BIJLAGE 3a

Algemene Rekenkamer

Methodologische toelichting analyse gegevens mbo-studenten

februari 2024

Gelijke kansen mbo: Methodologische toelichting opzet en uitvoering analyse gegevens mbo-studenten

Vooraf

De Algemene Rekenkamer heeft onderzoek gedaan naar het probleem dat studenten in het middelbaar beroepsonderwijs (mbo) niet allemaal evenveel kans hebben om tot goede onderwijsresultaten te komen.

De minister van OCW heeft de afgelopen jaren verschillende maatregelen genomen om de kansengelijkheid voor mbo-studenten te verbeteren. We wilden weten in hoeverre deze maatregelen – en de budgetten die de minister daarvoor inzet – daadwerkelijk bijdragen aan kansengelijkheid voor mbo-studenten.

We hebben daartoe niet alleen het beleid van de minister bestudeerd, maar ook gekeken naar de situatie in het mbo zelf. Zo hebben we gekeken naar de praktijk-situatie op de scholen. Daarvoor hebben we 7 mbo-scholen bezocht. Bij deze mbo-scholen hebben we gesprekken gevoerd met studenten, docenten, (beleids) medewerkers, bestuurders en werkgevers die stages en leerbanen bieden aan mbo-studenten. Daarnaast hebben we gegevens geanalyseerd die de Dienst Uitvoering Onderwijs (DUO) bijhoudt over de kenmerken van mbo-studenten en de wijze van instroom, doorstroom en uitstroom van deze studenten. Aan de hand van deze gegevens konden we vaststellen of er een relatie bestaat tussen de achtergrond-kenmerken en thuissituatie van studenten en hun onderwijsresultaten.

Het doel van onze analyse van DUO-gegevens was antwoord te geven op onderzoeksvraag 2 van het onderzoek naar gelijke kansen in het mbo: *In welke mate komt kansenongelijkheid van studenten op mbo-instellingen voor, welke factoren zijn hierop van invloed en wat doen mbo-instellingen hieraan op basis van hun kwaliteitsagenda's?*

In het onderzoek zijn we vertrokken vanuit de definitie die de minister van OCW hanteerde in verschillende beleidsdocumenten bij de start van ons onderzoek (begin 2022): er is sprake van gelijke kansen in het mbo als het onderwijssucces, kansen op de arbeidsmarkt en het vinden van een stage afhankelijk zijn van de capaciteiten en inzet van studenten en niet worden beïnvloed door hun achtergrond en afkomst, zoals het inkomen of de opleiding van ouders. In het geval dat achtergrond en afkomst wel een negatieve invloed hebben, dan is volgens de minister sprake van kansenongelijkheid (OCW, 2018).

In deze bijlage 3a geven we een uitgebreide toelichting op de wijze waarop we onze analyse van de DUO-gegevens hebben uitgevoerd. Hiermee willen we inzicht geven in de opzet en uitvoering van ons onderzoek, inclusief de beperkingen van onze analyse en de keuzes die we hebben gemaakt. Dit inzicht kan behulpzaam zijn voor toekomstige onderzoeken en monitoring van de resultaten van het gelijkekansenbeleid van de minister van OCW en van de besturen van mbo-instellingen.
De resultaten van onze analyse hebben we opgenomen in 2 rapporten. Deel 1 hebben we op 12 september 2023 uitgebracht, samen met een brief aan de MBO Raad. Deel 2 is op 14 februari 2024 uitgebracht samen met deze bijlage en een tabellenboek (bijlage 3b). In dit tabellenboek hebben we de cijfermatige uitkomsten van onze analyse van DUO-gegevens opgenomen.

Deze bijlage 3a is als volgt opgebouwd. In hoofdstuk 1 zetten we uiteen hoe we om zijn gegaan met het begrip (on)gelijke kansen. In hoofdstuk 2 gaan we in op de data (gegevens), die we hebben gebruikt in onze analyses. In hoofdstuk 3 lichten we onze analysemethode toe. We sluiten af met een overzicht van gebruikte afkortingen en literatuur.

1. Over het begrip gelijke kansen

1.1 Twee opvattingen over (on)gelijke kansen

Wanneer gesproken wordt over kansenongelijkheid kan dit op twee manieren opgevat worden. De eerste opvatting is dat achtergrondkenmerken überhaupt niet mogen uitmaken. Het idee hierachter is dat 'verklarende' factoren zoals bijvoorbeeld capaciteit, motivatie en vaardigheden ook het resultaat zijn van kansenongelijkheid. Eerder onderzoek geeft bijvoorbeeld aan dat het ervaren van (financiële) stress samenhangt met lagere cito-scores (Weel, Bussink en Koeman, 2023).
Een tweede opvatting is dat er best verschillen mogen zijn tussen groepen met verschillende achtergrondkenmerken, maar dat die verklaard zouden moeten worden door capaciteit. Bij deze tweede opvatting wordt capaciteit meer gezien als een vaststaand gegeven (bijvoorbeeld genetisch bepaald of aanleg) dan als het resultaat van eerdere kansenongelijkheid. In dit onderzoek testen we deze tweede opvatting. Niet omdat we per definitie vinden dat dit de juiste opvatting is, maar omdat op deze manier inzichtelijk wordt of achtergrondkenmerken uitmaken, ongeacht welke opvatting gehanteerd wordt. Bovendien past dit bij de opvatting van de minister van OCW over wanneer sprake is van kansenongelijkheid: wanneer niet de capaciteiten, maar achtergrondkenmerken van een leerling of student bepalend zijn voor de onderwijsresultaten.

