12 /25

WEGING VAN ENQUÊTES MET LEGE EN DUN BEZETTE STRATA

SIMULATIESTUDIE VOOR DE VERGELIJKING VAN WEGINGSPROCEDURES VOOR PANEL FRYSLÂN

CIJFERS / MENSEN / INZICHT

// Inhoudsopgave

// Samenvatting 3

// Inleiding 4

// Data 7

// Variabelen 9
Continu . . . 9
Binair . . . 9
Ordinaal . . . 10
Aanpassingen . . . 10

// Wegingsprocedures 11
Basisgewichten . . . 11
Poststratificatie . . . 11
Negeren van lege/te dun bezette strata ('naïeve' poststratificatie) . . . 12
Samenvoegen van lege/te dun bezette strata met andere strata . . . 12
Raking . . . 13
Begrenzing gewichten . . . 14

// Resultaten 16
Responspatronen in de simulaties . . . 16
Maten van schattingskwaliteit . . . 20
Continue variabele . . . 22
Binaire variabele . . . 30
Ordinale variabele . . . 37

// Conclusie 41

// Referenties 44

// Bijlage: R-Algorithme voor het samenvoegen van strata 45

// Samenvatting

Planbureau Fryslân onderhoudt sinds 2017 het burgerpanel 'Panel Fryslân' om onderzoek te doen naar het inwonersperspectief op maatschappelijke kwesties. Sommige groepen inwoners reageren in uitvragen aan dit panel relatief veel, en anderen relatief weinig. Om eventuele vertekeningen te corrigeren, worden resultaten statistisch gewogen. Vaak reageren sommige demografische subgroepen ('strata') echter nauwelijks of helemaal niet op een uitvraag. Hierdoor kunnen sommige wegingsprocedures niet correct worden toegepast.

In dit rapport vergelijken we manieren om antwoorden op vragenlijsten te wegen als correctie op non-respons. Hierbij kijken we specifiek naar een situatie waarin er vanuit sommige strata weinig of geen respons is. Om niet vast te zitten aan wiskundig hanteerbare wegingsprocedures voeren we een simulatiestudie uit. Daarin nemen we lage respons van bepaalde groepen op, en zetten we drie procedures naast elkaar die met weging corrigeren. Twee daarvan zijn aangepaste versies van poststratificatie, die groep voor groep corrigeert. Een derde procedure is raking, die stapsgewijs correctiegewichten schat op basis van de verdelingen van ieder van de groepskenmerken (weegfactoren) apart. We gebruiken de steekproefsamenstelling van vragenlijsten in Panel Fryslân om onze simulaties op basis van een relevant scenario uit te voeren. Wat betreft de uitkomsten kijken we zowel naar de systematische vertekening die gewichten van een bepaalde procedure geven, als naar de onzekerheid van individuele schattingen met behulp van die gewichten. We kijken kort naar de toepasbaarheid van gewichten in subgroepen.

Uiteindelijk heeft raking onze voorkeur boven de twee varianten van poststratificatie, omdat het een kleinere en wat voorspelbaardere vertekening geeft. Schattingen met behulp van raking worden echter minder consistent als er veel variatie binnen groepen is. En welke wegingsprocedure in een subgroep het beste is, is met onze analyse nog niet te zeggen.

Dit onderzoek is voor ons reden om over te gaan op raking als wegingsprocedure in analyses met data van 'Panel Fryslân'. Om te zien hoe afhankelijk onze resultaten zijn van onze simulatie-opzet en wat de beste wegingsprocedure is voor subgroepanalyses, is verder onderzoek nodig. De resultaten van de simulaties zijn, door dun bezette en lege strata te simuleren, relevant voor andere onderzoeksorganisaties die statistieken willen corrigeren op selectieve non-respons, maar te maken hebben met lege of dun bezette wegingsstrata.

// Inleiding

Eén van de hoofdbenaderingen in het onderzoek van Planbureau Fryslân is vragenlijstonderzoek met behulp van een burgerpanel. Het burgerpanel bestaat uit inwoners van de provincie Fryslân van achttien jaar en ouder, die binnen leeftijdsgroepen en regio's willekeurig uitgenodigd worden. Ongeveer zes keer per jaar wordt leden van het panel gevraagd een vragenlijst in te vullen over maatschappelijk relevante onderwerpen.[1]

Hoewel het aanschrijven van inwoners zo wordt uitgevoerd dat iedere inwoner in een leeftijdscategorie een even grote kans heeft om uitgenodigd te worden, is het nog steeds aan een individuele uitgenodigde inwoner om te beslissen of die meedoet aan het panel of niet. Op dezelfde manier hebben panelleden ook vrije keuze om wel of geen antwoord te geven op de enquêtes die naar panelleden worden gestuurd. Verschillen in de bereidheid om mee te doen aan het panel kunnen ervoor zorgen dat de achtergrond- en persoonskenmerken van panelleden verschilt van die van inwoners van Fryslân in het algemeen. Dat soort systematische verschillen tussen de bevolkingsopbouw van Fryslân en de samenstelling van ons panel zijn er (*Panel Fryslân Onderzoeksverantwoording*, 2023), net zoals bij veel andere enquête- en panelonderzoeken. Voor individuele enquêtes is het zo dat bepaalde subgroepen van de inwoners van Fryslân helemaal niet, of maar met een handjevol, reageren.

In theorie hoeft het geen probleem te zijn dat specifieke groepen meer of minder meedoen aan het panel, maar in de praktijk is dat vaak wel een probleem. Als groepen die stelselmatig vaker of minder vaak meedoen aan het panel verschillen in de manier waarop ze vragen in de enquête beantwoorden, dan krijgen we door de afwijkende samenstelling van ons panel vertekende uitkomsten bij de antwoorden op enquêtevragen. Er zijn dan gedeelde verklaringen tussen de bereidheid van mensen om te antwoorden op een enquête en de manier waarop ze die enquête invullen, waarbij de ongecorrigeerde enquête-uitkomsten een vertekend beeld geven van inwoners van Fryslân boven de 18 in het algemeen. Bij groepen die helemaal niet op de enquête reageren, gebeurt dit al helemaal: hun 'stem' komt helemaal niet terug in de enquête-uitkomsten. Daardoor kunnen we hun bijdrage aan het algemene beeld van Fryslân niet meenemen.

Een populaire methode om te corrigeren voor over- en ondervertegenwoordiging van groepen inwoners, die ook binnen Panel Fryslân wordt toegepast, is weging. Bij weging worden groepen die in een vragenlijst een groter aandeel vormen dan in de bevolking minder meegeteld, en groepen die een kleiner aandeel vormen extra zwaar meegeteld (Bethlehem, 2008). Bij Panel Fryslân worden die groepen gevormd door combinaties van de achtergrondkenmerken COROP-regio, geslacht, leeftijdscategorie en opleidingsniveau. Om terminologie consistent te houden met de wetenschappelijke literatuur over weging noem ik dat soort groepen in de rest van dit rapport 'strata' (enkelvoud stratum). Mannen met een havo-, vwo- of mbo-diploma van tussen de 35 en 49 in Noord Fryslân vormen dus een stratum, vrouwen zonder startkwalificatie tussen de 18 en 34 in Zuidwest Fryslân een ander stratum, enzovoorts. Door oververtegenwoordigde strata minder mee te tellen en ondervertegenwoordigde strata extra, trekt weging de afwijkingen tussen het panel en de bevolking van Fryslân boven de 18 die we kunnen zien, zo veel mogelijk recht.

Voor weging kunnen we gegevens gebruiken die van alle inwoners van Fryslân bekend zijn, zoals geslacht, leeftijd, opleiding en woonregio. Voor andere kenmerken, zoals maatschappelijke betrokkenheid, ontbreken zulke gegevens voor de hele bevolking. Daar kunnen we dus niet voor corrigeren. Maar we hopen dat kenmerken waarvan we de over- en onderrepresentatie niet direct kunnen corrigeren, zoals maatschappelijke betrokkenheid, samenhangen met de kenmerken waar we wel voor kunnen corrigeren (geslacht, leeftijd, opleiding en woonregio). Dan kunnen we veel van de vertekening uit onze resultaten halen ondanks het relatief kleine aantal kenmerken waar we op kunnen corrigeren.

De wegingsprocedure die theoretisch het meest recht-door-zee is, is poststratificatie (Bethlehem, 2008). Poststratificatie corrigeert stratum voor stratum voor de over- of ondervertegenwoordiging van dat stratum in de steekproef vergeleken met de populatie. Als het duidelijk is welke factoren zowel invloed hebben op antwoordgedrag als op de neiging om mee te doen aan Panel Fryslân en individuele paneluitvragen, én we die

[1] Zie voor meer informatie https://www.planbureaufryslan.nl/panelfryslan/

factoren als weegfactoren gebruiken, heffen we de vertekening precies op. Dan zijn de uitkomsten in de steekproef representatief voor de inwoners van Fryslân.

Het probleem met poststratificatie is dat het relatief strenge eisen stelt aan de hoeveelheid enquêtedeelnemers dat binnen ieder stratum valt (Bethlehem, 2008). Als strata leeg blijven, werkt de correctie überhaupt niet. Bij dunbezette strata presteert poststratificatie ook niet goed. In enquêtes van Panel Fryslân blijven strata regelmatig leeg, of zijn er maar een handjevol respondenten in een stratum. Vooral bij jongeren en mensen zonder startkwalificatie is dit het geval. Daarom moeten we een andere wegingsprocedure gebruiken dan 'zuivere' poststratificatie.

Eén benadering om de tekortkomingen van poststratificatie op te vangen is ad-hoc aanpassingen doen aan de procedure. Lege strata kunnen bijvoorbeeld genegeerd worden (Bethlehem, 2008, p. 22), ook al geeft dit in theorie verkeerde uitkomsten. Lege of onderbezette strata kunnen ook worden samengevoegd tot één breder stratum dat wel voldoende gevuld is.

In plaats van ad-hoc aanpassingen kunnen we ook voor een heel andere wegingsprocedure dan poststratificatie kiezen. Eén zo'n methode die veel andere grote sociaalwetenschappelijke studies gebruiken (bijvoorbeeld European Social Survey, 2014; European Values Study (EVS), 2022) is raking.[2] Raking is een algoritme dat gebruikt kan worden om te wegen als er onvoldoende populatiedata is voor strata, maar wel voldoende data voor de individuele kenmerken die samen de strata definiëren. Daarnaast is raking beter bestand tegen lege of weinig gevulde strata (Bethlehem, 2008, p. 33). Raking is nog steeds niet immuun voor vertekening door lege strata in de streekproef, omdat het algorithme de samenhang tussen de weegvariabelen gebruikt zoals ze in de steekproef te zien zijn (Založnik, 2011, pp. 205–211). Daarnaast moet raking uitgevoerd worden door een stapsgewijs algoritme, met door de gebruiker gezette stopvoorwaarden in plaats van een directe formule. Er is dan een (gecontroleerd) risico dat het algoritme toevallig op een verkeerde uitkomst uitkomt (Bethlehem, 2008, p. 34). Omdat raking gebruikelijk is voor het wegen van enquêtes, samenhang tussen weegvariabelen mee kan nemen en mogelijk beter om kan gaan met dun bezette strata, overwegen we het als belangrijkste alternatief voor ad-hoc aangepaste poststratificatie.

Wegingsprocedures zijn bedoeld om (systematische) verschillen tussen de inschatting van bijvoorbeeld een gemiddelde van een bepaalde variabele, bijvoorbeeld inkomen of tevredenheid met de woonomgeving, te corrigeren. Dat betekent dat zowel de schatting op basis van de steekproef als de 'echte' waarde in de populatie bekend moet zijn om te zien hoe goed een wegingsprocedure werkt. Echter, de reden waarom we een statistiek als een gemiddelde vanuit een steekproef willen schatten, is juist dat deze niet bekend is voor de populatie. Om wegingsprocedures te vergelijken op een manier die voor ons relevant is, moeten we dus zelf een populatie definiëren met een bepaalde waarde voor de uitkomst die we willen schatten, en kijken hoe goed schattingen met verschillende wegingsprocedures daar in de buurt komen.

Eén benadering is om wiskundig af te leiden wanneer welke wegingsprocedure het beste werkt, maar met een algoritme als raking is dit vrij ingewikkeld (Bethlehem, 2008; Založnik, 2011). Met een simulatiestudie benaderen we het probleem directer en technisch eenvoudiger door een plausibele populatie te maken met geconstrueerde variabelen, hier een steekproef met non-respons die lijkt op wat we in Panel Fryslân zien uit te trekken, en met ieder van de wegingsprocedures het gemiddelde voor de geconstrueerde variabele te schatten.

De onderzoeksvraag achter de simulatiestudie is:

Onder welke omstandigheden kunnen antwoorden op de uitvragen van Panel Fryslân met lege of onderbezette strata beter worden gecorrigeerd voor non-respons met aangepaste poststratificatie, en wanneer beter met raking?

[2] Raking is een dusdanig intuitief algorithme dat het vaak opnieuw is bedacht voor verschillende toepassingen (Založnik, 2011, p. 3). Daardoor heeft het algoritme een aantal verschillende namen. De meest gebruikelijke naam voor het achterliggende algorithme is iterative proportional fitting. 'Raking' komt het meest voor in onderzoek naar weging van enquêtedata, dus ik gebruik deze term.

Om de simulatieresultaten aan te passen op de context van Panel Fryslân, worden er synthetische kopieën van de Friese bevolking, het panel en een paneluitvraag gemaakt. Die synthetische data zijn gebaseerd op empirische data over de samenstelling van de Friese bevolking, de samenstelling van Panel Fryslân en respondenten op drie paneluitvragen. Om het effect van de wegingsprocedures systematisch te onderzoeken, worden er drie soorten kunstmatige toetsingsvariabelen (continu, binair en ordinaal) gemaakt, waar een aantal varianten van worden bekeken.

Omdat vertekenende non-respons en onderbezette of lege strata niet alleen bij Panel Fryslân voorkomen, kunnen de conclusies van dit onderzoek ook voor andere vragenlijstonderzoeken van belang zijn. Daarbij moet er natuurlijk een vertaalslag gemaakt worden. Maar het soort problemen waar bij Panel Fryslân voor gecorrigeerd moet worden, is waarschijnlijk vergelijkbaar met dat bij andere onderzoeken.

In de rest van dit document wordt eerst in detail de simulatiedata en experimentele opzet beschreven, gevolgd door de resultaten. Tot slot wordt besproken welke wegingsprocedure op basis van de simulaties het beste lijkt.

// Data

Bij simulatiemethoden is het niet makkelijk om te bepalen naar welke situaties ze te generaliseren zijn (Morris et al., 2019). In plaats van te proberen iets te zeggen over enquêteweging in het algemeen, probeert dit onderzoek daarom om de simulatie-opzet zo dicht mogelijk bij de situatie van Panel Fryslân te brengen. Hiervoor wordt gebruik gemaakt van CBS-data over de achtergrondkenmerken van de bevolking van Fryslân in 2022, data over de achtergrondkenmerken van deelnemers aan Panel Fryslân ten tijde van paneluitvragen, en data over de achtergrondkenmerken van respondenten op individuele paneluitvragen van Planbureau Fryslân. Deze data vormt de basis van de simulaties.

Als ijkpunt voor de samenstelling van panels en uitvragen gebruiken we paneluitvragen uit juni 2021, mei 2022 en april 2023. Nieuwe respondenten worden namelijk geworven in om de zoveel jaar plaatsvindende (her)wervingen, tot nu toe in 2016/2017 (Visser & Fernee, 2017), 2019 (*Onderzoeksverantwoording*, 2019), 2021 (*Onderzoeksverantwoording*, 2021), 2023 (*Panel Fryslân Onderzoeksverantwoording*, 2023) en 2025 (verantwoording nog te publiceren). Panel Fryslân kan alleen met deze herwerving aan nieuwe deelnemers komen, dus tussen herwervingen neemt het aantal deelnemers gestaag af. Met juni 2021, mei 2022 en april 2023 kijken we naar een uitvraag kort na de herwerving van 2021, een uitvraag midden tussen twee herwervingen in en een uitvraag vlak vóór de herwerving van 2023. Zo is het effect van verlies aan deelnemers over de tijd in beeld te brengen. Vooral omdat bepaalde groepen sneller uitvallen of langer in het panel blijven is dat van belang bij het specifieke probleem van lege en onderbezette strata.

De achtergrondkenmerken die we gebruiken zijn de COROP-regio waarin iemand woont, het geslacht dat iemand aangeeft te hebben[3], diens leeftijdscategorie en diens opleidingsniveau. Voor deze vier achtergrondkenmerken hebben we kruistabellen met het aantal inwoners/panelleden/respondenten per combinatie van achtergrondvariabelen (bijvoorbeeld het aantal vrouwen tussen de 18 en 34 jaar zonder startkwalificatie in Zuidwest Fryslân).

Om van de kruistabellen tot een gesimuleerde uitvraag te komen, maken we eerst een 'lange' populatiedataset. Dit is een datatabel waarin iedere rij één inwoner van Fryslân vertegenwoordigt, waarbij iedere rij kolommen heeft voor ieder van de achtergrondvariabelen. Voor iedere combinatie van achtergrondvariabelen (stratum) zijn er dan evenveel rijen in de lange dataset als het aantal inwoners met die combinatie van achtergrondkenmerken in de kruistabel.

Voordat het synthetische panel en de sythetische uitvraag worden gemaakt, voegen we eerst kunstmatige toetsvariabelen toe aan de 'lange' populatiedataset in nieuwe kolommen. Om de verschillende soorten variabelen die in paneluitvragen veel voorkomen goed te onderzoeken, kijken we naar drie soorten variabelen: Een continue variabele, een binaire variabele en een ordinale variabele. Ieder van de toetsvariabelen wordt bepaald door de achtergrondkenmerken van een inwoner en ruis (een foutenterm). Vervolgens kijken we naar de robuustheid van mijn resultaten door de variabelen op vier vrij combineerbare manieren aan te passen: 1) het toevoegen van een interactie tussen achtergrondvariabelen, 2) het niet-monotoon (niet in één richting lopend) maken van het verband tussen leeftijd en de uitkomsten, 3) het weghalen van het effect van COROP-regio en 4) het vergroten van de hoeveelheid ruis (meetfout) in de variabele. In de sectie "Variabelen" beschrijven we de basisvariabelen en de vier aanpassingen wiskundig.

Als de 'lange' populatiedataset af is, trekken we hier voor iedere simulatie een nieuw panel uit, en trekken we uit dat panel een uitvraag. Om het panel te trekken geven we iedere rij een kans om in het panel te komen gelijk aan het aantal panelleden met de combinatie van achtergrondvariabelen in die rij gedeeld door het aantal inwoners met die combinatie van achtergrondvariabelen. Als er bijvoorbeeld 10 vrouwen tussen de 18 en 34 jaar zonder startkwalificatie uit Zuidwest Fryslân lid waren van Panel Fryslân ten tijde van een uitvraag,

[3] We gaen er daarbij vanuit dat er maar twee geslachtscategorieën zijn, en negeren dus andere/niet-specifieke geslachten, omdat de CBS-data die wij gebruiken uitgaat van een tweedeling en het aantal mensen dat geen 'Man' of 'Vrouw' antwoordt in Panel Fryslân en paneluitvragen te klein is om statistisch verantwoord te gebruiken.

en er woonden toen 2000 vrouwen tussen de 18 en 34 jaar zonder startkwalificatie in Zuidwest Fryslân, dan is de kans van iemand uit dat stratum om onderdeel te zijn van het panel 0,5% ($\frac{10}{2000}$).

