	ntb	v	o	ZZ
a	o	o	o	v	v	o	ZZ	o
e	vh	o	v	v	o	v	o	ZZ



**Tabel 3.9: Wijken vooral inspecteurs die recent in dienst zijn af van de beslisregels?**

	n
ja	7
nee	22
onbekend	2

**Tabel 3.10: Inspecteurs die afweken van de beslisregels na duo fase**

	OP0	OP2	OP3	VS1	OR1	SKA1	EOS	bereken d eindoor deel
j	v	o	v	v	v	v	v	o
k	o	o	v	v	ntb	o	o	zz
l	o	o	v	v	v	v	v	o
n	v	o	v	v	v	v	v	o
p	o	o	v	v	v	v	v	o
i	o	o	v	v	v	v	v	o

Een tweede manier waarop de wijze van beoordelen kan verschillen, is doordat inspecteurs strenger of milder zijn op papier dan in het veld. We brengen hier een subtiele nuance aan. Merk op dat dit met name een probleem is wanneer inspecteurs onderling verschillen in de *mate* waarin ze afwijken van hun gebruikelijke wijze van beoordelen, omdat hierdoor de mate van ruis kan worden beïnvloed. Met andere woorden: als alle inspecteurs in gelijke mate strenger worden, is het effect op de mate van overeenstemming waarschijnlijk beperkt.

We zien allereerst in tabel 3.2 dat zowel voor de oorspronkelijke eindoordelen als bij eindoordelen op basis van de vignetten in de individuele fase precies 56% voldoende was. De aantallen zijn klein en de gelijke percentages betekenen uiteraard niet dat inspecteurs in het veld even streng zouden zijn als ze op papier zijn geweest. De beschreven scholen zijn beoordeeld door de vijf auteurs van de vignetten en dus mogelijk niet representatief voor wat de invullers van de vignetten in het veld hadden geoordeeld. Bovendien is inspecteurs verteld dat ze ervan uit mochten gaan dat er geen zwaarwegende contra-indicaties zijn als informatie ontbrak. Dit geeft mogelijk een bias in de richting van mildere oordelen.

Een derde manier waarop de wijze van beoordelen kan verschillen tussen veld en papier is doordat er op papier te weinig informatie was, waardoor er willekeur is ontstaan in het oordeelsproces. Om dit te voorkomen hebben constructeurs van de vignetten en een strategisch inspecteur voor afname van de studie beoordeeld of de vignetten voldoende informatie bevatten om tot een beoordeling te kunnen komen. Zij hebben aangegeven dat dit het geval is. Een empirische aanwijzing voor hoe groot dit probleem is geweest, kan worden verkregen door te kijken naar hoe vaak het niet lukte om een voorgegeven reden te kiezen voor het verschil van beoordeling van een duo. Bij ongeveer een kwart van de besproken vignetten werd een open antwoord gegeven. Als we antwoorden voor deze categorie 'anders' bekijken, dan zien we dat één keer werd genoemd dat 'cruciale informatie ontbreekt'. Zes keer werd genoemd dat informatie 'onduidelijk' of 'multi-interpretabel' was. In één geval (dus twee aparte duo's) betrof dit dezelfde standaard uit hetzelfde vignet. Dat zijn kleine aantallen.



### 3.2 Hoofdanalyse

We maken als gezegd een onderscheid tussen drie soorten uitkomsten: hoofduitkomsten, secundaire en exploratieve. We bespreken in deze paragraaf de hoofd- en secundaire uitkomsten.

#### 3.2.1 Primaire uitkomsten

##### 3.2.1.1 Keuzes in de databewerking

Voor schattingen van overeenstemming voor individuele oordelen volgen we de stappen voor *data cleaning* die we hebben voorgenomen in het pre-analyseplan. We filteren het negende en tiende vignet per inspecteur weg (omdat we geen oververtegenwoordiging van snelle werkers willen) en verwijderen ook inspecteurs die na augustus 2023 in dienst zijn gekomen (omdat ze nog in opleiding zijn). De oordelen op de standaard OP0 worden, alweer volgens het pre-analyseplan, niet meegenomen in de hoofdanalyse.

In het pre-analyseplan hebben we ook vastgelegd dat we voor de hoofdanalyse naar de overeenstemming in de individuele fase kijken. In eerste instantie ligt het meer voor de hand om primair naar de overeenstemming in de duo-fase te kijken. De overeenstemming na duo-oordelen benadert immers het dichtst de oordeelsvorming in het veld, omdat ook in de praktijk na duo-overleg wordt geoordeeld door inspecteurs. Het probleem met deze benadering is echter dat duo's niet alle standaarden hebben beoordeeld waar ze verschillend voor oordeelden in de individuele fase.

Merk op dat we voor de verwerking van de data in de tabel *aannemen* dat als het individuele oordeel gelijk was, dit oordeel het duo-oordeel is.

**Tabel 3.11: Hoe vaak geven duo's een duo-oordeel als ze in de individuele fase beiden geoordeeld hebben?**

	individueel oordeel	duo oordeel	n	percentage
standaardoordeel	verschilt	geen duo-oordeel	57	6
	verschilt	wel duo-oordeel	98	11
	zelfde	wel duo-oordeel	755	83
eindoordeel	verschilt	geen duo-oordeel	13	7
	verschilt	wel duo-oordeel	46	25
	zelfde	wel duo-oordeel	123	68

Als we de duo-oordelen zouden willen gebruiken hebben we twee opties.

1. We beperken ons tot Pa tussen duo's voor de gegeven duo-oordelen. Deze optie zou echter waarschijnlijk tot een *onderschatting* van de totale hoeveelheid overeenstemming leiden, omdat we dan die vignet-standaarden zouden selecteren waarvoor we ook in de duo-fase de grootste verschillen in mening zouden verwachten.
2. De tweede optie is om de gegeven duo-oordelen aan te vullen met standaardoordelen waar duo's het in de individuele fase over eens waren. Er blijven dan 57 standaardoordelen over waarover duo's een verschillend individueel oordeel gaven én niet tot een duo oordeel kwamen. Dat is 6% van de vignet-standaarden waarop beide duo-partners een oordeel gaven. Inspecteurs verschilden hier van oordeel; daarom beschouwen we deze vignet-standaarden als relatief ambigu.



Door deze relatief ambigue vignet-standaarden weg te laten *overschatten* we waarschijnlijk de overeenstemming. Er kan geargumenteed worden dat dit past binnen ons uitgangspunt om de bovengrens van overeenstemming te schatten. We laten echter het nadeel dat de vertekening zo groot kan worden dat onze conclusies aan scherpste verliezen zwaarder wegen.

Wij vinden de nadelen van beide opties dermate zwaarwegend dat we ervoor hebben gekozen om van de individuele oordelen onze hoofdmaat te maken. Deze keuze stelt ons bovendien in staat om uitspraken te doen met een veel grotere betrouwbaarheid en kan deze hoofdmaat beter vergeleken worden met vergelijkbare studies bij andere inspecties. We rapporteren daarom overeenstemming op individuele oordelen als hoofduitkomst en geven de bovengrens op basis van duo-oordelen zoals beschreven in optie 2 erbij voor de interpretatie.

### 3.2.1.2 Onzekerheid

Omdat we conclusies willen trekken over zowel de populatie van inspecteurs als de populatie van scholen die beoordeeld worden, volgen we Gwet (2021) in de berekening van de standaardfout. Deze bestaat eruit dat we de gebruikelijke standaardfout bij maten van overeenstemming, die geldt voor het aantal vignetten, aanvullen met een standaardfout voor het aantal inspecteurs. Deze laatste berekenen we met behulp van de *jackknife*-methode, door dus steeds één observatie weg te laten en de variantie opnieuw te bepalen.

### 3.2.1.3 Resultaten en duiding

In de onderstaande tabel geven we het percentage overeenstemming ( $P_a$ ) weer voor zowel eindoordeelen als beoordelingen van standaarden. We bespreken de maten die corrigeren voor toeval in de exploratieve analyse.

**Tabel 3.12: Schattingen proportie agreement in de individuele fase**

	<b>est agreement</b>	<b>confidence interval 95</b>
eindoordeel	0,65	c(0,56, 0,73)
standaardoordeel	0,82	c(0,78, 0,85)

We zien dat het percentage overeenstemming bij oordelen op standaarden groter is dan voor eindoordeelen. Dit is ook wat we verwachten. De eindoordeelen zijn een functie van de oordelen op standaarden. Deze functie bestaat uit de beslisregels uit het onderzoekskader. Als bijvoorbeeld precies één van de standaarden OP2, OP3, OR1 of VS1 Onvoldoende is, dan is het eindoordeel Onvoldoende. Uit deze structuur is intuïtief te zien dat er veel manieren zijn om tot een ander eindoordeel te komen, als er voor elke standaard een kleine kans bestaat dat een afwijkend oordeel wordt gegeven. De kleine kansen voor afwijkingen per standaard tellen zo op tot grote kansen voor afwijkende eindoordeelen. We illustreren dit met een rekenvoorbeeld in het kader hieronder



Stel dat van de vier standaarden OP2, OP3, OR1 of VS1 alleen OR1 objectief Onvoldoende is. Het objectieve eindoordeel is dan ook Onvoldoende, want precies één van de vier kernstandaarden is Onvoldoende. Stel verder dat elk objectief oordeel voor een standaard 90% kans heeft om correct geïdentificeerd te worden. We nemen ook aan dat standaarden onafhankelijk van elkaar beoordeeld worden.

Er zijn nu drie mogelijk eindoordeelen: Voldoende, Onvoldoende en Zeer zwak.

De kans op een Voldoende is:

$$P(\text{alle standaarden Voldoende}) = 0,10 * 0,90^3 = 7\%$$

De kans op een Onvoldoende is:

$$P(\text{alleen OR1 Onvoldoende of OR1 Voldoende en tenminste één andere standaard Onvoldoende}) = 0,90 * 0,90^3 + 0,10 * (1 - 0,90^3) = 68\%$$

De kans op Zeer zwak is:

$$P(\text{OR1 Onvoldoende en tenminste één van overige drie Onvoldoende}) = 0,90 * (1 - 0,90^3) = 24\%$$

Als we niet afronden tellen deze kansen op tot 100%. De kans op het – correcte – eindoordeel Onvoldoende is dus slechts 68%. Met andere woorden: hoewel elke standaard 90% kans heeft om correct geïdentificeerd te worden, is de kans op een correct eindoordeel slechts 68%.

#### 3.2.1.4 Bovengrens duo-oordelen

Zoals eerder aangekondigd berekenen we een (extra hoge) bovengrens aan het Pa voor de *duo-oordelen*.<sup>6</sup> We herhalen hiertoe twee belangrijke punten.

1. We kiezen ervoor om de duo-oordelen aan te vullen met individuele oordelen als duo-partners het eens waren in de individuele fase.
2. Er ontbreken standaardoordelen waar duo-partners het in de individuele fase oneens waren, maar die ze niet her-beoordeeld hebben

Om een (waarschijnlijk extra hoge) bovengrens aan de overeenstemming te schatten doen we de volgende aanname:

als inspecteurs de onder 2. genoemde vignet-standaarden *we/* zouden herbeoordelen, dan zou de overeenstemming voor deze (relatief ambigue) vignet-standaarden niet hoger zijn dan voor de overige vignet-standaarden

Deze analyse levert de volgende schatting van de bovengrens van Pa voor duo-oordelen op.

<sup>6</sup> Er zijn duo's gevormd binnen kantoren en niet binnen teams, zoals gebruikelijk is in de praktijk. Als inspecteurs binnen teams meer op elkaar lijken, dan zouden ze het sneller met elkaar eens worden als duo's worden gevormd binnen teams dan binnen kantoren. Maar merk op dat we niet Pa binnen duo's, maar Pa tussen duo's meten. Als je meer gelijkgestemde inspecteurs duo's laat vormen zullen de verschillen van oordelen juist minder afnemen door de duo-vorming, en zal de Pa dus naar verwachting lager zijn. Ook in dit opzicht zal de Pa van duo's (gevormd binnen kantoren en niet binnen teams) dus een bovengrens zijn.



**Tabel 3.13: Schattingen Pa aangevulde duo-oordelen**

	<b>est agreement</b>	<b>confidence interval 95</b>
eindoordelen	0,76	c(0,63, 0,88)
standaarden	0,89	c(0,84, 0,93)

### 3.2.1.5 Resultaten in relatie tot andere onderzoeken

Om de betekenis van de gevonden waarde voor Pa nader te duiden vergelijken we deze met studies bij andere inspecties. Hieronder vallen inspecties van het onderwijs in andere landen en inspecties voor aspecten van gezondheidszorg. In deze studies is steeds de Pa tussen inspecteurs (en niet tussen duo's) gemeten op standaardniveau. In bijlage III plaatsen we de relevante waarde uit onze studie (82%) tussen de gevonden waarden van andere inspecties. We zien dan dat de Pa er niet uit springt. De hogere waarden van Pa (oranje bolletjes) zijn gevonden in studies in Nederland en de Verenigde Staten, waarin onderwijsinspecteurs samen een school hebben bezocht, en waarbij inspecteurs samen de voedselveiligheid in restaurants onderzochten. Er zijn redenen om bij deze studies te twijfelen aan de onafhankelijkheid van de oordelen.

### 3.2.2 *Secundaire uitkomsten*

In de secundaire analyse proberen we te duiden waar de verschillen in oordelen vandaan komen. We beantwoorden hier de *secundaire onderzoeksvragen*:

1. Hoe groot is het verschil in overeenstemming tussen individuele oordelen en oordelen van duo's? (zie 3.2.2.1)
2. Is de overeenstemming groter binnen kantoren en binnen cohorten inspecteurs? (zie 3.2.2.2)
3. Welke redenen noemen duo's inspecteurs voor verschillen in individuele oordelen? (zie 3.2.2.3)

#### 3.2.2.1 Pre en post-duo

Als we de oordelen van inspecteurs uit de duo-fase vergelijken met de oordelen in de individuele fase, dan is het niet vanzelfsprekend dat er meer overeenstemming is in de duo-fase. Het zou immers kunnen dat de oordelen van inspecteurs in de duo-fase worden toegetrokken naar de oordelen van de meest extreme inspecteurs. In dat geval zal de overeenstemming tussen de duo's op de beoordeelde vignetten juist afnemen. Maar als inspecteurs door discussie naar het meerderheidsoordeel tenderen (of als inspecteurs met meerderheidsoordelen meer invloed hebben), dan zal de overeenstemming tussen duo's juist toenemen.

Het berekenen van deze toename is niet eenvoudig. We kunnen bijvoorbeeld geen duo-oordelen voor en na het overleg vergelijken. Er zijn immers geen duo-oordelen vóór de overlegfase. We kunnen uit de individuele oordelen van de toekomstige duo's uiteraard ook geen pre-duo-oordelen construeren, omdat latere duo-partners van oordeel kunnen verschillen. We moeten dus voor de pre-conditie de overeenstemming tussen *individuele inspecteurs* nemen en in de post-conditie de overeenstemming tussen *duo's*. En deze beide Pa met elkaar vergelijken.

Voor een zo eerlijk mogelijke vergelijking moeten we er verder rekening mee houden dat niet alle vignetten in gelijke mate beoordeeld zijn in de individuele en duo-fase. We bespraken dit eerder en gaven toen twee opties aan. De eerste is om ons te beperken tot de daadwerkelijk gegeven duo-oordelen en de tweede om deze aan te vullen met oordelen waarover inspecteurs het al eens waren. Ook hier kiezen we voor de tweede optie. (De resultaten voor de eerste optie zijn terug te vinden in de exploratieve analyses.)



We testen of er een verschil is tussen Pa in de individuele en duo-fase met een gepaarde t-test, waarbij we verschillen in Pa binnen vignetten (of vignet-standaarden) onderzoeken. Voor deze test conditioneren we dus op de deelnemende inspecteurs. Dat wil zeggen dat de test de steekproefonzekerheid van de in vignetten beschreven scholen reflecteert, maar niet de steekproefonzekerheid van de inspecteurs. Het aantal vrijheidsgraden is dan ook: het aantal vignetten – 1 (of het aantal vignet-standaarden – 1, bij standaarden). In de tabel hieronder zijn deze aantallen weergegeven in de kolom 'n vign'.

**Tabel 3.14: Verschil in oordelen tussen individuele en duo fase met aangevulde oordelen**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,67	0,76	0,08	1,08	0,30	16,00	60,00	31,00
standaarden	paired t test	0,85	0,89	0,04	1,49	0,14	80,00	60,00	31,00

We zien allereerst dat het aantal *beoordelaars* in de duo-fase 31 was (een duo geldt hier als een *beoordelaar*). Het aantal *beoordelaars* in de individuele fase (60) is ook lager dan het aantal deelnemende inspecteurs dat aan de criteria voldoet (65). Dit komt doordat niet iedere inspecteur in de duo-fase een partner heeft kunnen vinden die voor augustus 2023 is begonnen. Verder zien we dat het Pa nu in de individuele fase hoger is dan in de primaire analyse. Dit komt door de missende data. De individuele oordelen die ontbreken zijn gevallen waar duo-partners het oneens waren, maar die niet zijn besproken in de duo-fase. We mogen verwachten dat deze vignet-standaarden relatief ambigu zijn, aangezien men het daarover oneens was. Als we deze oordelen weglaten is het dus niet zo gek dat de Pa hoger ligt.

Het verschil in Pa voor standaarden tussen beide fases is niet statistisch significant (verschil is 0,04 met p-waarde 0,14 en  $df = 79$ ). Het verschil voor eindoordeelen behoort niet tot de secundair analyse, maar voegen we toe voor de volledigheid. Ook voor deze maat is het verschil niet statistisch significant.

### 3.2.2.2 Vergelijkingen van overeenstemming voor groepen inspecteurs

Inspecteurs maken onderdeel uit van groepen, die mogelijk invloed hebben op hoe ze oordelen. Omdat we ordinale uitkomsten meten is er geen eenduidige manier om variantie toe te schrijven tussen en binnen groepen. Maar als de overeenstemming binnen groepen hoger is dan tussen individuele inspecteurs, dan geeft dat aanleiding om te denken dat een deel van de ruis wordt verklaard doordat groepen inspecteurs van elkaar verschillen in hoe ze oordelen.

We onderscheiden twee relevante groepen inspecteurs. De eerste groepsdefinitie is die op basis van het kantoor waar inspecteurs werken. Het zou kunnen dat inspecteurs binnen een kantoor een hogere overeenstemming hebben, omdat ze vaker met elkaar overleggen over de werkwijze. Bovendien worden inspecteurs binnen een afdeling van een kantoor vooral aan elkaar gekoppeld bij kwaliteitsonderzoeken. Er is hierbij geen sprake van vaste koppels, maar juist van veel roulatie. Hierdoor kunnen inspecteurs elkaar ook beïnvloeden wat betreft de wijze van beoordelen. Er zijn dus redenen om te denken dat inspecteurs binnen kantoren meer zullen overeenstemmen dan als we inspecteurs landelijk met elkaar vergelijken.

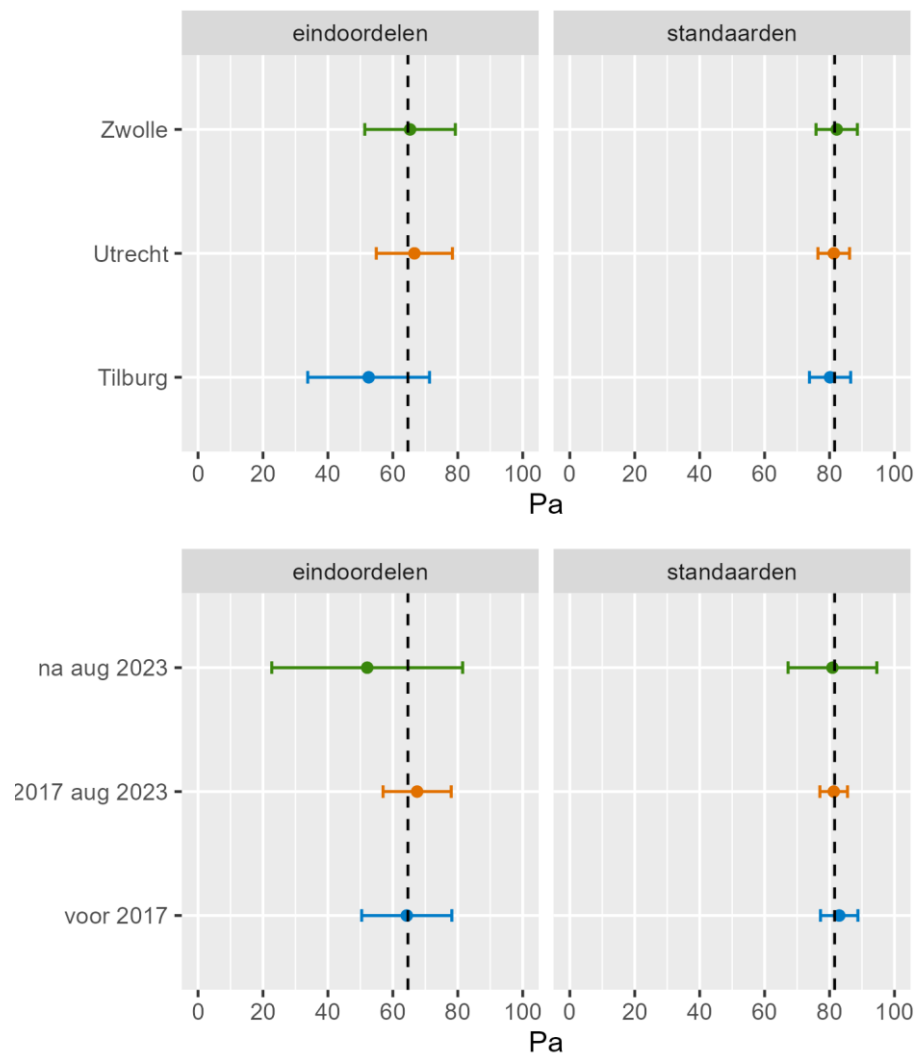
De tweede groepsdefinitie is op basis van het moment van indiensttreding als inspecteur. We hebben periodes opgedeeld met als cesuur de



ingrijpende wijziging van het Onderzoekskader in 2017. Inspecteurs die voor deze periode in dienst zijn getreden zijn opgeleid om meer gedetailleerde criteria te gebruiken bij de beoordeling van scholen. Hierdoor lijken ze mogelijk meer op elkaar dan op inspecteurs die na 2017 in dienst zijn getreden.

In de figuur hieronder zijn de individuele oordelen van inspecteurs weergegeven, om een betrouwbaarder beeld te krijgen. Op basis van deze oordelen is steeds binnen een groep het Pa berekend. De stippellijn geeft het Pa tussen alle inspecteurs weer, en is toegevoegd als referentie. We zien geen sterke afwijkingen in de mate van overeenstemming binnen groepen ten opzichte van een analyse waarin Pa binnen groepen en tussen groepen is gecombineerd (het Pa van de stippellijn). Ter informatie nemen we hier voor de cohorten ook de groep inspecteurs mee die na augustus 2023 zijn begonnen.

**Figuur 3.1: Pa binnen kantoren en binnen cohorten. De stippellijn geeft de Pa van alle inspecteurs weer**





### 3.2.2.3 Redenen voor gebrek aan overeenstemming

Hieronder geven we weer hoe vaak een reden is gegeven voor een verschil in beoordeling van een standaard. Deze redenen zijn door het onderzoeksteam opgesteld in samenwerking met een strategisch inspecteur. We geven hier alleen de redenen weer voor verschillen in de standaardoordelen Onvoldoende, Voldoende en Goed, omdat we voor de hoofdanalyse de standaardoordelen 'Voldoende met herstelopdracht' hebben weggelaten. In de paragraaf met alternatieve specificaties laten we zien welke redenen werden gegeven als we de oordelen 'Voldoende met herstelopdracht' als aparte categorie beschouwen.

Er zijn 97 duo-oordelen op standaarden gegeven door de 31 duo's. Hierbij zijn nieuwe inspecteurs en OP0 uiteraard niet meegenomen. Inspecteurs mochten meerdere redenen geven als verklaring voor een verschil in oordeel op een bepaalde standaard. Daardoor kan dezelfde vignet-standaard bij meerdere redenen ondergebracht zijn.

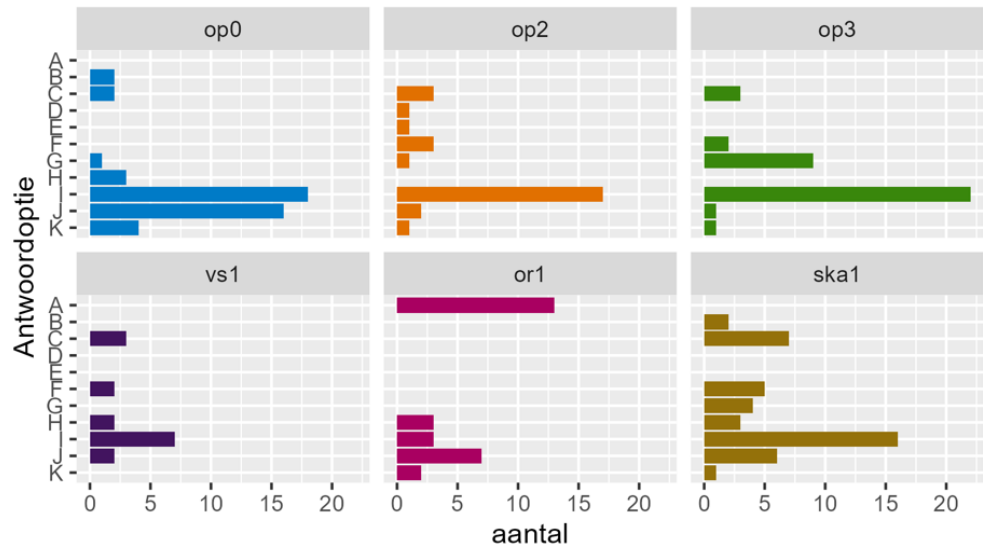
**Tabel 3.15: Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OP0 Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer (%) <sup>a</sup>
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) N.B.: 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.



**Figuur 3.2: Redenen voor verschillend oordeel per standaard**



- 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

### 3.3 Exploratieve analyse

#### 3.3.1 Verdiepende analyse van verschillen in eendoordelen

##### 3.3.1.1 Uitsplitsing van verschillen in eendoordelen

Voor de berekening van de Pa worden inspecteurs gekoppeld aan alle andere inspecteurs die hetzelfde vignet hebben beoordeeld. In de onderstaande tabel laten we zien welke mogelijke eendoordelen alle mogelijke koppels van twee inspecteurs hebben gegeven: Voldoende (v), Onvoldoende (o) en Zeer zwak (zz). Op deze manier kunnen we bijvoorbeeld zien dat inspecteurs het in 80% (64,9% + 13,6%) van de gevallen eens zijn over de grens tussen een positief (v) en negatief (o of zz) eendoordeel. Verder valt op dat bij 2,1% van de standaarden de ene inspecteur een zeer zwak en de andere een voldoende geeft.



