



Paper

Herziening gewichtsenregeling primair onderwijs-Fase I

Hanneke Posthumus

Bart Bakker

Jan van der Laan

Martine de Mooij

Sander Scholtus

Mersiha Tepic

Jeroen van den Tillaart

Nander de Vette

Oktober 2016

Inhoudsopgave

Samenvatting	5
1. Inleiding	7
Deel I: het gebruik van onderwijsregistraties	9
2. Onderwijsregistraties en imputatie	10
2.1 Achtergrond	10
2.2 Imputatiemethoden	12
2.3 Vergelijking imputatiemethoden	13
2.4 Simulatiestudie	15
Deel II: modelleren van onderwijsachterstanden	21
3. Effecten op de schoolprestatie	22
3.1 Inleiding	22
3.2 Schoolprestatie	23
3.3 Kindkenmerken	23
3.4 Gezinsinvloeden	26
3.5 Buurtkenmerken	28
3.6 Schoolkenmerken	28
4. Onderzoeksmethode	30
4.1 Inleiding	30
4.2 Data	30
4.3 Methode	35
5. Resultaten	39
5.1 Inleiding	39
5.2 Modellen met intelligentie	39
5.3 Modellen zonder intelligentie	42
6. Conclusies en vervolg	44
6.1 Conclusies	44
6.2 Vervolg	45
Bijlage 1. Gebruikte literatuur	46
Bijlage 2. Databronnen	50
2.1 COOL5-18	50
2.2 SSB	51
2.3 Leefbaarometer	53

Bijlage 3. Variabelen en operationalisering	54
Variabelen	54
Operationalisering	55
Bijlage 4. Logistische regressie-imputatie	60
Bijlage 5. Correctie van nscct-scores voor invloed sociale omgeving	62
Bijlage 6. Geschatte SEM-modellen met 50%-75%-correctie nscct	63
Bijlage 7. Fitmaten voor SEM-modellen en nscct-varianten	69
Bijlage 8. Stepwise SEM-modellen voor nscct-varianten	71
Bijlage 9. Modellen zonder intelligentie	75
Bijlage 10. Afgevalen records	81
Bijlage 11. Correlatiematrix variabelen in stepwise SEM-model	83
Bijlage 12. Begeleidingscommissie	85

Samenvatting

Nederland voert sinds de jaren zeventig beleid om de onderwijskansen voor kinderen uit achterstandsmilieus te vergroten. Het beleid is erop gericht onderwijsachterstanden onder kinderen (op de basisschool) ten gevolge van sociale, economische of culturele oorzaken zoveel mogelijk te voorkomen en om eenmaal opgelopen achterstanden te verkleinen. Een belangrijk onderdeel van het huidige onderwijsachterstandenbeleid vormt de zogenaamde 'gewichtenregeling'. Deze regeling bepaalt hoeveel geld een basisschool krijgt om onderwijsachterstanden weg te werken. De gewichtenregeling is momenteel volledig gebaseerd op het opleidingsniveau van de ouders van kinderen. Daarbij worden drie categorieën onderscheiden. Als beide ouders van leerlingen geen diploma in het voortgezet onderwijs bezitten, krijgen scholen 1,2 leerling extra gefinancierd. Als dat geldt voor één ouder en de andere ouder heeft maximaal vmbo, dan krijgen scholen 0,3 leerling extra gefinancierd. Alle overige leerlingen krijgen geen extra financiering vanuit het onderwijsachterstandenbeleid.

Het huidige beleid roept steeds meer vragen op: Is enkel het opleidingsniveau van ouders wel genoeg om onderwijsachterstanden in te schatten? Leidt het stijgende opleidingsniveau van ouders niet ten onrechte tot een krimpend budget voor onderwijsachterstanden? Bovendien ligt de administratie van de opleidingsniveaus van ouders bij de scholen zelf. Dat blijkt behoorlijk lastig en tijdrovend. Voor het Ministerie van Onderwijs, Cultuur en Wetenschap (OCW) is deze aanpak verder duur omdat de controle ervan erg arbeidsintensief is.

Om het onderwijsachterstandenbeleid te kunnen verbeteren, heeft OCW het Centraal Bureau voor de Statistiek (CBS) gevraagd een model te ontwikkelen dat de verwachte onderwijsachterstanden van leerlingen in het basisonderwijs beter berekent en dat minder arbeidsintensief is voor scholen en OCW. Met dit model worden onderwijsachterstanden zo goed mogelijk verklaard op basis van verschillende (combinaties van) kenmerken - van kinderen zelf, hun ouders, school en buurt - die in registraties bij CBS aanwezig zijn. In dit onderzoek spreken we van een onderwijsachterstand wanneer leerlingen door bepaalde omgevingskenmerken slechter presteren op school dan op basis van hun intelligentie kan worden verwacht. Dit wordt gemeten door de schoolprestaties van kinderen te corrigeren voor hun intelligentie. Als maat voor schoolprestatie worden de Cito-scores in groep 8 gebruikt.