Om de uitvraag te trekken heeft ieder lid van het zojuist getrokken panel vervolgens een kans om in de uitvraag te komen gelijk aan het aantal respondenten op een uitvraag met dezelfde combinatie van achtergrondkenmerken gedeeld door het aantal panelleden met die combinatie van achtergrondkenmerken in het echte panel ten tijde van die uitvraag. Als van de 10 vrouwen tussen de 18 en 34 jaar zonder startkwalificatie uit Zuidwest Fryslân die in het panel zaten bijvoorbeeld 2 op een uitvraag reageerden, is de kans van een gesimuleerde vrouw zonder startkwalificatie tussen de 18 en 34 uit Zuidwest Fryslân die lid is van Panel Fryslân om op een uitvraag te reageren 20%. Uiteindelijk voegen we voor iedere simulatie twee variabelen toe aan de 'lange' populatiedataset: de eerste geeft aan of het lid van de populatie in het panel zat voor een simulatie, de tweede of het lid van het panel ook in de uitvraag zat.

// Variabelen

Continu

Als eenvoudig uitgangspunt kijken we naar een continue variabele met de vergelijking:

$$
\begin{aligned}
Y &= \sqrt{\frac{1}{2}} \times \frac{D - \overline{D}}{SD(D)} + \epsilon \\
D &= 0.2 \times I(Leeftijd = \text{“35-49”}) + 0.4 \times I(Leeftijd = \text{“50-64”}) + 0.6 \times I(Leeftijd = \text{“65-74”}) \\
&\quad + 0.8 \times I(Leeftijd = \text{“75+”}) + 1.0 \times I(Geslacht = \text{“Man”}) + 0.2 \times I(COROP = \text{“Zuidwest”}) \\
&\quad + 0.4 \times I(COROP = \text{“Zuidoost”}) + 0.5 \times I(Opleidingsniveau = \text{“havo/vwo/mbo 2-4”}) \\
&\quad + 1.0 \times I(Opleidingsniveau = \text{“hbo/wo”}) \\
\epsilon &\sim \mathcal{N}(0, \frac{1}{2})
\end{aligned}
$$

Kort gezegd neemt Y hier toe met leeftijd, is het hoger bij mannen, is het in Zuidwest en Zuidoost Fryslân hoger dan in Noord Fryslân, en neemt het toe met opleidingsniveau.

De stap van D naar Y zorgt ervoor dat altijd duidelijk is hoeveel variatie er in de variabele zit en hoe die variatie verdeeld is over de foutenterm en het deterministische deel van de vergelijking. In die stap wordt D zo bewerkt dat het een gemiddelde van 0 en een variantie van $\frac{1}{2}$ heeft (dus een standaarddeviatie van $\sqrt{\frac{1}{2}}$), en voeg ik daar een foutenterm met een gemiddelde van 0 en een variantie van $\frac{1}{2}$ aan toe. Daardoor heeft Y altijd een gemiddelde van 0 en een variantie van 1 die evenveel voortkomt uit het deterministische deel als uit de foutenterm ($\mathbb{V}[D + \epsilon] = \mathbb{V}[D] + \mathbb{V}[\epsilon] + 2Cov(D, \epsilon)$ en hier geldt $Cov(D, \epsilon) = 0$). Bij het vergroten van de hoeveelheid ruis verandert dit, omdat er dan nog een variantie van 15 bij de totale variantie komt en de standaarddeviatie van de uiteindelijke variabele 4 wordt (zie Tabel 1).

Binair

De binaire variabele verschilt niet heel veel van de continue variabele. De enige veranderingen zijn dat de foutenverdeling geen normaalverdeling meer is maar een logistische verdeling, en dat een hogere D-waarde nu een hogere kans geeft dat $B = 1$, en niet een hogere Y-waarde.

$$
\begin{aligned}
B &= I(K > 0.5) \\
Logit(K) &= \sqrt{\frac{1}{2}} \times \frac{D - \overline{D}}{SD(D)} + \delta \\
\delta &\sim \mathcal{L}ogistic(0, \sqrt{\frac{1}{2} \times 3 \times \frac{1}{\pi^2}})
\end{aligned}
$$

Ordinaal

De ordinale variabele heeft ook weer hetzelfde deterministische deel D als Y, waar vervolgens een andere foutenverdeling en een transformatie bij komen. In dit geval is het:

$$O = \begin{cases} 1 & \text{als } L \text{ hoogstens 0.1 is} \\ 2 & \text{als } L \text{ tussen 0.1 en 0.2 ligt} \\ 3 & \text{als } L \text{ tussen 0.2 en 0.4 ligt} \\ 4 & \text{als } L \text{ tussen 0.4 en 0.8 ligt} \\ 5 & \text{als } L \text{ boven 0.8 ligt} \end{cases}$$

$$Logit(L) = \sqrt{\frac{1}{2}} \times \frac{D - \overline{D}}{SD(D)} + \psi$$

$$\psi \sim \mathcal{L}ogistic(0, \sqrt{\frac{1}{2} \times 3 \times \frac{1}{\pi^2}})$$

Bij de transformatie van L naar O houd ik de stapgroottes niet gelijk, omdat gelijke stapgroottes een min of meer symmetrische verdeling geven. Een symmetrische variabele ligt dichter bij het soort symmetrische variabelen waar veel statistische theorie van uitgaat, en is daarmee minder lastig voor statistische methodes om mee om te gaan. Om de wegingsprocedures een uitdagender testscenario te geven heb ik de stapgroottes gevarieerd, waardoor de ordinale variabele O linksscheef verdeeld is.

Aanpassingen

In Tabel 1 zijn de vier aanpassingen van de basisvariabelen wiskundig beschreven. De eerste aanpassing voegt een interactie tussen de achtergrondkenmerken geslacht en opleidingsniveau toe, de tweede zorgt ervoor dat de toetsvariabele niet meer stijgt met leeftijd, maar in de leeftijdscategorie van 65 tot en met 74 weer afneemt. De derde aanpassing haalt de invloed van COROP-regio weg zodat er op teveel variabelen gecorrigeerd wordt, en de vierde aanpassing zorgt ervoor dat een groter deel van de variatie in de basisvariabele niet wordt bepaald door achtergrondkenmerken maar willekeurig is.

Tabel 1: Tabel met formalisering van experimentele aanpassingen. $\mathcal{V}er$ staat voor $\mathcal{N}$ bij de continue variabele, $\mathcal{L}ogistic$ bij de binaire en ordinale variabele

Aanpassing	Formalisering
Interactie	$0.5 \times I(Geslacht =$ "Man"$) \times I(Opleidingsniveau =$ "havo/vwo/mbo 2-4"$) -$ $1 \times I(Geslacht =$ "Man"$) \times I(Opleidingsniveau =$ "hbo/wo"$)$
Niet-monotoon effect leeftijd	$0.6 \times I(Leeftijd =$ "35-49"$) + 0.8 \times I(Leeftijd =$ "50-64"$) + 0.4 \times I(Leeftijd =$ "65-74"$)$
Overbodige correctie	$-0.2 \times I(COROP =$ "Zuidwest Fryslân"$) - 0.4 \times I(COROP =$ "Zuidoost Fryslân"$)$
Veel ruis	$\eta \sim \mathcal{V}er(0, s)$ met een s waardoor $SD[D + \epsilon + \eta] = 4$

// Wegingsprocedures

Er zijn drie wegingsprocedures die we naast elkaar zetten: twee varianten op poststratificatie die uitgevoerd kunnen worden als er combinaties van weegvariabelen (strata) zijn waar maar weinig respondenten bij horen en raking. Voor poststratificatie moet eigenlijk een voldoende aantal respondenten per stratum beschikbaar zijn (Bethlehem, 2008, p. 14), maar omdat dit in Panel Fryslân niet consistent het geval is, kijken we hoe goed poststratificatie werkt als die geschonden aanname wordt genegeerd en wat er gebeurt als die met het samenvoegen van strata wordt omzeild. Bij alle wegingsprocedures passen we na het berekenen van gewichten een grenswaarde toe waarbij gewichten die die grenswaarde overschrijden naar beneden bijgesteld worden tot de grenswaarde. Daarnaast zijn alle drie de wegingsprocedures bedoeld ter correctie op non-respons, en worden vóór deze stap basisgewichten bepaald. Deze zijn af te leiden uit de uitnodigingskansen die in het steekproefontwerp van Panel Fryslân zijn ingebed (*Onderzoeksverantwoording*, 2019, 2021; *Panel Fryslân Onderzoeksverantwoording*, 2023; Visser & Fernee, 2017). De basisgewichten corrigeren voor ongelijke uitnodigingskansen en worden startgewichten die door de wegingsprocedure bijgesteld worden.

Basisgewichten

De basisgewichten voor Panel Fryslân (en enquête- en panelonderzoek meer algemeen) corrigeren ervoor dat uit sommige strata meer inwoners uitgenodigd worden dan anderen (Bethlehem, 2008, pp. 6–7), waarbij bepaalde groepen bij Panel Fryslân meer uitgenodigd worden om te compenseren voor beperkte reactie op uitnodigingen (*Onderzoeksverantwoording*, 2019, 2021; *Panel Fryslân Onderzoeksverantwoording*, 2023). Basisgewichten zijn voor iedereen in het steekproefkader (mensen die uitgenodigd kúnnen worden) direct af te leiden uit hun kans om uitgenodigd te worden. Als die kans π_i is, dan is het basisgewicht $\frac{1}{\pi_i}$. Als onder mannelijke inwoners van Noord Fryslân tussen 18 en 34 jaar bijvoorbeeld ongeveer één op de twintig een uitnodiging ontvangt voor het panel, dan wordt het basisgewicht van degenen die meedoen twintig.

Basisgewichten kunnen genormaliseerd worden zodat het gemiddelde basisgewicht 1 is maar groepen die vaker meedoen een groter gewicht hebben. Als er bijvoorbeeld tien panelleden zijn met een basisgewicht van 20 en tien met een basisgewicht van 40, dan rekent de normalisatie de basisgewichten van 20 om naar $\frac{2}{3}$, en die van 40 naar $1\frac{1}{3}$. Dan is het gemiddelde 1, en zijn de hogere gewichten nog steeds het dubbele van de lagere.

De basisgewichten worden gebruikt als uitgangspunt voor de herweging als correctie voor non-respons (Bethlehem, 2008, pp. 4–8). Het uiteindelijke gewicht van een respondent is het product van diens basisgewicht en diens non-responsgewicht.

Poststratificatie

Poststratificatie is de meest directe manier om te corrigeren voor afwijking tussen het aantal respondenten in een bepaald stratum en het aandeel van dat stratum in de populatie. Voor ieder stratum wordt de verhouding berekend tussen het aandeel respondenten in dat stratum en het aandeel populatieleden in dat stratum (Bethlehem, 2008). Respondenten krijgen een correctiegewicht gelijk aan het omgekeerde van die verhouding. Stel bijvoorbeeld dat er in een steekproef van de Friese bevolking vanaf 18 jaar non-respons is onder jonge vrouwen met opleidingsniveau havo, vwo of mbo, zodat ze maar 5 procent van de steekproef vormen, terwijl die groep (bijvoorbeeld) 15 procent van de bevolking van Fryslân vormt. Post-stratificatie geeft die jonge vrouwen met een havo-, vwo- of mbo-diploma in de steekproef een non-responscorrectiegewicht van 3. Daarmee dragen ze weer voor 15 procent bij aan geschatte statistieken op basis van de steekproef. Als het gemiddelde binnen ieder stratum zuiver is doordat de wegingsprocedure gebruik maakt van alle variabelen die zowel beïnvloeden of potentiële deelnemers deelnemen aan het panel/de uitvraag als beïnvloeden wat ze antwoorden in een uitvraag, heft dit de vertekening door non-respons exact op.

Doordat poststratificatie voor ieder stratum apart rekent, moeten er bij alle combinaties van weegvariabelen voldoende respondenten te vinden zijn in de data (één vuistregel is meer dan vijf, Bethlehem, 2008, p. 22). Als er helemaal geen respondenten in het stratum zitten, is het gewicht niet te berekenen. Bij een klein aantal respondenten is de schatting van de variantie binnen het stratum erg onnauwkeurig. Bij Panel Fryslân bevatten niet alle strata consistent voldoende respondenten, en zijn er ook lege strata (zie ook "Responspatronen in de simulaties"). We kijken naar twee strategieën om poststratificatie toch te kunnen gebruiken: het negeren van te kleine strata en het samenvoegen van te kleine strata met andere strata.

Negeren van lege/te dun bezette strata ('naïeve' poststratificatie)

Bij het negeren van te kleine strata worden de poststratificatieberekeningen onveranderd uitgevoerd. Strata die eigenlijk te klein zijn krijgen een gewicht alsof ze wel voldoende respondenten zouden bevatten, en lege strata komen niet in de steekproef voor, dus daarvoor hoeft geen gewicht berekend te worden. Dit maakt gebruik van het gegeven dat de berekening van gewichten voor een stratum 'blind' is ten aanzien van de andere strata. De berekening voor een individueel stratum gebruikt alleen gegevens over dat stratum in de steekproef en de populatie, en de totale omvang van de steekproef en de populatie. Hoewel het berekenen van poststratificatiegewichten bij lege strata dus niet de juiste gewichten geeft om tot goed gecorrigeerde populatieschattingen te komen (Bethlehem, 2008, p. 22), is het rekenkundig wel gewoon mogelijk.

Samenvoegen van lege/te dun bezette strata met andere strata

Bij het samenvoegen van cellen gebruiken we een algorithme dat eerst kijkt of alle strata ten minste vijf respondenten bevatten. Als dat niet zo is, dan wordt eerst de weegvariabele COROP-regio genegeerd. Zijn er vervolgens nog steeds te dun bezette strata, dan voegt het algorithme te dun bezette strata samen met de strata die het meest op dat stratum lijken. Om te bepalen welk stratum het meest lijkt op een ander stratum gebruiken de volgende regel:

1. Ieder stratum krijgt een 'plaats' in een denkbeeldige ruimte met drie dimensies: Een coördinaat gebaseerd op geslacht, een coördinaat gebaseerd op leeftijdscategorie en een coördinaat gebaseerd op opleidingsniveau.
2. Tussen ieder paar strata wordt de euclidische afstand (de afstand 'in vogelvlucht') berekend.
3. Ieder stratum dat te dun bezet is wordt gekoppeld aan het andere stratum dat in de denkbeeldige ruimte het dichtst bij dat stratum ligt. Te dun bezette of lege strata kunnen zowel aan vollere strata als aan andere te dun bezette of lege strata gekoppeld worden. Het dichtsbijzijnde andere stratum is altijd een ander stratum dat aan het uitgangsstratum grenst op ten minste één dimensie/weegvariabele. Diagonalen bij euclidische afstanden zijn namelijk altijd langer dan een rechte lijn langs één van de dimensies van de diagonaal[4], en strata die op meerdere dimensies tegelijkertijd verschillen liggen dus altijd verder van elkaar af dan strata die op één dimensie minder verschillen.

Nadat gekoppelde strata zijn samengevoegd, kijkt het algorithme of er nog lege of te dun bezette strata over zijn, en zolang dit zo is blijft het herhaald strata samenvoegen. Wanneer er geen te kleine strata meer zijn worden poststratificatiegewichten berekend aan de hand van de samengevoegde strata. In de bijlage laten wij de implementatie van het algorithme voor het samenvoegen in R zien.

Dit algorithme is gebaseerd op de praktijk bij Planbureau Fryslân om poststratificatie te gebruiken als correctie voor non-respons ondanks de aanwezigheid van lege of onderbezette strata. Bij lege of onderbezette

[4] Denk aan een bol die als een ballon groeit in een holle vorm met rechte hoeken, zoals een balk of een trappyramide. Welke vorm de holle vorm ook heeft, de bol komt eerst tegen een wand of naar binnen wijzende hoek aan, en komt daarna pas bij welke naar buiten wijzende hoek dan ook.

strata wordt eerst COROP-regio genegeerd, net als bij dit algorithme gebeurd. Als er dan nog steeds lege of onderbezette cellen zijn, worden deze cellen op basis van een informele inschatting samengevoegd met aangrenzende cellen. Het samenvoegingsalgorithme is een manier om dit geautomatiseerd en geformaliseerd na te bootsen.

Bij het bepalen van coördinaten krijgen geslacht, leeftijd en opleiding krijgen verschillende coëfficiënten, waarbij opleiding over alle categorieën samen de grootste coëfficiënten heeft, geslacht de kleinste, en leeftijd daar tussenin zit. Dus de cijfermatige verandering in de coördinaat voor opleiding als een respondent een havo-, vwo- of mbo-diploma heeft is in plaats van geen startkwalificatie te hebben, is bijvoorbeeld groter dan de verandering in de coördinaat voor geslacht tussen mannen en vrouwen. Dat betekent dat twee strata die in geslacht en leeftijd gelijk zijn, maar waarbij de ene bestaat uit inwoners van Fryslân zonder startkwalificatie en de ander uit inwoners met een havo-, vwo- of mbo-diploma, minder snel samengevoegd worden dan twee strata die in leeftijdscategorie en opleiding gelijk zijn, maar waarbij de ene bestaat uit mannen en de ander uit vrouwen. Zie de bijlage voor de exacte cijfers.

De coëfficiënten voor de weegvariabelen zijn een indirect oordeel over de verschillen binnen en tussen categorieën van die variabele. Weegvariabelen waarbij de toetsvariabele relatief veel verschilt tussen categorieën ten opzichte van verschillen binnen de categorieën zijn minder aantrekkelijk om samen te voegen, weegvariabelen met weinig verschil tussen categorieën aantrekkelijker (Bethlehem, 2008, pp. 15–16). Bij het samenvoegen van categorieën wordt binnen die samenvoeging niet meer gecorrigeerd met weging. Als de strata veel van elkaar verschillen, zijn er relatief grote verschillen waar idealiter wel voor gecorrigeerd zou worden, maar waarvoor dat niet meer gebeurt. Als de strata weinig van elkaar verschillen, maakt het wegvallen van correctie weinig uit. De grootte van de coëfficiënten vertaalt dat idee: een grotere coëfficiënt tussen twee categorieën maakt ze minder aantrekkelijk om samen te voegen en past dus bij variabelen met grotere verschillen tussen categorieën, bij een kleinere coëfficiënt is samenvoegen aantrekkelijker.

Welke variabelen door de verschillen tussen hun categorieën belangrijk zijn om te behouden bij weging hangt natuurlijk af van de variabele waarvoor statistieken geschat moeten worden. Bij gebruik van specifieke socialmediaplatforms kan leeftijd bijvoorbeeld best belangrijker zijn dan opleidingsniveau. En bij het spreken van Fries in huis is het als eerste negeren van COROP-regio waarschijnlijk een minder logische keuze dan het eerst negeren van geslacht. In dit onderzoek houden we daar geen rekening mee, en hebben we coëfficiënten gekozen die vrij breed bruikbaar zouden moeten zijn. Er zijn, zoals hierboven genoemd, makkelijk gevallen te bedenken waarin een andere keuze beter had gewerkt, maar de huidige waarden zijn in veel gevallen bruikbaar.

Raking

Raking, of multiplicatief wegen, is een manier om met verdelingen van de weegvariabelen in de populatie en de gezamelijke verdeling van de weegvariabelen in de steekproef gewichten te berekenen (Bethlehem, 2008, pp. 33–38; Založnik, 2011). Raking kan gebruik maken van doorgekruisde populatieverdelingen, maar heeft in principe alleen de marginale verdelingen van de weegvariabelen nodig. Doordat het niet per stratum vergelijkt tussen de populatie en de steekproef, heeft raking minder beperkingen door de beschikbaarheid van populatiedata, en kan het ook beter omgaan met onderbezette of lege cellen. Lege strata in de steekproef maken de weging nog steeds minder effectief, en als categorieën van weegvariabelen leeg blijven, functioneert het algorithme niet (Založnik, 2011). Maar zolang de categorieën van de individuele weegvariabelen voldoende gevuld zijn, is raking te gebruiken. Als methode om te vergelijken met de aangepaste poststratificatiemethoden, en als mogelijke vervanging van poststratificatie voor Panel Fryslân, kijken we dus ook naar de schattingsfout die raking geeft.