**Tabel 3.16: Alle vergelijkingen van soorten eindoordelen van alle mogelijke koppels van individuele inspecteurs**

	<b>percentage</b>
zelfde	64,9
o en zz	13,6
o en v	19,4
v en zz	2,1

We kunnen hetzelfde doen door naar de oordelen in de duo-fase te kijken. Hierbij nemen we weer aan dat als duo's in de individuele fase hetzelfde oordeel hebben gegeven, dit oordeel ook hun gemeenschappelijke oordeel zou zijn geweest. Aangezien de vignetten die ontbreken in deze analyse waarschijnlijk relatief ambigue waren – het individuele oordeel hierover verschilde, is dit dus weer een extra hoge bovengrens voor de overeenstemming. Nu is 91% (75,5% + 15%) het eens over de grens tussen een negatief (zz of o) en een positief (v) eindoordeel. Het is geruststellend om te zien dat de gesprekken in de duo-fase leiden tot een afname van situaties waarin inspecteurs tot een zeer verschillend eindoordeel komen. Zo neemt het percentage gevallen waarin zowel een voldoende als zeer zwak eindoordeel voor hetzelfde vignet wordt gegeven af van 2,1% in de individuele fase naar 0,4% in de duo-fase.

**Tabel 3.17: Alle vergelijkingen van soorten eindoordelen van alle mogelijke koppels van duo's**

	<b>percentage</b>
zelfde	75,5
o en zz	15,0
o en v	9,0
v en zz	0,4

Hieronder herhalen we deze analyse voor standaardoordelen, waarbij we eerst de individuele fase van de studie en vervolgens de duo-fase bespreken. o = Onvoldoende, v = Voldoende, g = Goed, ntb = Niet te beoordelen (enkel betrekking op OR1).

**Tabel 3.18: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van individuele inspecteurs**

	<b>percentage</b>
zelfde	81,6
o en v	13,5
v en ntb	4,3
v en g	0,5
o en ntb	0,1
o en g	0,0
g en ntb	0,0

**Tabel 3.19: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van duo's**

	<b>percentage</b>
zelfde	88,5
o en v	10,2
v en ntb	0,8
v en g	0,5
o en g	0,0
o en ntb	0,0
g en ntb	0,0



### 3.3.1.2 Bovengrens van het percentage gelijke eindoordelen

Tot nu toe hebben we onderzocht hoe vaak inspecteurs het met elkaar eens zijn. We kunnen ons ook afvragen hoe vaak inspecteurs het correcte oordeel geven. Aangezien we niet weten welk eindoordeel voor een bepaald vignet correct is, benaderen we deze vraag door eerst een andere vraag te stellen: of een inspecteur (of duo) bij een vignet het meerderheidsoordeel heeft gekozen. Er zijn nu twee mogelijkheden. Als we aannemen dat het meest gegeven oordeel het correcte oordeel is, dan kunnen we het percentage correcte oordelen bepalen. En als we aannemen dat het vaakst gekozen oordeel *niet* het correcte oordeel is, dan zal het werkelijke percentage correcte oordelen altijd lager zijn. Met andere woorden: het percentage meest gekozen oordelen is een bovengrens voor het percentage correcte oordelen. We geven het percentage inspecteurs dat per vignet het meerderheidsoordeel heeft gegeven voor eindoordelen en standaardoordelen in de tabellen hieronder.

**Tabel 3.20: Percentage inspecteurs dat het meerderheidsoordeel geeft voor eindoordelen**

	<b>schatting</b>	<b>standaardfout</b>
individueel	76,0	4,0
duo bovengrens	82,9	4,8

**Tabel 3.21: Percentage inspecteurs dat het meerderheidsoordeel geeft voor oordelen bij standaarden**

	<b>schatting</b>	<b>standaardfout</b>
individueel	88,0	1,5
duo bovengrens	92,9	1,4

We herhalen hier nog eens in welke zin dit een bovengrens is. Het daadwerkelijke percentage correcte eindoordelen kan lager zijn, omdat het meest gegeven oordeel niet altijd het correcte oordeel hoeft te zijn, omdat ontbrekende duo-oordelen de overeenstemming waarschijnlijk groter maken en omdat een vignetstudie de overeenstemming waarschijnlijk overschat omdat er minder informatie beschikbaar is voor de beoordelaar.

### 3.3.2 *Pa per standaard*

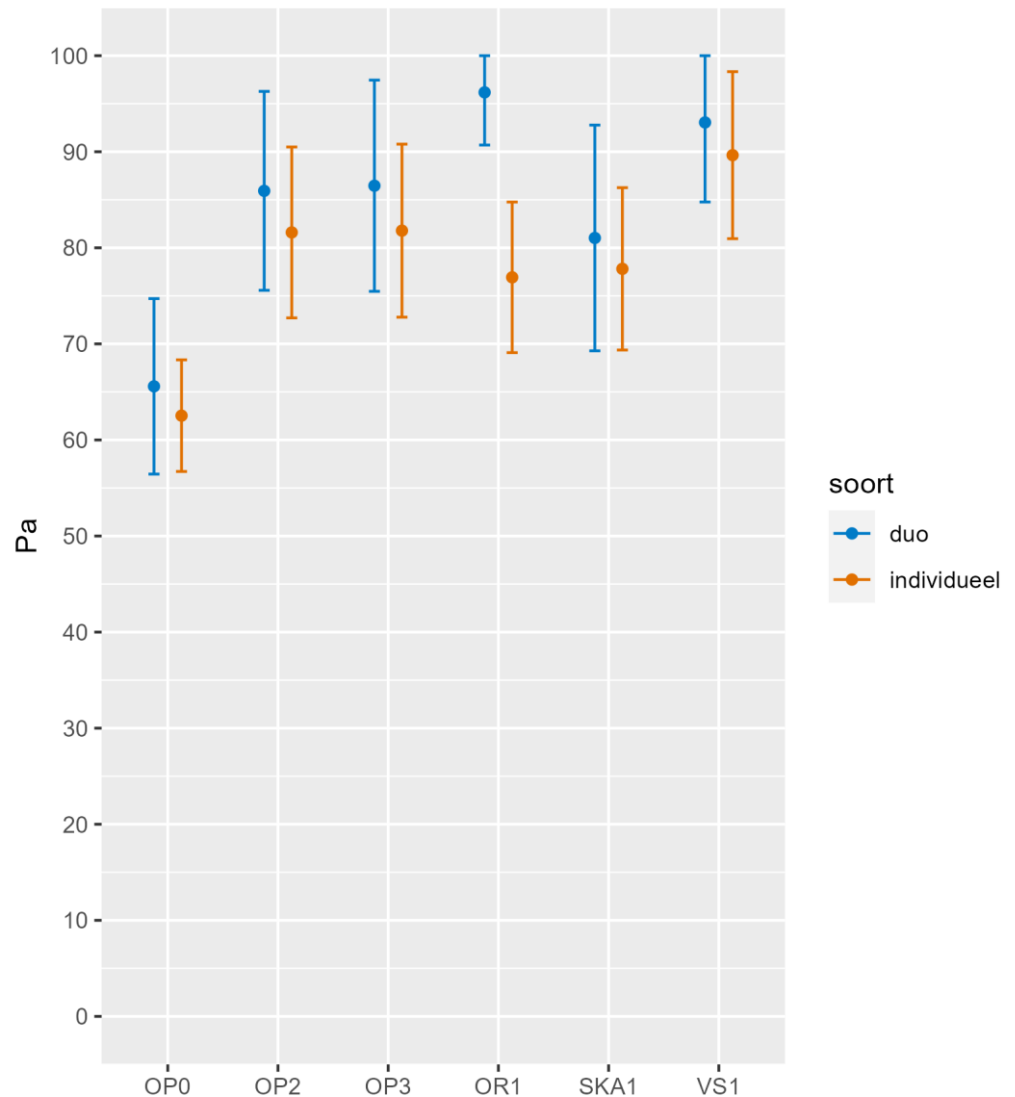
We kunnen  $P_a$  ook berekenen voor elk van de zes standaarden afzonderlijk. Hierbij moet worden opgemerkt dat niet iedere standaard zich even goed leent voor een vignetstudie. De standaard OP3 Pedagogisch-didactisch handelen, waarbij de inspecteur in de praktijk lesobservaties uitvoert, is nu eenmaal moeilijker te vangen in tekst dan de standaarden VS1 Veiligheid en OR1 Resultaten. Het zou daardoor kunnen dat we de overeenstemming voor OP3 sterker overschatten dan de overeenstemming voor VS1 en OR1. (Maar let wel: alleen *verschillen* in overschatting tussen standaarden zijn een probleem voor onze onderzoeksopzet, niet de overschatting op zich.)

De lage mate van overeenstemming voor de standaard OP0 Basisvaardigheden valt op. Dit is een nieuwe standaard, die in de praktijk ook nog niet wordt beoordeeld en waar inspecteurs nog weinig ervaring mee hebben. Wel worden er herstelopdrachten voor gegeven. Ook de relatief lage overeenstemming voor OR1 Resultaten in de individuele fase valt op, omdat heldere beslisregels aanwezig zijn voor de beoordeling van deze standaard. Bij de redenen voor een verschil van oordeel werd bij OR1 het vaakst opgemerkt: "alleen bij OR1 beslisregels anders toegepast". Dit betekent dat één van de inspecteurs zich heeft vergist bij het toepassen van de beslisregels in de individuele fase. In de figuur is te zien dat deze vergissingen in de duo-fase



veelal zijn gecorrigeerd en dat de mate van overstemming op deze standaard zoals verwacht het hoogst is.

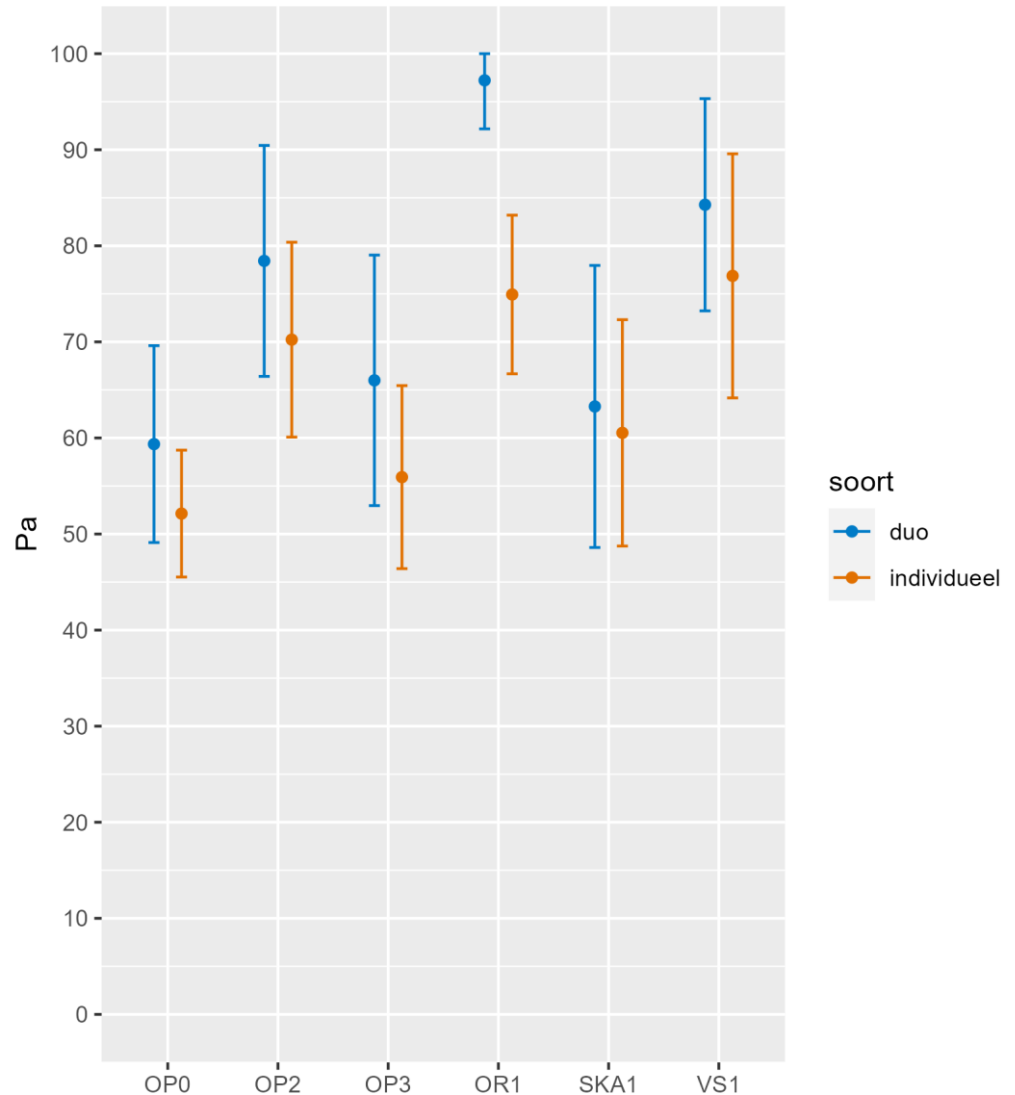
**Figuur 3.3: Pa van oordelen per standaard: Voldoende met herstelopdracht is gecodeerd als Voldoende**



We geven de Pa per standaard ook weer in de situatie waarin een Voldoende met herstelopdracht een aparte categorie is. Het is niet verwonderlijk dat de Pa lager is in deze situatie, omdat er meer manieren zijn waarop twee inspecteurs het niet eens kunnen zijn.



**Figuur 3.4: Pa van oordelen per standaard: Voldoende met herstelopdracht is een aparte categorie**



### 3.3.3 *Strengheid per inspecteur*

Verschillen in oordelen kunnen ook in termen van verschillen in strengheid van inspecteurs worden gezien. Met de strengheid van een inspecteur bedoelen we hoeveel procent van de oordelen voor standaarden als Onvoldoende is beoordeeld (of zou zijn beoordeeld). Strengheid is een *eigenschap van* een inspecteur, en daarmee een wezenlijk andere maat dan Pa, die bestaat *tussen* inspecteurs. Omdat we deze maat per inspecteur meten is de onzekerheid veel groter. Zeker als er weinig oordelen per inspecteur zijn gemeten hebben uitschieters een groot effect en zullen de schattingen van strengheid verder uit elkaar liggen in de steekproef dan in de populatie.

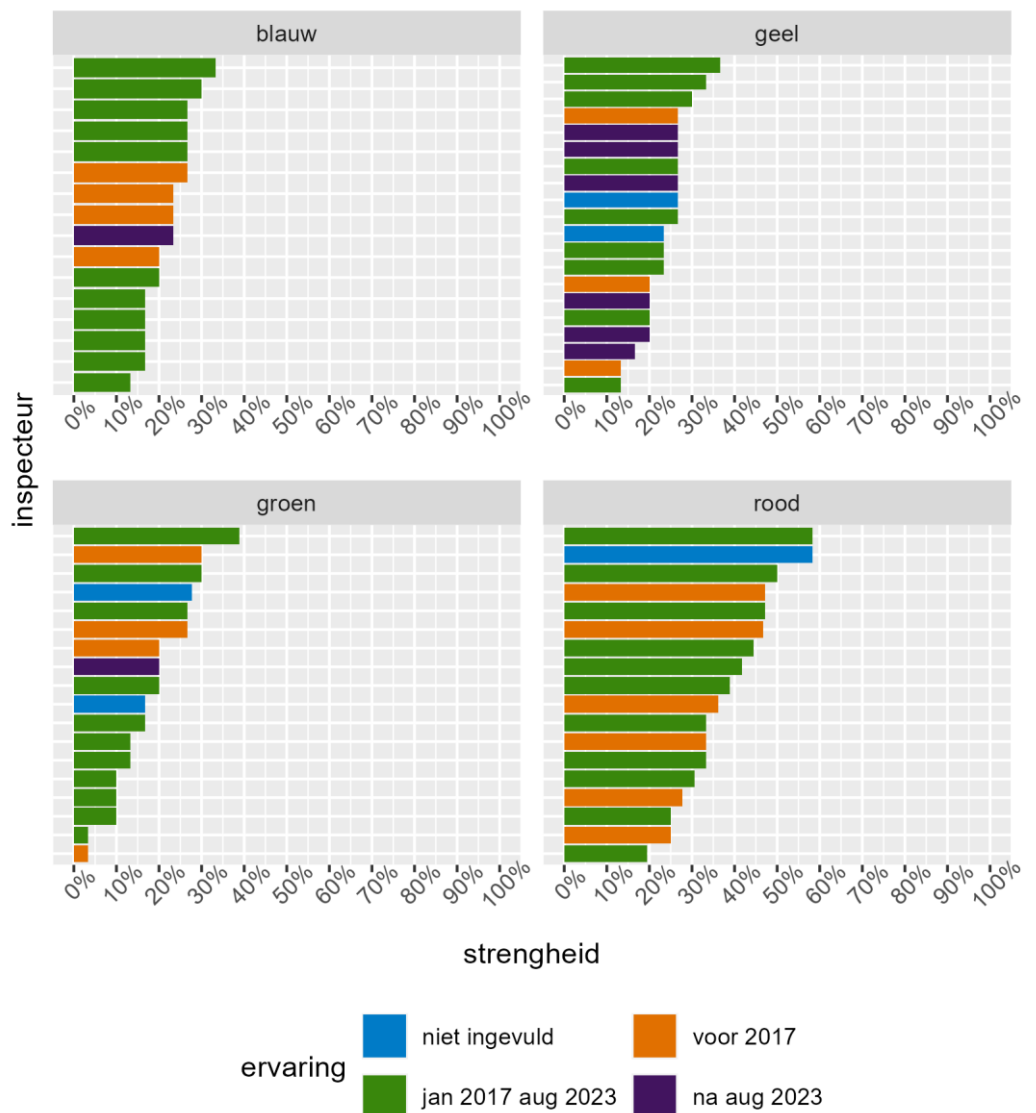
Er is slechts een beperkt aantal observaties per inspecteur, dus om de betrouwbaarheid zo groot mogelijk te maken gebruiken we oordelen op standaarden. We nemen hier ook OP0 en de antwoorden van snel werkende inspecteurs mee om, alweer, een zo betrouwbaar mogelijk beeld te krijgen.



### 3.3.3.1 De strengheid binnen boekjes

Om een eerste indruk te krijgen, vergelijken we de strengheid van alle inspecteurs die hetzelfde boekje hebben gekregen en dus ook dezelfde vignetten hebben beoordeeld. Omdat niet iedereen even snel werkte hebben we in deze stap de standaardoordelen op de eerste 5 vignetten genomen en inspecteurs weggefilterd die minder dan 5 vignetten volledig hebben beoordeeld.

**Figuur 3.5: Strengheid op standaarden per kleur boekje**



Alleen de eerste 5 vignetten zijn meegenomen. Inclusief OP0

We zien in de bovenstaande figuur ten eerste dat het uitmaakt welke scholen een inspecteur ziet voor de strengheid die we op naïeve wijze aan hem of haar zouden toeschrijven. Scholen in het rode boekje werden aanmerkelijk strenger beoordeeld dan scholen die beschreven werden in de andere boekjes. De mildste inspecteur die het rode boekje kreeg gaf bijna 20% van de



standaardbeschrijvingen een negatief oordeel tegenover bijna 60% voor de strengste.

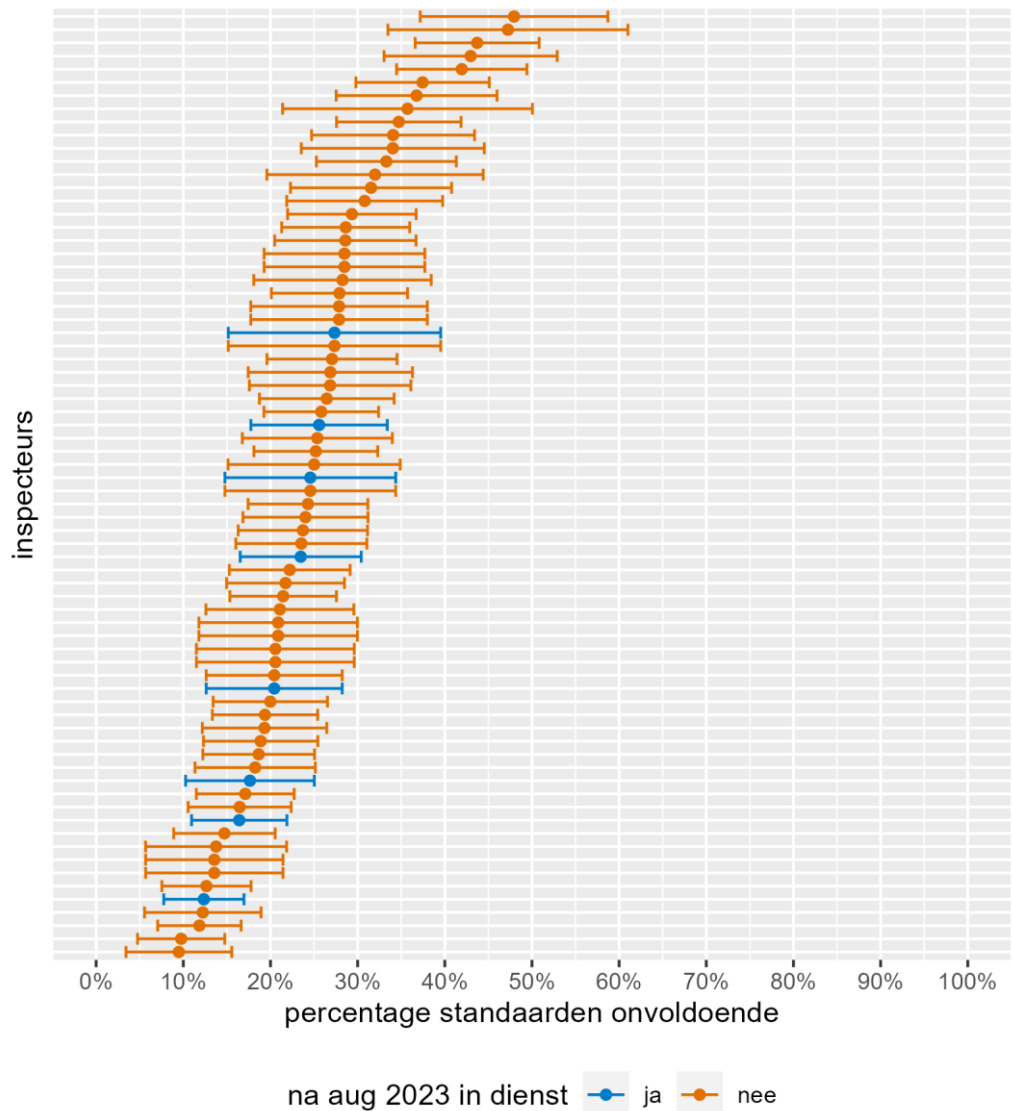
In deze grafiek gooien we echter veel informatie weg. Bovendien kunnen we inspecteurs tussen kleuren niet vergelijken, omdat de strengheidscore afhangt van welke vignetten je te zien krijgt. We kunnen op twee manieren alle gegeven oordelen voor de analyse benutten en tegelijk de strengheidsscores op dezelfde schaal krijgen. De eerste manier is met een regressiemodel. De tweede manier is door gebruik te maken van imputatie.

### 3.3.3.2 Strengheid volgens een regressiemodel

In het regressiemodel corrigeren we voor het feit dat sommige inspecteurs bijvoorbeeld vaker scholen met een lage kwaliteit beoordeelden door te corrigeren voor de gemiddelde score op de betreffende vignet-standaarden. Het model schat de kans per vignet-standaard op een onvoldoende en de kans op een onvoldoende per inspecteur. De coëfficiënt voor de inspecteursstrengheid in het model is echter van toepassing op één van de vignet-standaarden. Om de gemiddelde strengheid van een inspecteur te verkrijgen berekenen we op basis van het regressiemodel vervolgens de verwachte score voor elke vignet-standaard van een inspecteur. Door al deze scores te middelen krijgen we de verwachte strengheid over alle 96 vignet-standaarden. Merk op dat de strengheid van de inspecteur dus is geschat op basis van de vignetten die hij of zij heeft beoordeeld, gecorrigeerd voor de gemiddelde score van die vignetten. We berekenen deze gemiddelde scores met het R *emmeans* package, dat ons de bijbehorende standaardfouten geeft. We kunnen op deze wijze dus ook de onzekerheid van onze schattingen uitdrukken.



**Figuur 3.6: Gemiddelde strengheid per inspecteur: op basis van een logistisch regressiemodel**



inclusief OP0 en alle gescoorde vignet-standaarden

Bij de interpretatie van deze grafiek moet er rekening mee worden gehouden dat als we meer observaties zouden hebben per inspecteur, de uitschieters naar verwachting naar het gemiddelde zouden opschuiven.

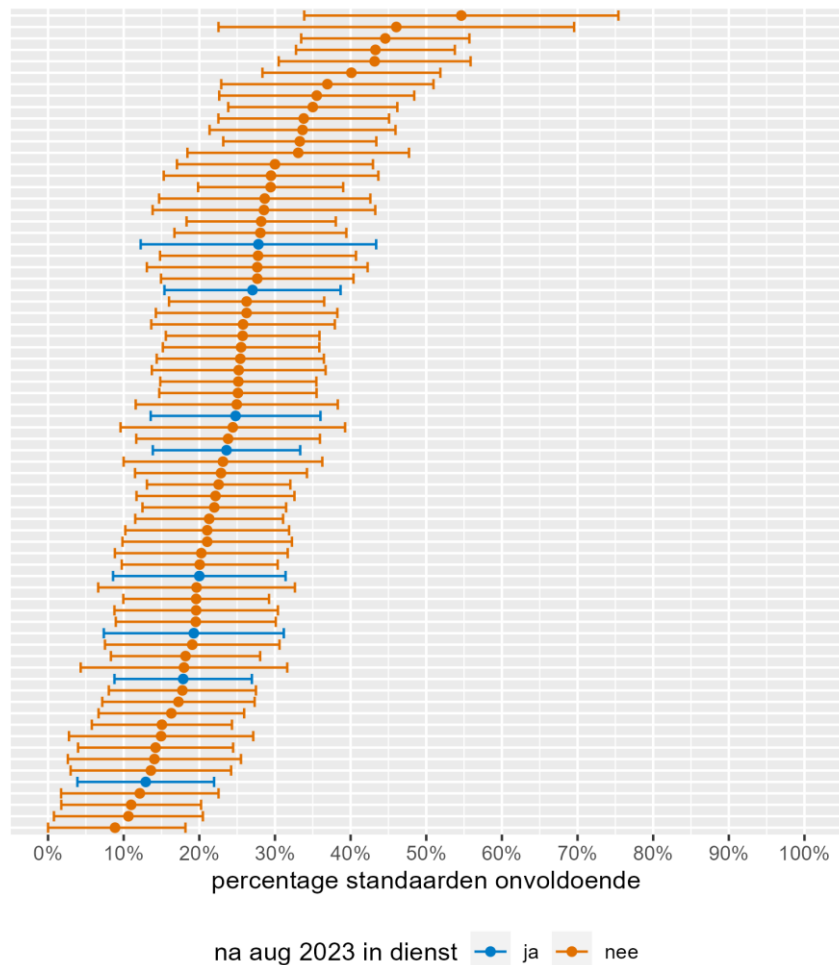
Met het regressiemodel schatten we hoe een inspecteur had geoordeeld op alle vignet-standaarden. De aanname die we daarbij doen is dat een inspecteur voor niet-geobserveerde vignet-standaarden net zoveel verschil in strengheid als het gemiddelde van alle inspecteurs. De betrouwbaarheidsintervallen verschillen in breedte vanwege zowel de grootte van de puntschatting als het gegeven dat inspecteurs verschillende aantallen vignet-standaarden hebben beoordeeld tijdens de afname.