1.2 Achtergrondkenmerken mbo-studenten

Onder achtergrondkenmerken verstaan we in dit onderzoek: geslacht, migratie-achtergrond en sociaaleconomische status van ouders (waaronder opleidingsniveau). Er is sprake van kansen*gelijkheid* als de verschillen tussen onderwijsresultaten van leerlingen en studenten variëren met hun capaciteiten, talenten en vaardigheden en dit verband niet afhankelijk is van hun achtergrond-kenmerken (OCW, 2018). Wanneer we in deze bijlage spreken over de resultaten van onze analyses, spreken we over verschillen in onderwijsresultaten en niet over kansengelijkheid. Om te kunnen spreken over kansen(on)gelijkheid, moet de causaliteit van de relaties zijn onderzocht en moeten alle bepalende factoren voor kansen(on)gelijkheid zijn meegenomen. We spreken in dit onderzoek over verschillen in onderwijsresultaten. Dat komt omdat we niet alle factoren hebben kunnen meenemen, die op basis van eerder uitgevoerde onderzoeken bekend zijn. En omdat we in ons onderzoek de causaliteit niet hebben kunnen onderzoeken. Wij zijn niet op de hoogte van een onderzoek dat kansengelijkheid op het mbo in Nederland op een vergelijkbare manier heeft onderzocht. Ook is er naar ons weten geen dashboard dat kansen(on)gelijkheid monitort, zoals door de minister is gedefinieerd.

2. Data

2.1 Schoolcarrière en achtergrondkenmerken

De dataset die in dit onderzoek centraal staat, is afkomstig van de Dienst Uitvoering Onderwijs (DUO). DUO bezit informatie over de volledige schoolcarrière van *alle* mbo-studenten die tussen 2015 en 2021 ingeschreven stonden. Bijvoorbeeld gegevens over de in- en uitschrijfdatum, de instelling en locatie waar studenten ingeschreven staan, het niveau en type opleiding (bol/bbl en sector). Daarnaast heeft DUO ook informatie over onderwijsprestaties, zoals de centrale eindtoetsscore in het primair onderwijs, het behalen van een diploma, doorstroom naar een hoger niveau, en het voortijdig verlaten van het onderwijs. Tot slot bezit DUO ook demografische gegevens (geslacht, leeftijd en migratieachtergrond van de student) en gegevens die iets zeggen over hun sociaaleconomische status (opleidingsniveau ouders). Migratieachtergrond is één van de achtergrondkenmerken van studenten waar het beleid gericht op het bevorderen van kansengelijkheid zich op richt. Met het meenemen van migratieachtergrond (eerste- en tweedegeneratie) in onze analyses willen we de feitelijke situatie in kaart brengen en toetsen of evenredigheid in onderwijsuitkomsten behaald zijn. Wel moet rekening worden gehouden dat eventuele verschillen naar migratieachtergrond terug te voeren kunnen zijn naar bijvoorbeeld verschillen in het huishoudinkomen. Doordat deze gegevens niet beschikbaar waren bij DUO, hebben we deze kenmerken niet mee kunnen nemen. In totaal hebben we data over 1.359.242 mbo-studenten uit de periode 2015-2021.

In februari 2022 heeft DUO de eerste data aan de Algemene Rekenkamer voor dit onderzoek geleverd. Vervolgens heeft DUO de data in vier leveringen aangeleverd. In april 2022 is informatie over voortijdig schoolverlaten (vsv) aangeleverd. Op verzoek van de Algemene Rekenkamer is in april 2022 een extra jaar (2009) toegevoegd aan de informatie over de centrale eindtoetsscores. Op verzoek van de Algemene Rekenkamer zijn in juli 2022 gegevens over het opleidingsniveau van ouders nagestuurd. In december 2022 heeft DUO de instructie geleverd over het juist gebruiken van de vsv-gegevens.

De data zijn aangeleverd in een ‘breed format’. Dit betekent dat alle gegevens van alle jaren van één student op één rij staan. In tabel 1 is een fictieve dataset te zien als voorbeeld. Studenten hebben waarden op bepaalde variabelen in een bepaald jaar wanneer zij in dat jaar stonden ingeschreven op het mbo. Wanneer zij niet stonden ingeschreven in dat jaar, hebben zij dus geen waarden op de variabelen in dat jaar.

Tabel 1 *Fictief voorbeeld van een dataset in breed format*

ID	Leeftijd 2016	Diploma 2016	Geslacht 2016	Leeftijd 2017	Diploma 2017	...	Leeftijd 2020
1	20	Ja	Man				
2				18	Nee		21
3	16	Nee	Vrouw	17	Nee		

2.2 Kwaliteitsagenda's per mbo-instelling

De minister heeft in 2018 bestuurlijke afspraken gemaakt met de MBO Raad. In deze afspraken heeft de minister ook drie landelijke speerpunten opgenomen, waaronder het speerpunt gelijke kansen. De besturen van mbo-scholen moesten vervolgens elk voor zich de bestuurlijke afspraken uitwerken in een kwaliteitsagenda voor de periode 2019-2022. Op basis van deze kwaliteitsagenda zouden zij ook in aanmerking komen voor het extra geld dat de minister hiervoor beschikbaar had gesteld.