Raking werkt door het cyclisch variabele voor variabele bijstellen van gewichten om de afwijking tussen de steekproef en de populatie op die variabele te corrigeren (Barthélemy & Suesse, 2018; Bethlehem, 2008, pp. 33–38; Forthomme et al., z.d.). Het bijstellen op de ene variabele verstoort de bijstelling op de voorgaande

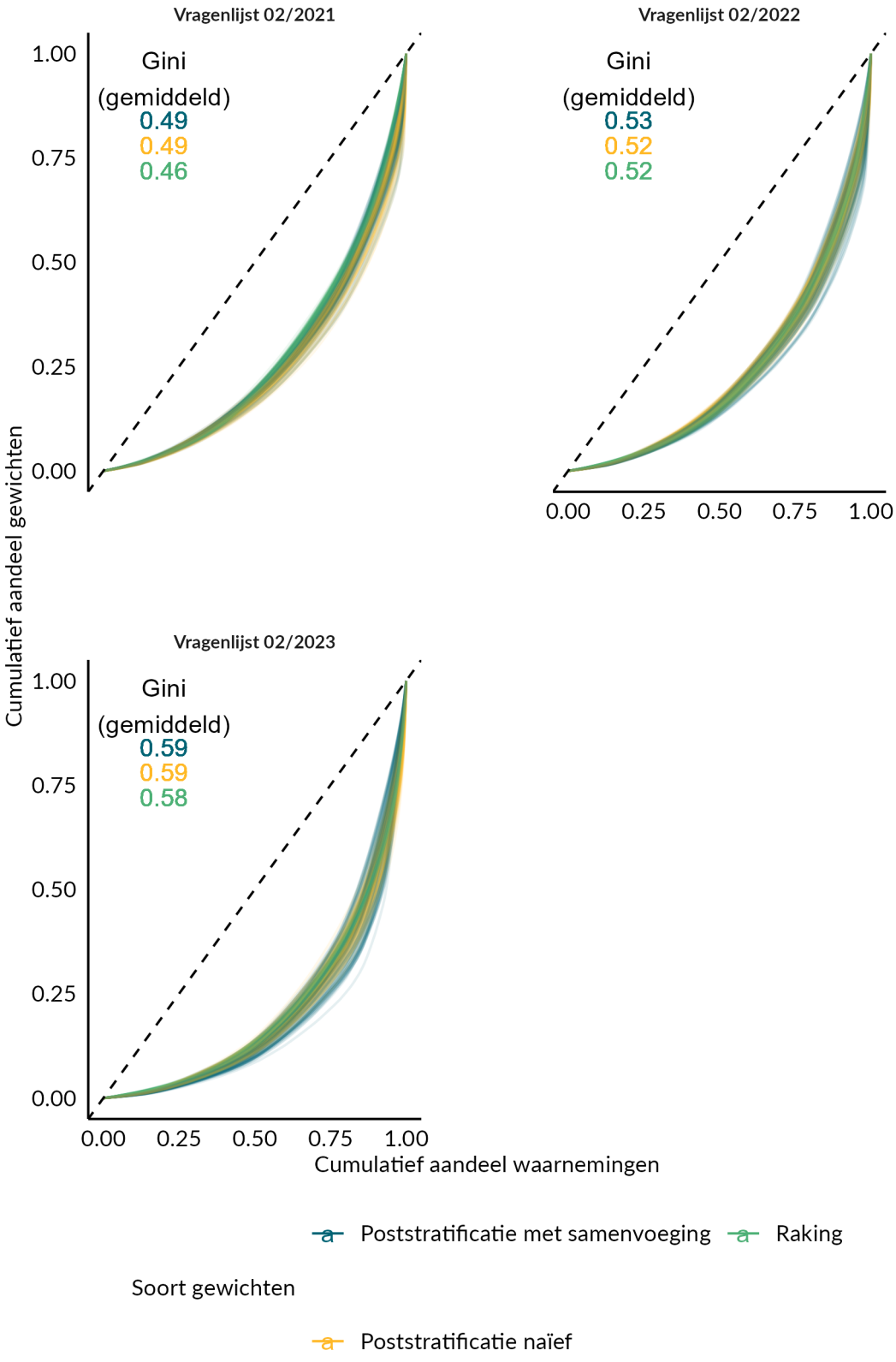
variabelen, maar omdat het verband tussen de weegvariabelen dat in de steekproef te zien is bewaard moet worden, doet het die niet helemaal teniet. De cyclus herhaalt zich totdat de gewichten door een nieuwe bijstelling nauwelijks nog veranderen (of het algorithme al door een groot aantal cycli is gegaan). Doordat iedere volgende bijstelling de gewichten van de vorige gebruikt, komt het algorithme bij een set gewichten die voor alle individuele variabelen zorgt voor een gewogen steekproefverdeling die zo nauw als de gebruiker wil aansluit op de populatieverdeling en de verbanden in de steekproef behoudt.

Begrenzing gewichten

In de praktijk van enquête-onderzoek worden gewichten meestal begrensd om de bijdrage van observaties aan schattingen niet te ongelijk te maken. Grote verschillen in gewichten kunnen de schattingsfout namelijk groter maken (Potter & Zheng, 2015). Tegelijkertijd komt met te strenge begrenzing de vertekening die de gewichten moeten corrigeren weer terug (Potter & Zheng, 2015). Veel onderzoeken werken met een begrenzing van 3 tot 4 keer het gemiddelde of mediane gewicht (Kerckhove et al., 2014). De European Social Survey hanteert bijvoorbeeld een maximumgewicht van 4 (European Social Survey, 2014). Andere studies kijken naar vaste delen van de verdeling, de European Values Study brengt de 2,5% grootste gewichten bijvoorbeeld terug tot de 97,5de percentiel (European Values Study (EVS), 2022). In het geval van de European Values Study is het maximale gewicht hiermee ongeveer twintig keer het gemiddelde gewicht (eigen berekening op EVS/WVS, 2024). Naast vuistregels over wat een te groot gewicht is, is er enig theoretisch onderzoek gedaan naar de kwaliteit van verschillende soorten begrenzingsprocedures (Kerckhove et al., 2014; Potter & Zheng, 2015). Het systematisch toepassen van die theorie op Panel Fryslân is aan later onderzoek.

Om vanuit de intuïtie dat het begrenzen van gewichten te ongelijke weging voorkomt een vergelijking te maken tussen de wegingsprocedures, laat Figuur 1 Lorenzcurven en Ginicoëfficiënten voor de gewichten uit alle drie de procedures zien voor alle simulaties met alle vragenlijsten. Voor de vragenlijst uit mei 2022 en april 2023 lijken de Lorenzcurven redelijk door elkaar heen te lopen, en lijkt de ene wegingsprocedure niet systematisch voor meer of minder ongelijkheid in gewichten te zorgen dan de ander. Bij de vragenlijst uit juni 2021 lijkt de ongelijkheid in gewichten bij raking meestal kleiner te zijn dan bij de twee soorten poststratificatie. De ongelijkheid van gewichten verkregen met raking is in alle vragenlijsten een stuk constanter dan voor de twee vormen van poststratificatie, vooral bij poststratificatie met samenvoeging verschuiven en vervormen de Lorenzcurven van simulatie tot simulatie meer.

Uitendelijk hebben we ervoor gekozen om gewichten groter dan twintig keer het gemiddelde gewicht te begrenzen tot twintig keer het gemiddelde. De toevalligheden in het antwoordgedrag van respondent die meer dan twintig keer zwaarder meewegen dan gemiddeld krijgen zoveel invloed op de resultaten dat het niet verantwoord lijkt om dat gewicht te behouden. Met een grens van twintig zitten we in de buurt van de grens die de European Values Study in de praktijk trekt. Aangezien het begrenzen van gewichten soms meer kwaad dan goed kan doen, zijn we relatief voorzichtig (Potter & Zheng, 2015). De begrenzing raakt in de simulaties 0.11% van de ‘naïeve’ poststratificatiegewichten, 0.031% van de poststratificatiegewichten bij het samenvoegen van wegingsstrata, en 0.0054% van de rakinggewichten. Dat het percentage begrensde gewichten bij raking veel kleiner is dan bij de andere wegingsprocedures ligt er misschien aan dat de ongelijkheid in gewichten in de bovenkant van de verdeling regelmatig kleiner is voor raking dan voor de twee vormen van poststratificatie.



Figuur 1: Lorenzcurven van de gewichten uit de simulaties op basis van de drie vragenlijsten. *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.

// Resultaten

Voor het interpreteren van de resultaten kijken we eerst naar de bezetting van de combinaties van achtergrondvariabelen in de drie gesimuleerde paneluitvragen. Voor 'correcte' poststratificatie moeten namelijk respondenten beschikbaar zijn voor iedere combinatie van achtergrondvariabelen die ook in de populatie voorkomt (Bethlehem, 2008, p. 14). Daarom bekijken we hoe vaak in de simulaties poststratificatie is toegepast terwijl dit eigenlijk niet correct is. Vervolgens introduceren we de maat waarmee we de zuiverheid van de schattingen op basis van de wegingsprocedures heb gemeten, en de maat voor de niet-systematische schattingsfouten (de variantie van de schatters). Daarna gaan we door de resultaten voor de continue, binaire en ordinale toetsvariabele. Bij het analyseren van de toetsvariabelen gebruiken we 95%-betrouwbaarheidsintervallen (Morris et al., 2019, p. 2086). Om een betrouwbaarheidsinterval te krijgen dat klein genoeg is in verhouding tot de vertekeningsmaat draaien we 100 simulaties per paneluitvraag (Morris et al., 2019, p. 2088).[5]

Responspatronen in de simulaties

In Tabel 2 is voor de vragenlijst uit juni 2021 bij iedere combinatie van achtergrondvariabelen te zien in hoeveel simulaties (welk percentage van de simulaties) geen respondenten die combinatie van achtergrondvariabelen hadden of minder dan vijf respondenten in dat stratum zaten. Vooral inwoners zonder startkwalificatie onder de vijftig ontbreken vaak in de gesimuleerde uitvragen. Vrouwen ontbreken daarbij vaker dan mannen. Vrouwen van 18 tot en met 34 jaar zonder startkwalificatie in Zuidoost Fryslân missen altijd, aangezien er geen respondenten uit deze groep voorkwamen in de uitvraag waar de simulaties op gebaseerd zijn. Dat betekent dat 100 van de 100 simulaties ten minste één leeg stratum hadden, waardoor de aannames van poststratificatie nooit opgaan.

Tabel 3 laat percentages niet of weinig geobserveerde combinaties van achtergrondvariabelen zien voor simulaties op basis van de uitvraag van mei 2022. Hier is het aantal lege of onvoldoende gevulde categorieën toegenomen door de uitval van respondenten. Het zijn nog steeds jonge inwoners zonder startkwalificatie die het vaakst ontbreken in de steekproef. In 100 van de 100 simulaties was er ten minste één lege cel. Ook hier gingen de aannames van poststratificatie dus nooit op.

Zoals Tabel 4 laat zien, zijn de niet-geobserveerde combinaties van achtergrondvariabelen bij simulaties gebaseerd op de uitvraag van april 2023 vooral jonge inwoners zonder startkwalificatie. In 100 van de 100 simulaties was er ten minste één lege cel. Net als bij de andere twee vragenlijsten was poststratificatie hier dus nooit zonder aanpassing bruikbaar geweest.

[5] **Rekenvoorbeeld:** de variantie (en dus standaarddeviatie) van de continue doelvariabele is 1. De variantie van een ongewogen gemiddelde in een steekproef in de orde van 2000 deelnemers is dan in de orde $1/2000$. De standaardfout van de geschatte vertekening is $\sqrt{\frac{1}{n_{sim}} \times \frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2} = \sqrt{\frac{1}{n_{sim}} \times \widehat{\mathbb{V}[\hat{\theta}]}}$ (Morris et al., 2019, p. 2086). Met $\widehat{\mathbb{V}[\hat{\theta}]} \approx \frac{1}{2000}$ is de standaardfout dan in de orde $\sqrt{\frac{1}{100} \times \frac{1}{2000}} = \sqrt{\frac{1}{200000}} \approx 0{,}002$. De standaardfouten voor gewogen gemiddelden zullen groter zijn, maar dit geeft wel een uitgangspunt.

Tabel 2: Aantal simulaties op vragenlijst 2/2021 (juni 2021) waarin een combinatie van achtergrondvariabelen niet/minder dan vijf keer voorkomt.

COROP	Geslacht	Leeftijd	Basisonderwijs, vmbo, mbo 1	Havo, vwo, mbo 2-4	Hbo, wo
Noord Fryslân	Mannen	18-34	/31		
		35-49	/1		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	9/77		
		35-49	/19		
		50-64			
		65-74			
		75+		/20	
Zuidwest Fryslân	Mannen	18-34	17/92		
		35-49	1/37		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	100/100	/2	
		35-49	6/77		
		50-64			
		65-74			
		75+		2/40	/13
Zuidoost Fryslân	Mannen	18-34	9/87		
		35-49	1/63		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	13/94		
		35-49	22/96		
		50-64	/1		
		65-74			
		75+		/10	

Noot: Data op basis van honderd simulaties door auteur, gecalibreerd op de panelsamenstelling van Panel Fryslân ten tijde van de uitvraag. Waarde voor de schuine streep geeft het aantal simulaties waar het stratum leeg was aan, waarde achter de schuine streep het aantal simulaties waarin er minder dan vijf respondenten in het stratum zaten. Nullen worden voor de overzichtelijkheid niet getoond.

Tabel 3: Aantal simulaties op vragenlijst 2/2022 (mei 2022) waarin een combinatie van achtergrondvariabelen niet/minder dan vijf keer voorkomt.

COROP	Geslacht	Leeftijd	Basisonderwijs, vmbo, mbo 1	Havo, vwo, mbo 2-4	Hbo, wo
Noord Fryslân	Mannen	18-34	4/85		
		35-49	1/20		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	100/100		
		35-49	3/71		
		50-64			
		65-74			
		75+		/1	
Zuidwest Fryslân	Mannen	18-34	100/100	/14	/7
		35-49	38/99	/1	
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	37/100	/47	/9
		35-49	26/100	/9	
		50-64	/1		
		65-74			
		75+		/11	
Zuidoost Fryslân	Mannen	18-34	34/99	/3	/5
		35-49	/44		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	36/100	/1	/1
		35-49	1/58		
		50-64	/1		
		65-74			
		75+		/5	

Noot: Data op basis van honderd simulaties door auteur, gecalibreerd op de panelsamenstelling van Panel Fryslân ten tijde van de uitvraag. Waarde voor de schuine streep geeft het aantal simulaties waar het stratum leeg was aan, waarde achter de schuine streep het aantal simulaties waarin er minder dan vijf respondenten in het stratum zaten. Nullen worden voor de overzichtelijkheid niet getoond.

Tabel 4: Aantal simulaties op vragenlijst 2/2023 (april 2023) waarin een combinatie van achtergrondvariabelen niet/minder dan vijf keer voorkomt.

COROP	Geslacht	Leeftijd	Basisonderwijs, vmbo, mbo 1	Havo, vwo, mbo 2-4	Hbo, wo
Noord Fryslân	Mannen	18-34	11/95		
		35-49	/36		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	100/100		
		35-49	7/80	/1	
		50-64			
		65-74			
		75+			
Zuidwest Fryslân	Mannen	18-34	100/100	3/70	/14
		35-49	35/100	/2	/1
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	37/99	39/98	/16
		35-49	100/100	1/28	/5
		50-64			
		65-74			
		75+	/1	/1	
Zuidoost Fryslân	Mannen	18-34	35/99	1/44	/10
		35-49	/41		
		50-64			
		65-74			
		75+			
	Vrouwen	18-34	34/100	8/80	/12
		35-49	31/100	/21	
		50-64	/4		
		65-74		/1	
		75+		/10	

Noot: Data op basis van honderd simulaties door auteur, gecalibreerd op de panelsamenstelling van Panel Fryslân ten tijde van de uitvraag. Waarde voor de schuine streep geeft het aantal simulaties waar het stratum leeg was aan, waarde achter de schuine streep het aantal simulaties waarin er minder dan vijf respondenten in het stratum zaten. Nullen worden voor de overzichtelijkheid niet getoond.

Maten van schattingskwaliteit

Om de vertekening van geschatte gemiddelden van de continue en binaire variabelen te meten, gebruiken we de absolute gemiddelde afwijking tussen een schatting en de parameter die het moet schatten:

$$\widehat{|Bias|} = \left| \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i - \theta \right|$$

Waarbij B staat voor het aantal simulaties ($B = 100$, in ons geval), θ voor de te schatten parameter en $\hat{\theta}_i$ voor de schatting in de $i$de simulatie. Varianten van absolute gemiddelde afwijking worden meer gebruikt bij simulatieonderzoek in statistische methoden (Robbins et al., 2019; bijvoorbeeld Rueda et al., 2023; Yang et al., 2024; zie ook Morris et al., 2019). Daarbij hebben we ervoor gekozen om geen relatieve maat te gebruiken, omdat de continue variabele een modelgemiddelde van 0 heeft en relatieve maten dus erg onstabiel zijn (Morris et al., 2019, p. 2086). Daarnaast passen we de omvorming naar absolute waarden toe ná het berekenen van het gemiddelde, omdat anders zuiverheid en betrouwbaarheid van een schatter door elkaar gaan lopen (Morris et al., 2019, pp. 2086–2087). Als de individuele schattingsfouten worden omgezet naar absolute waarden en daarna pas gemiddeld worden, dan heeft ook een zuivere schatter een absolute gemiddelde afwijking groter dan 0. Schattingsfouten kunnen elkaar dan namelijk niet meer uitmiddelen.

Naast de zuiverheid van een schatter is ook de schattingsfout van of variatie in een schatter van belang (James et al., 2021, pp. 33–36; Morris et al., 2019, p. 2077). Een schatter die gemiddeld over veel schattingen zuiver is, maar waarvan er van uitgegaan moet worden dat een individuele schatting ver van de te schatten parameter af kan liggen, geeft met de ene schatting die in de praktijk mogelijk is nog steeds geen betrouwbaar beeld van de populatie. Daarom kijken we naast zuiverheid ook naar de empirische standaardfout van de geschatte gemiddelden per wegingsprocedure:

$$\widehat{EmpSE} = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i - \bar{\hat{\theta}})^2}$$

De notatie is hetzelfde als die voor de absolute gemiddelde afwijking, met de nieuwe term $\bar{\hat{\theta}}$ voor het gemiddelde van de $B = 100$ schattingen voor een wegingsprocedure voor een bepaalde variant van de toetsvariabele. De empirische standaardfout geeft aan hoe ver een schatting gemiddeld van de gemiddelde schatting af ligt. Dit is de schattingsfout die niet aan onzuiverheid is toe te schrijven. Bij een zuivere schatter is dit te vergelijken met de gemiddelde afwijking van de parameter, al wordt hier gekwadrateert in plaats van het gemiddelde genomen om aan te sluiten op de kleinste-kwadratenmethode die grotere afwijkingen zwaarder rekent en een centraal onderdeel vormt van de klassieke statistiek.