### 3.3.3.3 Strengheid volgens imputatie

Een alternatieve manier om de strengheid van inspecteurs te schatten is door de niet beoordeelde vignet-standaarden te imputeren. Dat wil zeggen dat we schatten hoe een inspecteur had geoordeeld op alle 96 vignet-standaarden, door gebruik te maken van informatie van inspecteurs met overlappende oordelen die de ontbrekende vignet-standaard wel hebben beoordeeld. We hebben hiertoe met het R *mice* package de methode logistische regressie gebruikt en twintig geïmputeerde datasets gemaakt. De schattingen verschillen per imputatieronde doordat voor elke ronde de missende waarden initieel worden ingevuld met random trekkingen uit de populatie van waarden uit een kolom. Deze waarden worden gedurende meerdere iteratie vervangen door schattingen op basis van regressies waarbij informatie uit andere kolommen wordt gebruikt. De betrouwbaarheidsintervallen in de onderstaande grafiek reflecteren nu niet alleen meer de steekproeffout, maar ook de imputatie-onzekerheid. Een breder betrouwbaarheidsinterval geeft in dit geval dus ook weer dat er meer onzekerheid is over de imputatie voor een bepaalde inspecteur (de geïmputeerde schattingen liepen voor deze inspecteur dus meer uiteen).

**Figuur 3.7: Gemiddelde strengheid per inspecteur: na imputatie**

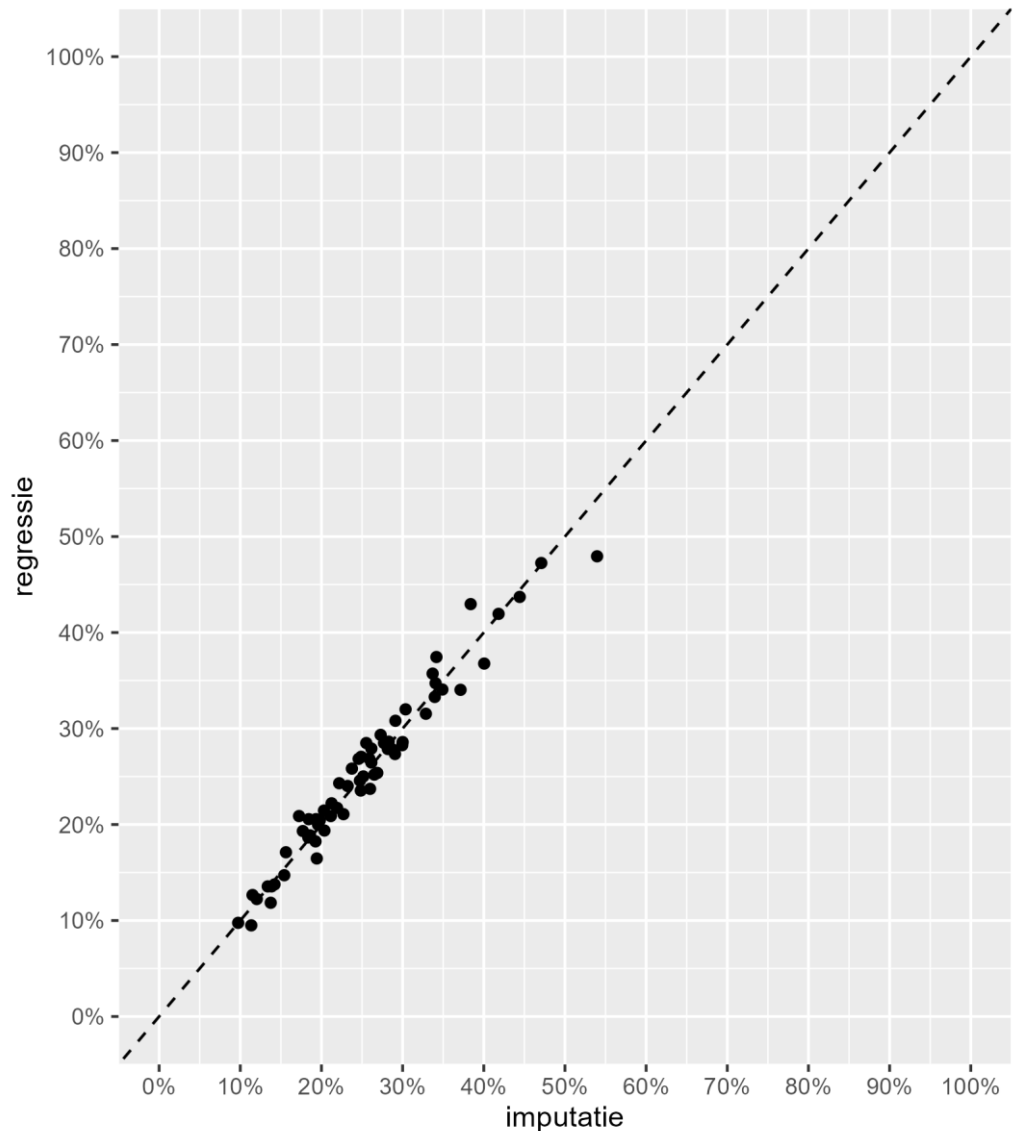


inclusief OP0 en alle gescoorde vignet-standaarden



Om de verschillen tussen het regressiemodel en de imputatie te vergelijken, plotten we de schattingen van beide modellen in dezelfde grafiek. De correlatie tussen beide methodes is met 0,98 zeer sterk. Dit geeft vertrouwen in de puntschattingen.

**Figuur 3.8: Vergelijking imputatie en regressie: De correlatie is 0.98**



Op basis van de imputatie kunnen we op een eerlijke manier verschillen in strengheid tussen kantoren in beeld brengen. In de tabel hieronder is te zien dat er geen aanwijzingen zijn voor grote verschillen.

**Tabel 3.22: Verschillen in strengheid tussen kantoren**

	perc Onvoldoende	standaardfout	95% BI
Tilburg	25,7	1,5	(22,7, 28,7)
Utrecht	24,2	1,1	(22,1, 26,3)
Zwolle	25,2	1,2	(22,8, 27,6)



In de onderstaande tabel is te zien dat hetzelfde geldt voor verschillen tussen cohorten inspecteurs.

**Tabel 3.23: Verschillen in strengheid tussen cohorten**

	perc Onvoldoende	standaardfout	95% BI
niet ingevuld	32,3	2,8	(26,7, 37,9)
voor 2017	24,6	1,4	(22,0, 27,3)
jan 2017 aug 2023	24,7	0,9	(22,9, 26,5)
na aug 2023	21,7	1,8	(18,1, 25,3)

### 3.3.4 *Alternatieve specificaties*

In bovenstaande analyse zijn veel keuzes gemaakt, die we voor het grootste deel hebben vastgelegd in een pre-analyseplan. Voor de volledigheid geven we in deze paragraaf de resultaten voor een beperkt aantal alternatieve keuzes.

#### 3.3.4.1 Toevalscorrectie

We hebben er voor gekozen om in onze hoofduitkomsten niet te corrigeren voor toevalsovereenstemming, omdat de methodes die dit beogen te doen, geplaagd worden door paradoxen en omdat ze er niet in slagen om op een geloofwaardige manier een rol aan toeval toe te schrijven. De intuïtie voor het wel toepassen van deze correctie is dat beoordelaars die verschillende impliciete beslisregels volgen, het op ordinale of nominale maten vaak eens zullen zijn, omdat ze om verschillende redenen op hetzelfde oordeel uitkomen. (Voor continue maten speelt dit probleem niet, omdat bij een gedetailleerde schaal de kans op dezelfde uitkomst verwaarloosbaar klein is.) Het probleem is dus echter dat ze toeval niet goed kunnen modelleren.

Wij bespreken hier twee methodes, Fleiss Kappa en Gwet's AC1, omdat dit respectievelijk een veelgebruikte en robuuste maat is voor overeenstemming. AC1 is robuust in de zin dat de maat relatief immuun is voor paradoxen. Bij Fleiss Kappa wordt per categorie bepaald wat de kans is dat deze categorie wordt gekozen. Deze kans is de frequentie van een categorie over alle vignetten en beoordelaars. De kans op toevalsovereenstemming voor alle oordelen is gedefinieerd als de som van de kans op overeenstemming per categorie. Een probleem bij deze maat is dat oprechte overeenstemming overschat wordt als een bepaalde categorie vanwege dezelfde impliciete beslisregels vaak gekozen wordt. Gwet's AC1 probeert daarvoor te corrigeren door de toevalsovereenstemming te beperken tot vignetten die moeilijk te scoren zijn.

**Tabel 3.24: Schattingen overeenstemming volgens drie methoden**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,65	c(0,56, 0,73)
	standaardoordeel	0,82	c(0,78, 0,85)
Fleiss Kappa	eindoordeel	0,44	c(0,31, 0,57)
	standaardoordeel	0,60	c(0,52, 0,68)
Gwet's AC1	eindoordeel	0,48	c(0,35, 0,62)
	standaardoordeel	0,78	c(0,73, 0,83)

#### 3.3.4.2 Overeenstemming inclusief Voldoende met herstelopdracht

We kunnen onze hoofduitkomsten ook berekenen voor het geval 'Voldoende met herstelopdracht' als een aparte categorie wordt beschouwd. Dit geeft uiteraard alleen voor standaardoordelen andere resultaten. Hierbij hebben we gekeken



naar de individuele oordelen. We zien dat de mate van overeenstemming daalt van 82% (zie tabel 3.24) naar 68% (tabel 3.25).

**Tabel 3.25: Schattingen overeenstemming volgens drie methoden, waarbij Voldoende met herstelopdracht een aparte categorie is**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,65	c(0,56, 0,73)
	standaardoordeel	0,68	c(0,63, 0,73)
Fleiss Kappa	eindoordeel	0,44	c(0,31, 0,57)
	standaardoordeel	0,52	c(0,45, 0,59)
Gwet's AC1	eindoordeel	0,48	c(0,35, 0,62)
	standaardoordeel	0,61	c(0,55, 0,67)

Eindoordelen zijn ongewijzigd door coderen 'voldoende met herstel' en zijn toegevoegd als referentie

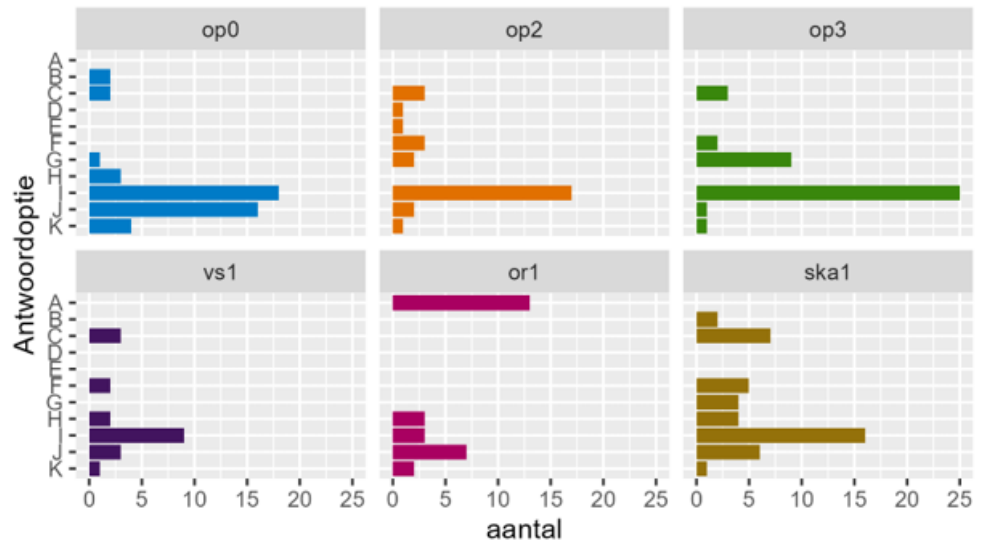
We geven hier ook de redenen voor afwijkende oordelen in de duo fase waarbij we het oordeel herstel meenemen.

**Tabel 3.26: Redenen voor afwijkende oordelen per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is (exclusief OP0 Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer <sup>a</sup>
Elementen uit het afwegingskader verschillend gewogen	88
Afwegingskader verschillend geïnterpreteerd	35
Contextinformatie school anders gewogen	18
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	16
Beslisregel OR1 anders toegepast	13
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school')	12
Informatie in het vignet over het hoofd gezien	12
Handleiding met afwegingskader wel/niet gebruikt	10
Kenmerken leerlingenpopulatie anders gewogen	4
Contextinformatie bestuur anders gewogen	1
Toezichthistorie anders gewogen	1
Anders	26



**Figuur 3.9: Redenen voor verschillend oordeel per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is**



- 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

### 3.3.4.3 Verschil Pa individueel-duo zonder aanvullen

In de secundaire analyse (paragraaf 3.2.2) beschreven we het verschil tussen individuele oordelen en duo-oordelen waarbij we duo-oordelen aanvullen met individuele oordelen waarover inspecteurs het eens waren. Hieronder beschrijven we de resultaten van de vergelijking als we *niet* aanvullen.

**Tabel 3.27: Verschil in oordelen tussen individuele en duo fase als duo oordelen niet worden aangevuld**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,58	0,73	0,15	1,81	0,09	16,00	56,00	31,00
standaarden	paired t test	0,77	0,85	0,08	2,24	0,03	80,00	56,00	31,00

We zien in de bovenstaande tabel in kolom 'pa indiv' dat de individuele overeenstemming zowel op eindoordeelen als standaarden lager is dan als we naar alle individuele oordelen kijken. De reden hiervoor is dat we hier kijken naar een deelverzameling van alle oordelen in de individuele fase: alleen die vignet-standaarden waar duo's in de duo-fase over hebben geoordeeld, zijn meegenomen. Dit zijn de vignet-standaarden waar duo's het in de individuele

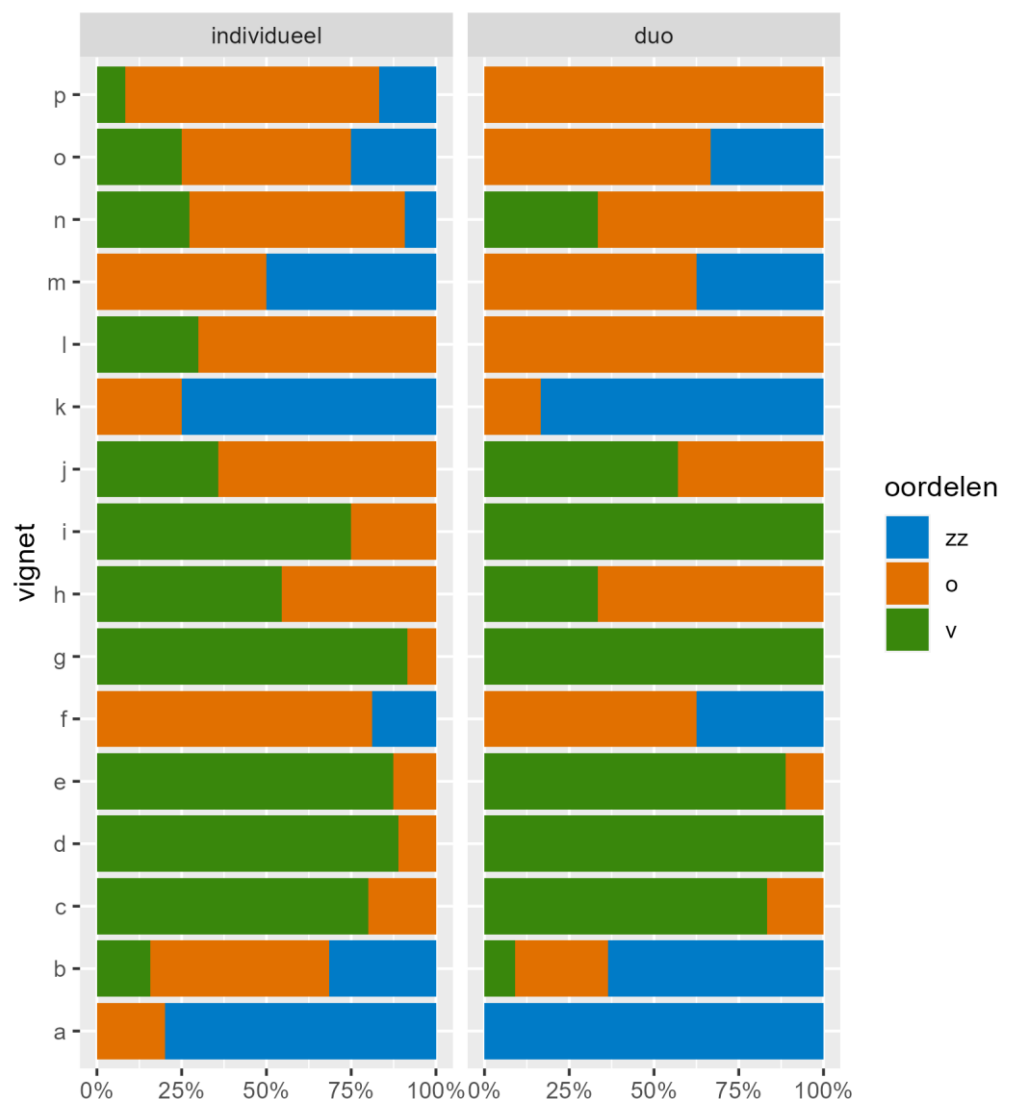


fase niet eens waren. We mogen daarom aannemen dat dit relatief ambigue vignetten waren. Het is dan ook te verwachten dat de overeenstemming op deze standaarden lager is in de individuele fase. Bovendien is te verwachten dat we de toename in overeenstemming na duo-overleg hier overschatten. Voor casussen waar al overeenstemming was, zou de overeenstemming immers niet toenemen, waardoor het hier gevonden effect zou worden gedempt.

Met deze kanttekeningen in het achterhoofd zien we een significante toename van de overeenstemming voor de bewuste vignet-standaarden van 8 procentpunt. De toename in Pa voor eindoordelen is niet significant, hoewel de effectschatting groter is.

Als we kijken naar de verschillen in de verhoudingen van soorten eindoordelen per vignet, dan is goed te zien dat duo-overleg niet altijd leidt tot meer overeenstemming tussen beoordelaars. Zie bijvoorbeeld de vignetten j en f in onderstaande Figuur 3.10, waar het meerderheidsoordeel juist kleiner wordt.

**Figuur 3.10: Verdeling van eindoordelen per fase van de studie: alleen eindoordelen die zijn gewijzigd in duo-fase**





3.3.4.4 Pa voor standaarden met onvoldoende, herstel en voldoende als categorieën

**Tabel 3.28: Schattingen proportie agreement per standaard met Voldoende met herstelopdracht als aparte categorie**

	est_agreeme nt	standaardfou t totaal	confidence_in terval 95	soort_oordeel
individueel	0,68	0,03	0,63 - 0,73	standaarden
duo	0,78	0,03	0,72 - 0,84	standaarden



## 4 Sector (v)so

In de volgende hoofdstukken geven we dezelfde tabellen en grafieken weer als bij po. Voor toelichting bij de tabellen en grafieken verwijzen we naar hoofdstuk 3 over po. We zullen in de navolgende hoofdstukken alleen bijzonderheden eruit lichten. Bij de sector (v)so is het object van toezicht de school.

### 4.1 Diagnostische data-analyse

#### 4.1.1 Respons en representativiteit

De doelpopulatie bestond voor de sector (v)so uit 25 inspecteurs. Hiervan hebben er 21 deelgenomen aan de studie.

**Tabel 4.1: Aantal respondenten per kleur**

	<b>n</b>
blauw	4
geel	6
groen	6
rood	5

**Tabel 4.2: historische- en vignetoordelen in percentages**

	<b>Zeer zwak</b>	<b>Onvoldoende</b>	<b>Voldoende</b>
KO/SKO 23	4	18	78
vignetten	6	25	69
beoordeelde vignetten	12	35	54

**Tabel 4.3: aantal volledig ingevulde vignetten**

<b>ingevulde_vignetten</b>	<b>n</b>	<b>cumulatieve_n</b>	<b>perc</b>
10	3	3	14
9	4	7	33
7	3	10	48
6	5	15	71
5	2	17	81
4	3	20	95
3	1	21	100

**Tabel 4.4: response vignetten per indiensttreding**

	<b>response (%)</b>	<b>n</b>
1e inspecteur	71	14
2e inspecteur	64	5
recent in dienst	85	2

**Tabel 4.5: response vignetten per kleur boekje**

	<b>response (%)</b>	<b>n</b>
blauw	62	4
geel	66	6
groen	85	6
rood	64	5



#### 4.1.2 Validiteit

**Tabel 4.6: Inspecteurs die afweken van de beslisregels in de individuele fase**

	vignet	OP0	OP2	OP3	VS1	OR1	SKA	EOS	bereke nd_ein oorde el
133	m	o	h	h	h	h	o	o	v
	p	o	o	h	h	v	o	o	ZZ
322	d	o	o	v	v	o	v	ZZ	o
	e	v	v	v	v	o	h	o	v
	f	v	h	o	v	v	v	v	o
425	p	o	h	v	v	h	o	o	v
417	m	o	v	o	h	h	o	o	ZZ
	n	h	v	v	v	o	v	o	v
	d	o	v	h	h	o	v	o	v

**Tabel 4.7: Wijken vooral inspecteurs die recent in dienst zijn af van de beslisregels?**

	n
ja	3
nee	6

**Tabel 4.8: Inspecteurs die afweken van de beslisregels na duo-fase**

	OP0	OP2	OP3	VS1	OR1	SKA1	EOS	bereken d_eindo oordeel
m	o	o	v	v	o	o	o	ZZ

## 4.2 Hoofdanalyse

### 4.2.1 Primaire uitkomsten

**Tabel 4.9: Hoe vaak geven duo's een duo-oordeel als ze in de individuele fase beiden geoordeeld hebben?**

	individueel oordeel	duo oordeel	n	percentage
standaardoordeel	verschilt	geen duo- oordeel	20	8
	verschilt	wel duo- oordeel	26	11
	zelfde	wel duo- oordeel	194	81
eindoordeel	verschilt	geen duo- oordeel	3	6
	verschilt	wel duo- oordeel	12	25
	zelfde	wel duo- oordeel	33	69

**Tabel 4.10: Schattingen proportie agreement in de individuele fase**

	est agreement	confidence interval 95
eindoordeel	0,72	c(0,59, 0,85)
standaardoordeel	0,85	c(0,80, 0,89)



**Tabel 4.11: Schattingen Pa aangevulde duo-oordelen**

	est agreement	confidence interval 95
eindoordelen	0,77	c(0,53, 1,00)
standaarden	0,89	c(0,80, 0,98)

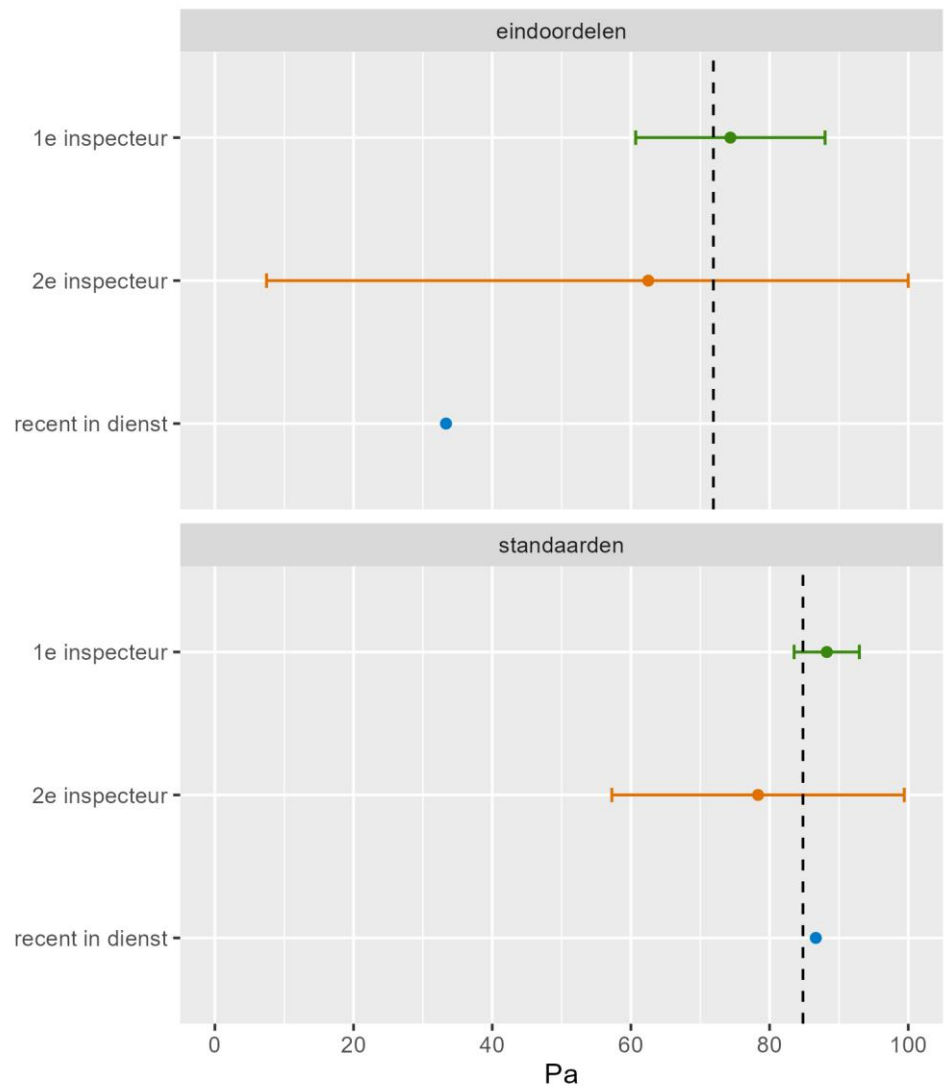
4.2.2

*Secundaire uitkomsten*

**Tabel 4.12: Verschil in oordelen tussen individuele en duo fase met aangevulde oordelen**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,75	0,77	0,02	0,16	0,87	16,00	18,00	9,00
standaarden	paired t test	0,89	0,89	0,00	0,03	0,97	96,00	18,00	9,00

**Figuur 4.1: Pa binnen cohorten. De stippelijijn geeft de Pa van alle inspecteurs weer**



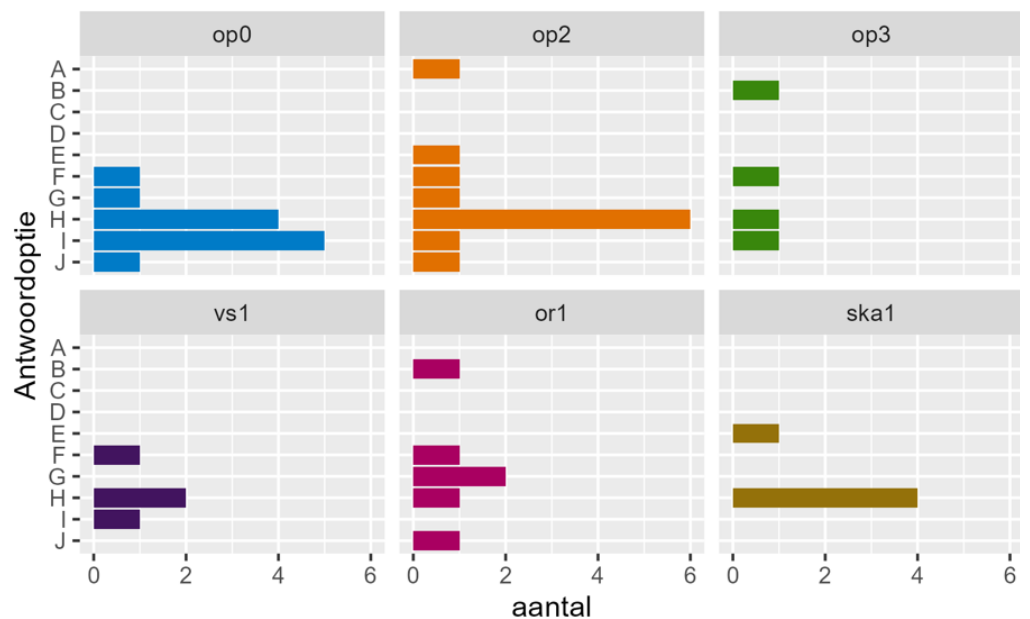


**Tabel 4.14: Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OP0 Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer (%) <sup>a</sup>
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%)

<sup>a</sup>) N.B.: 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.