Alle mbo-besturen hebben een kwaliteitsagenda 2019-2022 opgesteld. Deze kwaliteitsagenda's zijn openbaar beschikbaar via de Rijksoverheid website en geven onder andere inzicht in de:

- doelen, maatregelen en acties voor het bevorderen van gelijke kansen van mbo-studenten;
- verdeling van het kwaliteitsbudget over de landelijke speerpunten waaronder gelijke kansen;
- monitoring van de voortgang.

In eerste instantie wilden we de kwaliteitsagenda's meenemen in onze analyses om de effecten van de beleidsmaatregelen te kunnen meten. Om meerdere redenen was dit niet mogelijk. Allereerst bleek het niet mogelijk om de periode voor het invoeren van de kwaliteitsagenda's te analyseren, door missende gegevens over de locatie van de school waar studenten onderwijs volgen. Daarnaast bleek uit interviews met medewerkers van mbo-instellingen dat er geen eenduidig moment te bepalen is waarop de kwaliteitsagenda's "effect" zouden moeten hebben. Deze informatie hadden we wel nodig om het in onze analyses te kunnen modelleren. Tot slot bleek uit interviews dat de kwaliteitsagenda's vaak niet goed weergeven wat een instelling precies doet om gelijke kansen te bevorderen. In hoofdstuk 5 van het rapport Gelijke kansen mbo deel II gaan we hier verder op in.

2.3 Variabelen

Onderwijsresultaten

We hebben drie soorten onderwijsresultaten geanalyseerd in dit onderzoek. Allereerst is gekeken naar *doorstroom*. Met doorstroom wordt bedoeld dat een student in jaar *t* een diploma haalt en in het daaropvolgend jaar (*t*+1) ingeschreven staat op een hoger onderwijsniveau. Dat kan een hoger niveau op het mbo zijn, maar ook doorstroom naar het hbo. Deze studenten zijn afgezet tegen de groep die in jaar *t* wel een diploma heeft behaald, maar in jaar *t*+1 niet ingeschreven staat op een hoger niveau. We hebben niet onderzocht of de doorstroom 'succesvol' was, d.w.z. of bij het doorstromen naar een hoger niveau ook een propedeuse of diploma is behaald.

Daarnaast is *voortijdig schoolverlaten* geanalyseerd. Deze variabele is aangeleverd door DUO. Of iemand een voortijdig schoolverlater (vsv'er) is, wordt bepaald door de inschrijving van de student in jaar *t* te vergelijken met de inschrijving van de student in het daaropvolgend jaar (*t*+1). Een student die het onderwijs in jaar *t*+1 heeft verlaten zonder een startkwalificatie (minstens een diploma op mbo niveau 2 of een havo of vwo diploma), wordt aangemerkt als nieuwe vsv'er. Hierop gelden wel enkele uitzonderingen. Zo kan een student wel het onderwijs verlaten zonder start-kwalificatie, maar niet worden aangemerkt als vsv'er als een student: ouder is dan 22, nieuwkomer is die minder dan één jaar in Nederland is, geen vaste woonplaats heeft of in het buitenland woont, een entree diploma heeft (niveau 1) én minstens 12 uur per week werkt (DUO, 2022).

Tot slot hebben we geprobeerd het behalen van een diploma behalen te analyseren. Samengevat bleek het analyseren van behaalde diploma's niet mogelijk, omdat er geen vergelijkingsgroep is. Dit komt omdat er geen duidelijke nominale studieduur is op het mbo. We overwogen om de nominale studieduur op 10 jaar te zetten. Dit bleek onhaalbaar, omdat de locatiedata die nodig zijn voor onze analyses pas beschikbaar zijn vanaf 2019. In 2019 is de nominale studieduur van 10 jaar voor studenten die vanaf 2019 studeren nog niet voorbij, en is er dus geen vergelijkingsgroep.

Studentkenmerken

In de analyses nemen we ook een aantal studentkenmerken mee, namelijk **geslacht** (laatst bekende geslacht zoals bekend bij DUO, onderverdeeld in vrouw en man), **leeftijd** (op 1 augustus voor een betreffend jaar) en **migratieachtergrond** (aangeleverd door DUO, onderverdeeld in niet-westerse migratieachtergrond, westerse migratie-achtergrond en geen migratieachtergrond). Iemand heeft volgens de definitie van DUO een migratieachtergrond als hijzelf en/of één of beiden van zijn ouders in een ander land dan Nederland geboren is.

Ook nemen we het (laatst bekende) **opleidingsniveau van de ouders** mee. Dit blijkt uit eerder onderzoek namelijk een factor van invloed te zijn op de onderwijsresultaten van studenten (SER, 2021; SCP, 2014). Bij DUO staat het opleidingsniveau van de ouders van de studenten niet geregistreerd. Wel is bekend of één of beiden ouders maximaal het opleidingsniveau vmbo-kader hebben gevolgd. Onze variabele bevat dus twee categorieën: ouders hebben maximaal het opleidingsniveau vmbo-kader of ouders hebben een hoger opleidingsniveau dan vmbo-kader. Omdat er van een groot aantal studenten geen gegevens bekend zijn over het opleidingsniveau van de ouders, voegen we ook de categorie 'onbekend' toe.