Voor de ordinale variabele wordt de verdeling over alle categorieën geschat (iedere categorie krijgt een geschat aandeel), dus hier is één afwijking niet te gebruiken. In plaats van een absolute gemiddelde afwijking gebruik ik daarom de gemiddelde (euclidische) afstand tussen de set geschatte proporties en de modelproporties:

$$\widehat{Afstand} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{\sum_{j=1}^{m} (\hat{p}_{ji} - p_j)^2}$$

Waarbij m het aantal proporties is (hier 5), p_j de modelproportie van categorie j, en $\hat{p}_{ji}$ de geschatte proportie voor categorie j in simulatie i. Dit is ook een absolute maat, de *afstand* tussen twee dingen is nooit negatief, en is toe te passen op een set variabelen ongeacht het aantal variabelen. Afstand gedeeld door het

aantal dimensies zou in schaal dichter bij absolute gemiddelde afstand voor één proportie in de buurt zitten, maar omdat we niet tussen typen variabelen vergelijken is dat niet relevant.

Een nadeel aan afstand als foutmaat is dat het zowel vertekening ook schattingsfout meeneemt, waardoor we die twee niet zo mooi kunnen scheiden als we bij de absolute gemiddelde schattingsfout en empirische standaardfout konden doen (Morris et al., 2019, pp. 2086–2087). Dit is van belang bij de interpretatie, omdat de resultaten voor de ordinale toetsvariabele dus bij elkaar voegen 1) of de schatting systematisch hoger of lager is dan de parameter, en 2) of individuele schattingen door toevallige invloeden van de parameter af komen te liggen.

Continue variabele

Per toetsvariabele gaan we per vragenlijst door de resultaten. Figuur 2 laat de absolute gemiddelde schattingsfout over de honderd simulaties zien per methode om het gemiddelde van de continue toetsvariabele te schatten bij de simulaties gebaseerd op de uitvraag uit juni 2021. Hier lijkt raking vrij consistent minstens even zuivere schatters te geven als de twee vormen van poststratificatie, en zijn alle gewogen schattingen zuiverder dan de ongewogen schattingen. Als veel van de aanpassingen aan de basisvariabele tegelijkertijd actief zijn, neemt de achterstand van de twee varianten van poststratificatie ten opzichte van raking af. Als er een interactie is tussen twee weegvariabelen, neemt de onzuiverheid van de ongewogen schatting duidelijk af. In dit geval betekent de interactie dat de effecten van het hebben van een hbo- of wo-diploma en een man zijn niet meer optellen, terwijl die van man zijn en een havo-, vwo- of mbo-diploma hebben elkaar versterken. Respondenten zonder startkwalificatie en vrouwen zijn in de gesimuleerde steekproeven relatief sterk ondervertegenwoordigd, terwijl mannen en vooral respondenten met een hbo- of wo-diploma oververtegenwoordigd zijn. Het kan zijn dat het verkleinen van het verschil tussen vrouwen en respondenten zonder startkwalificatie en mannen en respondenten met een hbo- of wo-diploma die ondervertegenwoordiging een minder groot probleem maakt. Dan zou de ongewogen schatting zuiverder worden.

De verschillen tussen de drie wegingsprocedures zijn klein en onzeker, en raking lijkt het in ieder geval niet slechter te doen dan de andere wegingsprocedures.

De resultaten in Figuur 3 laten de empirische standaardfout per wegingsprocedure en variant van de continue toetsvariabele zien voor de simulaties gebaseerd op de uitvraag uit juni 2021. Wat direct duidelijk te zien is, en in de ordening van de figuur wordt benadrukt, is dat het voor de empirische standaardfout van de wegingsprocedures veel uitmaakt of individuele afwijkingen van het stratumgemiddelde groot of klein zijn, en er dus veel of weinig ruis is. Maar de verschillen tussen de wegingsprocedures blijven ongeveer hetzelfde, en ook de ongewogen schatting wordt minder betrouwbaar. De andere bijzonderheden aan de toetsvariabele maken dan weinig uit. Nu betekent meer ruis dat de variantie van de variabele groter wordt, en de standaardfout is meestal, heel kort door de bocht, de variantie van een variabele waarmee geschat wordt gedeeld door een getal dat groter is naarmate er meer respondenten zijn. Dus dat schatters bij meer ruis minder betrouwbaar worden, ligt voor de hand.

Deze eerste resultaten voor empirische standaardfouten lijken ook een goede plek om er aandacht aan te besteden dat de ongewogen schattingen een kleinere empirische standaardfout hebben dan de gewogen schattingen. Het is over het algemeen zo dat weging standaardfouten groter maakt. Omdat de ongewogen schattingen verreweg de grootste vertekening hebben is het echter geen serieuze optie om in plaats van een vorm van weging een ongewogen schatting te gebruiken. De standaardfout van de ongewogen schatting geeft wel een referentiepunt om te kijken hoeveel groter weging de standaardfouten maakt. En die toename lijkt bij alle wegingsprocedures mee te vallen.

Figuur 4 laat de zuiverheid per wegingsprocedure en variant van de continue toetsvariabele zien voor simulaties op de uitvraag uit mei 2022. Ook hier is raking altijd minstens even goed als de andere schatters, maar neemt het voordeel iets af naarmate er meer complicaties aan de toetsvariabele zijn. Ook is de ongewogen schatting weer zuiveder als de effectenm van opleiding en geslacht van elkaar afhangen.

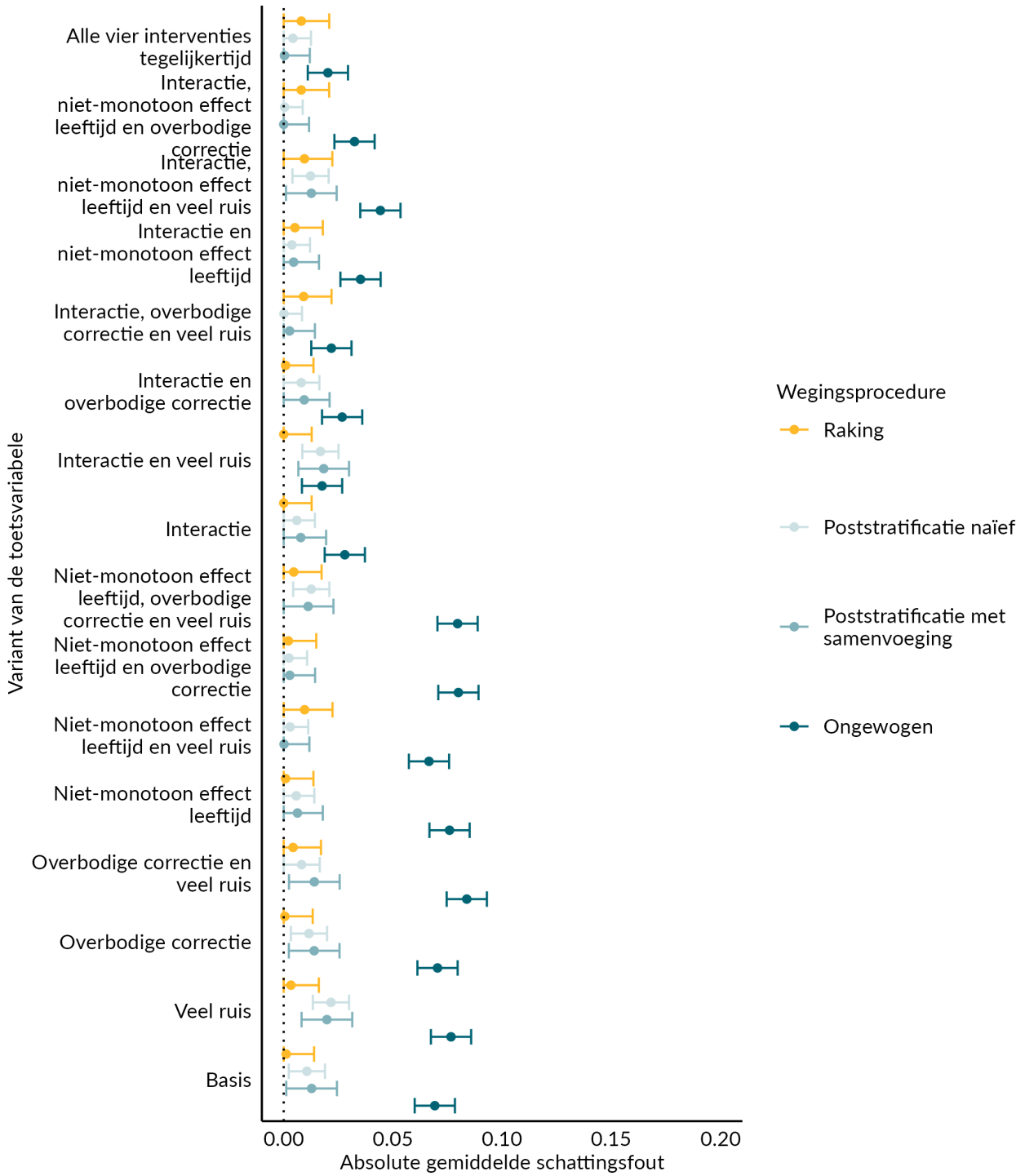
In Figuur 5 staat de empirische standaardfout van iedere wegingsprocedure en variant van de continue toetsvariabele voor simulaties op basis van de uitvraag uit mei 2022. Ook hier heeft de hoeveelheid ruis in de variabele de meeste invloed. Raking is bij beperkte ruis ongeveer even betrouwbaar als de twee varianten van poststratificatie. Bij meer ruis wordt raking net wat onbetrouwbaarder dan de andere wegingsprocedures, wat in Figuur 3 minder sterk te zien was.

Figuur 6 geeft de vergelijking van zuiverheid van schattingen bij simulaties op basis van vragenlijst uit april 2023. Meer dan bij de andere uitvragen is raking hier de meest zuivere schatter, terwijl de vertekening van zowel de ongewogen als met poststratificatie gewogen schattingen groter is dan bij de andere twee vragenlijsten. In dit geval was het echter ook wat moeilijker te beoordelen hoe zuiver raking was, te zien aan de grotere

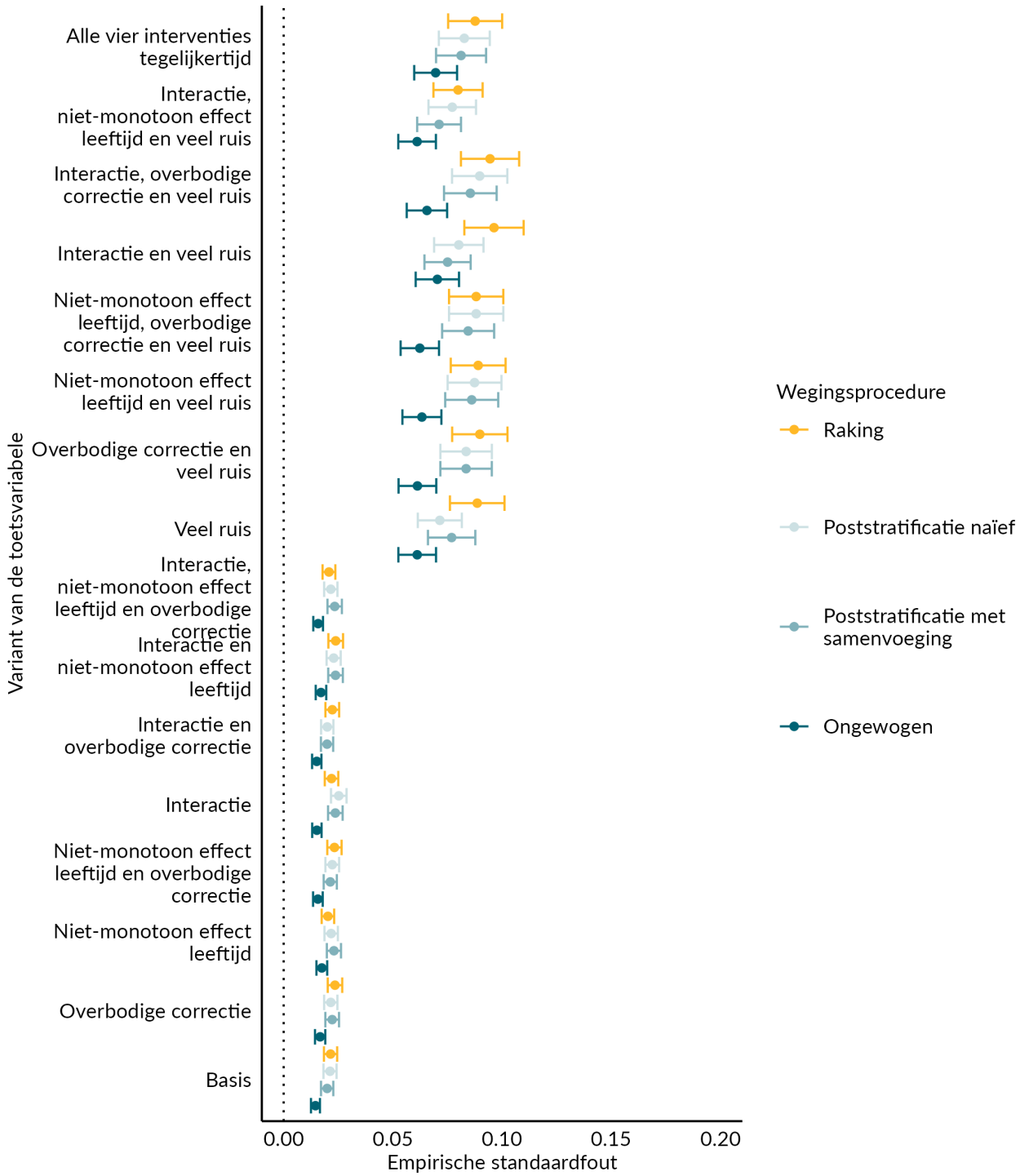
betrouwbaarheidsintervallen. Voor de ongewogen schatting maakt het weer uit of er interactie tussen de weegvariabelen is. Het verschil in zuiverheid tussen raking en de twee vormen van poststratificatie neemt nog steeds af naarmate er meer aanpassingen aan de toetsvariabele zijn.

De resultaten in Figuur 7 laten weer zien hoe betrouwbaar de schattingen met verschillende wegingsprocedures voor verschillende toetsvariabelen zijn, en hier is de hoeveelheid ruis in de toetsvariabele weer de belangrijkste invloed. Net als in Figuur 5 is raking minder betrouwbaar dan de vormen van poststratificatie als er meer ruis is, en is dat sterker als er meer ruis is.

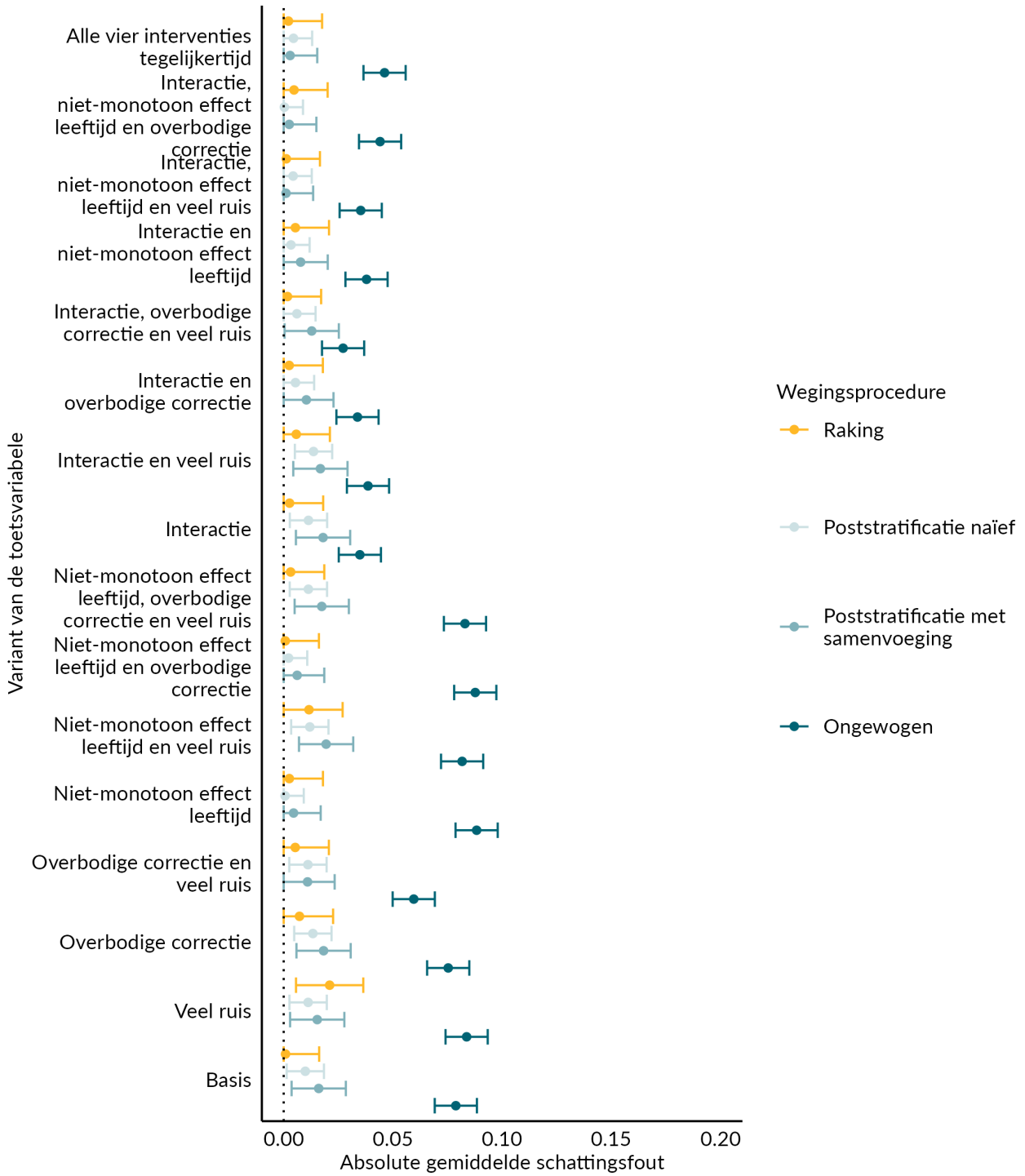
Gezien over de drie vragenlijsten samen levert raking de meest zuivere schattingen op, of schattingen die daar in de buurt zitten. Dat is bij alle varianten van de toetsvariabele en alle uitvragen het geval. De winst ten opzichte van de twee vormen van poststratificatie wordt echter kleiner naarmate er meer onderdelen als interactie tussen weegvariabelen, ruis in de toetsvariabele, een niet-monotoon verband en overbodige correctie zijn. Raking geeft als de variabele wat meer toevallige variatie binnen strata heeft minder betrouwbare resultaten, en als naarmate er meer onderbezette of lege strata zijn wordt dat onderscheid duidelijker.