**Figuur 4.2: Redenen voor verschillend oordeel per standaard**



- 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



### 4.3 Exploratieve analyse

#### 4.3.1 *Verdiepende analyse van verschillen in eendoordelen*

**Tabel 4.15: Alle vergelijkingen van soorten eendoordelen van alle mogelijke koppels van individuele inspecteurs**

	percentage
zelfde	71,9
o en zz	8,9
o en v	15,7
v en zz	3,5

**Tabel 4.16: Alle vergelijkingen van soorten eendoordelen van alle mogelijke koppels van duo's**

	percentage
zelfde	76,7
o en zz	8,9
o en v	10,0
v en zz	4,4

**Tabel 4.17: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van individuele inspecteurs**

	percentage
zelfde	84,8
o en v	10,6
v en g	4,6
o en g	0,0
o en ntb	0,0
v en ntb	0,0
g en ntb	0,0

**Tabel 4.18: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van duo's**

	percentage
zelfde	89,3
o en v	5,9
v en g	4,2
o en ntb	0,7
o en g	0,0
v en ntb	0,0
g en ntb	0,0

**Tabel 4.19: Percentage inspecteurs dat het meerderheidsoordeel geeft voor eendoordelen**

	schatting	standaardfout
individueel	79,8	4,8
duo bovengrens	88,9	4,7

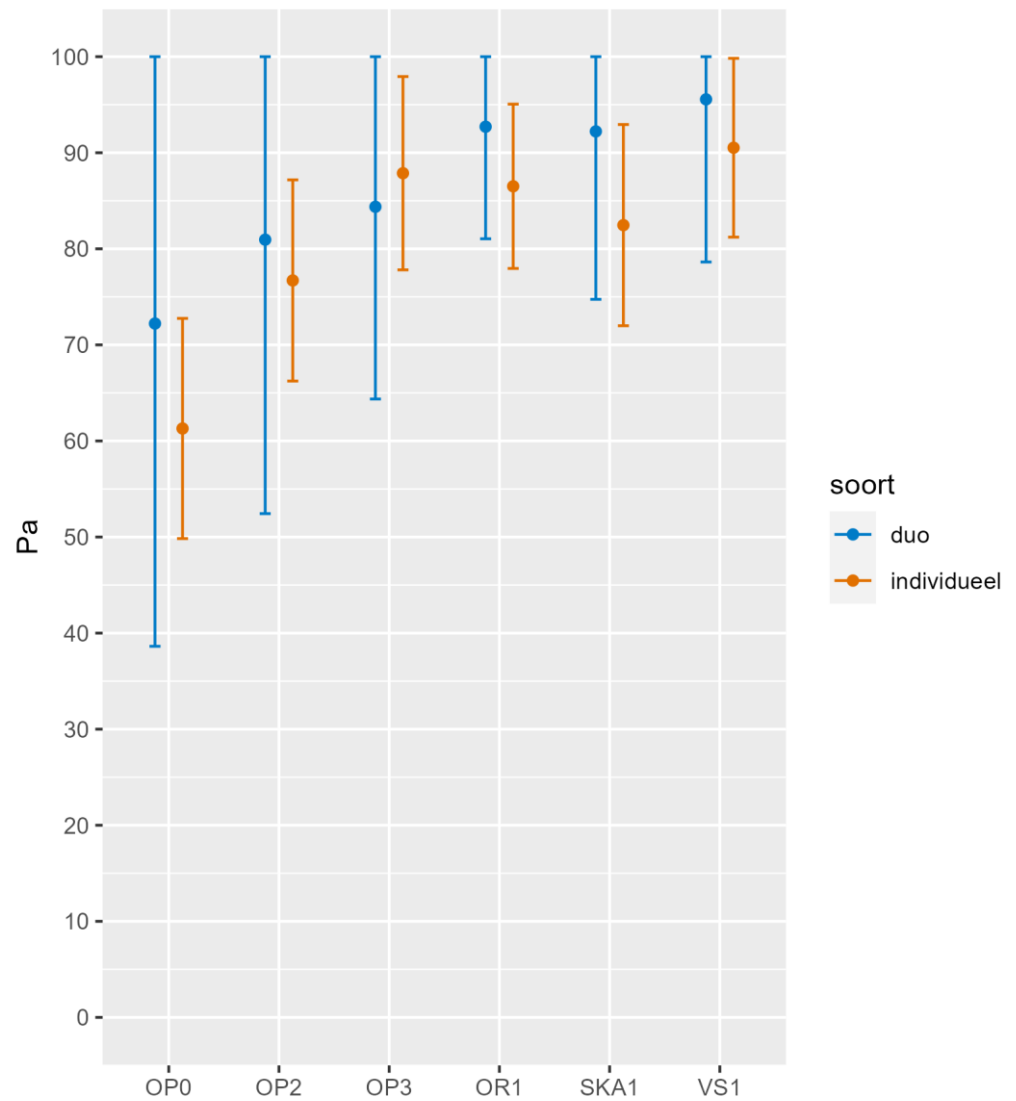
**Tabel 4.20: Percentage inspecteurs dat het meerderheidsoordeel geeft voor oordelen bij standaarden**

	schatting	standaardfout
individueel	90,9	1,4
duo bovengrens	94,5	1,5



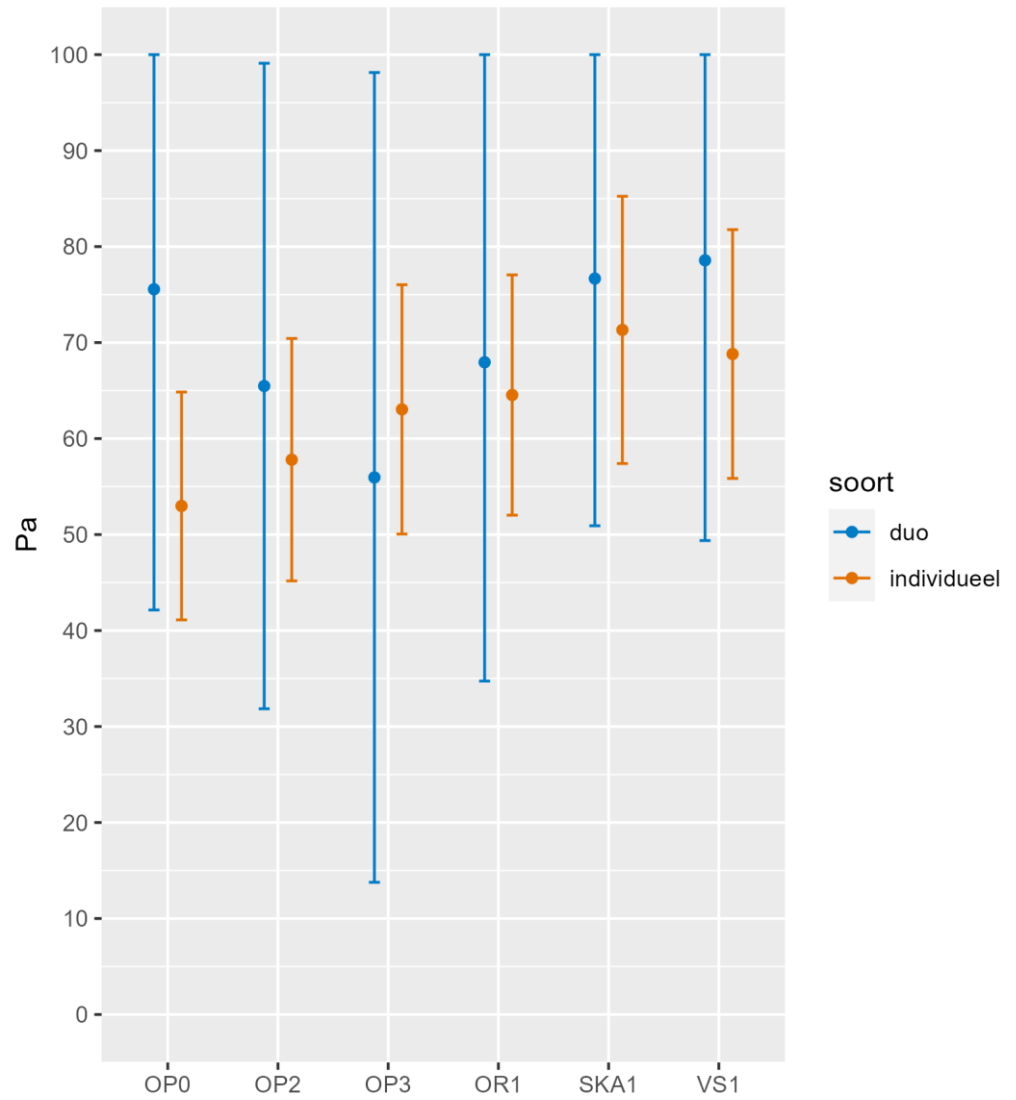
4.3.2 *Pa per standaard*

**Figuur 4.3: Pa van oordelen per standaard: Voldoende met herstelopdracht is gecodeerd als Voldoende**





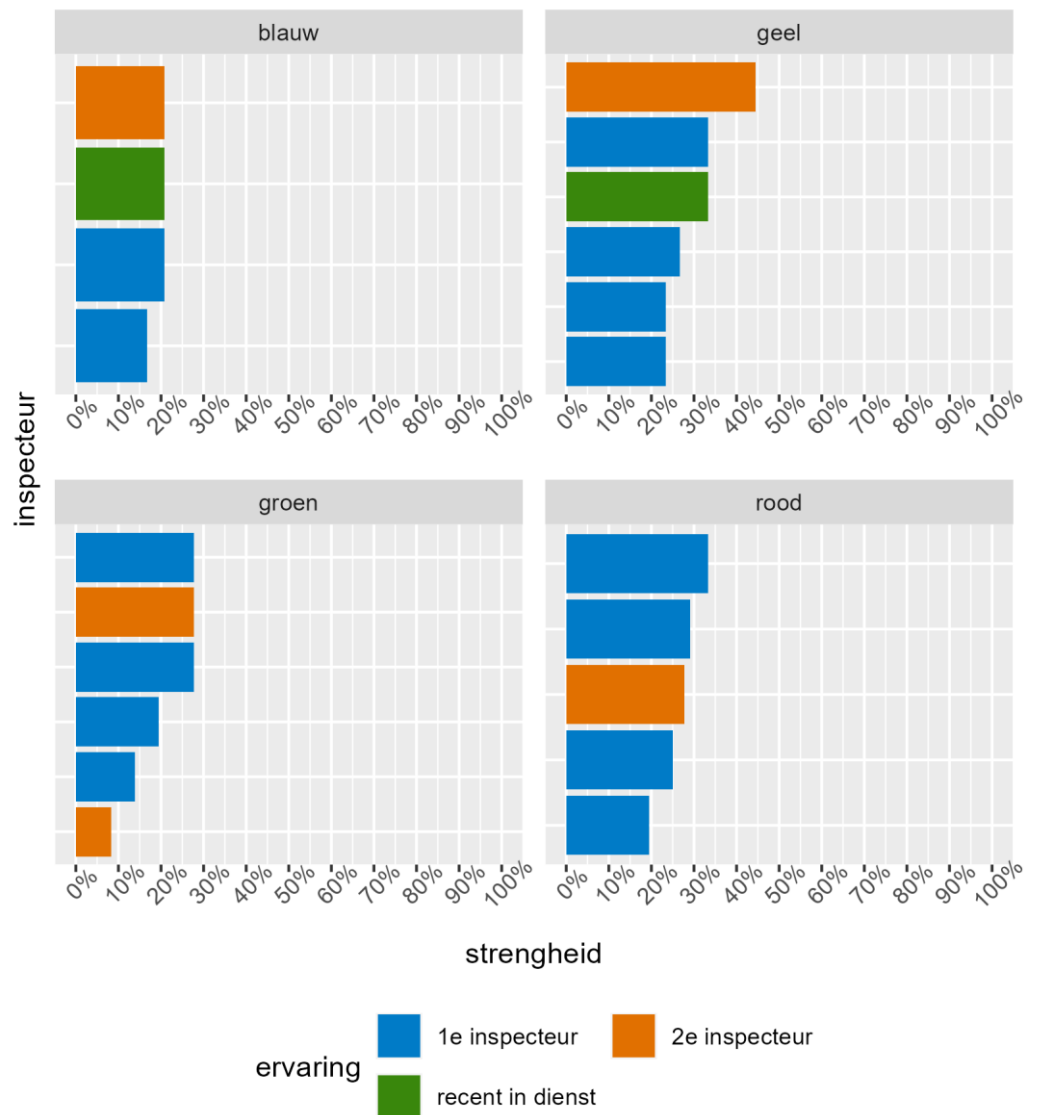
**Figuur 4.4: Pa van oordelen per standaard: Voldoende met herstelopdracht is een aparte categorie**





4.3.3 *Strengheid per inspecteur*

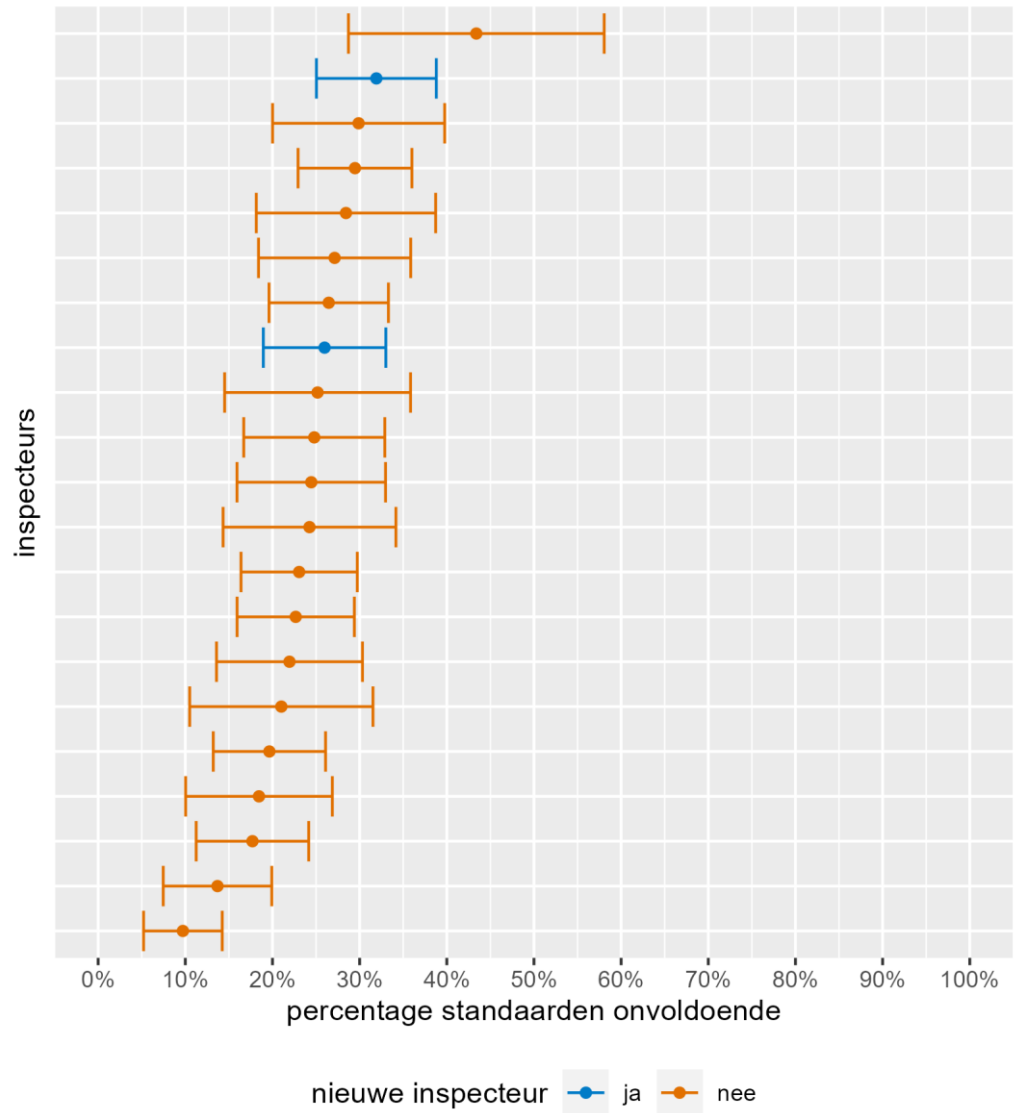
**Figuur 4.5: Strengheid op standaarden per kleur boekje**



Alleen de eerste 5 vignetten zijn meegenomen. Inclusief OP0



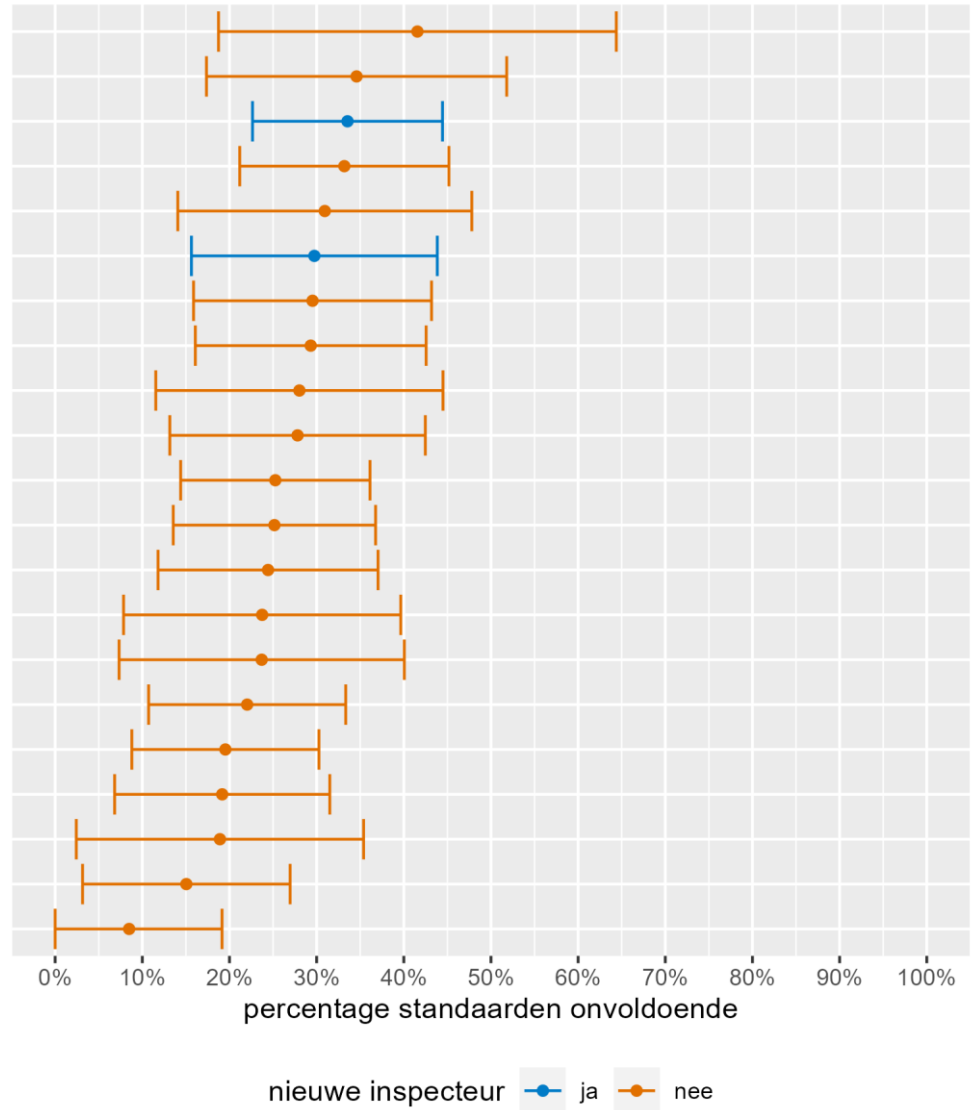
**Figuur 4.6: Gemiddelde strengheid per inspecteur: op basis van een logistisch regressiemodel**



inclusief OP0 en alle gescoorde vignet-standaarden



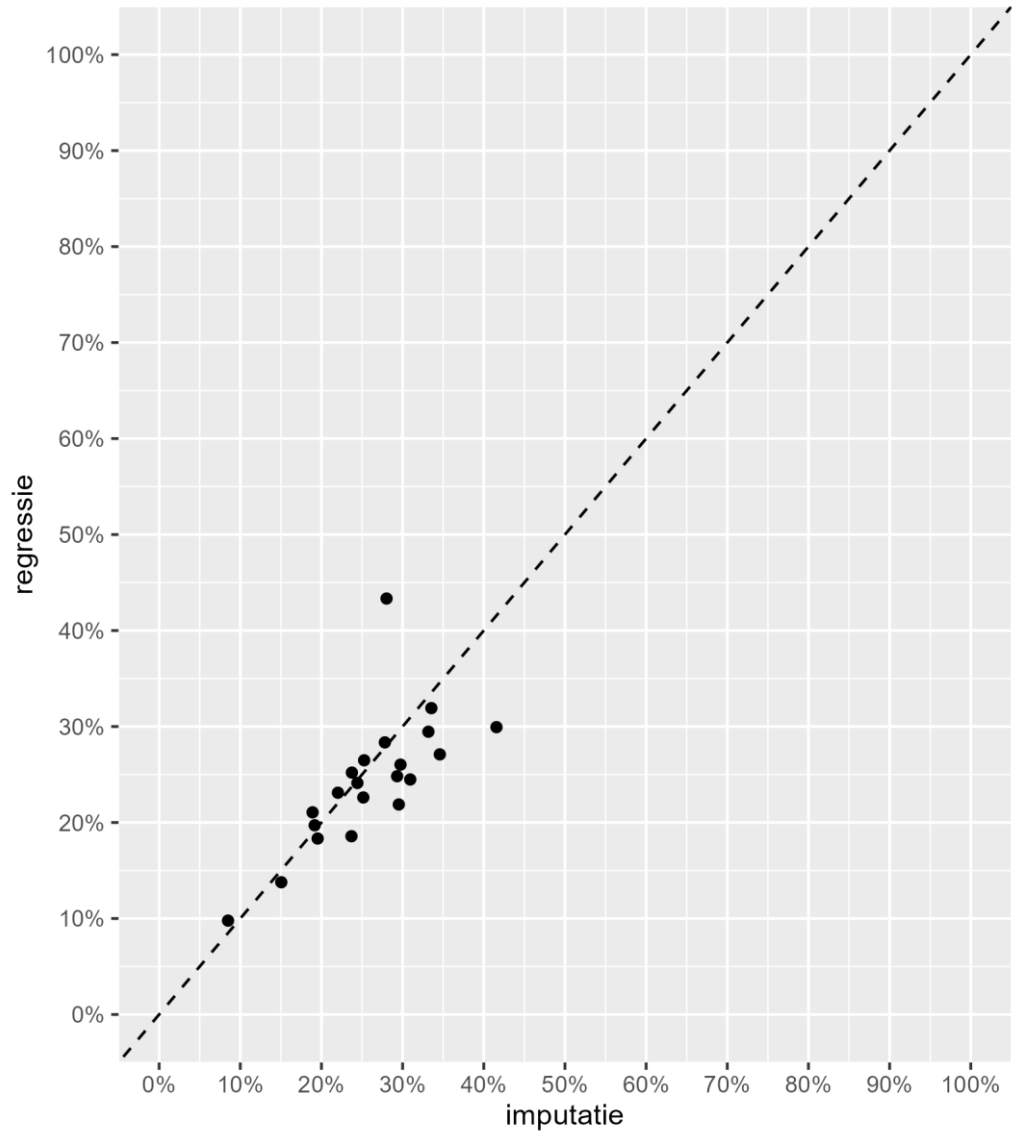
**Figuur 4.7: Gemiddelde strengheid per inspecteur: na imputatie**



inclusief OP0 en alle gescoorde vignet-standaarden



**Figuur 4.8: Vergelijking imputatie en regressie: De correlatie is 0.72**



**Tabel 4.21: Verschillen in strengheid tussen cohorten**

	perc Onvoldoende	standaardfout	95% BI
recent in dienst	31,6	4,7	(22,4, 40,8)
2e inspecteur	25,4	3,4	(18,8, 32,0)
1e inspecteur	25,3	1,8	(21,6, 28,9)



4.3.4 *Alternatieve specificaties*

**Tabel 4.22: Schattingen overeenstemming volgens drie methoden**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,72	c(0,59, 0,85)
	standaardoordeel	0,85	c(0,80, 0,89)
Fleiss Kappa	eindoordeel	0,49	c(0,30, 0,68)
	standaardoordeel	0,52	c(0,38, 0,66)
Gwet's AC1	eindoordeel	0,61	c(0,41, 0,82)
	standaardoordeel	0,82	c(0,76, 0,88)

**Tabel 4.23: Schattingen overeenstemming volgens drie methoden, waarbij Voldoende met herstelopdracht een aparte categorie is**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,72	c(0,59, 0,85)
	standaardoordeel	0,65	c(0,59, 0,71)
Fleiss Kappa	eindoordeel	0,49	c(0,30, 0,68)
	standaardoordeel	0,41	c(0,31, 0,51)
Gwet's AC1	eindoordeel	0,61	c(0,41, 0,82)
	standaardoordeel	0,57	c(0,49, 0,64)

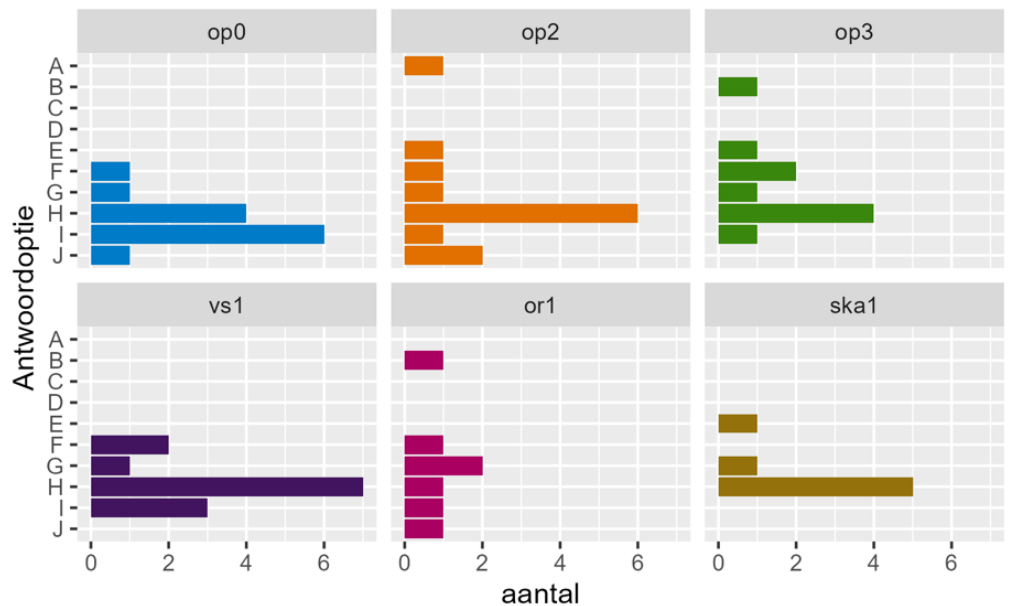
Eindoordelen zijn ongewijzigd door coderen 'voldoende met herstel' en zijn toegevoegd als referentie