Tot slot nemen we in de studentkenmerken ook de capaciteitsmaat **citoscore** mee. We nemen alleen de centrale eindexamen (cito) scores mee. Dit heeft meerdere redenen. Allereerst is voor studenten die de eindtoets voor 2014 hebben gemaakt niet bekend welke eindtoets zij hebben afgelegd. Het bereik van de citotoets (een minimum 500 tot een maximum van 550) overlapt als enige eindtoets niet met het bereik van andere eindtoetsen. Daarom kan voor studenten met een eindtoetsscore tussen de 500 en 550 met zekerheid worden vastgesteld dat zij de citotoets hebben afgelegd. Ten tweede gebruiken we alleen de citoscores omdat verschillende eindtoetssoorten niet met elkaar te vergelijken zijn. De cito is in onze onderzoeksperiode veruit de meest gebruikte eindtoetssoort. Door te kiezen voor de citoscore verliezen we in verhouding met andere toetssoorten zo min mogelijk informatie over zo min mogelijk studenten. Omdat de citoscores voor 2014 niet werden gestandaardiseerd, voegen we ook toe in welk **jaar** of welke periode een student de toets heeft afgelegd om te corrigeren voor jaarlijkse verschillen in de normering.

Studiekenmerken

Drie studiekenmerken hebben we meegenomen in de analyses, namelijk:

- het **niveau** van de mbo-opleiding die de student heeft gevolgd in jaar *t* op basis van de tabel met codes van mbo-opleidingen, de Crebotabel van SBB;
- of een student **bol of bb**l doet;
- de **sector** van de opleiding die de student in jaar *t* heeft gevolgd op basis van de sectorkamers van SBB en onderverdeeld in vier categorieën, namelijk: *zakelijke dienstverlening & handel* [sectorkamer zakelijke dienstverlening en veiligheid en sectorkamer handel], *techniek en ict* [sectorkamer techniek en gebouwde omgeving en sectorkamer creatieve industrie en ICT], *zorg en sport* [sectorkamer zorg, welzijn en sport] en *overig* [bovensectoraal, entree, sectorkamer onbekend, sectorkamer mobiliteit, transport, logistiek en maritiem, sectorkamer specialistisch vakmanschap en sectorkamer voedsel, groep en gastvrijheid].

2.4 Omgaan met missende waarden

Er zijn twee variabelen waarop we voor een aanzienlijk deel van de studenten geen waarde hebben, namelijk *citoscore* en *opleidingsniveau van de ouders*. De missende waarden zijn niet willekeurig (niet *missing completely at random*). De standaard-methode om met missende waarden om te gaan, *listwise deletion*, zorgt dus mogelijk voor *biased* resultaten. Dit betekent dat de uitkomsten van de analyses met *listwise deletion* mogelijk afwijken van de "werkelijke" uitkomsten die je zou vinden wanneer je volledige data hebt (of er sprake is van missing *completely at random*). Daarnaast zorgt *listwise deletion* ervoor dat er niet meer geanalyseerd wordt met populatiedata, omdat een aanzienlijk deel van de studenten niet wordt meegenomen in de analyses. Dit zorgt ervoor dat we de uitkomsten van onze regressieanalyses niet onderling kunnen vergelijken. *Listwise deletion* heeft dus niet onze voorkeur.

Er zijn verschillende manieren om met missende data om te gaan (zie bijvoorbeeld Kang, 2013; Van Buuren & Groothuis-Oudshoorn, 2011). Om complexiteitsredenen hebben we ervoor gekozen om de *missing indicator method* toe te passen. Bij deze methode worden missende observaties op een vaste waarde gezet en een extra dummy indicator wordt toegevoegd om aan te geven dat het gaat om een missende waarde. Hierdoor kunnen alle studenten worden meegenomen in de analyses, waardoor de hele populatie geanalyseerd wordt en de resultaten van de modellen onderling vergeleken kunnen worden. Het nadeel van deze methode is dat de afwijkingen (*bias*) van de analyse afhangen van de grootte van de (onbekende) correlatie tussen de ontbrekende informatie en de andere variabelen. Dit betekent dat, afhankelijk van de oorzaak van de ontbrekende data in dit onderzoek, de relatie tussen de onafhankelijke variabelen en doorstroom/vsv kan worden overschat of onderschat. We kunnen niet met zekerheid zeggen hoe groot het effect van deze afwijking is.

2.5 Betrouwbaarheid en beperkingen data en analyses

De data zijn afkomstig uit verschillende bronbestanden: DUO mbo 1-cijfer, DUO ho 1-cijfer, DUO po 1-cijfer, DUO vsv-bestand, CBS APCG, Crebo-tabel SBB en data van DUO die buiten de andere bronbestanden vallen. Een deel van de data is op geaggregeerd niveau openbaar gemaakt op o.a. de website DUO Open onderwijsdata.