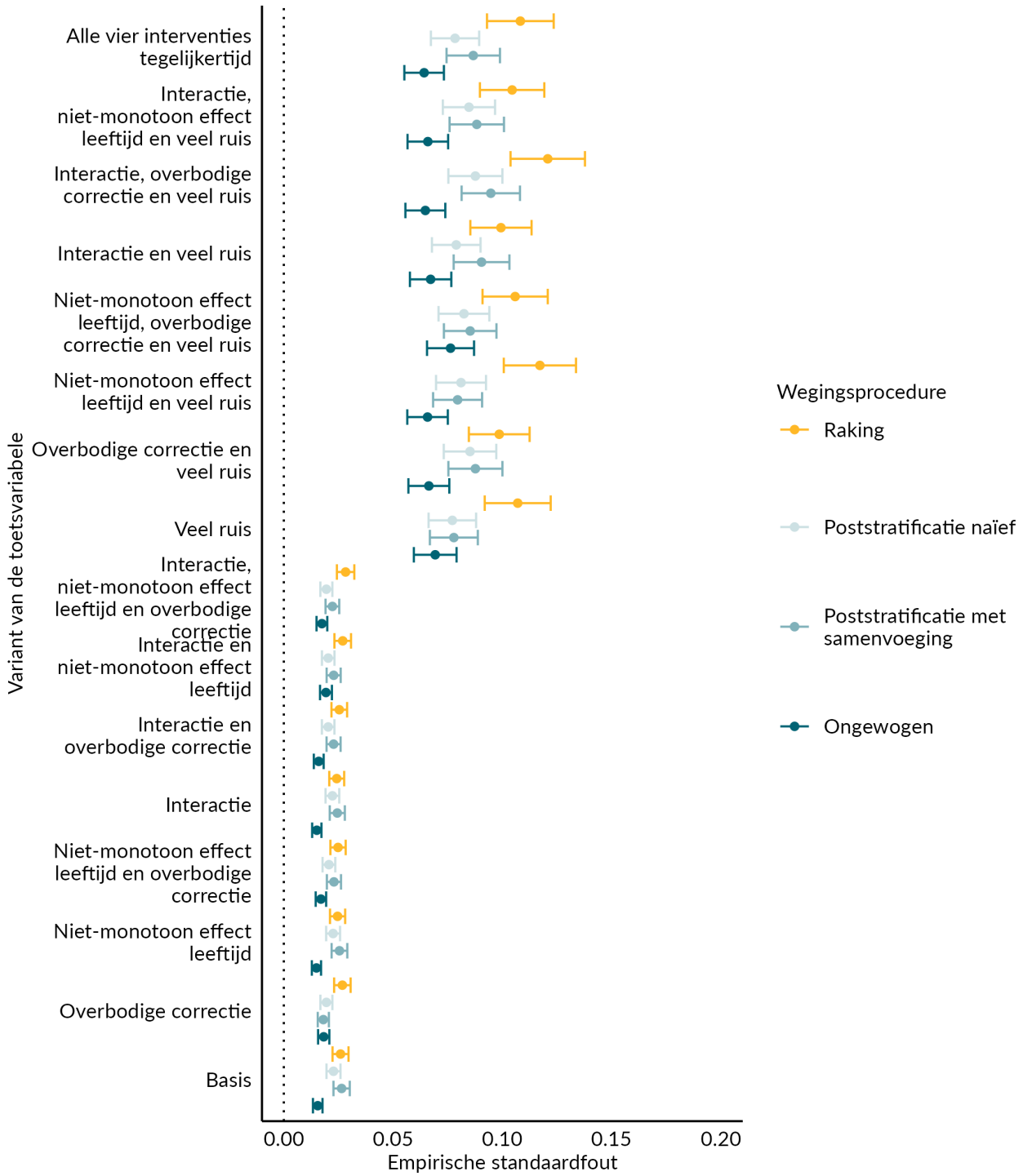
Figuur 2: Absolute gemiddelde schattingsfouten bij vragenlijst 2/2021 (juni 2021). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



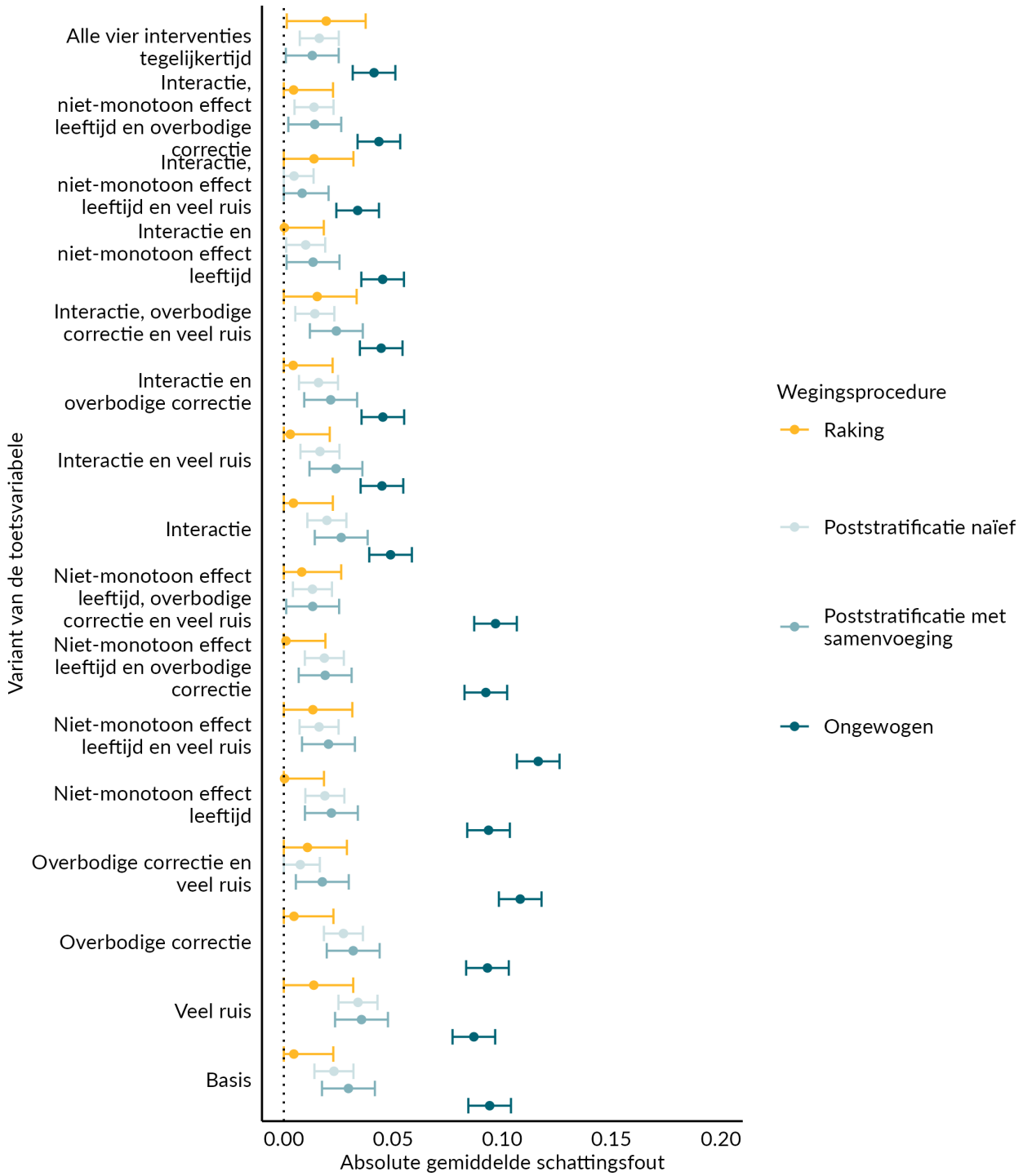
Figuur 3: Empirische standaardfouten van de schattingen bij vragenlijst 2/2021 (juni 2021). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



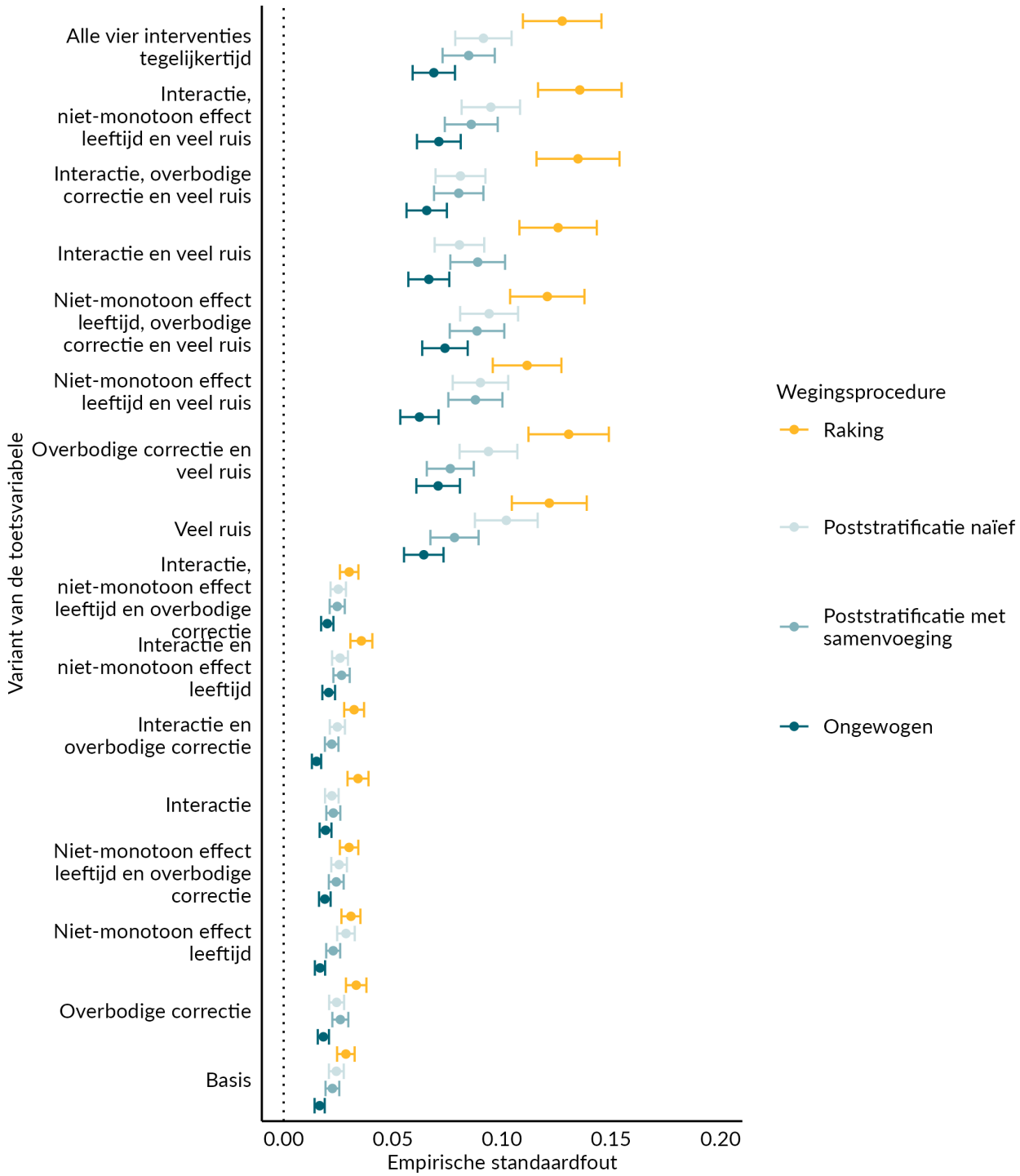
Figuur 4: Absolute gemiddelde schattingsfouten bij vragenlijst 2/2022 (mei 2022). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 5: Empirische standaardfouten van de schattingen bij vragenlijst 2/2022 (mei 2022). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 6: Absolute gemiddelde schattingsfouten bij vragenlijst 2/2023 (april 2023). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 7: Empirische standaardfouten van de schattingen bij vragenlijst 2/2023 (april 2023). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.

Binaire variabele

Voor de schattingen van de proportie positieve antwoorden op de binaire toetsvariabele in simulaties gebaseerd op de uitvraag uit juni 2021 staan de resultaten in Figuur 8. Welke wegingsprocedure ook toegepast wordt, in alle gevallen geeft weging een duidelijk zuiverdere schatting dan een ongewogen schatter. De ongewogen schatting wordt hier niet meer beïnvloed door een interactie tussen opleidingsniveau en geslacht, in plaats daarvan maakt meer ruis de ongewogen schatting betrouwbaarder. Waarschijnlijk komt dit doordat systematische verschillen waar de ongewogen schatting niet voor corrigeert minder belangrijk worden, en toevallige verschillen belangrijker. Ten opzichte van de keuze tussen een gewogen en een ongewogen schatting maakt de keuze voor een bepaalde wegingsprocedure weinig uit. Raking geeft bij 'eenvoudige' toetsvariabelen zuiverdere schattingen dan de twee vormen van poststratificatie, maar het verschil wordt kleiner als meer van de aanpassingen aan de basisvariabele in het spel zijn.

De empirische standaardfouten van de schattingen op de binaire toetsvariabele in simulaties gebaseerd op de uitvraag uit juni 2021 staan in Figuur 9. De verschillen tussen wegingsprocedures lijken voor alle toetsvariabelen klein. Raking is wat minder betrouwbaar dan de andere wegingsprocedures, maar de verschillen zijn klein.

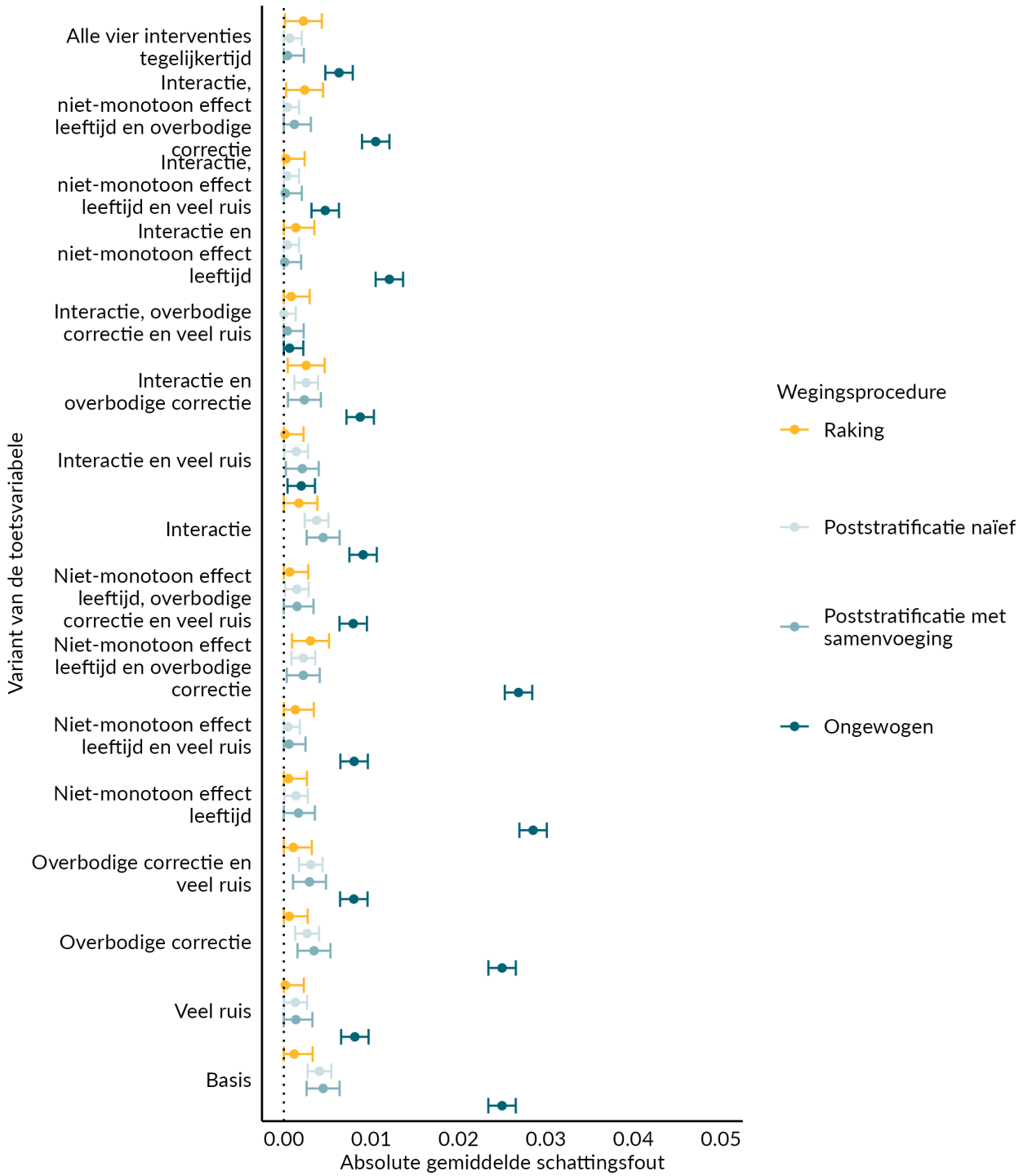
De resultaten bij de binaire toetsvariabele voor de vragenlijst van mei 2022 staan in Figuur 10. Hier lijken de resultaten erg veel op die voor de vragenlijst van juni 2021. Raking geeft meestal de zuiverste schatting. Vooral voor de basisvariabele doet raking het duidelijk beter dan de varianten van poststratificatie. En wegen geeft duidelijk zuiverdere schatters dan niet wegen, maar het verschil wordt kleiner als er veel ruis is.

Figuur 11 toont empirische standaardfouten per wegingsprocedure en variant van de binaire toetsvariable uit simulaties op basis van de uitvraag van mei 2022. Ook hier zijn de resultaten een wat minder duidelijke versie van de resultaten voor de continue variabele op dezelfde uitvraag. Raking is dus iets minder betrouwbaar dan de andere wegingsprocedures.

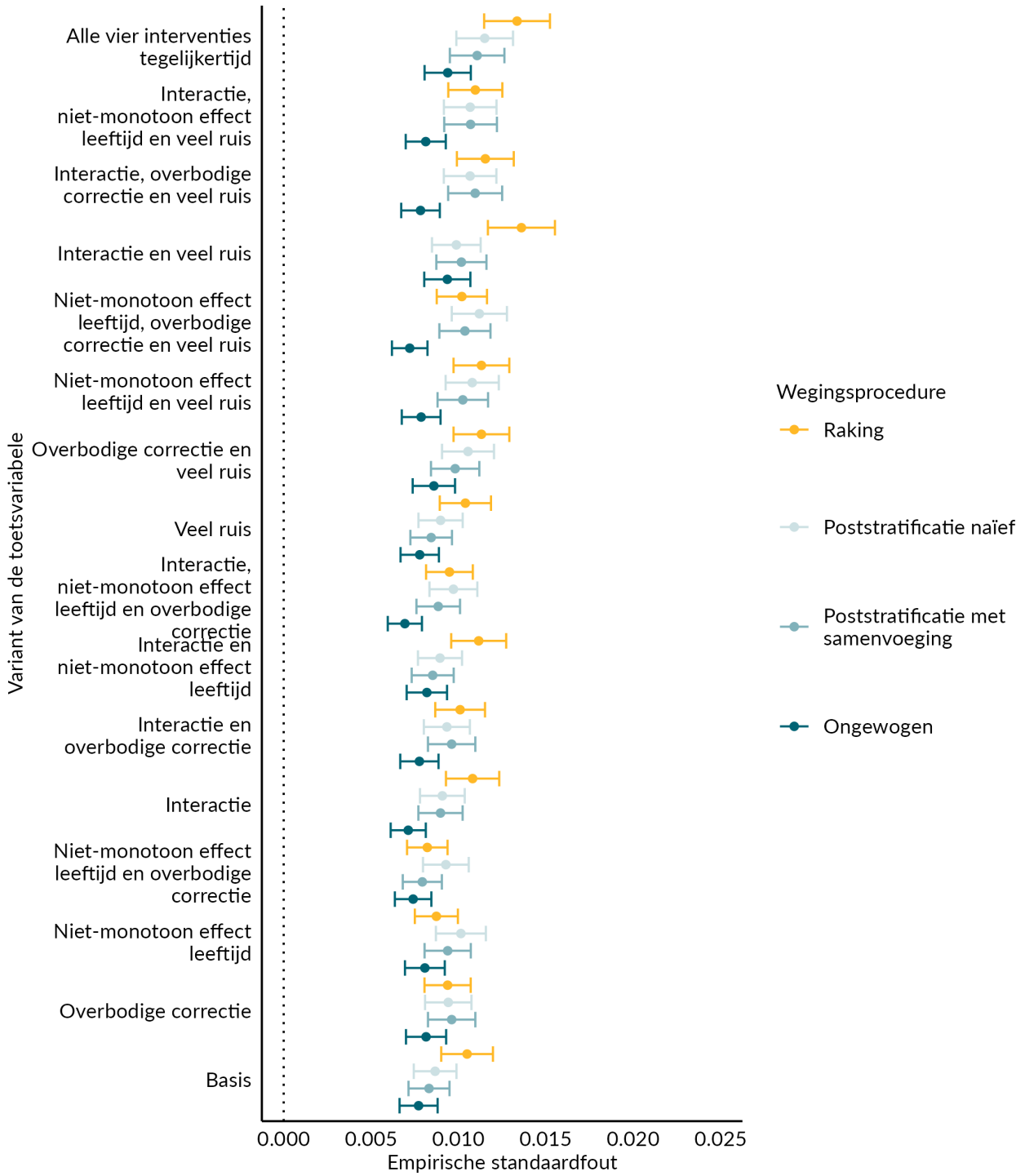
In Figuur 12 staan de resultaten voor de binaire toetsvariabele bij de vragenlijst van april 2023. Over het geheel is er niet veel nieuws te zien in deze resultaten, maar vergelijkend met de andere figuren is te zien dat het voordeel van raking ten opzichte van de twee vormen van poststratificatie duidelijker wordt naarmate de uitvraag waarvan het responsgedrag wordt gesimuleerd verder van de herwerving af ligt, en er dus meer onderbezette en lege strata zijn.