**Tabel 4.24: Redenen voor afwijkende oordelen per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is (inclusief OP0 Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer <sup>a</sup>
Elementen uit het afwegingskader verschillend gewogen	27
Afwegingskader verschillend geïnterpreteerd	12
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	7
Informatie in het vignet over het hoofd gezien	7
Handleiding met afwegingskader wel/niet gebruikt	4
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school')	3
Contextinformatie school anders gewogen	2
Kenmerken leerlingenpopulatie anders gewogen	1
Contextinformatie bestuur anders gewogen	0
Toezichthistorie anders gewogen	0
Anders	8



**Figuur 4.9: Redenen voor verschillend oordeel per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is**



- 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

#### 4.3.4.1 Duo oordelen niet aanvullen

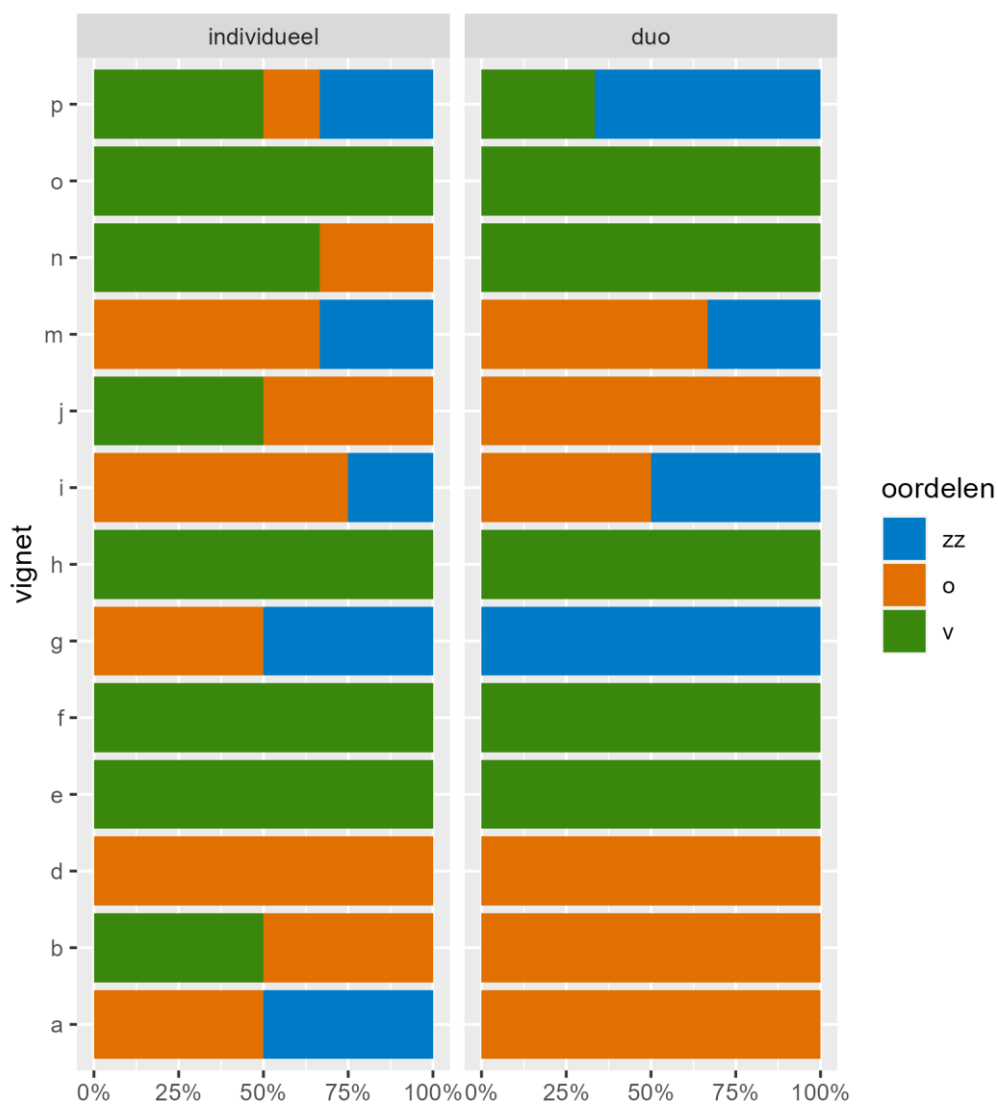
**Tabel 4.25: Verschil in oordelen tussen individuele en duo fase als duo oordelen niet worden aangevuld**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,52	0,42	-0,10	-0,37	0,71	16,00	16,00	9,00
standaarden	paired t test	0,64	0,74	0,09	0,50	0,62	96,00	16,00	9,00



#### 4.3.4.2 Veranderingen in eindoordelen

**Figuur 4.9: Verdeling van eindoordelen per fase van de studie: alleen eindoordelen die zijn gewijzigd in duo-fase**



#### 4.3.4.3 Pa voor standaarden met Onvoldoende, Voldoende met herstelopdracht en Voldoende als categorieën

**Tabel 4.26: Schattingen proportie agreement per standaard met Voldoende met herstelopdracht als aparte categorie**

	est_agreem ent	standaardfou t_totaal	confidence_in terval_95	soort_oordeel
individueel	0,65	0,03	0,59 - 0,71	standaarden
duo	0,69	0,08	0,54 - 0,84	standaarden



## 5 Sector vo

Bij de sector vo is het object van toezicht de afdeling. Van de 16 vignetten van afdelingen bleek dat op één afdeling geen eindoordeel is uitgesproken of kan worden uitgesproken. Deze afdeling behoorde daarmee niet tot de doelpopulatie van afdelingen en is buiten alle analyses gehouden (behalve de diagnostische data-analyse). Bij de afname van de studie in de sector vo bleek dat veel inspecteurs alleen konden deelnemen aan de individuele fase. Hierdoor is het aantal duo-oordelen relatief klein.

### 5.1 Diagnostische data-analyse

#### 5.1.1 Respons en representativiteit

Bij de sector vo behoorden 55 inspecteurs tot de doelpopulatie, hiervan namen er 48 deel aan de studie.

**Tabel 5.1: Aantal respondenten per kleur**

	n
blauw	13
geel	11
groen	11
rood	13

**Tabel 5.2: historische- en vignetoordelen in percentages**

	Zeer zwak	Onvoldoende	Voldoende	Geen oordeel
KO/SKO 23	2	34	64	0
vignetten	6	38	50	6
beoordeelde vignetten	12	41	46	2

**Tabel 5.3: aantal volledig ingevulde vignetten**

ingevulde vignetten	n	cumulatieve_n	perc
10	26	26	54
9	2	28	58
8	6	34	71
7	4	38	79
6	5	43	90
5	4	47	98
2	1	48	100

**Tabel 5.4: Respons vignetten per kleur boekje**

	response (%)
blauw	86
geel	94
groen	82
rood	85

**Tabel 5.5: Respons vignetten per indiensttreding**

	response (%)
in dienst per maart 2024 of daarna	78
3 jaar terug tot maart 2024	86
langer dan 3 jaar	94



5.1.2 Validiteit

**Tabel 5.6: Inspecteurs die afweken van de beslisregels in de individuele fase**

	vignet	OP0	OP2	OP3	VS1	OR1	SKA1	EOS	bereken d eindoor deel
310	j	o	o	v	v	o	v	o	zz
129	k	o	o	o	h	ntb	v	o	zz
418	f	h	o	v	v	o	o	o	zz
	i	o	v	v	h	v	v	o	v
	k	h	o	o	o	v	v	zz	o
2	a	h	h	o	o	v	o	zz	o
	e	o	o	h	o	ntb	v	o	zz
325	a	o	h	o	v	v	v	v	o
410	k	o	o	o	v	ntb	v	o	zz
121	j	o	h	h	o	o	v	o	zz
	k	o	o	o	v	ntb	v	o	zz
231	n	o	v	v	o	v	o	v	o
224	b	o	v	v	v	v	v	o	v
109	k	o	o	o	v	ntb	v	o	zz
404	k	o	o	o	h	ntb	v	o	zz
130	n	o	h	v	o	v	o	v	o
	o	o	h	o	o	v	v	zz	o
415	k	o	o	o	h	ntb	v	o	zz
	l	v	h	v	v	v	v	o	v
425	k	o	o	o	v	ntb	v	o	zz
314	c	o	o	o	h	v	o	zz	o
	d	o	v	h	o	v	h	v	o
	g	h	o	v	v	v	o	v	o
131	k	o	o	o	v	ntb	v	o	zz
431	j	h	o	h	v	o	v	o	zz
220	m	o	o	v	v	v	v	v	o
	n	o	v	v	o	v	o	v	o
	p	o	o	v	v	v	v	v	o
	e	o	o	v	v	ntb	v	v	o
126	k	o	o	o	v	ntb	v	o	zz

**Tabel 5.7: Wijken vooral inspecteurs die recent in dienst zijn af van de beslisregels?**

	n
ja	14
nee	13
onbekend	3

**Tabel 5.8: Inspecteurs die afweken van de beslisregels na duo fase**

	OP0	OP2	OP3	VS1	OR1	SKA1	EOS	bereken d eindoor deel
e	o	o	v	o	ntb	v	o	zz



## 5.2 Hoofdanalyse

### 5.2.1 Primaire uitkomsten

**Tabel 5.9: Hoe vaak geven duo's een duo-oordeel als ze in de individuele fase beiden geoordeeld hebben?**

	individueel oordeel	duo oordeel	n	percentage
standaardoordeel	verschilt	geen duo-oordeel	20	5
	verschilt	wel duo-oordeel	26	6
	zelfde	wel duo-oordeel	354	88
eindoordeel	verschilt	geen duo-oordeel	3	4
	verschilt	wel duo-oordeel	15	19
	zelfde	wel duo-oordeel	62	78

**Tabel 5.10: Schattingen proportie agreement in de individuele fase**

	est agreement	confidence interval 95
eindoordeel	0,81	c(0,71, 0,90)
standaardoordeel	0,89	c(0,85, 0,92)

**Tabel 5.11: Schattingen Pa aangevulde duo-oordelen**

	est agreement	confidence interval 95
eindoordeelen	0,82	c(0,66, 0,97)
standaarden	0,94	c(0,89, 0,99)

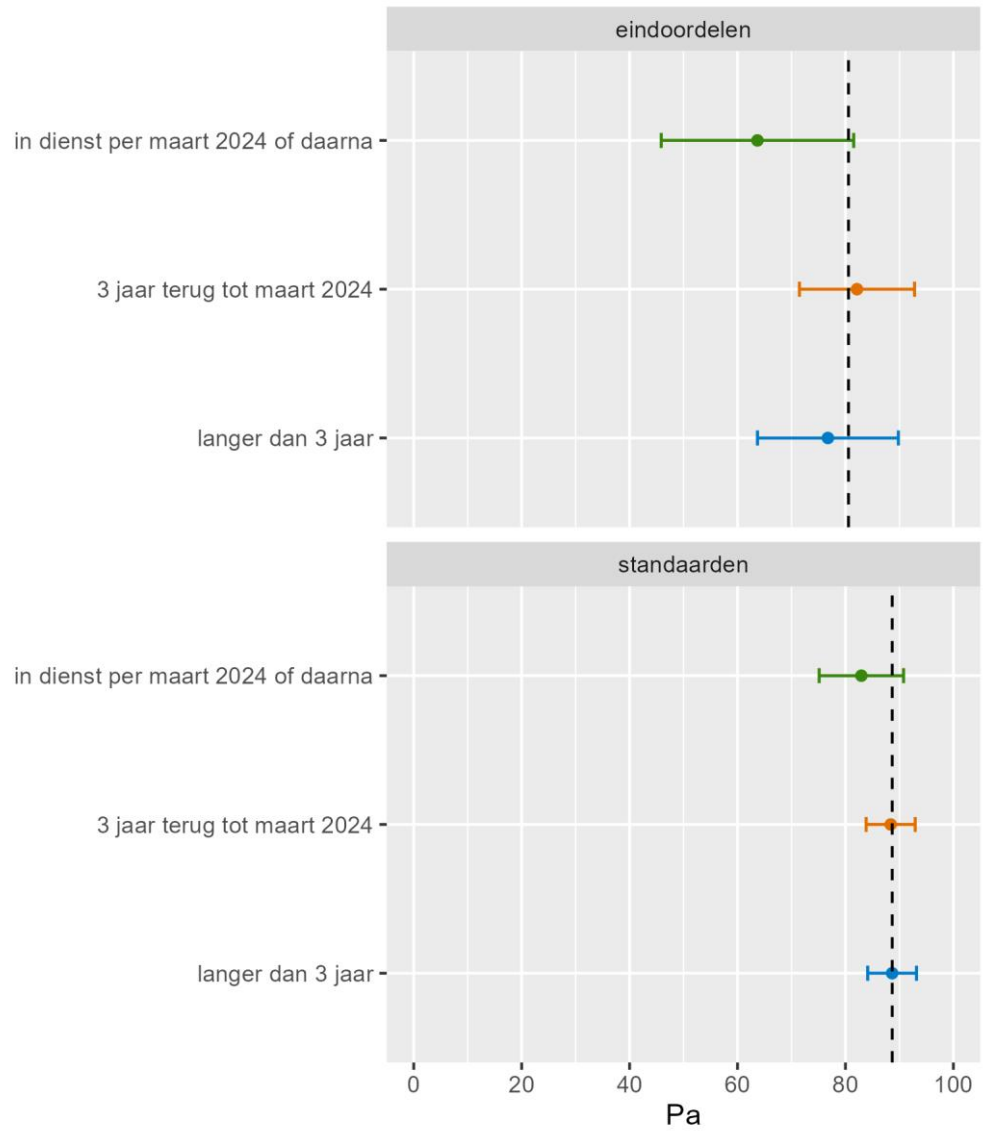
### 5.2.2 Secundaire uitkomsten

**Tabel 5.12: Verschil in oordelen tussen individuele en duo fase met aangevulde oordelen**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,82	0,82	-0,01	-0,10	0,92	15,00	24,00	13,00
standaarden	paired t test	0,92	0,94	0,02	0,74	0,46	75,00	24,00	13,00



**Figuur 5.1: Pa binnen cohorten. De stippellijn geeft de Pa van alle inspecteurs weer**



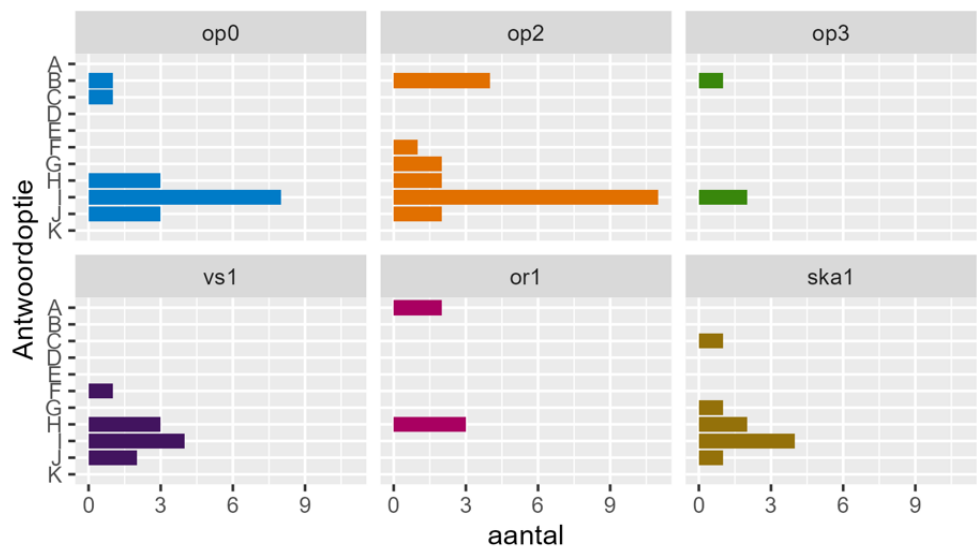


**Tabel 5.14: Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden (exclusief OP0 Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer (%) <sup>a</sup>
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%)

<sup>a</sup>) N.B.: 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.

**Figuur 5.2: Redenen voor verschillend oordeel per standaard**



- 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



### 5.3 Exploratieve analyse

#### 5.3.1 *Verdiepende analyse van verschillen in eendoordelen*

**Tabel 5.15: Alle vergelijkingen van soorten eendoordelen van alle mogelijke koppels van individuele inspecteurs**

	percentage
zelfde	80,6
o en zz	5,5
o en v	13,8
v en zz	0,1

**Tabel 5.16: Alle vergelijkingen van soorten eendoordelen van alle mogelijke koppels van duo's**

	percentage
zelfde	81,6
o en zz	4,0
o en v	14,4
v en zz	0,0

**Tabel 5.17: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van individuele inspecteurs**

	percentage
zelfde	88,7
o en v	9,5
v en g	1,2
v en ntb	0,6
o en g	0,0
o en ntb	0,0
g en ntb	0,0

**Tabel 5.18: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van duo's**

	percentage
zelfde	93,7
o en v	5,6
v en ntb	0,7
o en g	0,0
o en ntb	0,0
v en g	0,0
g en ntb	0,0

**Tabel 5.19: Percentage inspecteurs dat het meerderheidsoordeel geeft voor eendoordelen**

	schatting	standaardfout
individueel	87,2	3,4
duo bovengrens	89,6	3,8

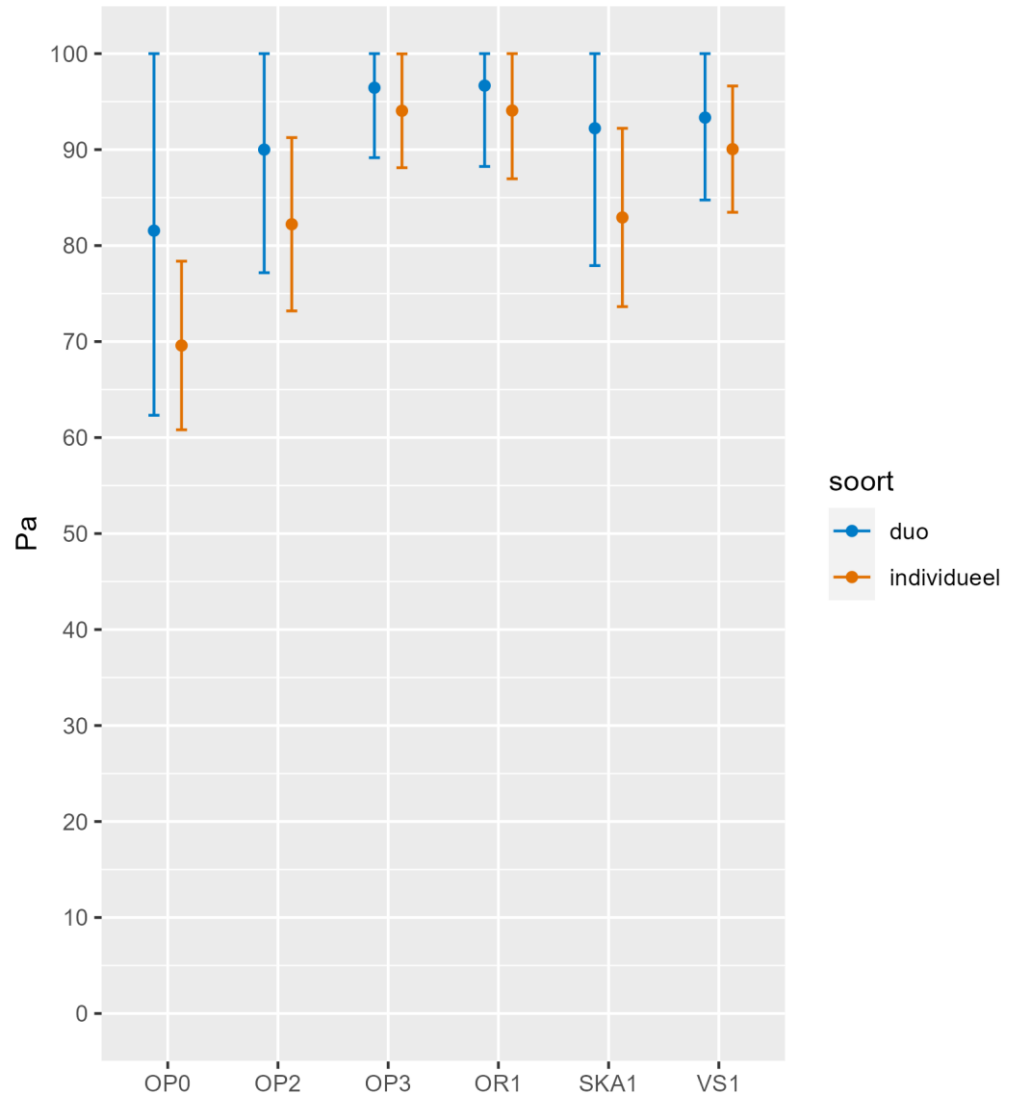
**Tabel 5.20: Percentage inspecteurs dat het meerderheidsoordeel geeft voor oordelen bij standaarden**

	schatting	standaardfout
individueel	92,9	1,3
duo bovengrens	96,8	1,0



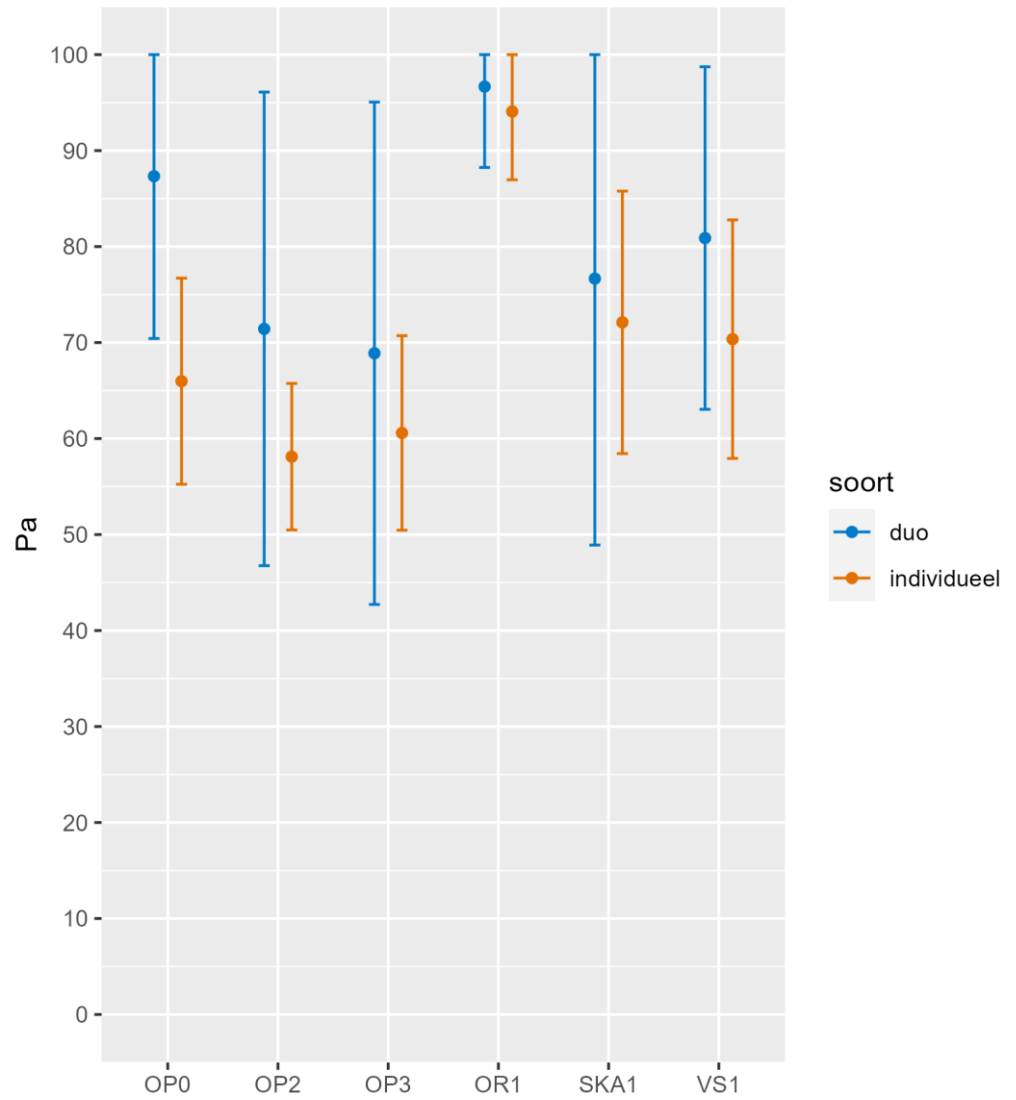
5.3.2 Pa per standaard

**Figuur 5.3: Pa van oordelen per standaard: Voldoende met herstelopdracht is gecodeerd als Voldoende**





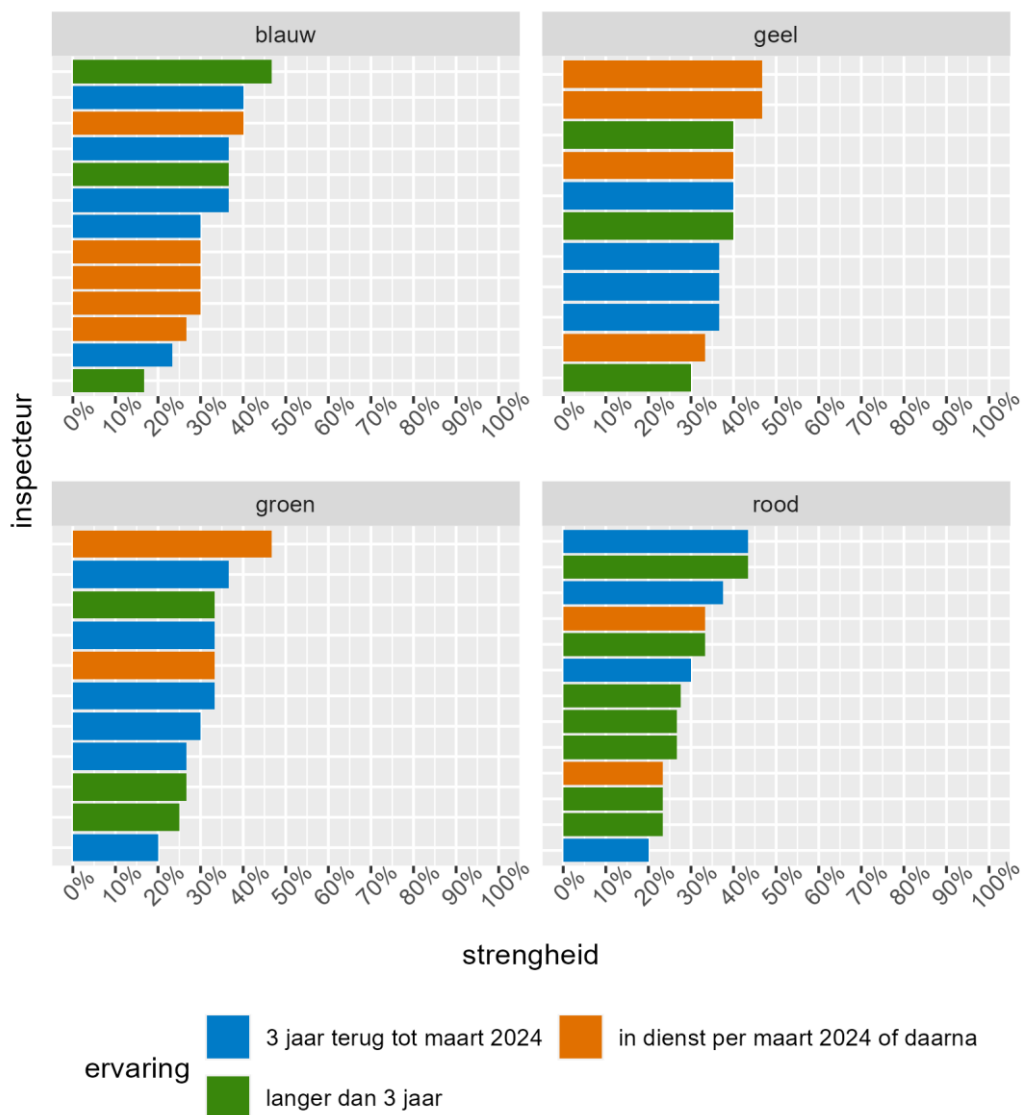
**Figuur 5.4: Pa van oordelen per standaard: Voldoende met herstelopdracht is een aparte categorie**