Hoewel dit onderzoek vernieuwende elementen heeft, heeft het onderzoek ook belangrijke beperkingen. Omdat we met data van DUO werken, zijn we gebonden aan de (achtergrond)kenmerken die DUO registreert. Hierin is bijvoorbeeld wel informatie over de migratieachtergrond van de studenten beschikbaar op individueel niveau, maar is geen informatie beschikbaar over andere belangrijke kenmerken, zoals

gedetailleerde gegevens over het inkomen van de ouders en andere kenmerken die wijzen op de socialeeconomische status van een student en zijn of haar gezin. Ook bevat de data geen informatie over kenmerken die het effect van deze achtergrondkenmerken kunnen verklaren: denk aan het hebben van schulden, psychosociale problemen, gezondheidsproblemen of de gezinssamenstelling. Tevens was dit laatste ook niet het doel van dit onderzoek: het doel was om te monitoren of er ongelijkheid plaats vindt op basis van achtergrond: volgens de minister is dit onwenselijk. In een studie van CBS blijkt bijvoorbeeld dat migratieachtergrond weinig effect heeft op de kans om vsv'er te worden, als er ook gecontroleerd wordt voor dit soort kenmerken (CBS, 2022). In deze studie kunnen we dus enkel aantonen of studenten met bepaalde achtergrondkenmerken van elkaar verschillen in onderwijsresultaten én of dit verschil verdwijnt als je capaciteit meeneemt. Het zegt echter zeker niet dat het achtergrondkenmerk de oorzaak is van het onderwijsresultaat. Waarom eventuele verschillen bestaan en wat dit verklaart, valt buiten de scope van deze data-analyse. Op basis van door ons bestudeerde literatuur en ons onderzoek in de praktijk noemen we in ons rapport Gelijke kansen mbo deel II wel een aantal logische verklaringen voor de samenhang die we tussen de kenmerken van studenten en de onderwijsresultaten hebben aangetroffen.

Voor dit onderzoek kijken we ook naar het effect van het achtergrondkenmerk 'migratieachtergrond' op onderwijsresultaten. Rondom het gebruik van dit kenmerk en het uitsplitsen van cijfers naar migratieachtergrond worden terecht veel vragen gesteld. Het CBS heeft hierom een afwegingkader opgesteld op basis waarvan zij voor zichzelf bepalen of het meenemen van migratieachtergrond gegrond is en gerechtvaardigd wordt. Het is van belang om deze keuze bewust te maken. Het meenemen van migratieachtergrond kan er namelijk voor zorgen dat waargenomen verschillen tussen herkomstgroepen impliciet worden toegeschreven aan culturele verschillen. Ook kan het bij individuen en groepen gevoelens van uitsluiting en stigmatisering oproepen. Aan de andere kant heeft het meenemen van dit kenmerk ook voordelen. Door een cijfermatige onderbouwing kunnen discussies worden gevoerd op basis van feiten en niet op basis van onderbuikgevoelens of (voor) oordelen. Een bewuste afweging hierover maken is dus belangrijk. In het afwegingskader wordt dit gedaan op basis of het legitiem, functioneel, proportioneel en subsidiariaat is. Wij komen tot de conclusie dat voor dit onderzoek het meenemen van migratieachtergrond gerechtvaardigd is. Een belangrijke reden hiervoor is dat het relevant wordt geacht voor monitoring van beleid, bijvoorbeeld met betrekking tot het realiseren van gelijke kansen, ongeacht migratieachtergrond.

De data heeft verder nog een aantal beperkingen:

- Voor 2019 is er geen informatie bekend over de onderwijslocatie van de student.
- De onderwijslocatiecodes zijn geanonimiseerd; het is voor ons niet mogelijk om op basis van de data te achterhalen om welke schoollocatie het gaat. Dit had van meerwaarde kunnen zijn voor het kwalitatieve deel van het onderzoek.
- Het opleidingsniveau van de ouder(s) is onderverdeeld in 'lager opgeleid' (max vmbo-kader) of daarboven. Dat is geen ideale maat voor opleidingsniveau, omdat je alle variatie mist in het opleidingsniveau van ouders.
- Gegevens over het inkomen van de ouder(s) van de student zijn niet beschikbaar op huishoudniveau, maar alleen op viercijferig postcodegebied. Om deze reden kon het inkomen van de ouders niet meegenomen worden in ons onderzoek.
- De migratieachtergrond van de student is ingedeeld in 'geen', 'westers' of 'niet-westers'. We volgen hierbij de definitie van het CBS (www.CBS.nl). Ook dit is geen ideale indeling, binnen zowel westers, maar zeker niet-westers is zeer veel variatie mogelijk. Een specifiekere indeling was echter niet mogelijk, omdat de groepen per afkomstland dan zeer klein zouden worden. Omdat onze modellen al complex zijn en convergentieproblemen hebben, hebben we deze driedeling gebruikt. In dit onderzoek hadden we daarnaast de verwachting dat een niet-westerse migratie-achtergrond zou kunnen leiden tot bijvoorbeeld vooroordelen (SER, 2021; Elffers, 2020; Van Leest et al. 2020, Timmermans et al. 2018), daarom is er toch voor gekozen deze indeling te handhaven. Voor meer nauwkeurig onderzoek raden we aan om wel een uitsplitsing te maken naar specifieke herkomstlanden en indien mogelijk ook naar generatie.
- Studenten die twee studies tegelijk doen, staan maar voor één van de studies ingeschreven (hoofdinschrijving). Wanneer zij een diploma behalen voor de nevenopleiding, is dat niet te zien in de data.
- Er zijn studenten waarvan we wel weten dat ze studeren, maar die (grotendeels) in onze dataset ontbreken. Het kan voorkomen dat studenten zich na 1 oktober inschrijven voor een studie en dan dat jaar een diploma halen. Deze studenten hebben dan wel een waarde op diploma behalen, maar niet op de variabelen gerelateerd aan achtergrond en opleiding. Dit komt, omdat die variabelen worden vastgelegd op de peildatum van 1 oktober. Omdat we niks kunnen zeggen over deze studenten, behalve dat ze een diploma halen, nemen we ze niet mee in de analyse. Dit ging in 2020 om ongeveer 1,5% van alle studenten die ingeschreven stonden.
- Studenten die in jaar *t* na 1 oktober stonden ingeschreven en in jaar *t*+1 zich voor 1 oktober hebben uitgeschreven, zijn niet opgenomen in de dataset.
- We kunnen voor studenten die in één jaar twee diploma's behalen van twee verschillende niveaus niet uit de data opmaken dat deze studenten zijn door-gestroomd. In de data wordt alleen het hoogste niveau benoemd.