De empirische standaardfouten uit simulaties op basis van de uitvraag uit april 2021 in Figuur 13 zetten de wegingsprocedures weer iets scherper tegen elkaar af, en laten raking er weer negatiever uitkomen. Raking is minder betrouwbaar dan de andere wegingsprocedures, en dit patroon is redelijk consistent met verschillende hoeveelheden ruis.

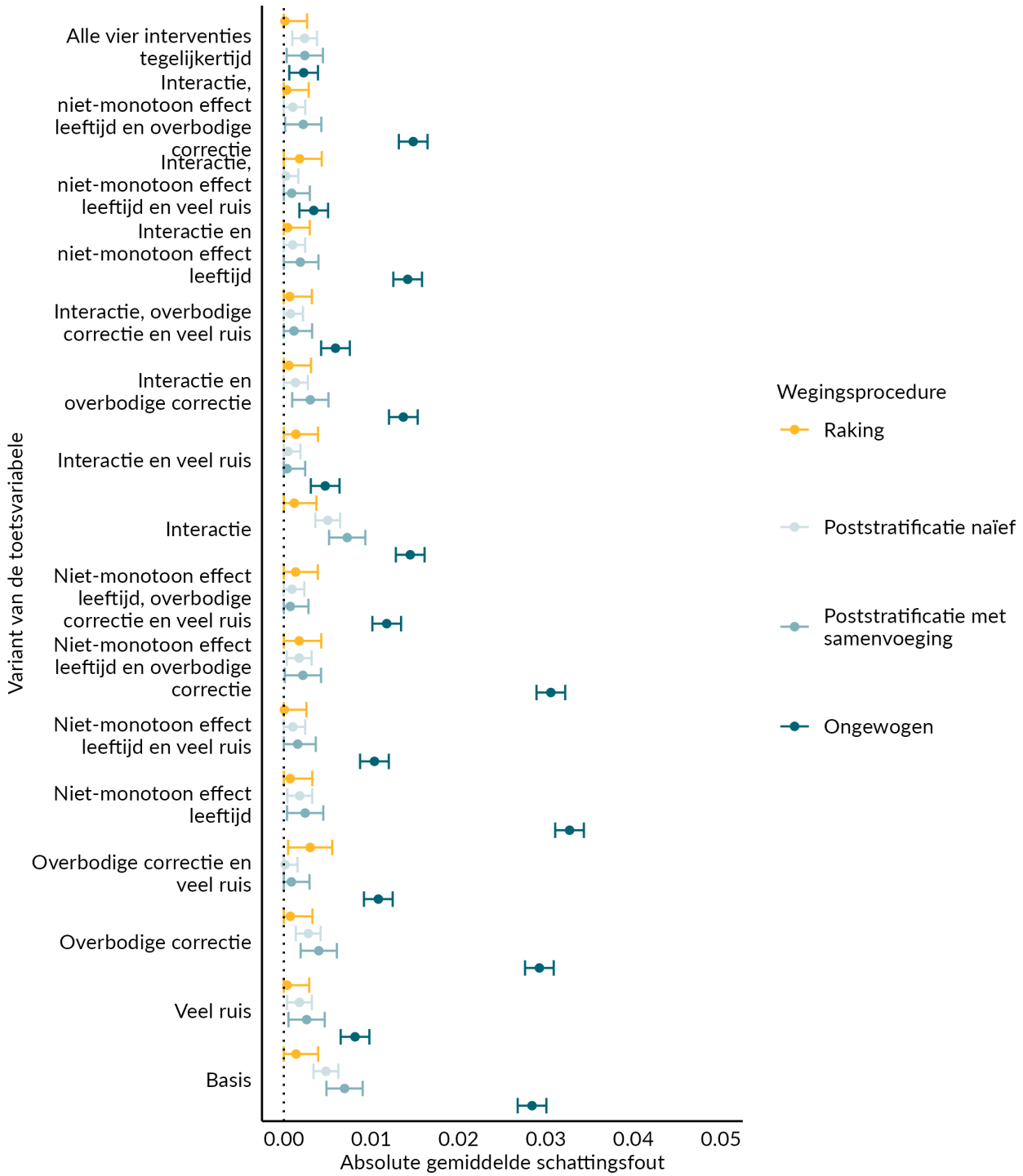
Het beeld voor de binaire variabele past bij dat voor de continue variabele. Raking is vrij consistent minstens even zuiver als de twee vormen van poststratificatie, en het voordeel neemt toe naarmate er meer lege of onderbezette strata komen. De betrouwbaarheid van raking is wel iets lager of duidelijk lager.



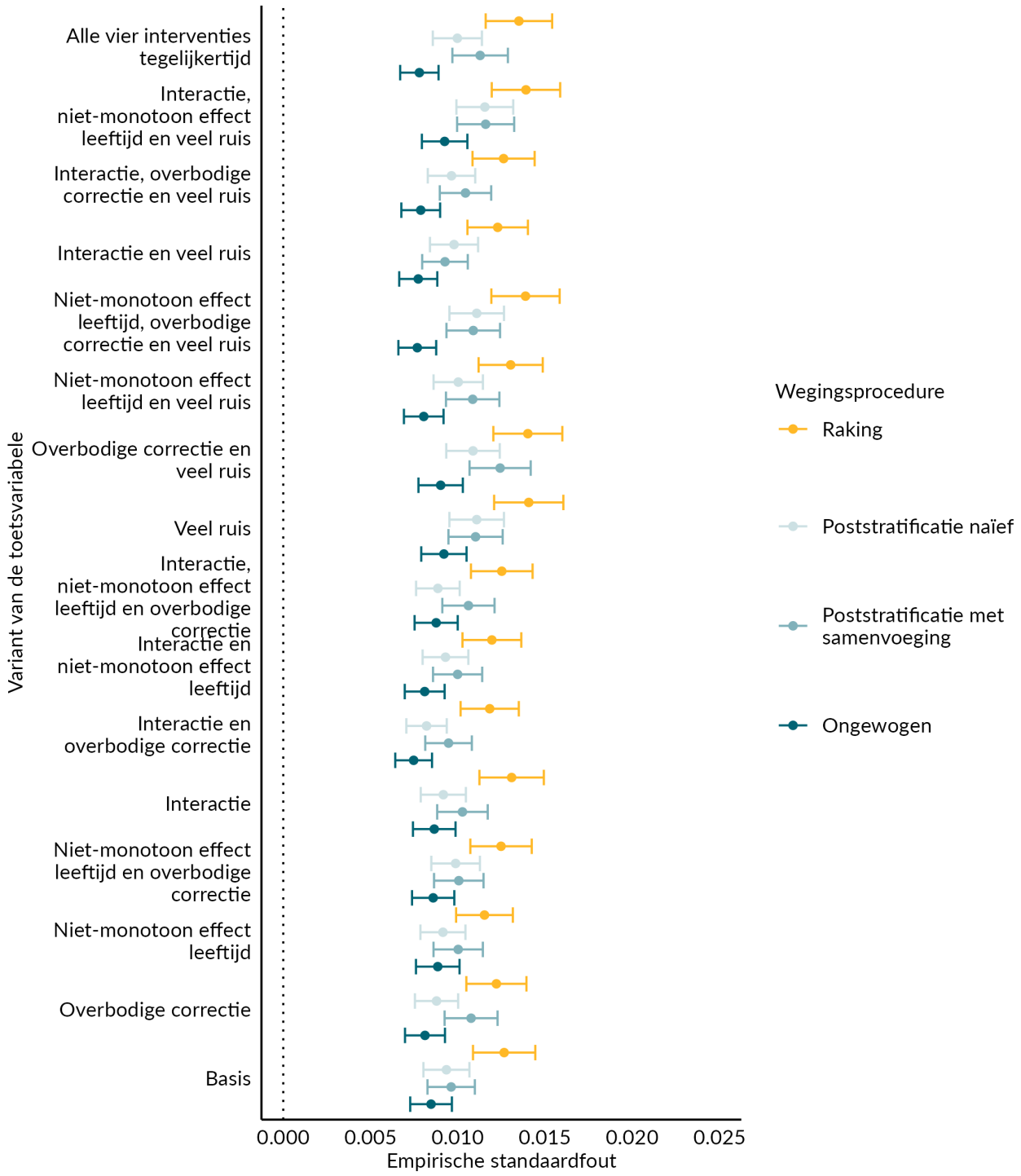
Figuur 8: Absolute gemiddelde schattingsfouten bij vragenlijst 2/2021 (juni 2021). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



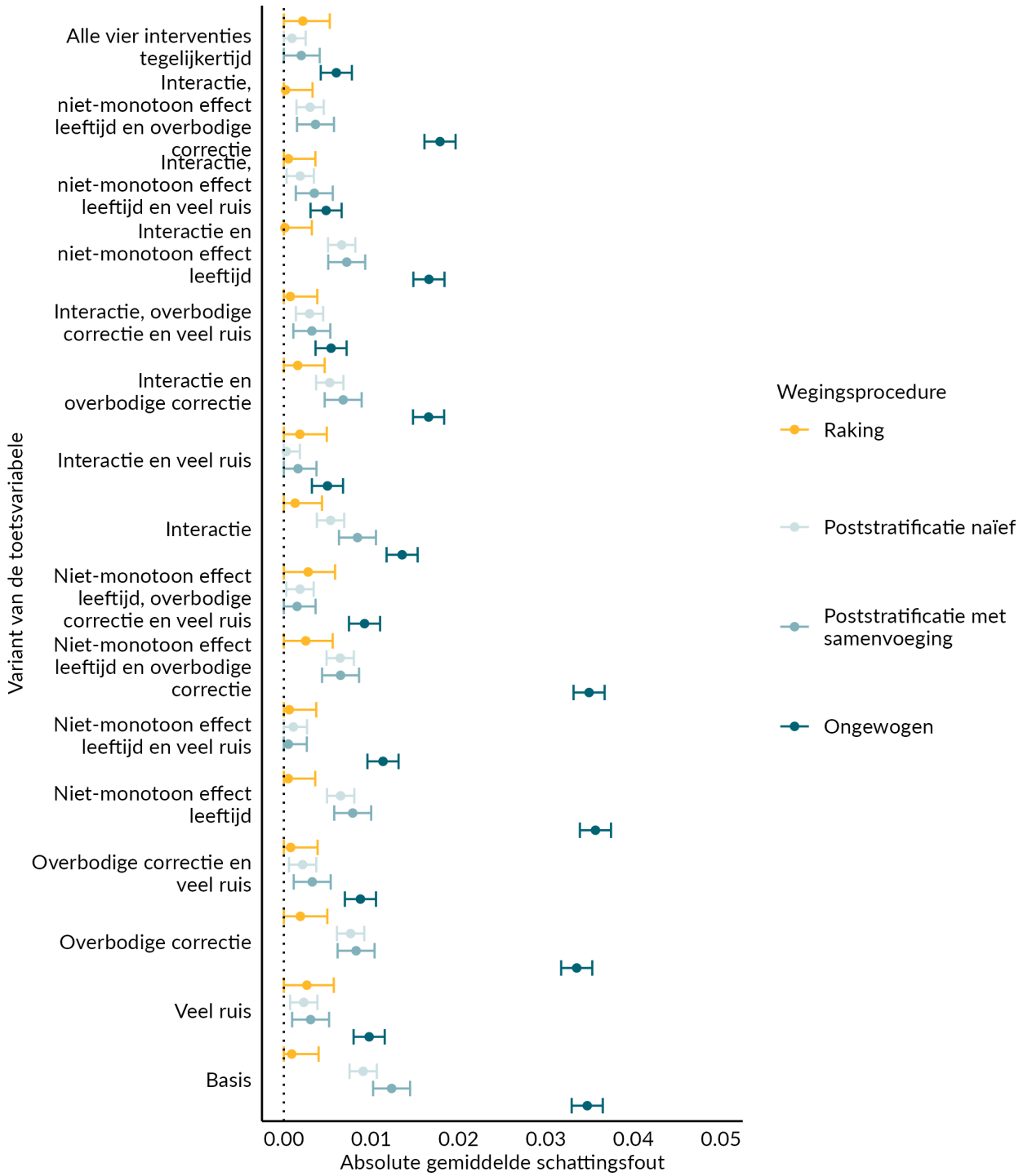
Figuur 9: Empirische standaardfouten van de schattingen bij vragenlijst 2/2021 (juni 2021). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



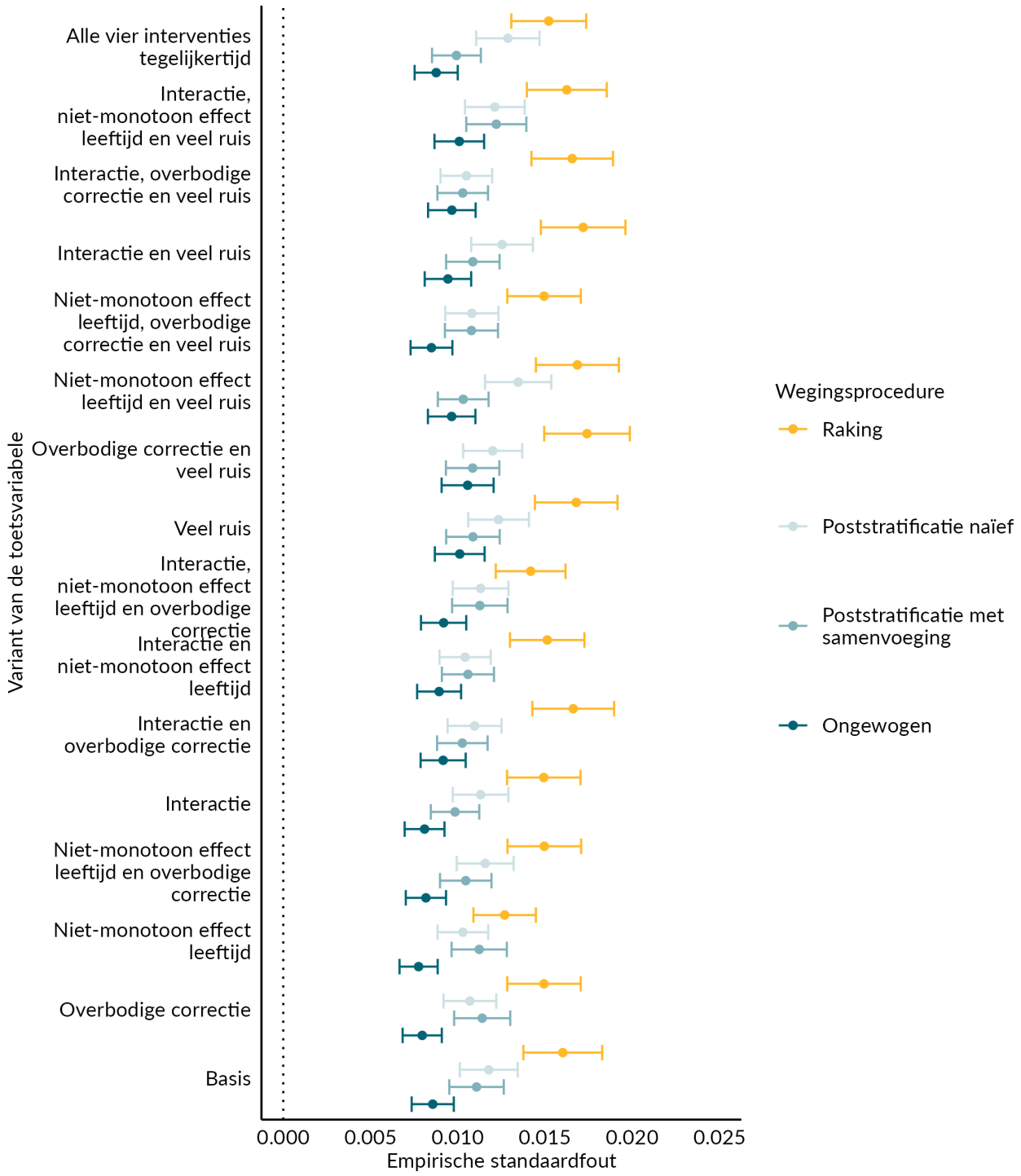
Figuur 10: Absolute gemiddelde schattingsfouten bij vragenlijst 2/2022 (mei 2022). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 11: Empirische standaardfouten van de schattingen bij vragenlijst 2/2022 (mei 2022). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 12: Absolute gemiddelde schattingsfouten bij vragenlijst 2/2023 (april 2023). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 13: Empirische standaardfouten van de schattingen bij vragenlijst 2/2023 (april 2023). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.