5.3.3 *Strengheid per inspecteur*

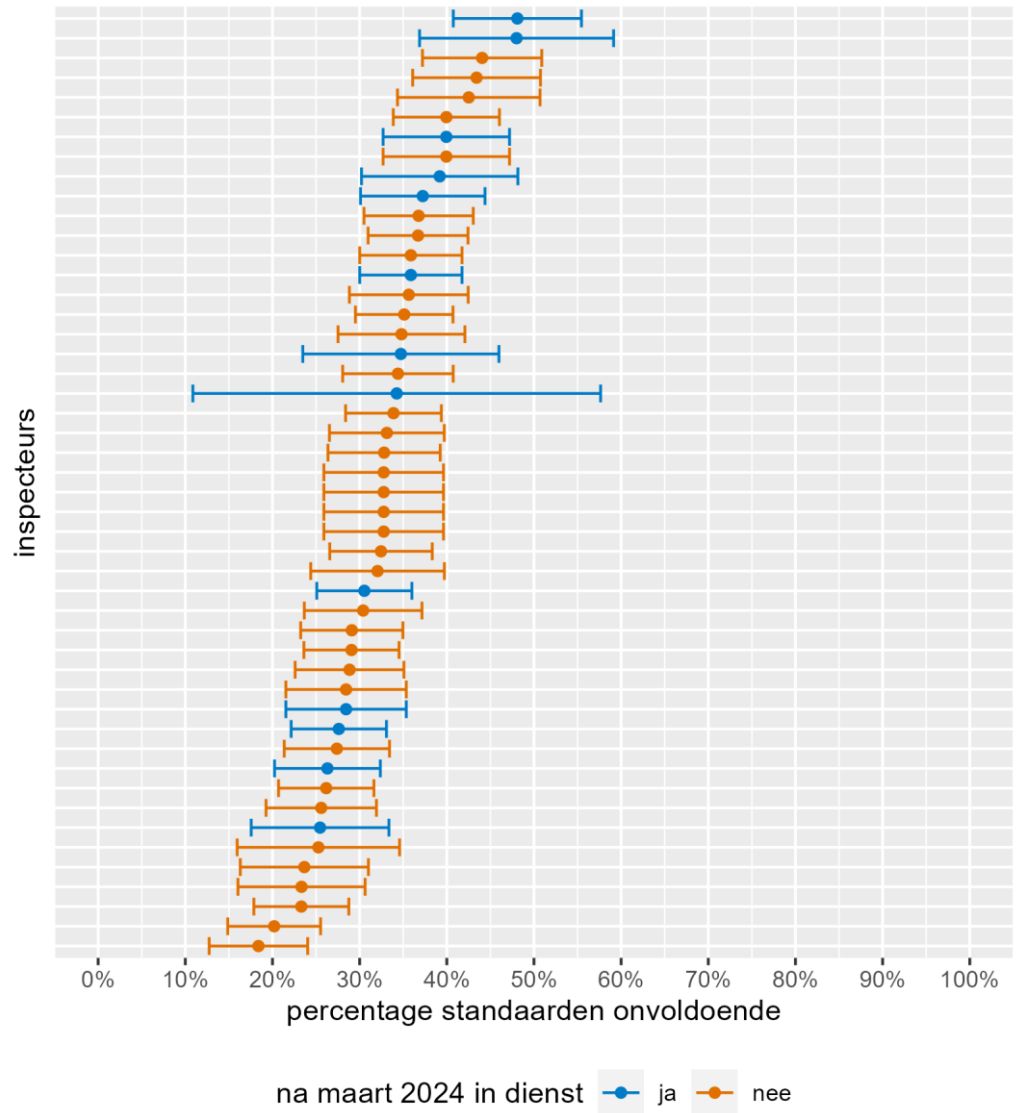
**Figuur 5.5: Strengheid op standaarden per kleur boekje**



Alleen de eerste 5 vignetten zijn meegenomen. Inclusief OP0



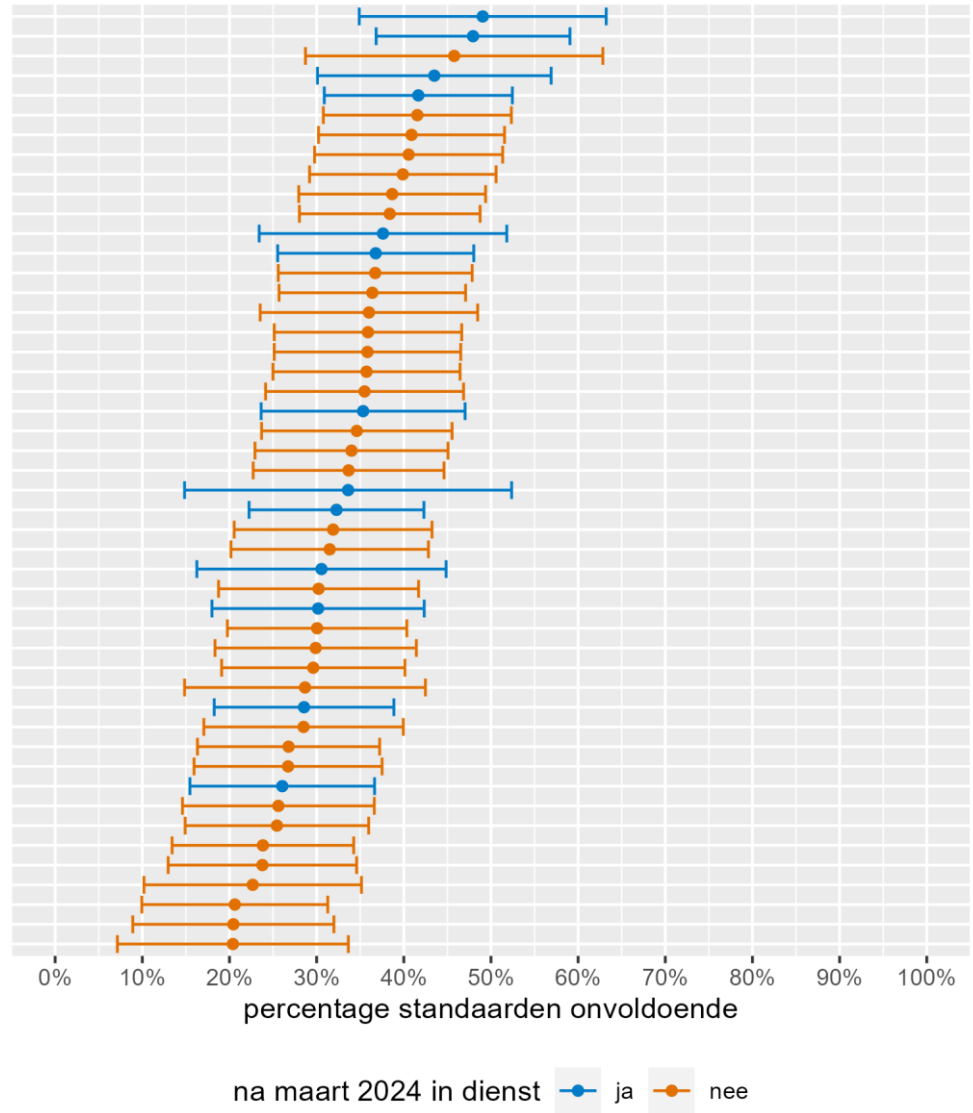
**Figuur 5.6: Gemiddelde strengheid per inspecteur: op basis van een logistisch regressiemodel**



inclusief OP0 en alle gescoorde vignet-standaarden



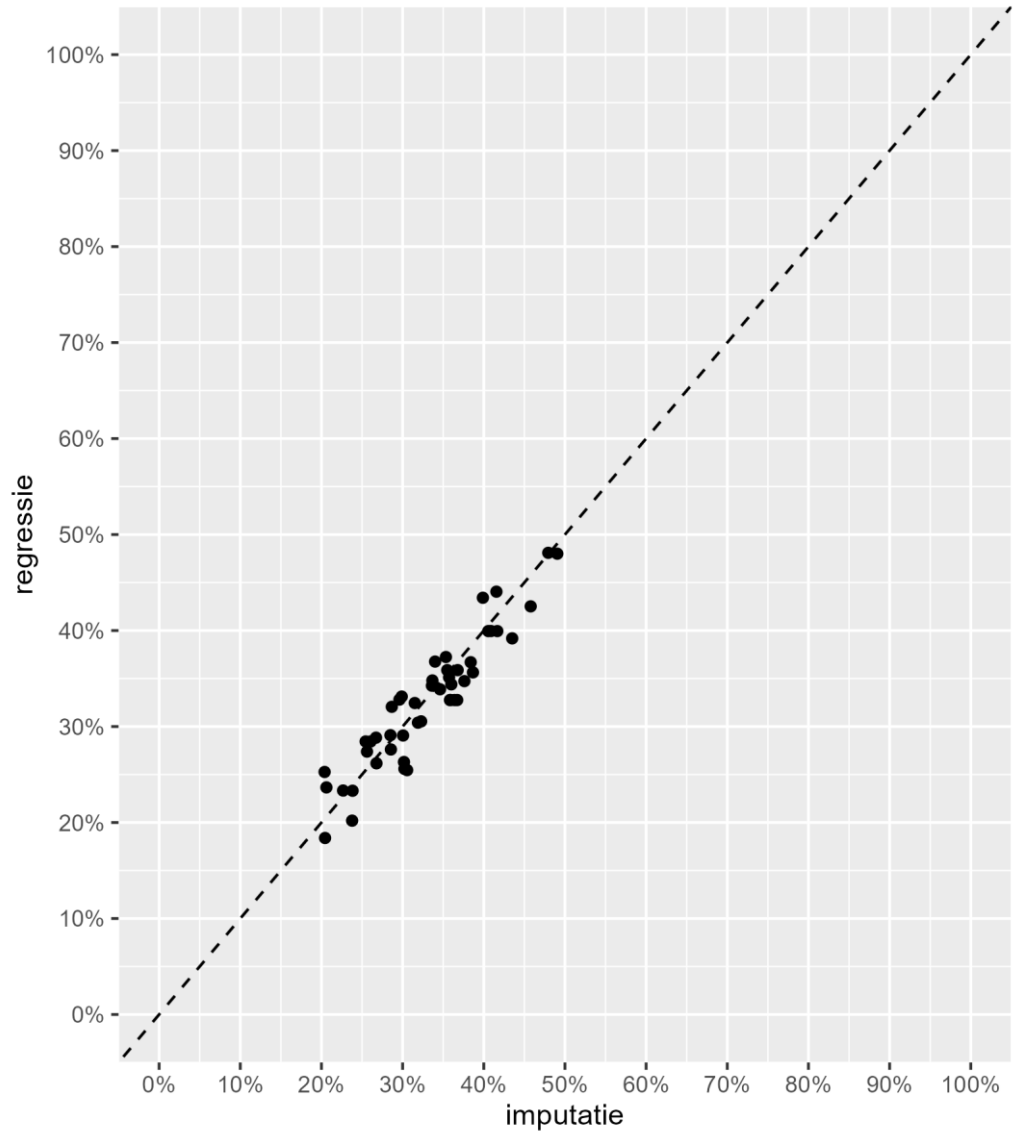
**Figuur 5.7: Gemiddelde strengheid per inspecteur: na imputatie**



inclusief OPO en alle gescoorde vignet-standaarden



**Figuur 5.8: Vergelijking imputatie en regressie: De correlatie is 0.93**



**Tabel 5.21: Verschillen in strengheid tussen cohorten**

	perc Onvoldoende	standaardfout	95% BI
in dienst per maart 2024 of daarna	36,4	2,0	(32,5, 40,3)
3 jaar terug tot maart 2024	32,0	1,3	(29,4, 34,6)
langer dan 3 jaar	31,8	1,5	(28,8, 34,8)

**Tabel 5.22: Verschillen in strengheid tussen cohorten**

	perc Onvoldoende	standaardfout	95% BI
in dienst per maart 2024 of daarna	36,4	2,0	(32,5, 40,3)
3 jaar terug tot maart 2024	32,0	1,3	(29,4, 34,6)
langer dan 3 jaar	31,8	1,5	(28,8, 34,8)



5.3.4 Alternatieve specificaties

**Tabel 5.23: Schattingen overeenstemming volgens drie methoden**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,81	c(0,71, 0,90)
	standaardoordeel	0,89	c(0,85, 0,92)
Fleiss Kappa	eindoordeel	0,66	c(0,50, 0,82)
	standaardoordeel	0,69	c(0,60, 0,79)
Gwet's AC1	eindoordeel	0,73	c(0,59, 0,87)
	standaardoordeel	0,87	c(0,83, 0,91)

**Tabel 5.24: Schattingen overeenstemming volgens drie methoden, waarbij Voldoende met herstelopdracht een aparte categorie is**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,81	c(0,71, 0,90)
	standaardoordeel	0,71	c(0,66, 0,77)
Fleiss Kappa	eindoordeel	0,66	c(0,50, 0,82)
	standaardoordeel	0,52	c(0,44, 0,60)
Gwet's AC1	eindoordeel	0,73	c(0,59, 0,87)
	standaardoordeel	0,66	c(0,59, 0,73)

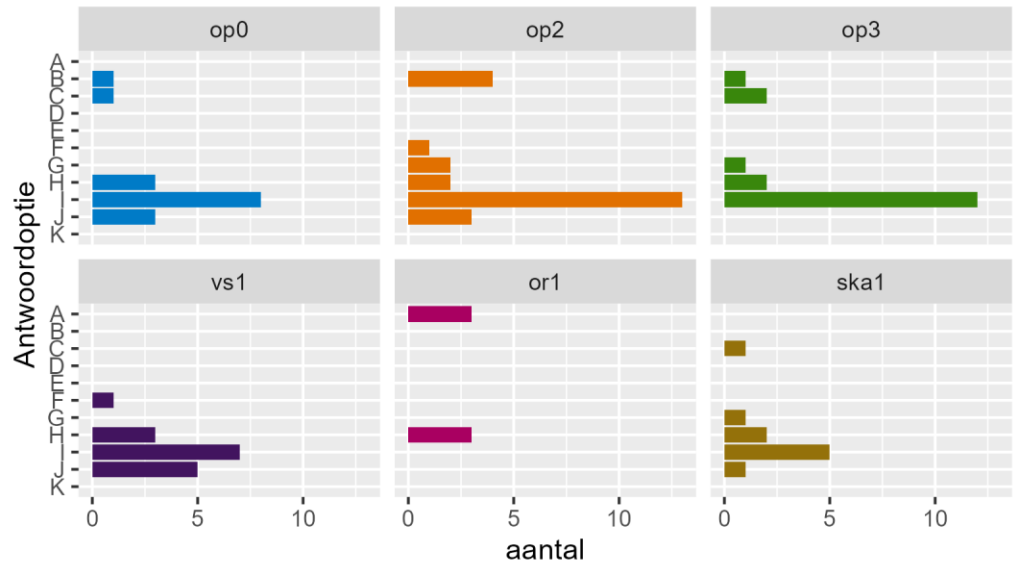
Eindoordelen zijn ongewijzigd door coderen 'voldoende met herstel' en zijn toegevoegd als referentie

**Tabel 5.25: Redenen voor afwijkende oordelen per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is (inclusief OP0 Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer <sup>a</sup>
Elementen uit het afwegingskader verschillend gewogen	45
Informatie in het vignet over het hoofd gezien	15
Afwegingskader verschillend geïnterpreteerd	12
Kenmerken leerlingenpopulatie anders gewogen	6
Contextinformatie afdeling anders gewogen	4
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	4
Beslisregel OR1 anders toegepast	3
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke school')	2
Handleiding met afwegingskader wel/niet gebruikt	0
Toezichthistorie anders gewogen	0
Contextinformatie bestuur anders gewogen	0
Anders	7



**Figuur 5.9: Redenen voor verschillend oordeel per standaard waarbij Voldoende met herstelopdracht een aparte categorie is**



- 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 gebruik

#### 5.3.4.1 Duo oordelen niet aanvullen

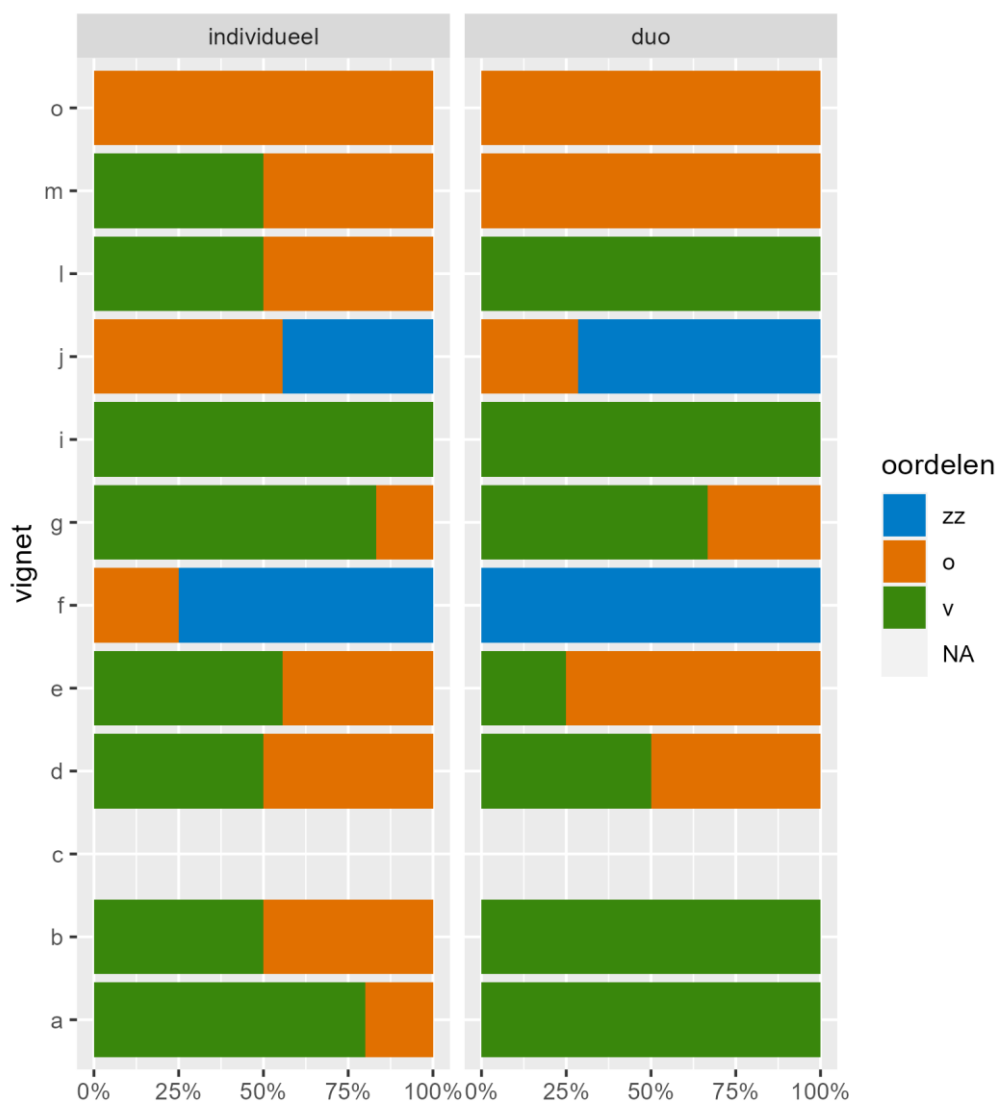
**Tabel 5.26: Verschil in oordelen tussen individuele en duo fase als duo oordelen niet worden aangevuld**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,37	0,68	0,31	1,10	0,29	15,00	26,00	13,00
standaarden	paired t test	0,66	0,76	0,10	0,50	0,62	75,00	26,00	13,00



5.3.4.2 Veranderingen in eindoordelen

**Figuur 5.10: Verdeling van eindoordelen per fase van de studie: alleen eindoordelen die zijn gewijzigd in duo-fase (vignet c werd niet gewijzigd)**



5.3.4.3 Pa voor standaarden met Onvoldoende, Voldoende met herstelopdracht en Voldoende als categorieën

**Tabel 5.27: Schattingen proportie agreement per standaard met Voldoende met herstelopdracht als aparte categorie**

	est_agreement	standaardfout_totaal	confidence_interval_95	soort_oordeel
individueel	0,71	0,03	c(0,66, 0,77)	standaarden
duo	0,79	0,05	c(0,68, 0,90)	standaarden



## 6 Sector mbo

Bij de sector mbo is het object van toezicht de opleiding binnen een instelling. Er zijn bekostigde en niet bekostigde opleidingen, die marginaal verschillend worden beoordeeld. Met deze verschillen is overigens rekening gehouden bij de analyse van de afwijking van de beslisregels. Een afwijking van de beslisregels als alleen OR1 Onvoldoende is, betekent in deze sector niet per se dat de inspecteur zich heeft vergist bij het toepassen van de beslisregel. Het onderzoekskader van mbo biedt in dit geval namelijk expliciet de ruimte om in bijzondere situaties af te wijken van de beslisregels voor het eindoordeel en toch een Voldoende eindoordeel te verstrekken.

### 6.1 Diagnostische data-analyse

#### 6.1.1 Respons en representativiteit

Bij de sector mbo behoorden 38 inspecteurs tot de doelpopulatie, hiervan namen er 31 deel aan de studie.

**Tabel 6.1: Aantal respondenten per kleur**

	n
blauw	7
geel	8
groen	8
rood	8

**Tabel 6.2: historische- en vignetoordelen in percentages**

	Zeer zwak	Onvoldoende	Voldoende	Geen oordeel
KO/SKO 24	2	48	50	0
historische oordelen vignetten	0	50	50	0
beoordeelde vignetten	1	53	46	1

**Tabel 6.3: aantal volledig ingevulde vignetten**

ingevulde vignetten	n	cumulatieve_n	perc
10	16	16	52
9	3	19	61
8	2	21	68
7	3	24	77
6	1	25	81
5	4	29	94
4	1	30	97
3	1	31	100

**Tabel 6.4: Respons vignetten per kleur boekje**

	response (%)
blauw	68
geel	82
groen	90
rood	91



**Tabel 6.5: Respons vignetten per indiensttreding**

	response (%)
Dit schooljaar begonnen	63
Aan minder dan vijf sko's deelgenomen	92
Aan vijf of meer sko's deelgenomen	87

**Tabel 6.6: Respons vignetten per kleur boekje, zonder nieuwe inspecteurs**

	response (%)
blauw	71
geel	93
groen	98
rood	91

6.1.2 Validiteit

**Tabel 6.7: Inspecteurs die afweken van de beslisregels in de individuele fase**

	vignette	OP0	OP2	OP3	OP5	VS1	BA1	BA2	OR1	SKA2	EOS	bereken eindoordeel
409	O	vh	vh	v	v	v	vh	v	v	v	o	v
202	G	v	v	o	o	o	o	v	v	o	zz	o
234	G	o	v	o	o	o	o	o	v	o	zz	o
118	O	v	v	o	o	v	o	v	g	v	zz	o
413	P	niet onderzocht	v	o	v	v	niet onderzocht	niet onderzocht	v	v	v	o
	D	v	v	o	v	v	v	v	v	v	v	o
419	D	v	v	o	v	v	v	v	v	v	v	o
205	L	niet onderzocht	o	v	v	v	v	v	niet onderzocht	v	v	o
212	N	v	v	v	v	v	v	v	o	v	v	o
112	N	v	v	v	v	v	v	v	o	v	v	o
134	K	v	v	v	v	v	vh	v	o	v	v	o
	L	niet onderzocht	o	v	v	v	v	v	niet onderzocht	v	v	o
318	D	v	v	o	v	vh	g	vh	v	o	v	o
113	N	v	v	v	v	v	v	v	o	v	v	o

**Tabel 6.8: Wijken vooral inspecteurs die recent in dienst zijn af van de beslisregels?**

	n
ja	1
nee	13

**Tabel 6.9: Inspecteurs die afweken van de beslisregels na duo fase**

	OP2	OP3	OP5	VS1	BA1	BA2	OR1	SKA2	EOS	bereken eindoordeel
D	v	o	v	v	v	v	v	v	v	o



## 6.2 Hoofdanalyse

### 6.2.1 Primaire uitkomsten

**Tabel 6.10: Hoe vaak geven duo's een duo-oordeel als ze in de individuele fase beiden geoordeeld hebben?**

	individueel oordeel	duo oordeel	n	percentage
standaardoordeel	verschilt	geen duo-oordeel	7	1
	verschilt	wel duo-oordeel	53	9
	zelfde	wel duo-oordeel	510	89
eindoordeel	verschilt	geen duo-oordeel	4	5
	verschilt	wel duo-oordeel	13	17
	zelfde	wel duo-oordeel	58	77

**Tabel 6.11: Schattingen proportie agreement in de individuele fase**

	est agreement	confidence interval 95
eindoordeel	0,76	c(0,65, 0,87)
standaardoordeel	0,89	c(0,85, 0,92)

**Tabel 6.12: Schattingen Pa aangevulde duo-oordelen**

	est agreement	confidence interval 95
eindoordeelen	0,80	c(0,64, 0,95)
standaarden	0,95	c(0,91, 0,99)