- We gebruiken de citoscore als een proxymaat voor capaciteit. De citotoets meet echter niet alleen capaciteit, maar ook zaken als beheersing van de Nederlandse taal. Daarnaast is het mogelijk dat er ook sprake is van kansenongelijkheid in de citoscores (bijvoorbeeld omdat sommige ouders bijles of cito-training voor hun kinderen regelen). Echter, de citoscore is de beste proxy voor capaciteit die in de DUO data beschikbaar is. In het schooladvies van leraren kan bijvoorbeeld een bias zitten (Elffers, 2020; Timmermans et al. 2018). De Inspectie van het Onderwijs heeft in de *Staat van het Onderwijs* hier sinds 2016 ook op gewezen.

3. Analyse-aanpak

3.1 Beschrijvende analyses

Voor de beschrijvende statistiek zijn de volgende selecties gemaakt:

- Alleen studenten die bol of bbl deden zijn meegenomen. De studenten die examenkandidaat waren (die als extraneus een opleiding afronden) zijn dus niet meegenomen. We hebben hiervoor gekozen omdat examenkandidaten geen lessen mogen volgen, maar alleen toetsen mogen maken. We verwachten dat mechanismen die te maken hebben met het doorstromen naar een hoger niveau of het worden van een voortijdig schoolverlater anders zijn dan voor “reguliere” studenten, omdat zij dus geen typisch schooltraject op het mbo hebben. Omdat er weinig studenten examenkandidaat zijn, zal het uitsluiten van deze studenten de resultaten minimaal beïnvloeden.
- Studenten met een missende waarde op migratieachtergrond zijn niet meegenomen in de analyses die kijken naar migratieachtergrond. Dit gaat om een klein aantal studenten (ongeveer 0.5% van de studenten per jaar).

Voor de figuren hebben we daarnaast nog de volgende selecties gemaakt:

- We hebben alleen studenten meegenomen die tussen de 14-25 jaar oud zijn. Dit sluit aan bij de selectie van studenten voor de multivariate analyses.
- De analyses zijn uitgevoerd voor studenten die gestart zijn met hun studie in de jaren 2015 tot en met 2021. Wanneer we kijken naar doorstroom en vsv gaat dit over de jaren 2016-2021, omdat je altijd kijkt naar *t*+1 (schooljaar + 1 jaar later). Wanneer we naar doorstroom en vsv kijken, gebruiken we de studie-informatie (niveau opleiding, sector opleiding etc.) en achtergrondkenmerken van het voorgaande schooljaar (schooljaar ‘*t*’).

Voor specifieke analyse zijn aanvullende selectiekeuzes gemaakt (zie 3.2).

Bij de beschrijvende analyses zijn per jaar eerst frequentietabellen gemaakt. Dit is gedaan voor achtergrondkenmerken (migratieachtergrond, geslacht, leeftijd, opleidingsniveau ouders), voor studiekenmerken (niveau, bol of bbl en sectorkamer) en onderwijsresultaten (citoscore, doorstroom, vsv en diploma behalen). Vervolgens hebben we kruistabellen gemaakt waarbij de we de variabele migratieachtergrond afzetten tegen de andere achtergrondkenmerken en de studiekenmerken. Hier is alleen voor migratieachtergrond gekozen, omdat dat de verklarende factor is in onze analyse. Ten slotte is gekeken hoe de onderwijsresultaten verschillen voor studenten zonder migratieachtergrond, diegene met een westerse migratieachtergrond en degene met een niet-westerse migratieachtergrond.

3.2 Selectie studenten voor multivariate analyses

Voor de multivariate analyses (zie analysetechniek) zijn de volgende selecties gemaakt:

- Alleen studenten die bij doorstroom minimaal 14 jaar oud waren en maximaal 25 jaar oud waren zijn meegenomen in de analyses. Dit is het overgrote deel van de studentenpopulatie (zo'n 85% van de studenten).
- Alleen studenten die tot de startpopulatie behoren zijn meegenomen in de analyses voor voortijdig schoolverlaten. We volgen hierbij de definitie van DUO (DUO, 2022; DUO, z.d.).
- Alleen studenten die bol of bbl deden zijn meegenomen. De studenten die examenkandidaat waren (die als exrtraneus een opleiding afronden) zijn dus niet meegenomen. We hebben hiervoor gekozen omdat examenkandidaten geen lessen mogen volgen, maar alleen toetsen mogen maken. We verwachten dat mechanismen die te maken hebben met het doorstromen naar een hoger niveau of het worden van een voortijdig schoolverlater anders zijn dan voor "reguliere" studenten, omdat zij dus geen typisch schooltraject op het mbo hebben. Omdat er weinig studenten examenkandidaat zijn, zal het uitsluiten van deze studenten de resultaten minimaal beïnvloeden.
- De analyses zijn alleen uitgevoerd voor de jaren 2020 en 2021 (dat betekent: is iemand doorgestroomd of vsv'er geworden in het schooljaar 2020). In deze analyses gebruiken we de studie-informatie (niveau opleiding, sector opleiding, etc.) van het voorgaande schooljaar. Voor 2019 is er geen informatie bekend over de onderwijslocatie. Omdat we zien dat deze informatie niet weggelaten kan worden, kunnen we alleen de jaren 2020 en 2021 analyseren (waarbij we dus data gebruiken over de studiejaren 2019 en 2020).
- Studenten met een missende waarde op migratieachtergrond zijn niet meegenomen. Dit gaat om een erg klein aantal studenten (in 2020 gaat het na het filteren op de andere keuzes om 25 studenten).

3.3 Analysetechniek

In dit onderzoek zijn multilevel logistische regressieanalyses uitgevoerd om de samenhang tussen de studentkenmerken en andere kenmerken enerzijds (de onafhankelijke variabelen) en de kans op studie-uitkomsten anderzijds te onderzoeken. Er zijn twee studie-uitkomsten geanalyseerd, namelijk *doorstroom naar een hoger niveau en voortijdig schoolverlaten.* In de analyses zijn studenten genest in scholen (locaties) en scholen genest in instellingen. Op deze manier houden we rekening met eventuele clustering binnen schoollocaties en instellingen.

Bij logistische regressieanalyse wordt de kansverhouding op een gebeurtenis voorspeld op basis van één of meer onafhankelijke variabelen. De kansverhouding wil zeggen: de kans op een gebeurtenis in verhouding tot de kans dat die gebeurtenis niet plaatsvindt. Deze methode wordt gebruikt om binaire uitkomstmaten te analyseren.

De modellen worden stapsgewijs opgebouwd. Doordat we populatiedata gebruiken en niet met een steekproef uit de populatie werken, kunnen we de resultaten van de modellen voor een bepaalde uitkomstmaat in een bepaald jaar met elkaar vergelijken. Daarnaast zijn bij het gebruik van populatiedata significantieniveaus niet relevant om te analyseren. De verschillen tussen studenten die we zien in onze analyses, zijn verschillen in de populatie van interesse. Of een statistisch effect relevant is, bepalen we daarom niet op basis van significantie maar op basis van de effectgrootte van de verschillende variabelen. De resultaten kunnen niet gegeneraliseerd worden naar andere populaties. We kunnen de uitkomsten tussen verschillende uitkomst-variabelen en/of jaren niet kwantitatief met elkaar vergelijken. Dit kan wel in kwalitatieve zin. De onderstaande tabel laat zien welke modellen er per uitkomst-maat zijn geanalyseerd.

Tabel 2 *De opbouw van de modellen per uitkomstmaat en jaar*

Model	Doorstroom 2020	Doorstroom 2021	Vsv 2020	Vsv 2021
Model 0: Intercept only *Geen onafhankelijke variabelen*	X	X	X	X
Model 1a: Individuele kenmerken *Migratieachtergrond en opleidingsniveau ouders*	X	X	X	X
Model 1b: Individuele kenmerken *Migratieachtergrond, opleidingsniveau ouders, geslacht student, niveau opleiding student, leeftijd student, migratieachtergrond student, sector opleiding student en bbl of bol*	X	X	X	X
Model 1c: Individuele kenmerken *Migratieachtergrond, opleidingsniveau ouders, geslacht student, niveau opleiding student, leeftijd student, migratieachtergrond student, sector opleiding student, bbl of bol en citoscore (en jaar)*	X	X	X	X
Model 2: random slope migratieachtergrond zonder interactie-term en maatregelen instellingsniveau *Migratieachtergrond, opleidingsniveau ouders, geslacht student, niveau opleiding student, leeftijd student, migratieachtergrond student, sector opleiding student, bbl of bol, citoscore (en jaar) en random slopes voor migratieachtergrond op locatieniveau*	X	X	X	X

Multilevel logistische regressies geven geen proportie verklaarde variantie (R^2). Wel bestaan er verschillende pseudo R^2-maten die vergelijkbaar zijn met de R^2 uit lineaire regressie analyse. We gebruiken (de conditionele) Nakagawa's pseudo R^2 maat (Nakagawa & Schielzeth, 2012; Nakagawa, Johnsonn & Schielzeth, 2017).

De interpretatie van de modellen in *average marginal effects.* In het kort vertelt een *average marginal effect* hoe een afhankelijke variabele gemiddeld verandert wanneer een specifieke onafhankelijke variabele verandert. Alle andere variabelen worden hierbij constant gehouden.[1]

1. Zie voor meer informatie het college van Perraillon (2019) over Interpreting Model Estimates: Marginal Effects.

3.4 Robustness checks

Het toevoegen van cito, de missing indicator én het jaar waarin iemand een citotoets heeft afgelegd, zorgt voor een *convergence error.* Dit komt vaak voor bij het schatten van grote, complexe modellen en kan zorgen voor incorrecte resultaten. De *optimizer* geeft aan dat het model wel is geconvergeerd. We hebben *robustness checks* uitgevoerd om te kijken of de *convergence error* die door cito wordt 'veroorzaakt' zorgt voor incorrecte schattingen van de andere variabelen.