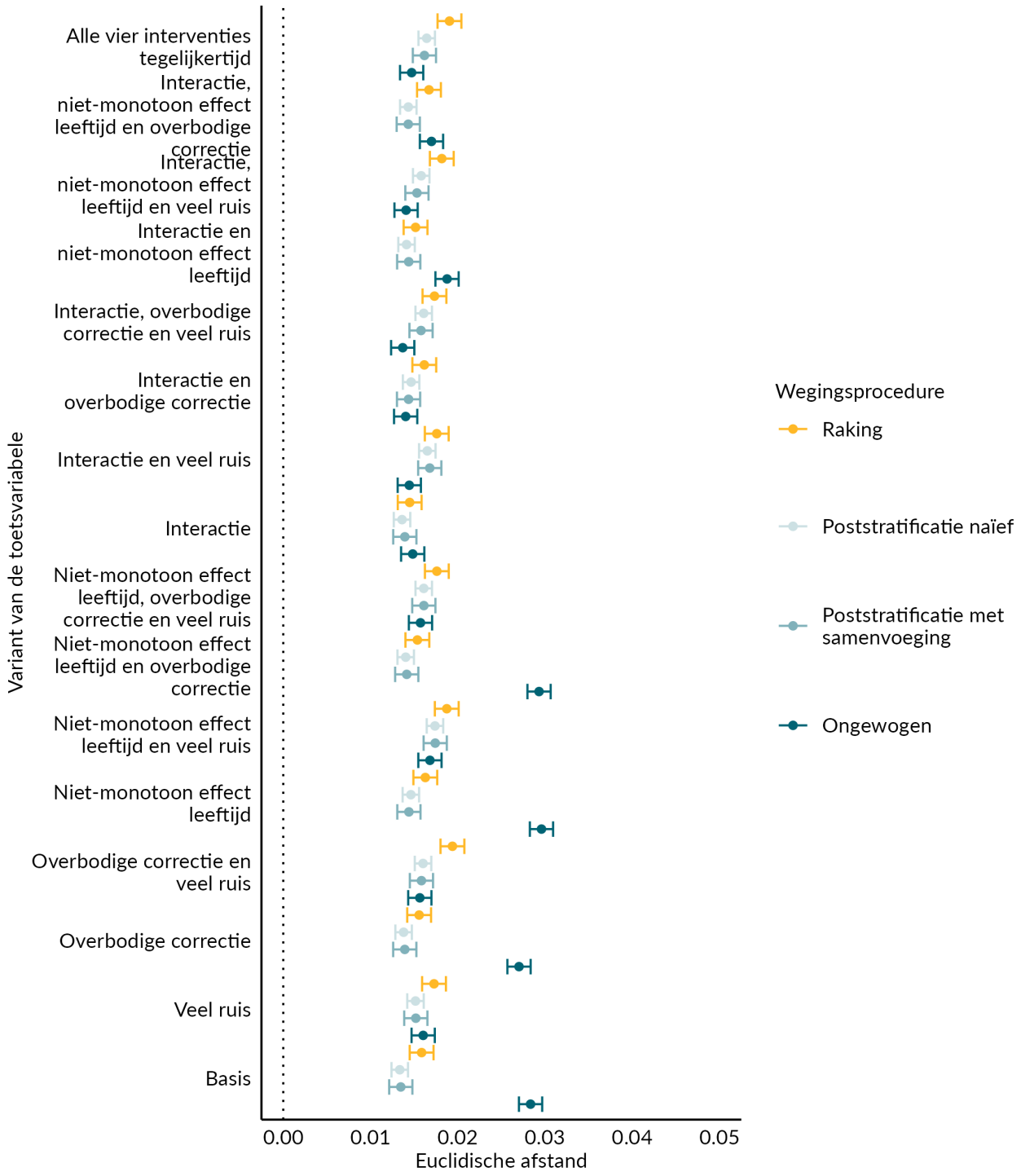
Ordinale variabele

De resultaten voor de ordinale toetsvariabele bij de vragenlijst uit juni 2021 staan in Figuur 14. Raking zorgt meestal voor een wat grotere totale fout dan de twee vormen van poststratificatie. Opvallend genoeg zijn ongewogen schattingen ongeveer even goed als gewogen schattingen als er veel ruis is. Dit komt waarschijnlijk omdat veel ruis betekent dat de systematische verschillen waar weging voor corrigeert minder belangrijk zijn terwijl toevallige verschillen belangrijker zijn. En over het algemeen hebben gewogen schattingen meer toevallige schattingsvariatie dan ongewogen schattingen.

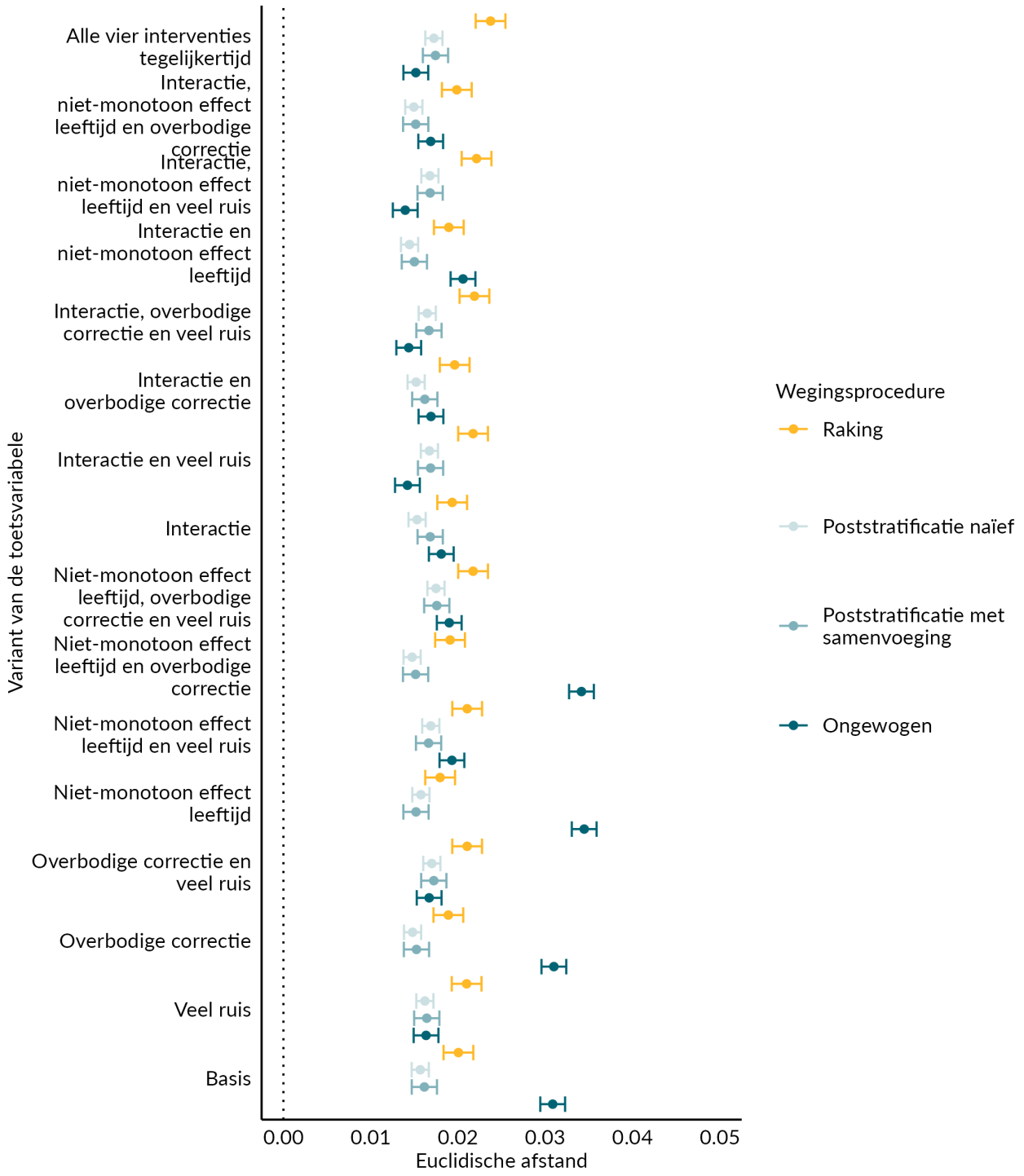
Voor de ordinale toetsvariabele en de vragenlijst uit mei 2022 staat de vergelijking tussen wegingsprocedures in Figuur 15. Raking lijkt het over het geheel wat minder goed te doen dan de twee vormen van poststratificatie, en de verschillen zijn wat groter dan bij de vragenlijst van juni 2021.

Bij de resultaten op basis van het responsgedrag voor de uitvraag uit april 2023 in Figuur 16 komt raking opnieuw iets minder positief uit de vergelijking dan de twee vormen van poststratificatie. Raking heeft bijna altijd een grotere totale schattingsfout.

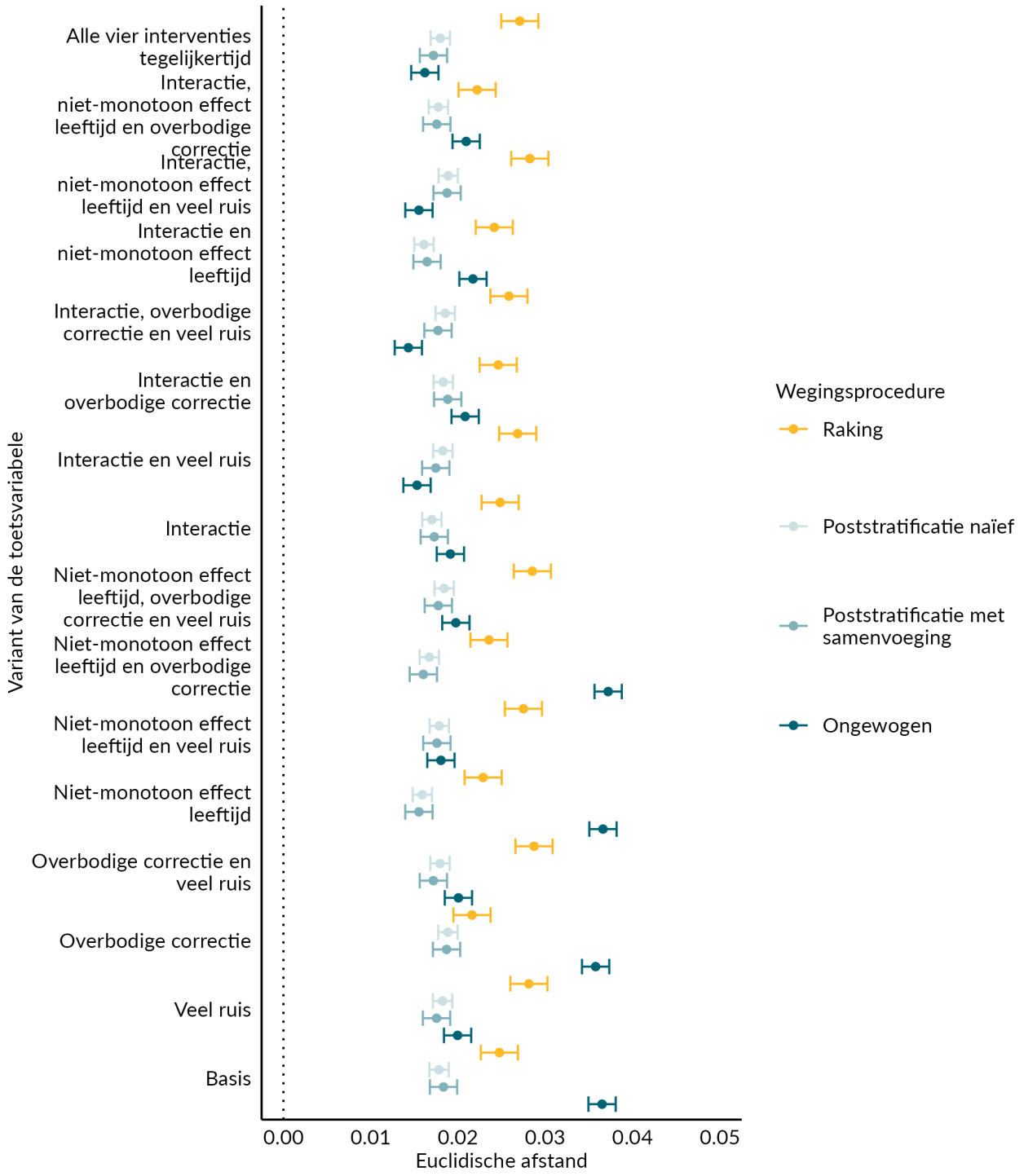
De vergelijking op schattingsfout bij de ordinale toetsvariabele is het minst overtuigend van de drie toetsvariabelen, maar in combinatie met de twee andere toetsvariabelen geeft het wel een beeld van de prestaties van de wegingsprocedures vanuit het oogpunt van zowel zuiverheid van de gemiddelde schatting als betrouwbaarheid van een individuele schatting. Bij een maat die naast systematische ook toevallige schattingsfouten meeneemt is het voordeel van raking namelijk een stuk minder overtuigend dan bij een maat die alleen systematische schattingsfouten rekent.



Figuur 14: Schattingsfouten in afstand bij vragenlijst 2/2021 (juni 2021). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 15: Schattingsfouten in afstand bij vragenlijst 2/2022 (mei 2022). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.



Figuur 16: Schattingsfouten in afstand bij vragenlijst 2/2023 (april 2023). *Noot:* Data gemaakt door auteur op basis van 100 gesimuleerde panels met uitvragen. Richtgetallen voor de omvang en samenstelling van gesimuleerde steekproeven zijn gebaseerd op de uitvraagdata en de samenstelling van het panel ten tijde van de uitvraag. De populatiesamenstelling is gebaseerd op een CBS maatwerktabel voor 2022.

// Conclusie

Het panelonderzoek voor Panel Fryslân heeft te maken met over- of ondervertegenwoordiging van groepen waarvan te verwachten valt dat ze ook kenmerkend antwoordgedrag hebben ten opzichte van andere groepen. Daardoor is het nodig om de data van vragenlijsten met herweging te corrigeren. Ons onderzoek heeft echter ook de uitdaging dat sommige van de strata waar respondenten voor herweging over verdeeld worden nauwelijks respondenten bevatten, bijvoorbeeld omdat er maar weinig jongeren en inwoners van Fryslân zonder startkwalificatie meedoen aan Panel Fryslân. Dit betekent dat de klassieke wegingsprocedure poststratificatie niet zomaar toegepast kan worden. Er zijn alternatieve wegingsprocedures, maar in de literatuur vonden wij geen onderzoek terug dat wegingsprocedures naast elkaar zet en op prestaties vergelijkt.

Om wegingsprocedures te vergelijken die ons niet beperkt tot wiskundig hanteerbare wegingsprocedures, en toegespitst is op onze situatie, hebben we een simulatiestudie uitgevoerd om wegingsprocedures te vergelijken. Voor populatieparameters van verschillende soorten variabelen met een variatie aan eigenaardigheden en over drie vragenlijsten die verschillen in de mate waarin strata gevuld zijn doet raking het als wegingsprocedure in veel gevallen ongeveer even goed als of iets beter dan twee strategieën om poststratificatie toe te passen als de aannames van poststratificatie niet voldaan zijn. Raking geeft meestal minstens even zuivere of zuiverdere schattingenn. Vooral als er meer lege of onderbezette strata zijn is raking zuiverder dan de twee aangepaste vormen van poststratificatie.

Een beperking aan het gebruik van raking is dat het relatief slechter gaat presteren bij variabelen waar meer 'ruis' in voorkomt. Deze 'ruis' is in praktische gevallen de variatie in de uitkomstvariabele binnen strata. In onze simulaties deed een wegingsprocedure het goed als de verhouding tussen variatie binnen strata en variatie binnen strata ongeveer 1:1 is, en slechter als de verhouding 1:15 is. Met een schetsmatig onderzoek leek de verhouding in echte data uit Panel Fryslân meestal rond de 1:10 te liggen, met 1:5 als lage verhouding en 1:15 als hoge verhouding.

De resultaten van deze simulaties laten niet zonder twijfel zien wat de beste wegingsprocedure is. De vertekening van een met raking gewogen schatting lijkt minder toe te nemen dan die van een schatting met de twee aangepaste vormen van poststratificatie als er meer lege of onderbezette strata zijn. Tegelijkertijd wordt raking sneller minder betrouwbaar te worden als er meer lege en onderbezette strata zijn.

Vooral vanwege de zuiverheid van raking lijkt het de moeite waard om poststratificatie met het samenvoegen van cellen te vervangen door raking als standaard wegingsprocedure voor Panel Fryslân. Raking zorgt wel voor iets minder betrouwbare schattingen, maar de vertekening werd kleiner. Raking is ook de standaard in een aantal grote sociaalwetenschappelijk studies (European Social Survey, 2014; European Values Study (EVS), 2022), wat betekent dat we door raking te gebruiken ook aansluiten op een standaard in vragenlijstonderzoek. In de resultaten voor de ordinale variabelen was te zien dat de totale fout van raking regelmatig groter was dan de totale fout van de andere wegingsprocedures. De totale fout van een ongewogen schatting was bij toetsvariabelen met veel aanpassingen echter vaak kleiner dan die van de gewogen schattingen, hier speelt ook de betrouwbaarheid weer mee. Een ongewogen schatting is bijna per definitie betrouwbaarder (minder gevoelig voor toevallige variatie) dan een gewogen schatting, maar dat betekent niet dat het de moeite waard is om de vertekening te negeren. Op dezelfde manier accepteren we een minder betrouwbare wegingsprocedure om de vertekening te verminderen.

Het effect van ruis op de betrouwbaarheid van raking is een aandachtspunt, omdat hier dus de zwakte van de wegingsprocedure zit. Ruis zorgde in de simulaties voor verschillen in waarden voor de toetsvariabelen die niet met de wegingsfactoren te maken hebben. Dat respondenten uit hetzelfde stratum verschillende waarden hebben, komt dus door de ruis. Als de variatie in een variabele binnen strata erg groot is, dan zal raking schattingen geven die gemiddeld over veel schattingen in de buurt van de werkelijke waarde komen, maar die individueel vrij ver naast de werkelijke waarde kunnen zitten. En in de praktijk valt er maar één schatting te maken, en zijn vertekening en onbetrouwbaarheid niet uit elkaar te houden. Het is dus belangrijk om ook zicht te houden op de niet-systematische variatie, en die te beperken waar dat kan. De meest voor de

hand liggende manier om dit te doen is om vragen in vragenlijsten zo op te stellen dat er minder verschillen in antwoorden ontstaan doordat mensen een vraag verschillend interpreteren of iets niet goed in kunnen schatten. Vragen gebruiken waarvan de betrouwbaarheid en validiteit in eerder onderzoek is aangetoond helpt daarbij.

De simulaties waarop deze conclusie bouwt zijn aangepast op de situatie in Planbureau Fryslân, en testen verschillende schattingsprocedures onder een redelijk breed spectrum aan omstandigheden. De simulaties gebruiken echter noodzakelijk specifieke aannames over de variabelen die van belang zijn voor de schattingen, hoe de steekproef tot stand komt, en hoe de variabelen er precies uitzien. Om de conclusies nog een slag robuuster te maken, zou het verstandig zijn om in ieder geval een sensitiviteitsanalyse te doen van de simulaties om te zien onder welke omstandigheden de conclusies anders uit zouden vallen en waarom dit gebeurt.

Voor een sensitiviteitsanalyse zouden de volgende sets modelparameters de belangrijkste aandachtspunten zijn:

- **Verband tussen weegfactoren en toetsvariabelen:** De weegfactoren hebben plausibel uitziende maar arbitraire coëfficiënten gekregen die ze met de toetsvariabelen verbinden. Zo hebben we gespecificeerd dat de continue basisvariabele voor 35- tot en met 49-jarigen 0,2 punten hoger is dan voor 18- tot en met 34-jarigen, voor 50- tot en met 64-jarigen 0,4 punten hoger, voor 65- tot en met 74-jarigen 0,6 punten hoger en voor 75+-ers 0,8 punten hoger. Maar in plaats van een gelijke stap tussen iedere leeftijdscategorie (terwijl sommige leeftijdscategorieën breder zijn dan andere) zou het ook een gelijke relatieve toename kunnen zijn (van 0,2 naar 0,4 naar 0,8 naar 1,6). En als we een interactie introduceren, is dat nu een interactie tussen geslacht en opleidingsniveau. In plaats daarvan had er ook een interactie tussen leeftijd en opleidingsniveau kunnen zijn. Voor deze simulaties hebben we gekozen voor een combinatie van coëfficiënten en verbanden die ons inhoudelijk relevant en plausibel leken, maar voor een goede analyse van de resultaten zou het goed zijn om systematischer door de mogelijkheden te gaan en de invloed van die keuzes in beeld te krijgen.
- **Ruisverhouding:** Een andere groep coëfficiënten in de toetsvariabelen zijn de factoren die bepalen hoeveel van de variantie in de toetsvariabele ruis is, en hoeveel van de variantie door variatie in de weegfactoren komt. Voor de continue basisvariabele wordt dit geregeld door de factor $\frac{1}{2}$ in Y en de variantie van $\frac{1}{2}$ voor ϵ, en als veel ruis als omstandigheid wordt toegevoegd komt de variantie van 15 voor η hier nog bij. Voor de basisvariabele is de verhouding tussen variantie door variatie in de weegfactoren en ruis 1:1, voor de continue variabele met veel ruis is de verhouding 1:15, maar in variabelen die we gebruiken ligt het met 1:5 tot en met 1:15 tussen die twee in. De ruisverhouding systematisch variëren lijkt dus een belangrijke stap om te zien hoe relevant de sprong in schattingsonnauwkeurigheid is die raking maakt als er sprake is van veel ruis.
- **Coëfficiënten van de stratasamenvoeging:** Voor onze vertaling van de weegmethode die Planbureau Fryslân tot de doorvoering van raking gebruikte hebben we op basis van sociaalwetenschappelijke intuïtie coëfficiënten bepaald die strata op basis van hun achtergrondkenmerken dichter bij elkaar of verder van elkaar af plaatsen. Zoals wij echter al aangaven, kan het belang van verschillende weegdimensies van doelvariabele tot doelvariabele verschillen. En ook hier is het goed om te zien hoe de prestaties van poststratificatie met samenvoeging van strata beter of slechter wordt als de verhoudingen tussen de coëfficiënten veranderen.