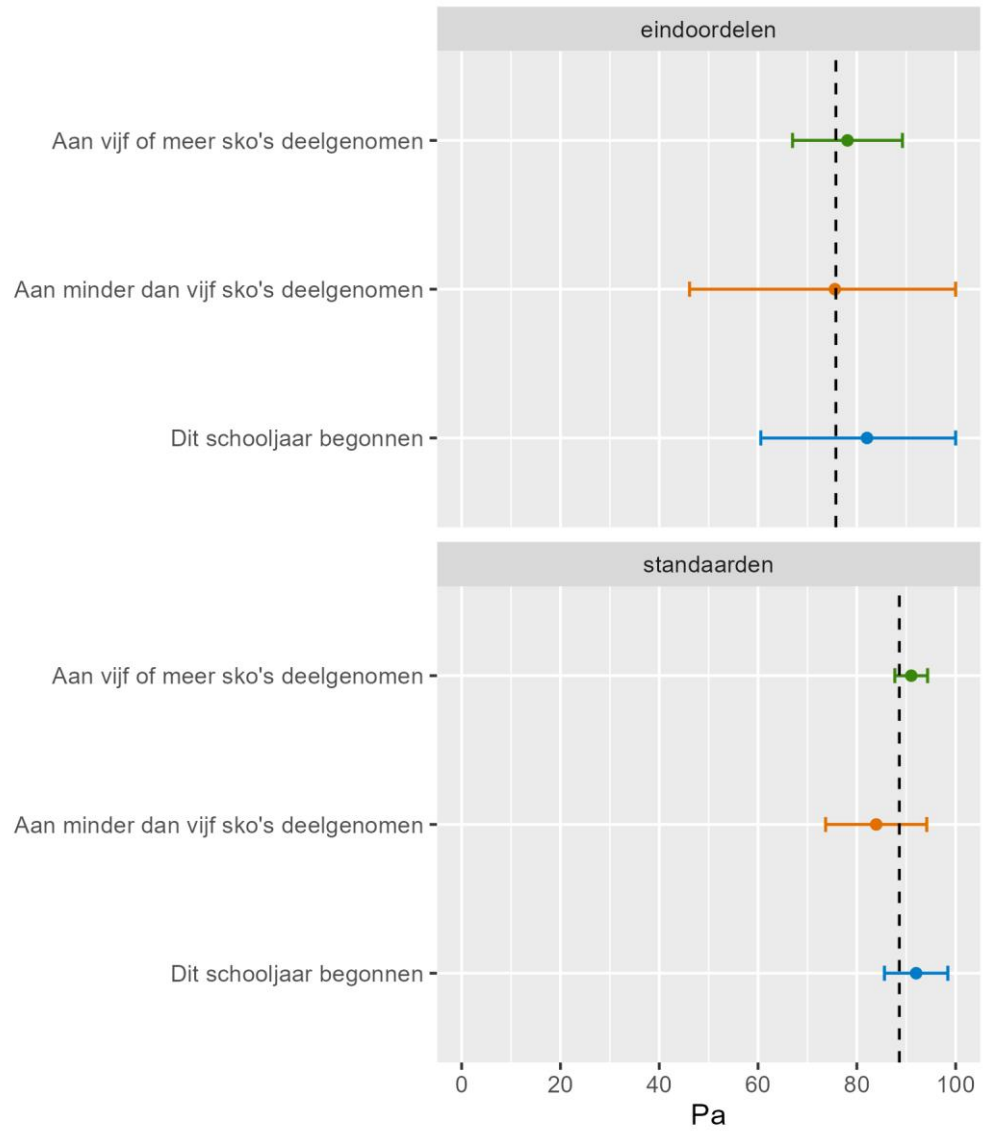
### 6.2.2 Secundaire uitkomsten

**Tabel 6.13: Verschil in oordelen tussen individuele en duo fase met aangevulde oordelen**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,79	0,80	0,01	0,15	0,88	16,00	20,00	10,00
standaarden	paired t test	0,90	0,95	0,05	2,54	0,01	122,00	20,00	10,00



**Figuur 6.1: Pa binnen cohorten. De stippellijn geeft de Pa van alle inspecteurs weer**



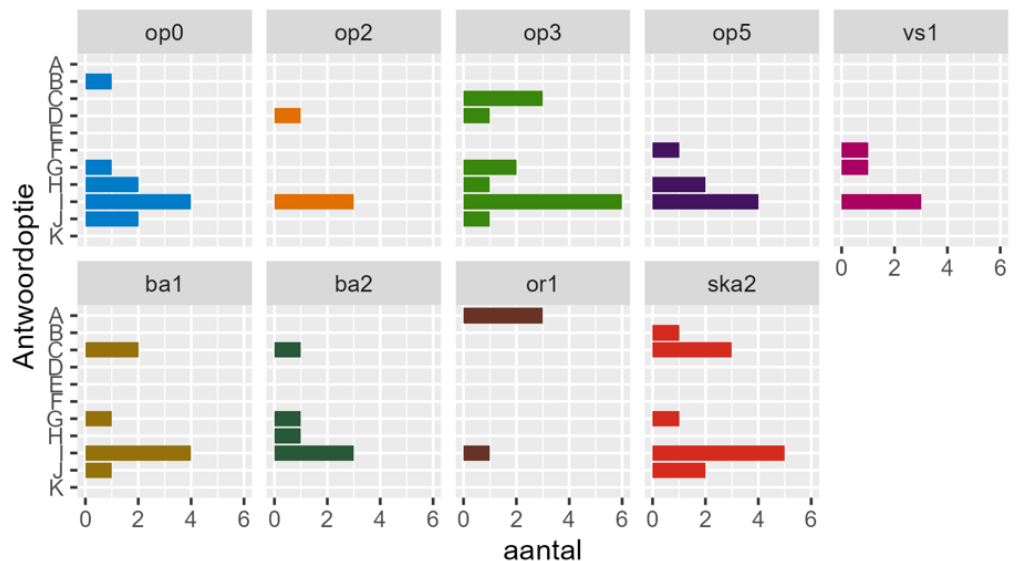


**Tabel 6.15: Zelfgerapporteerde verklaringen van duo's voor het verschillend oordelen op standaarden exclusief OPO Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer (%) <sup>a</sup>
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%)

<sup>a</sup>) N.B.: 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.

**Figuur 6.2: Redenen voor verschillend oordeel per standaard**



- 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



### 6.3 Exploratieve analyse

#### 6.3.1 *Verdiepende analyse van verschillen in eendoordelen*

**Tabel 6.16: Alle vergelijkingen van soorten eendoordelen van alle mogelijke koppels van individuele inspecteurs (nm = eendoordeel niet mogelijk)**

	percentage
zelfde	75,8
o en zz	1,4
o en v	19,9
v en zz	0,8
nm en zz	0,0
nm en o	1,2
nm en v	1,0

**Tabel 6.17: Alle vergelijkingen van soorten eendoordelen van alle mogelijke koppels van duo's**

	percentage
zelfde	79,6
o en zz	2,5
o en v	17,9
v en zz	0,0

**Tabel 6.18: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van individuele inspecteurs**

	percentage
zelfde	84,0
o en v	9,0
v en g	1,3
o en g	0,0
o en ntb	0,0
v en ntb	0,0
g en ntb	0,0

**Tabel 6.19: Alle vergelijkingen van soorten standaardoordelen van alle mogelijke koppels van duo's**

	percentage
zelfde	94,9
o en v	4,8
v en g	0,3
o en g	0,0
o en ntb	0,0
v en ntb	0,0
g en ntb	0,0

**Tabel 6.20: Percentage inspecteurs dat het meerderheidsoordeel geeft voor eendoordelen**

	schatting	standaardfout
individueel	85,2	3,8
duo bovengrens	87,3	4,5

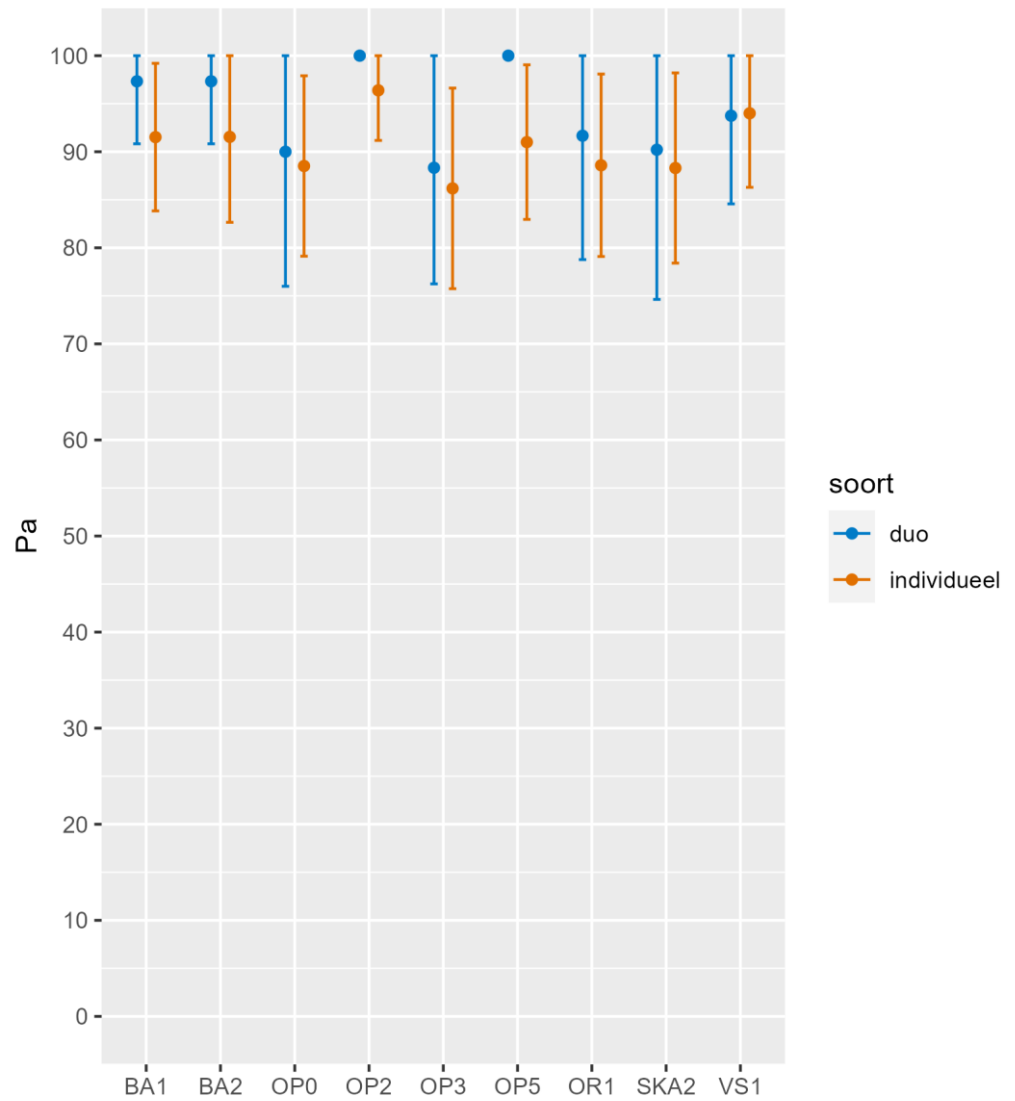
**Tabel 6.21: Percentage inspecteurs dat het meerderheidsoordeel geeft voor oordelen bij standaarden**

	schatting	standaardfout
individueel	93,2	1,0
duo bovengrens	97,2	0,8



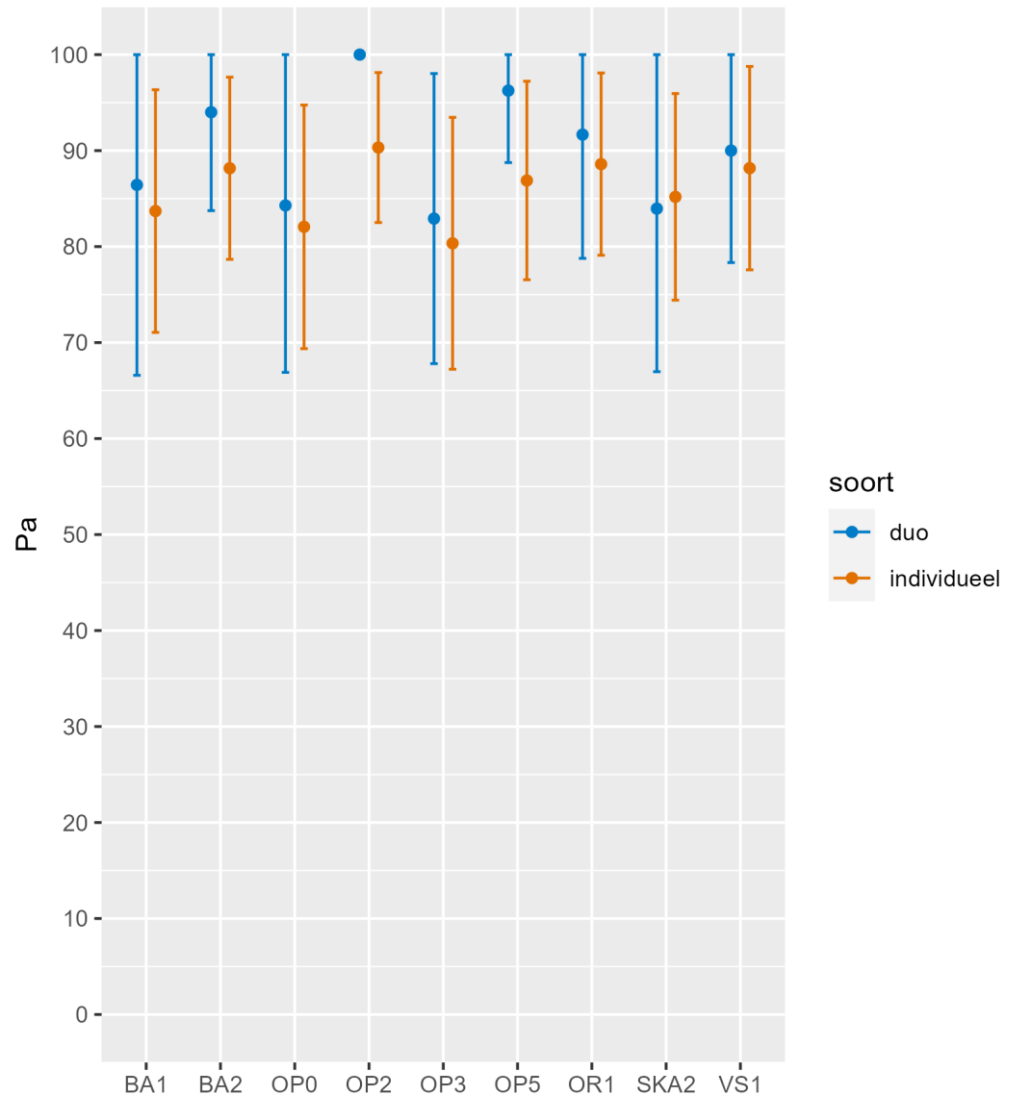
6.3.2 Pa per standaard

**Figuur 6.3: Pa van oordelen per standaard: Voldoende met herstelopdracht is gecodeerd als Voldoende**





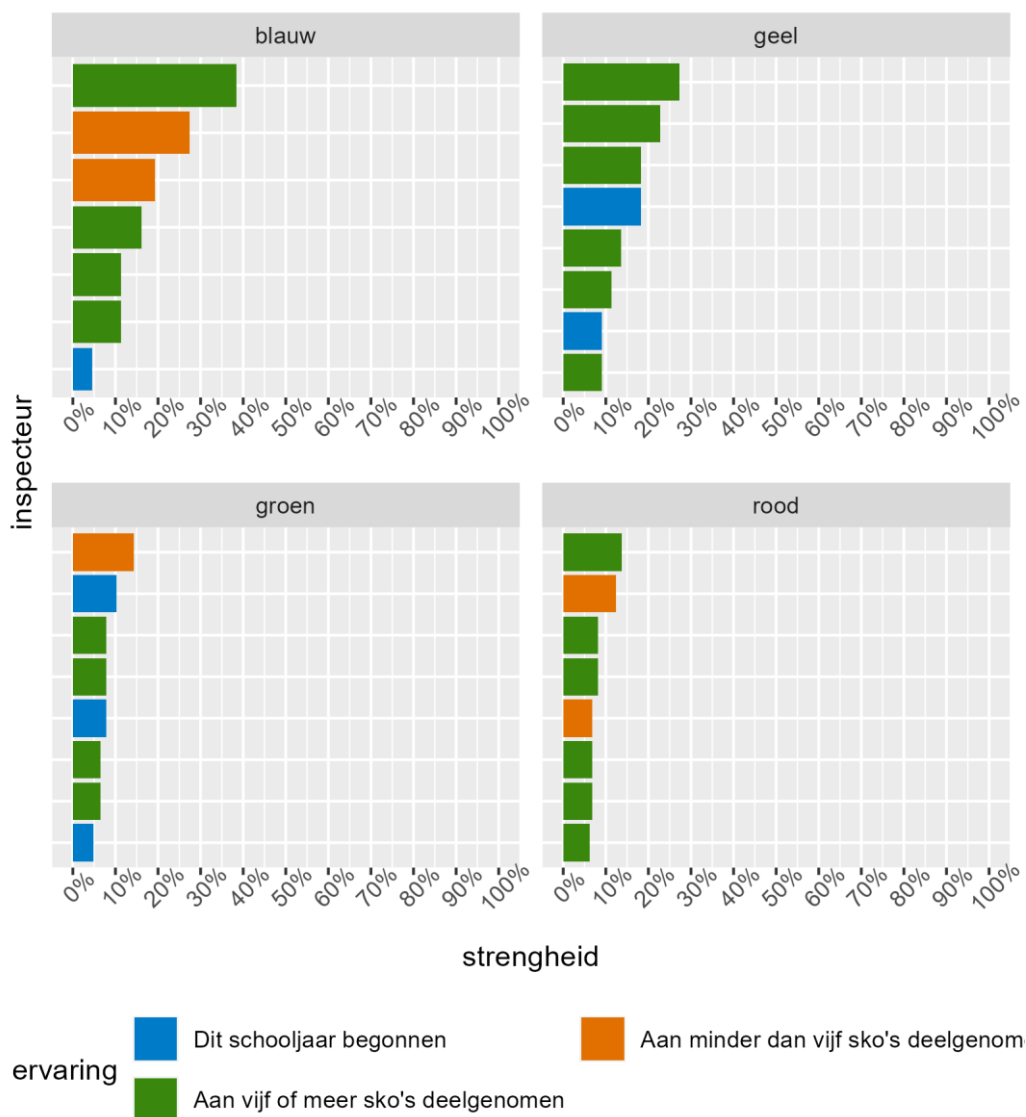
**Figuur 6.4: Pa van oordelen per standaard: Voldoende met herstelopdracht is een aparte categorie**





6.3.3 *Strengheid per inspecteur*

**Figuur 6.5: Strengheid op standaarden per kleur boekje**

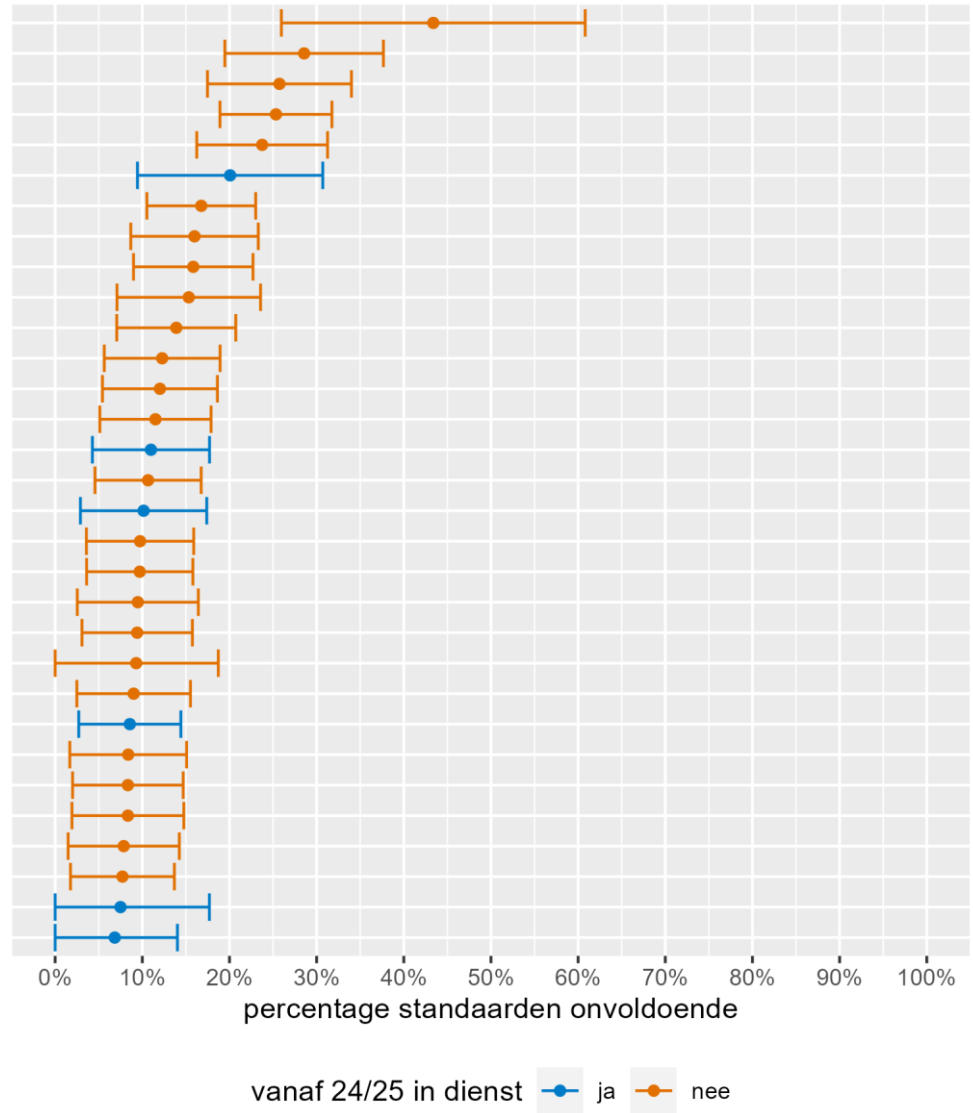


Alleen de eerst 5 vignetten zijn meegenomen. Inclusief OP0





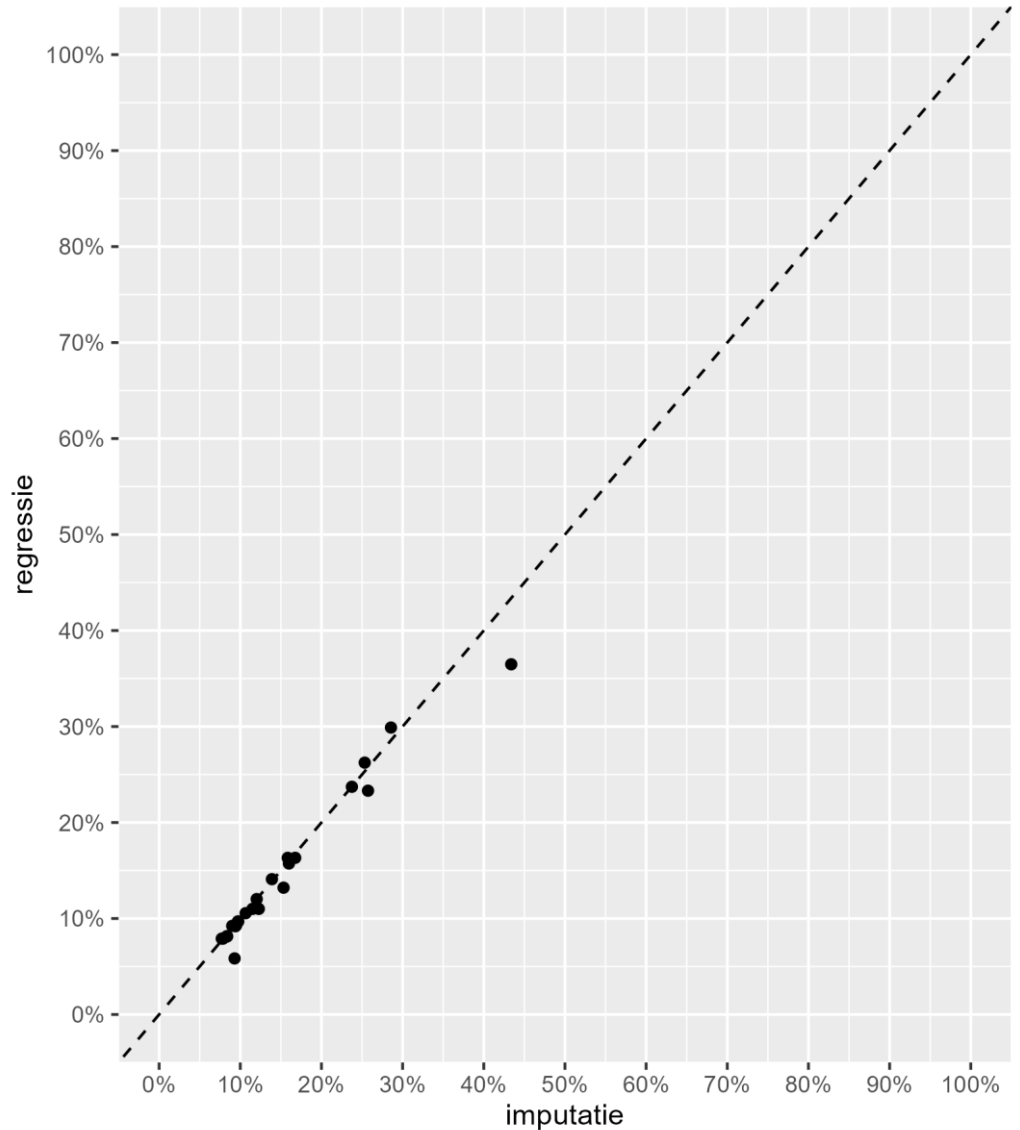
**Figuur 6.7: Gemiddelde strengheid per inspecteur: na imputatie**



inclusief OP0 en alle gescoorde vignet-standaarden



**Figuur 6.8: Vergelijking imputatie en regressie: De correlatie is 0.98**



**Tabel 6.22: Verschillen in strengheid tussen cohorten**

	<b>perc Onvoldoende</b>	<b>standaardfout</b>	<b>95% BI</b>
Dit schooljaar begonnen	10,7	1,9	(7,0, 14,4)
Aan minder dan vijf sko's deelgenomen	16,1	1,7	(12,7, 19,5)
Aan vijf of meer sko's deelgenomen	14,4	1,0	(12,4, 16,3)

**Tabel 6.23: Verschillen in strengheid tussen cohorten**

	<b>perc Onvoldoende</b>	<b>standaardfout</b>	<b>95% BI</b>
Dit schooljaar begonnen	10,7	1,9	(7,0, 14,4)
Aan minder dan vijf sko's deelgenomen	16,1	1,7	(12,7, 19,5)
Aan vijf of meer sko's deelgenomen	14,4	1,0	(12,4, 16,3)



6.3.4 Alternatieve specificaties

**Tabel 6.24: Schattingen overeenstemming volgens drie methoden**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,76	c(0,65, 0,87)
	standaardoordeel	0,89	c(0,85, 0,92)
Fleiss Kappa	eindoordeel	0,53	c(0,32, 0,73)
	standaardoordeel	0,56	c(0,43, 0,70)
Gwet's AC1	eindoordeel	0,71	c(0,57, 0,85)
	standaardoordeel	0,87	c(0,83, 0,91)

**Tabel 6.25: Schattingen overeenstemming volgens drie methoden, waarbij Voldoende met herstelopdracht een aparte categorie is**

	oordeel	est agreement	confidence interval 95
Percent agreement	eindoordeel	0,76	c(0,65, 0,87)
	standaardoordeel	0,84	c(0,80, 0,88)
Fleiss Kappa	eindoordeel	0,53	c(0,32, 0,73)
	standaardoordeel	0,49	c(0,36, 0,61)
Gwet's AC1	eindoordeel	0,71	c(0,57, 0,85)
	standaardoordeel	0,82	c(0,78, 0,87)