We hebben model 2 opnieuw geschat, maar dan zonder de variabelen cito, het jaar van de citotoets en de missing indicator van cito. Dit model convergeert zonder problemen en de coëfficiënten van de variabelen blijven nagenoeg gelijk. De *missing indicator method* die wordt gebruikt voor de missende waarden op cito, lijkt geen bias te veroorzaken in de schattingen van de relaties tussen de andere variabelen en de uitkomstmaten.

6. Gebruikte afkortingen

APCG	Armoedeprobleemcumulatiegebied
AR	Algemene Rekenkamer
BBL	Beroeps Opleidende Leerweg
BOL	Beroeps Begeleidende Leerweg
CBS	Centraal Bureau voor de Statistiek
CITO	Centraal Instituut voor Toets Ontwikkeling
DUO	Dienst Uitvoering Onderwijs
MBO	Middelbaar Beroepsonderwijs
OCW	Onderwijs Cultuur en Wetenschap
PO	Primair Onderwijs
SBB	Stichting Beroepsonderwijs en Bedrijfsleven
VO	Voortgezet Onderwijs
VSV	Voortijdig Schoolverlaten
VMBO	Voortgezet Middelbaar Beroepsonderwijs

7. Literatuur

CBS (zonder datum). Wat verstaat het CBS onder een allochtoon? Beschikbaar via: https://www.cbs.nl/nl-nl/faq/specifiek/wat-verstaat-het-cbs-onder-een-allochtoon-. Geraadpleegd op 04-01-2023.

CBS (2022). Schoolloopbanen van de tweede generatie. Beschikbaar via: https://longreads.cbs.nl/integratie-en-samenleven-2022/schoolloopbanen-van-de-tweede-generatie/ . Geraadpleegd op 29-03-2023.

DUO (2022). *Instructies data*. Interne documentatie

DUO (zonder datum). Rapportages: rapportages startpopulatie. Beschikbaar via: https://duo.nl/zakelijk/verzuim/verzuim/rapportages.jsp. Geraadpleegd op 04-02-2023.

Elffers, L. (2020). Eindtoets: gelijke kansen. Beschikbaar via: https://didactiefonline.nl/blog/louise-elffers/eindtoets-gelijke-kansen. Geraadpleegd op 29-03-2023.

Kang, H. (2013). *The prevention and handling of the missing data*. Korean journal of anesthesiology, 64(5), 402-406.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). *The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded*. Journal of the Royal Society Interface, 14(134), 20170213.

Nakagawa, S., & Schielzeth, H. (2013). *A general and simple method for obtaining R2 from generalized linear mixed-effects models*. Methods in ecology and evolution, 4(2), 133-142.

OCW (2018). Minister van OCW, Brief dd. 28 augustus 2018 over de Aanpak van gelijke kansen en stagediscriminatie in het mbo, Tweede Kamer, vergaderjaar 2017-2018, 31 524, nr. 374.

Parraillon, M.C. (2019). Interpreting Model Estimates: Marginal Effects. Beschikbaar via: https://clas.ucdenver.edu/marcelo-perraillon/sites/default/files/attached-files/perraillon_marginal_effects_lecture_lisbon.pdf. Geraadpleegd op 12-12-2022.

SCP (2014). De achterstand van autochtone doelgroepleerlingen. Beschikbaar via: https://repository.ubn.ru.nl/bitstream/handle/2066/211549/rapport-r2018.pdf. Geraadpleegd op 29-03-2023

SER (2021). Gelijke kansen in het onderwijs – structureel investeren in kansengelijkheid voor iedereen. Beschikbaar via: https://www.ser.nl/-/media/ser/downloads/adviezen/2021/gelijke-kansen-in-onderwijs.pdf. Geraadpleegd op 29-03-2023.

Timmermans et al. (2018). Track recommendation bias: Gender, migration background and SES bias over a 20-year period in the Dutch context. Beschikbaar via: https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1002/berj.3470. Geraadpleegd op 12-12-2022.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *mice: Multivariate imputation by chained equations in R*. Journal of statistical software, 45, 1-67.

Van Leest et al. (2020). Eindtoets en gelijke kansen. Een vergelijking voor en na. Beschikbaar via: https://didactiefonline.nl/artikel/eindtoets-en-gelijke-kansen-een-vergelijking-voor-en-na. Geraadpleegd op 29-03-2023.

Weel, ter B., Bussink, H. en Koeman, N. (2023). Kansenongelijkheid in Nederland. SEO *economisch onderzoek*.

Meer informatie

Op **www.rekenkamer.nl** vindt u informatie over de strategie, het werkprogramma en alle gepubliceerde onderzoeken van de Algemene Rekenkamer. Voor meer en/of andere inlichtingen kunt u zich wenden tot voorlichting@rekenkamer.nl.

** Aan de inhoud van deze factsheet kunnen geen rechten worden ontleend.*