Naast deze drie sets zijn de meeste parameters in de simulaties gecalibreerd op empirische data, en dus minder relevant voor gevoeligheidsanalyse. De manier waarop de simulaties verlopen is echter wel heel specifiek vastgelegd, en het zou goed zijn om de gevoeligheidsanalyse uit te breiden tot een robuustheidsanalyse waarbij de resultaten van dit model worden vergeleken met modellen die conceptueel nagenoeg gelijk zijn, maar bijvoorbeeld het trekken van steekproeven anders programmeren of een andere implementatie van het door raking gebruikte iterative proportional fitting-algorithme inzetten. Al dit soort veranderingen zouden op zich

niet uit moeten maken, dus checken of de simulatieresultaten geen eigenheden van de R-implementatie bevatten is een goede check.

Naast de robuustheid van de resultaten voor de onderliggende aannames, is een andere waardevolle richting om de resultaten in door te ontwikkelen om te kijken waardoor gewichten voor een analyse op de steekproef als geheel wel of niet voor subgroepanalyses toegepast kunnen worden. Bij Planbureau Fryslân worden er in analyses vaak uitsplitsingen gedaan op groepen, en om dat te kunnen doen moet niet alleen de algemene schatting accuraat zijn, maar moeten subgroepsschattingen ook zuiver en betrouwbaar zijn. In een korte analyse van schattingen voor subgroepen zag ik dat de ene wegingsprocedures soms beter werkt, en soms een andere, maar waarom dit zo is heb ik nog niet terug kunnen zien. Daar is dus verdiepend onderzoek naar nodig. Raking lijkt door uitval sneller toenemende schattingsvariatie te hebben dan de vormen van poststratificatie. De vertekening van schattingen met raking is dan weer vrij stabiel, terwijl de vertekening bij varianten van poststratificatie bij lage uitval meer verschilt van raking dan bij hoge uitval. Soms is een variant van poststratificatie daarbij in het voordeel, soms raking. Een verdiepingsslag in het onderzoek van subgroepschattingen op basis van voor de steekproef als geheel geschatte gewichten lijkt dus belangrijk.

Dit rapport vergelijkt tot slot drie wegingsprocedures, waarvan twee varianten van één en dezelfde basismethode zijn. Een voor de hand liggend alternatief is lineair wegen (Bethlehem, 2008, pp. 22–32). Dit heb ik hier niet overwogen, omdat die methode niet automatisch alle interacties tussen weegvariabelen meeneemt terwijl dat wel goed lijkt om te behouden. Maar bij het uitbouwen van deze analyse kan het waardevol zijn om meer wegingsprocedures te onderzoeken. Ten opzichte van een gevoeligheidsanalyse en het onderzoeken van subgroepschattingen heeft dit echter een lagere prioriteit zolang er geen sterke aanwijzingen zijn dat een bepaalde wegingsprocedure nog een stuk beter werkt.

Uit deze simulatiestudie naar drie wegingsprocedures komt raking naar voren als procedure die iets minder stabiele maar zuiverdere schattingen geeft als er sprake is van lege of dun bezette strata. Door een op empirische data gecalibreerde simulatie uit te voeren hebben we wegingsprocedures kunnen vergelijken in de theoretisch lastige maar praktisch relevante situatie dat bepaalde wegingstrata dun bezet of leeg zijn. Daarbij hebben we een breed scala aan variabelen onderzocht. Hiermee bouwen we voort op eerder werk dat het nut en de beperkingen van verschillende wegingsprocedures verkent, maar ze meestal niet naast elkaar zet, vooral niet als ze onder realistische omstandigheden worden getest (e.g., Bethlehem, 2008; Založnik, 2011). Ons werk is nog een stuk te verdiepen. Maar met deze eerste vergelijking leveren we een bijdrage aan beter opgezet (toegepast) onderzoek bij een veelvoorkomend probleem van dun bezette strata dat tot nu toe weinig aandacht kreeg.

// Referenties

Barthélemy, J., & Suesse, T. (2018). **mipfp** : An *R* Package for Multidimensional Array Fitting and Simulating Multivariate Bernoulli Distributions. *Journal of Statistical Software*, *86*. https://doi.org/10.18637/jss.v086.c02

Bethlehem, J. (2008). *Wegen als correctie voor non-respons* (08005). Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/onze-diensten/methoden/statistische-methoden/throughput/throughput/wegen-als-correctie-voor-non-respons

European Social Survey. (2014). *Documentation of ESS Post-Stratification Weights*. https://www.europeansocialsurvey.org/sites/default/files/2023-06/ESS_post_stratification_weights_documentation.pdf

European Values Study (EVS). (2022). *European Values Study (EVS) 2017: Weighting Data ; documentation update related to the final data release, May 2022* (2022/09; pp. 20 S.). https://doi.org/10.21241/SSOAR.70113.V2

EVS/WVS. (2024). *Joint EVS/WVS 2017-2022 Dataset (Joint EVS/WVS)Joint EVS/WVS 2017-2022 Dataset (Joint EVS/WVS)* (Versie 5.0.0). GESIS. https://doi.org/10.4232/1.14320

Forthomme, D., harisbal, & mattswoon. (z.d.). *ipfn* [Software]. https://github.com/Dirguis/ipfn

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2de dr.). Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Kerckhove, W. V. de, Mohadjer, L., & Krenzke, T. (2014). A Weight-Trimming Approach to Achieve a Comparable Increase to Bias Across Countries in the Programme for the International Assessment of Adult Competencies. *Proceedings of the American Statistical Association, Section on Survey Research Methods*. https://ww2.amstat.org/meetings/proceedings/2014/data/presinfo80433.cfm

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Onderzoeksverantwoording: Steekproef en werving van een representatief internetpanel. (2019). Fries Sociaal Planbureau. https://www.planbureaufryslan.nl/wp-content/uploads/2019/07/FSP_PF_verantwoording_2punt0-losse-pagina-DEF.pdf

Onderzoeksverantwoording: Steekproef en werving van een representatief internetpanel. (2021). Fries Sociaal Planbureau. https://www.planbureaufryslan.nl/wp-content/uploads/2022/06/FSP-2022-Panel-verantwoording-DEF.pdf

Panel Fryslân Onderzoeksverantwoording: Steekproef en werving van een representatief internetpanel. (2023). Planbureau Fryslân. https://www.planbureaufryslan.nl/wp-content/uploads/2024/01/Verantwoording-herwerving-panel-2023.pdf

Potter, F., & Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 2707–2719. https://ww2.amstat.org/meetings/proceedings/2015/data/presinfo124114.cfm

Robbins, M. W., Ghosh-Dastidar, B., & Ramchand, R. (2019). *Blending of Probability and Non-Probability Samples: Applications to a Survey of Military Caregivers*. arXiv. https://doi.org/10.48550/ARXIV.1908.04217

Rueda, M. D. M., Pasadas-del-Amo, S., Rodríguez, B. C., Castro-Martín, L., & Ferri-García, R. (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, *65*(2), 2200035. https://doi.org/10.1002/bimj.202200035

Visser, M., & Fernee, H. (2017). *Onderzoeksverantwoording Panel Fryslân: Steekproef en werving van een representatief internetpanel*. Fries Sociaal Planbureau. https://www.planbureaufryslan.nl/wp-content/uploads/2019/01/onderzoeksverantwoording_panel_fryslan_0.pdf

Yang, M., Mulrow, E., Huang, C.-M., & Herring-Nathan, E. (2024, oktober 11). Evaluating Alternative Estimation Methods using Probability and Nonprobability Samples. *2024 JSM Proceedings*. Joint Statistical Meetings. https://doi.org/10.5281/ZENODO.13921709

Založnik, M. (2011). *Iterative Proportional Fitting - Theoretical Synthesis and Practical Limitations* [Phdthesis, University of Liverpool]. http://rgdoi.net/10.13140/2.1.2480.9923

// Bijlage: R-Algorithme voor het samenvoegen van strata

```r
# Bepaal of iedere cel minstens 5 respondenten bevat
vereenvoudiging_nodig <-
  any(respons_gestratificeerd$aantal_respondenten < 5) |
  any(is.na(dekkingsgewichten$dekkingsgewicht))

# Als er in één of meer cellen onvoldoende respondenten zijn, negeer dan
# eerst COROP-regio als achtergrondvariabele
if (vereenvoudiging_nodig) {

  # Haal de opsplitsing naar COROP uit de tabel met de populatiesamenstelling
  # naar de achtergrondvariabelen
  synthetisch_fryslan_summary_vereenvoudigd <- synthetisch_fryslan_summary |>
    group_by(geslacht, opleiding, leeftijd) |>
    summarise(corop = paste(sort(unique(corop)), collapse = ", "),
              aantal_inwoners = sum(aantal_inwoners)) |>
    ungroup()

  # Haal de opsplitsing naar COROP uit de tabel met de respons
  # naar de achtergrondvariabelen
  respons_gestratificeerd_vereenvoudigd <- respons_gestratificeerd |>
    group_by(geslacht, opleiding, leeftijd) |>
    summarise(
      corop = paste(sort(unique(corop)), collapse = ", "),
      aantal_respondenten = sum(aantal_respondenten)
    ) |>
    ungroup()

  # Bereken de nieuwe dekkingsgewichten
  dekkingsgewichten_PF <-
    left_join(
      synthetisch_fryslan_summary_vereenvoudigd,
      respons_gestratificeerd_vereenvoudigd,
      by = join_by(corop, geslacht, opleiding, leeftijd)
    ) |>
    mutate(
      dekkingsgewicht_PF = (aantal_inwoners / sum(aantal_inwoners))/
        (aantal_respondenten/sum(aantal_respondenten, na.rm = TRUE))
    ) |>
    select(corop, geslacht, opleiding, leeftijd, dekkingsgewicht_PF)

  # Kijk of iedere cel nu minstens vijf respondenten bevat
  verdere_vereenvoudiging_nodig <-
    any(respons_gestratificeerd_vereenvoudigd$aantal_respondenten < 5) |
    any(is.na(dekkingsgewichten_PF$dekkingsgewicht_PF))

  # Zo niet, voeg dan gericht cellen samen
  while (verdere_vereenvoudiging_nodig) {
    if (verbose) {print("Voeg verdere cellen samen in weegtabel...")}
```

```
# Maak een logische filter voor cellen die verder moeten worden
# aangepast
correctie_nodig <-
  respons_gestratificeerd_vereenvoudigd$aantal_respondenten < 5 |
  is.na(dekkingsgewichten_PF$dekkingsgewicht_PF)

# Geef iedere cel in de weegtabel een positie om afstanden tot andere
# cellen mee te berekenen. Gebruik grepl en een deling om cellen die een
# samenvoeging zijn van meerdere categorieën de gemiddelde positie van
# die categorieën te geven

# Afstanden tussen leeftijdscategorieën zijn van categorie tot
# categorie niet zo groot, maar tellen snel op
positie_L <- (
  0 * grepl("18-34", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    1 * grepl("35-49", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    2 * grepl("50-64", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    3 * grepl("65-74", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    4 * grepl("75+", respons_gestratificeerd_vereenvoudigd$leeftijd)
)/(
  grepl("18-34", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    grepl("35-49", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    grepl("50-64", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    grepl("65-74", respons_gestratificeerd_vereenvoudigd$leeftijd) +
    grepl("75+", respons_gestratificeerd_vereenvoudigd$leeftijd)
)

# Geslacht is geen bijzonder belangrijke variabele om te blijven
# onderscheiden
positie_G <- (
  0 * grepl("Man", respons_gestratificeerd_vereenvoudigd$geslacht) +
    2 * grepl("Vrouw", respons_gestratificeerd_vereenvoudigd$geslacht)
)/(
  grepl("Man", respons_gestratificeerd_vereenvoudigd$geslacht) +
    grepl("Vrouw", respons_gestratificeerd_vereenvoudigd$geslacht)
)

# Posities naar opleiding verschillen relatief veel, dus het samenvoegen
# van opleidingscategorieën is relatief onaantrekkelijk
positie_O <- (
  0 * grepl(
    "laag opgeleid",
    respons_gestratificeerd_vereenvoudigd$opleiding
  ) +
    3 * grepl(
      "middelbaar opgeleid",
      respons_gestratificeerd_vereenvoudigd$opleiding
    ) +
    5 * grepl(
      "hoog opgeleid",
```

```
      respons_gestratificeerd_vereenvoudigd$opleiding
    )
)/(
  grepl(
    "laag opgeleid",
    respons_gestratificeerd_vereenvoudigd$opleiding
  ) +
    grepl(
      "middelbaar opgeleid",
      respons_gestratificeerd_vereenvoudigd$opleiding
    ) +
    grepl(
      "hoog opgeleid",
      respons_gestratificeerd_vereenvoudigd$opleiding
    )
)

# Bereken de afstanden
onderlinge_afstanden <- as.matrix(
  dist(positie_L) +
    dist(positie_G) +
    dist(positie_O)
)

# Zet de aftanden van cellen tot zichzelf op een arbitrair groot getal
diag(onderlinge_afstanden) <- Inf

# Bereken voor iedere cel de meest gelijkende andere cel
dichtsbijzijnde_cellen <- apply(
  onderlinge_afstanden,
  MARGIN = 1,
  FUN = which.min
)

# Geef iedere cel een arbitrair nummer om op samen te voegen
respons_gestratificeerd_vereenvoudigd$groep <-
  1:nrow(respons_gestratificeerd_vereenvoudigd)

# Geef cellen die samengevoegd worden een nieuw groepsnummer gelijk aan
# het kleinste van hun groepsnummers
samenvoegnummer <- pmin(
  respons_gestratificeerd_vereenvoudigd$groep[
    which(correctie_nodig)
  ],
  respons_gestratificeerd_vereenvoudigd$groep[
    dichtsbijzijnde_cellen[which(correctie_nodig)]
  ]
)
respons_gestratificeerd_vereenvoudigd$groep[
  dichtsbijzijnde_cellen[which(correctie_nodig)]
```

```
] <- samenvoegnummer
respons_gestratificeerd_vereenvoudigd$groep[
  which(correctie_nodig)
] <- samenvoegnummer

# Voeg cellen samen op groep en bereken nieuwe dekkingsgewichten voor
# de groepen
dekkingsgewichten_PF <- left_join(
  synthetisch_fryslan_summary_vereenvoudigd,
  respons_gestratificeerd_vereenvoudigd,
  by = join_by(
    corop,
    geslacht,
    opleiding,
    leeftijd
  )
) |>
  group_by(groep) |>
  summarise(
    corop = paste(sort(unique(corop)), collapse = ", "),
    geslacht = paste(sort(unique(geslacht)), collapse = ", "),
    opleiding = paste(sort(unique(opleiding)), collapse = ", "),
    leeftijd = paste(sort(unique(leeftijd)), collapse = ", "),
    aantal_inwoners = sum(aantal_inwoners),
    aantal_respondenten = sum(aantal_respondenten, na.rm = TRUE)
  ) |>
  mutate(
    dekkingsgewicht_PF = (aantal_inwoners / sum(aantal_inwoners))/
      (aantal_respondenten/sum(aantal_respondenten, na.rm = TRUE)
      )
  ) |>
  select(corop, geslacht, opleiding, leeftijd, dekkingsgewicht_PF)

# Update de populatietabel om populatieomvang per nieuwe groep te geven
synthetisch_fryslan_summary_vereenvoudigd <-
  left_join(synthetisch_fryslan_summary_vereenvoudigd,
            respons_gestratificeerd_vereenvoudigd,
            by = join_by(corop,
                         geslacht,
                         opleiding,
                         leeftijd)
  ) |>
  group_by(groep) |>
  summarise(
    corop = paste(unique(corop), collapse = ", "),
    geslacht = paste(unique(geslacht), collapse = ", "),
    opleiding = paste(unique(opleiding), collapse = ", "),
    leeftijd = paste(unique(leeftijd), collapse = ", "),
    aantal_inwoners = sum(aantal_inwoners)) |>
  select(corop, geslacht, opleiding, leeftijd, aantal_inwoners)
```

```
    # Voeg cellen samen in de responssamenvatting om te controleren of er
    # nog meer cellen moeten worden samengevoegd
    respons_gestratificeerd_vereenvoudigd <-
      respons_gestratificeerd_vereenvoudigd |>
      group_by(groep) |>
      summarise(
        corop = paste(unique(corop), collapse = ", "),
        geslacht = paste(unique(geslacht), collapse = ", "),
        opleiding = paste(unique(opleiding), collapse = ", "),
        leeftijd = paste(unique(leeftijd), collapse = ", "),
        aantal_respondenten = sum(aantal_respondenten, na.rm = TRUE)
      )

    # Controleer of het samenvoegen van cellen kan stoppen, of dat er nog
    # een ronde van samenvoegingen moet komen
    verdere_vereenvoudiging_nodig <-
      any(respons_gestratificeerd_vereenvoudigd$aantal_respondenten < 5) |
      any(is.na(dekkingsgewichten_PF$dekkingsgewicht_PF))
  }

  # Breng het dataframe met dekkingsgewichten weer in een vorm
  # waar iedere combinatie van achtergrondvariabelen een eigen cel heeft
  # om het makkelijk aan de uitvraag te koppelen
  dekkingsgewichten_PF <- dekkingsgewichten_PF |>
    mutate(corop = str_split(corop, pattern = ", "),
           geslacht = str_split(geslacht, pattern = ", "),
           leeftijd = str_split(leeftijd, pattern = ", "),
           opleiding = str_split(opleiding, pattern = ", ")) |>
    unnest_longer("corop") |>
    unnest_longer("geslacht") |>
    unnest_longer("leeftijd") |>
    unnest_longer("opleiding")
}
```

COLOFON

“Weging van enquêtes met lege en dun bezette strata”, is een uitgave van Planbureau Fryslân.

De samenvatting van dit rapport staat op de website van Planbureau Fryslân.

Auteurs
Siebren Kooistra

Met medewerking van
Mertan Aydogan, Chaïm la Roi, Riemer Wiersma

Eindredactie
Chaïm la Roi

Planbureau Fryslân
Prins Hendrikstraat 8
8911 BK Leeuwarden
(058) 234 85 00
info@planbureaufryslan.nl
www.planbureaufryslan.nl

Planbureau Fryslân wordt gesubsidieerd door de provincie Fryslân.