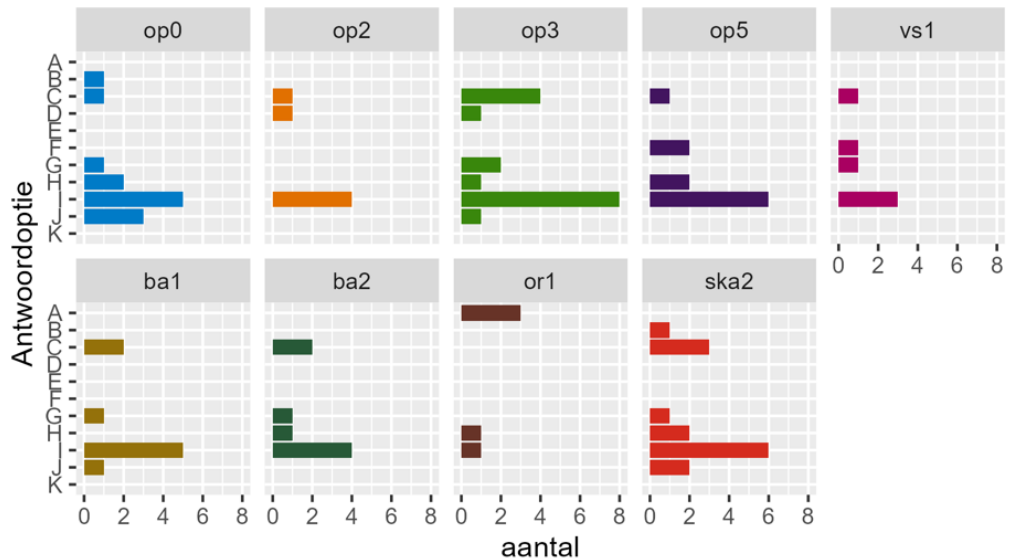
Eindoordelen zijn ongewijzigd door coderen 'voldoende met herstel' en zijn toegevoegd als referentie

**Tabel 6.26: Redenen voor afwijkende oordelen per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is (inclusief OPO Basisvaardigheden)**

Reden voor afwijkend oordeel	Aantal keer <sup>a</sup>
Elementen uit de handreiking verschillend gewogen	42
Contextinformatie opleiding anders gewogen	15
Informatie in het vignet over het hoofd gezien	9
Oordeel op andere standaard meegewogen (voorkomen doortikeffect)	7
Handreiking verschillend geïnterpreteerd	7
Beslisregel OR1 anders toegepast	3
Beredeneerd geoordeeld met oog op effect van toezicht ('dit is geen Onvoldoende of Zeer zwakke opleiding')	3
Contextinformatie bestuur anders gewogen	2
Kenmerken studentenpopulatie anders gewogen	2
Handleiding met handreiking wel/niet gebruikt	0
Toezichthistorie anders gewogen	0
Anders	15



**Figuur 6.9: Redenen voor verschillend oordeel per standaard, waarbij Voldoende met herstelopdracht een aparte categorie is**



- 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

#### 6.3.4.1 Duo oordelen niet aanvullen

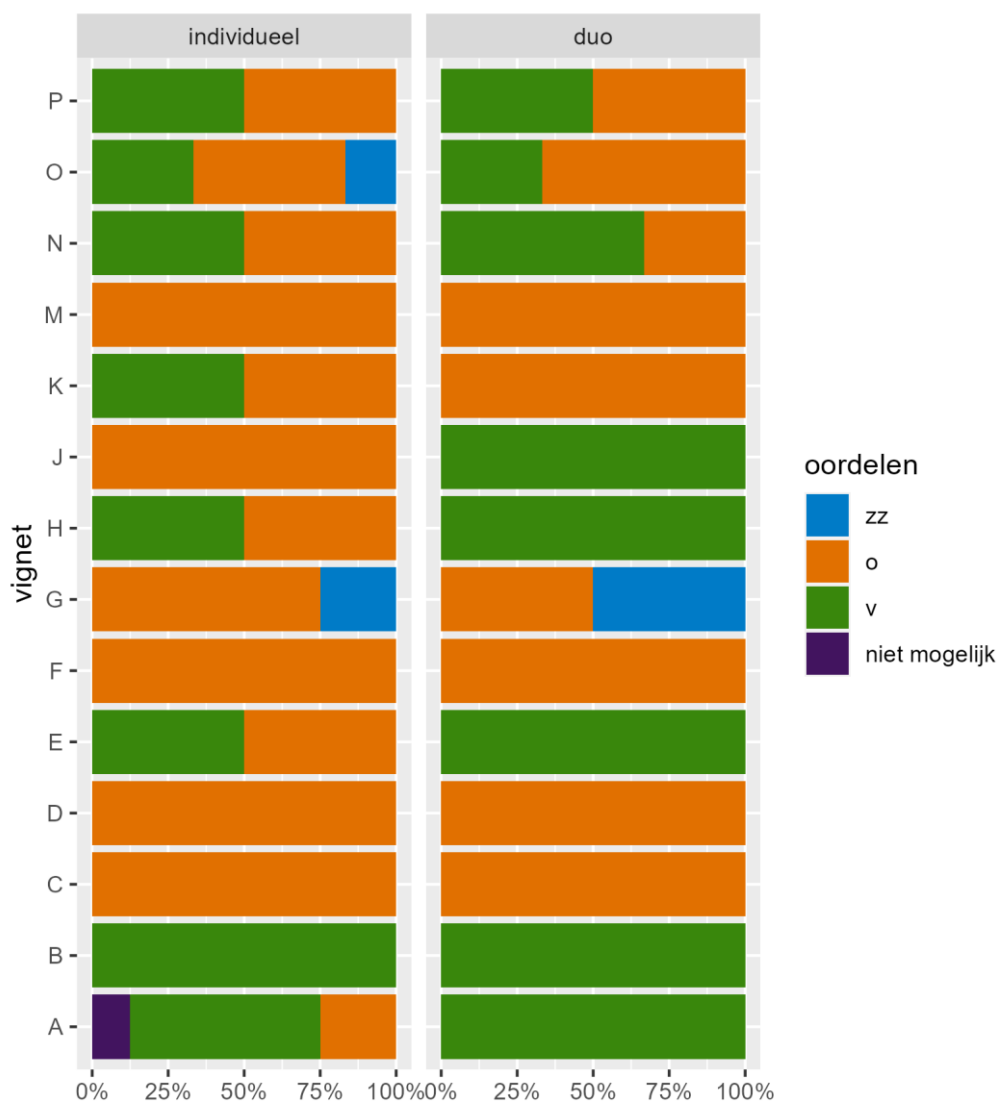
**Tabel 6.27: Verschil in oordelen tussen individuele en duo fase als duo oordelen niet worden aangevuld**

	test	pa individ	pa duo	verschil	t.stat	p waarde	n vign	n indiv	n duo
eindoordeel	paired t test	0,56	0,44	-0,11	-0,43	0,67	16,00	20,00	10,00
standaarden	paired t test	0,38	0,64	0,26	1,47	0,15	122,00	20,00	10,00



6.3.4.2 Veranderingen in eindoordelen

**Figuur 6.10: Verdeling van eindoordelen per fase van de studie: alleen eindoordelen die zijn gewijzigd in duo-fase**



6.3.4.3 Pa voor standaarden met Onvoldoende, Voldoende met herstelopdracht en Voldoende als categorieën

**Tabel 6.28: Schattingen proportie agreement per standaard met Voldoende met herstelopdracht als aparte categorie**

	est_agreement	standaardfout_totaal	confidence_interval_95	soort_oordeel
individueel	0,84	0,02	c(0,80, 0,88)	standaarden
duo	0,91	0,02	c(0,86, 0,95)	standaarden



## 7 Referenties

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

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282.

Rau, G. & Shih, Y. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, 53, 101026.

Tong et al. (2020). The determination of appropriate coefficient indices for inter-rater reliability: Using classroom observation instruments as fidelity measures in large-scale randomized research. *International Journal of Educational Research*, 99, 101514.

Zhao X, Liu JS, & Deng K. (2023). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association*. 36(1), 419–480. Doi:10.1080/23808985.2013.11679142



## Bijlage I

Vignetten worden toegewezen aan blokken door middel van een stratificatieschema dat rekening houdt met zowel de auteur van de vignetten als de historische oordelen. We beginnen met de auteurs. We kunnen de vignetten verdelen in drie groepen van vier vignetten per auteur en een vierde groep met vignetten van de overgebleven twee auteurs. Bovendien hebben we 16 vignetten, die we posities 1, 2, ... 16 geven. Om de auteurs en historische oordelen gelijkmatig over de blokken te verspreiden, definiëren we vier segmenten, waarbij het eerste segment vignetten 1, 2, 3, 4 bevat, het tweede 5, 6, 7, 8 enzovoort. We wijzen nu willekeurig een auteursgroep toe aan elk element van de segmenten, zodat bijvoorbeeld auteur 1 kan worden toegewezen aan het eerste element van elk segment. Vervolgens zorgen we voor de stratificatie van historische oordelen. In elke groep van auteurs hebben we minstens één vignette dat Onvoldoende of Zeer zwak is. We wijzen deze vier vignetten willekeurig toe aan de segmenten, waar ze de elementnummers van de auteurs zullen innemen. In elke auteursgroep hebben we nog drie vignetten over. Deze worden willekeurig toegewezen aan de overgebleven segmenten van de auteurs-elementen.



## Bijlage II

In deze bijlage bespreken we eerder onderzoek naar interbeoordelaarsbetrouwbaarheid (IBB) door onze eigen inspectie, gevolgd door onderzoek door vergelijkbare (internationale) inspecties. Aansluitend betrekken we de beoordeling van leerlingprestaties door leraren, en de beoordeling van leraarprestaties, omdat dit net als ons inspectiewerk professionele beoordeling van kwaliteit betreft in de onderwijscontext.

In 2007 en 2011 namen we de IBB van onze eigen onderwijsinspecteurs onder de loep. In twee vergelijkbare veldexperimenten bezochten veelal verschillende inspecteurs primair onderwijs, in duo's, 60 scholen in 2007 en 52 scholen in 2011. De vraag was of twee inspecteurs onafhankelijk van elkaar tot dezelfde oordelen komen. Anders dan normaal was de opdracht om niet met elkaar te communiceren tijdens het onderzoek. In 2007 werden per onderzoek 42 kwaliteitscriteria beoordeeld, in 2011 waren dat er 50. Elk criterium werd zowel op een vierpuntsschaal (slecht; onvoldoende; voldoende; goed) als op een geaggregeerde tweepuntsschaal (onvoldoende; voldoende) geanalyseerd. Het gemiddelde percentage overeenkomst op de vierpuntsschaal was 94,4% (range: 86-100%) in 2007 en 93,8% (range: 84-100%) in 2011. Voor de tweepuntsschaal betrof dit 95,9% (range: 86-100%) in 2007 en 95,2% (range: 84-100%) in 2011.

Deze twee onderzoeken laten een hoge IBB zien, maar hierbij plaatsen we enkele kanttekeningen. Ten eerste, raadpleging van een in 2007 en 2011 deelnemende inspecteur suggereert dat wel degelijk sprake was van onderlinge beïnvloeding. Zo vond af en toe toch overleg plaats. Ook zou uit de mimiek en de vragen die de collega-inspecteur ter plaatse stelt aan de school af te leiden zijn hoe deze denkt over de kwaliteit van de school. Enige mate van beïnvloeding is dus niet uit te sluiten. De data uit 2011 toont overigens dat een aanzienlijk deel van de duo's op alle 50 kwaliteitscriteria exact overeenkomstig oordeelden. Dit is niet onmogelijk maar wel opmerkelijk. Ten tweede toont de spreiding van oordelen op de kwaliteitscriteria en de afgegeven toezichtarrangementen ('eindoordelen' op schoolniveau) aan dat een relatief homogene groep scholen werd bezocht. Dat wil zeggen, in tegenstelling tot onze huidige aanpak waarbij we relatief veel vignetten van risicoscholen hadden, voldeden de destijds onderzochte scholen in de regel aan de basiskwaliteit. Dit uitte zich in een relatief groot aandeel oordelen 'voldoende'. Het is niet ondenkbaar dat de IBB lager uitvalt wanneer scholen worden beschouwd met meer risicovolle situaties en met meer serieuze gevolgen van beoordeling. Ten derde, in beide studies werd de situatie waarin van beide inspecteurs een score ontbrak ('missing') beschouwd als 'overeenkomstig oordeel'. Hoe vaak die situatie zich voordeed, is niet bekend. Deze drie kanttekeningen samengenomen, betekenen dat de gerapporteerde cijfers hoogstwaarschijnlijk een overschatting zijn van het werkelijke percentage overeenkomst.

De Engelse onderwijsinspectie, Ofsted, voerde in schooljaar 2015-2016 een vergelijkbaar veldexperiment uit. Zij hielden de IBB tegen het licht tijdens zogeheten *short inspections*<sup>7</sup> in het primair onderwijs. De oordelen in kwestie betroffen hier telkens één eindoordeel per school. In 22 van de 24 geanalyseerde

<sup>7</sup> Ofsted (2017). *Do two inspectors inspecting the same school make consistent decisions?* Ofsted research report 170004.



schoolbezoeken kwamen de inspecteurs tot hetzelfde eindoordeel, omgerekend een percentage overeenkomst van 92%. Kanttekeningen die de auteurs zelf plaatsten waren de relatief kleine sample en de specifieke context van *short inspections*. Met dit type inspectie houdt Ofsted een vinger aan de pols bij scholen met het vigerende oordeel *Good* (dit betreft ongeveer 70% van de Engelse scholen en instellingen; respectievelijk krijgen de resterende 5%, 10% en 15% het oordeel *Inadequate*, *Requires Improvement* en *Outstanding*<sup>8</sup>). Het ging dus om een relatief homogene subgroep van scholen. Ook het soort inspectie was specifiek: een beperkt deel van het onderzoekskader werd aangewend om te komen tot een van drie mogelijke uitkomsten: blijft 'goed'; een gedegen oordeel vergt een *full inspection*; omzetten 'goed' naar 'uitstekend' vergt een *full inspection*. Het uitgevoerde type onderzoek kent dus minder observaties, sub-oordelen en (af)wegingen dan een regulier inspectieonderzoek. Tot slot is ook in dit onderzoek de vraag of de duo's werkelijk onafhankelijk opereerden. Bij vier van de bezochte scholen liep een onafhankelijk observator mee om na te gaan of er sprake was van onderlinge beïnvloeding. Dat inspecteurs het in twee gevallen oneens waren en dat de onafhankelijk observator op één van de vier scholen onderlinge beïnvloeding constateerde, gaf de auteurs aanleiding te concluderen dat inspecteurs onafhankelijk van elkaar oordeelden. Het is echter niet uit te sluiten, en allicht aannemelijker, dat ook tijdens enkele schoolbezoeken waar geen observator meeliep sprake was van onderlinge beïnvloeding. Samengevat, het Engelse veldexperiment suggereert een hoge mate van overeenkomst, maar de kanttekeningen wijzen ook hier op een potentiële overschatting.

In 2019 nam Ofsted een specifiek onderdeel van haar werkwijze onder de loep, *workbook scrutiny*, ofwel de doorgronding van werkboeken in het primair onderwijs<sup>9</sup>. Hiermee maken inspecteurs een inschatting van de kwaliteit van curriculumimplementatie. Negen inspecteurs beoordeelden onafhankelijk van elkaar werkboeken van leerlingen uit drie verschillende scholen, op vier aspecten: 'incrementele opbouw van kennis'; 'dekking van onderwerpen'; 'voortgang ten opzichte van het startpunt'; 'gelegenheid tot oefening lesstof'. Hiertoe vulden zij telkens een vijfpuntsschaal in (1=minimaal, 5=maximaal). De berekende kappa-waarden voor de vier kwaliteitsaspecten lagen tussen 0,38 en 0,49. Ofsted concludeerde dat de IBB voor drie van de vier aspecten in orde was en beriep zich daarbij op de breed aangehaalde (maar discutabele) interpretatie van kappa-waarden<sup>10</sup>: 'gering' (0,0-0,20); 'matig' (0,21-0,40); 'redelijk' (0,41-0,60); 'voldoende tot goed' (0,61-0,80); 'bijna perfect' (0,81-1). Het is opvallend dat de betrouwbaarheid in deze studie beduidend lager ligt dan in Ofsted's veldexperiment uit 2017 (kappa = 0,80)<sup>11</sup>. De kleine steekproeven van beide studies beperken de betrouwbaarheid en onderlinge vergelijking van resultaten. Het is moeilijk te verklaren hoe de overeenstemming kleiner kan zijn op een onderdeel van schoolkwaliteit (*workbook scrutiny*) dan voor het globale oordeel over de school. Dat kan alleen als afwijkingen per onderdeel elkaar opheffen. Het zou kunnen dat in de praktijk de intuïtie over de globale kwaliteit van een school vaker gelijk is tussen inspecteurs, en dat deze leidend is bij de interpretatie van een onderdeel zoals *workbook scrutiny*. Maar als deze overeenkomstige intuïtie blijikbaar niet gebaseerd is op de onderdelen die formeel beoordeeld worden (daar lijkt immer sprake van meer verschil van mening) rijst wel de vraag op

8 Inspectie van het Onderwijs (2023) *Internationale Toezichtscan Stimulerend Toezicht en Oordelen* (intern document).

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

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

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



basis van welke aspecten van schoolkwaliteit deze dan wel ontstaat. En waarom deze aspecten dan niet formeel beoordeeld worden. Tot slot is het niet ondenkbaar dat onafhankelijke beoordeling beter geborgd was in het onderzoek naar *workbook scrutiny*. De resultaten reflecteren in dat geval een meer accurate benadering van het werkelijke percentage overeenkomst.

Aanvullend inzicht komt uit onderzoek naar IBB in professionele contexten waar andere inspecties een rol spelen. Zo toont een meta-analyse uit 2012 het effect van drie typen interventies gericht op verbetering van de IBB van professionals in de gezondheidszorg<sup>12</sup>. Op een pre- en post-test maakten deskundigen (para)medische beoordelingen (bijv. de diagnose van enkelfracturen). De gemiddelde overeenkomst op de pre-test was volgens de klassieke interpretatie 'matig', met een kappa-waarde van 0,31. Interventies gericht op het vergroten van de IBB gaven enige verbetering: de gemiddelde overeenkomst op de post-test was 0,43, oftewel 'redelijk'. Aanscherping van het diagnostische instrument had het grootste effect, gevolgd door een combinatie van deze interventie met aanvullende training van de clinicus. Ook enkel training sorteerde effect, maar nog iets minder. De auteurs merkten naast andere methodologische beperkingen op dat de controlegroepen zonder interventie eveneens verbetering over tijd lieten zien. Dit suggereert dat de gerapporteerde stijging in IBB een optimistische voorstelling van zaken geeft.

In 2016 werd gerapporteerd over de IBB van *food safety* inspecteurs werkzaam bij de Amerikaanse equivalent van de Voedsel- en Warenautoriteit<sup>13</sup>. Normaal gesproken gaan deze inspecteurs alleen op pad en toetsen zij restaurants op naleving van voedselveiligheidsnormen. In dit gerandomiseerde onderzoek met controle groep werd nagegaan of een periode van 4 maanden *peer review* – het in duo's afleggen van inspecties en het bespreken van (tegenstrijdige) bevindingen en leerpunten, de IBB ten goede komt. In de interventiegroep werden 378 duo-bezoeken afgelegd, wekelijks door nieuwe duo's om blootstelling aan verschillende inspecteursstijlen te maximaliseren. In de controle groep legden inspecteurs op reguliere individuele wijze restaurantbezoeken af. Telkens werden 52 zogeheten *code items* gecheckt, eenduidige indicatoren zoals 'juiste temperatuur koeling' en 'juiste bereidingstijd'. De resultaten toonden dat naarmate de interventie vorderde, ook de mate van overeenstemming steeg. Volledige overeenstemming over alle 52 items steeg van 35% gemiddeld over de eerste 6 weken van de interventie naar 51% gemiddeld over de laatste zes weken van de interventie. Het gedeelte van de 52 items waar duo's het over eens waren steeg in diezelfde periode van gemiddeld 96,4% naar 97,5% (Bijlage E<sup>14</sup>). Bij deze hoge IBB-getallen merken we op dat de duo's aan het einde van de dag weliswaar onafhankelijk hun oordelen noteerden, maar gedurende de inspectie elkaar continue schaduwden en samenwerkten. De interventie leidde er bovendien toe dat inspecteurs na de interventie tijdens hun reguliere, individuele restaurantbezoeken significant meer grove normschendingen detecteerden in vergelijking met de controlegroep. Deze laatste bevinding betreft overigens een persoonskenmerk, namelijk strengheid, en geen tussenpersoonskenmerk, zoals overeenstemming.

<sup>12</sup> Tuijn SM, Janssens FJG, Robben PBM, 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.

<sup>13</sup> Ho, D.E. (2016). Does peer review work? An experiment of experimentalism. *Stanford Law Review*, 69, 1.

<sup>14</sup> Ibidem.



Vergelijkbaar onderzoek onder inspecteurs van de Ugandese geneesmiddelenautoriteit toont minder gunstige cijfers<sup>15</sup>. Vier willekeurig samengestelde duo's beoordeelden elk twee overheidsinstellingen voor gezondheidszorg op drie zorgniveaus, wat optelde tot een totaal van 24 inspecties. Per inspectie werden 67 indicatoren onafhankelijk gescoord op basis van documentanalyse, observaties en aparte gesprekken met (dezelfde) patiënten. Naast 42 objectieve indicatoren waren er 25 zogeheten subjectieve indicatoren waar persoonlijke inschatting een rol speelt. Indicatoren doorslaggevend voor certificering (80%) werden gescoord op een driepuntsschaal (acceptabel; verbetering nodig; onacceptabel), de overige indicatoren (20%) dichotoom (ja vs. nee). Het mediane percentage overeenstemming voor alle indicatoren samen was 71%. Voor 31 van de 67 indicatoren gold een percentage overeenkomst van 75% of hoger, voor de onderzoekers indicatief voor 'adequate overeenstemming'. Over indicatoren met consequenties voor certificering was meer overeenstemming (range 60,0–87,5%) dan over overige indicatoren (range: 50,0–85,7%). En, over objectieve indicatoren was meer overeenstemming (range: 64,3–87,5%) dan over subjectieve indicatoren (range: 38,1–78,6%).

In een Brits onderzoek naar de kwaliteit van ziekenhuizen gespecialiseerd in acute zorg beoordeelden inspecteurs een tiental vignetten gebaseerd op bestaande inspectierapporten<sup>16</sup>. In de praktijk leggen multidisciplinaire teams van drie tot vijf inspecteurs bezoeken af. In het onderzoek gaven 268 inspecteurs (*response rate* 65%) in een online omgeving, en op individuele basis, per vignet één beoordeling van een kwaliteitsdomein op een vierpuntsschaal (onvoldoende; verbetering nodig; goed; uitstekend). Vervolgens werden 'teambeoordelingen' gesimuleerd op de verkregen data. De resultaten toonden voor individuele scores een gemiddeld percentage overeenstemming van 61% (range: 49–87%). Gesimuleerde data toonden meer overeenstemming bij teams van drie (gemiddeld 70%) en vijf inspecteurs (gemiddeld 73%), al waren deze stijgingen niet statistisch significant. De onderzoekers plaatsen zelf als kanttekeningen bij de externe validiteit van de resultaten dat de vignetten eenduidiger, abstracter en in mindere mate voorzien waren van context dan praktijksituaties.

Tot slot zijn er inzichten over de IBB van leraren die het werk van leerlingen beoordelen, en van observatoren die lerarenprestaties beoordelen. In Engels onderzoek ging men na in hoeverre de beoordeling van examenvragen door examinatoren overeenkomt met een vooraf bepaalde 'correcte beoordeling'<sup>17</sup>. Gesloten vragen (bv. *multiple choice*) gingen zoals verwacht gepaard met hoge IBB: rekening houdend met een vooraf bepaalde bandbreedte lag het percentage overeenkomst tussen 92,7% en 99,6%. Voor meer complexe open vragen (essay) lag dit beduidend lager, tussen 41,7% en 67,1%. Vergelijkbaar Brits onderzoek naar de beoordeling van examens toont, met een geaccepteerde bandbreedte van 10% rond het eindcijfer, een percentage overeenkomst variërend per item van 54,3% tot 97,4%<sup>18</sup>. Tot slot, Amerikaanse onderzoekers rapporteerden in 2012 over een review van studies naar de beoordeling van

15 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.

16 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.

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

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



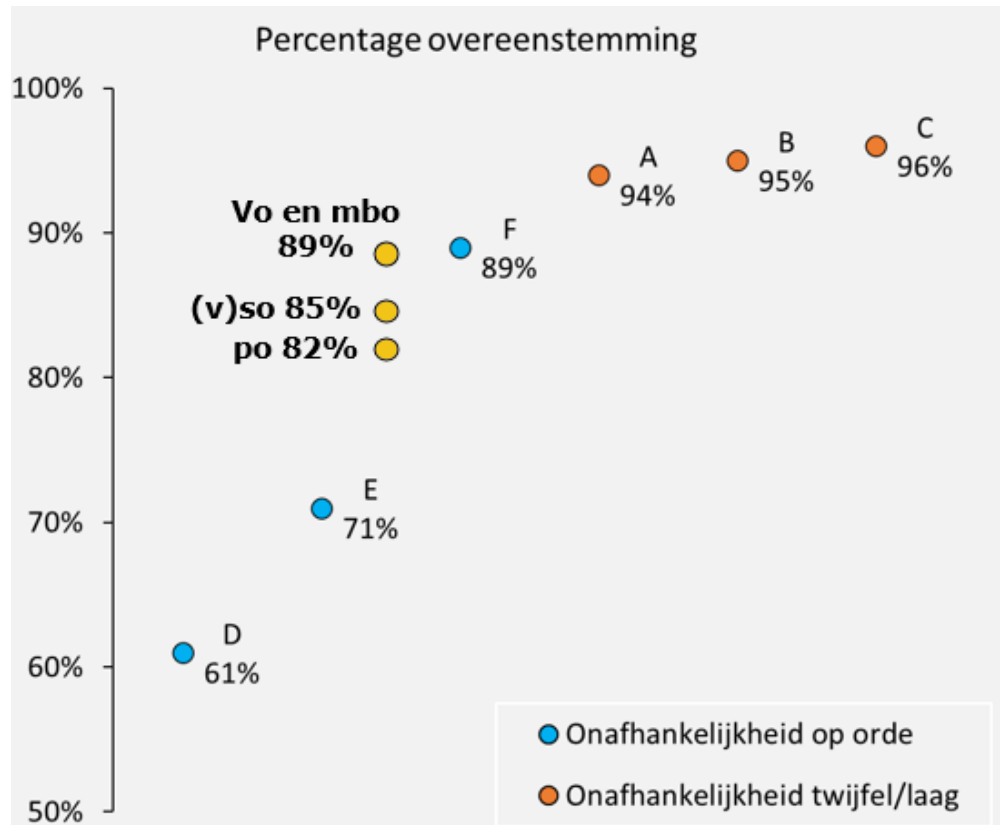
lerarenprestaties<sup>19</sup>. Het gemiddelde percentage overeenkomst tussen beoordelaars, beschouwd over 18 studies, was 70%; de gemiddelde kappa-waarde gebaseerd op zes studies was 0,54.

---

<sup>19</sup> 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.



### Bijlage III



- A) Inspectie scholen voor po in Nederland: interne verslaglegging (2011).
- B) Inspectie scholen voor po in Nederland: interne verslaglegging (2007).
- C) Inspectie restaurants in de VS: Ho, D.E. (2016). Does peer review work? An experiment of experimentalism. *Stanford Law Review*, 69, 1.
- D) Inspectie ziekenhuizen in Engeland: 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.
- E) Gezondheidscentra in Uganda: 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.
- F) Beoordeling verpleegzorg in de VS: Mor, V. et al. (2003). Inter-rater reliability of nursing home quality indicators in the U.S. *BMC Health Services Research*, 3.