Het onderzoek laat zien dat het opleidingsniveau van de ouders een belangrijk kenmerk is en blijft om onderwijsachterstanden te bepalen. Uit de CBS-registraties blijkt echter dat de informatie over het opleidingsniveau voor 40% van de ouders ontbreekt. Hier is rekening mee gehouden. Omdat bij de gewichtenregeling financiële middelen op schoolniveau verdeeld worden, is het niet belangrijk dat we voor iedere ouder het juiste opleidingsniveau weten, maar dat we per school de juiste samenstelling van de opleidingsniveaus van ouders kunnen berekenen. In het huidige onderzoek is het gelukt om deze samenstelling voor scholen met meer dan 40 leerlingen voldoende betrouwbaar en zonder noemenswaardige vertekening te berekenen. Dit betekent dat het niet langer nodig is om het opleidingsniveau van ouders door scholen zelf te laten invullen en vervolgens te controleren.

Verder blijkt uit het huidige onderzoek dat naast het opleidingsniveau van de ouders, er ook andere kenmerken zijn die invloed hebben op schoolprestaties van kinderen. Door behalve met het opleidingsniveau van de ouders eveneens rekening te houden met de herkomst en verblijfsduur van ouders, of zij in de schuldsanering zitten en het gemiddelde opleidingsniveau van moeders op school, kunnen we de verwachte onderwijsachterstand het beste bepalen. Als ook de intelligentie van leerlingen wordt meegenomen, waardoor we voor hun leerpotentie corrigeren, wordt meer dan 40 procent van de verschillen in Cito-scores in groep 8 verklaard.

Naast het beste passende model, heeft OCW verschillende andere modellen laten onderzoeken die soms eveneens een behoorlijke verklaringskracht blijken te hebben. Van deze modellen is het model met daarin herkomst en opleiding het meest geschikt om onderwijsprestaties te verklaren.

Om tot een breed gedragen en kwalitatief hoogwaardige indicator te komen, zijn CBS en OCW tijdens het gehele onderzoeksproces geadviseerd door een begeleidingscommissie van deskundigen uit de praktijk en wetenschap. Op basis van de uitkomsten van het onderzoek en de adviezen van de begeleidingscommissie, beslist het ministerie van OCW op basis van welke modellen zij de gewichtenregeling mogelijk willen herzien. CBS zal, in een vervolgonderzoek, deze modellen vertalen naar de verwachte onderwijsachterstanden per school. Daarna zal het Ministerie van OCW een keuze maken voor de invulling van het toekomstige onderwijsachterstandenbeleid.

1. Inleiding

Nederland voert sinds de jaren zeventig beleid om de onderwijskansen voor kinderen uit achterstandsmilieus te vergroten. Het beleid is erop gericht onderwijsachterstanden onder (basisschool)leerlingen ten gevolge van sociale, economische of culturele oorzaken zoveel mogelijk te voorkomen en om eenmaal opgelopen achterstanden te verminderen. Belangrijke onderdelen van het huidige onderwijsachterstandenbeleid vormen de zogenaamde 'gewichtenregeling' en 'impulsregeling'. Deze regelingen bepalen hoeveel geld een basisschool krijgt om onderwijsachterstanden weg te werken.

Het beleid is de afgelopen decennia een aantal keer van vorm veranderd. De indicator voor achterstanden was in eerste instantie het beroep van de vader. Het beleid was er toen vooral op gericht autochtone 'arbeiderskinderen' extra te stimuleren. Later werden ook de kenmerken 'opleidingsniveau van de ouders' en 'etnische herkomst' onderdeel van het beleid. De huidige gewichtenregeling is volledig gebaseerd op het opleidingsniveau van de ouders (Kloprugge en de Wit, 2015). Scholen die in een postcodegebied liggen waar het gemiddelde inkomensniveau laag is en de werkloosheid hoog (de zogenaamde impulsgebieden), krijgen daarnaast voor elke gewichtenleerling extra budget toegekend.

Het huidige beleid roept steeds vaker vragen op. Is enkel het opleidingsniveau van ouders wel de beste schatter van onderwijsachterstanden, of zijn er ook meer of andere kenmerken die een belangrijke rol spelen? Leidt het stijgende opleidingsniveau van ouders niet ten onrechte tot een krimp budget voor onderwijsachterstanden? En is de huidige regeling niet onnodig tijdrovend en duur in haar uitvoering? De administratie van de opleidingsniveaus van ouders en de vertaling naar gewichten ligt momenteel bij de scholen zelf. Dat blijkt behoorlijk lastig en tijdrovend voor ze. Ook voor het ministerie is het duur omdat de controle ervan voor scholen die veel gewichtenleerlingen hebben arbeidsintensief is. Daar komt bij dat uit eerder onderzoek van het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) blijkt dat circa 30 procent van de leerlinggewichten niet kan worden onderbouwd, omdat het gewicht niet correspondeert met informatie uit de ouderverklaring of omdat informatie ontbreekt die nodig is om het gewicht vast te stellen. Dit leidt ertoe dat het beschikbare budget niet altijd bij de juiste scholen terecht komt.¹

Om het onderwijsachterstandenbeleid te kunnen verbeteren, heeft OCW het Centraal Bureau voor de Statistiek (CBS) gevraagd om op basis van kenmerken die in registraties bij CBS aanwezig zijn een indicator samen te stellen die onderwijsachterstanden in het primair onderwijs beter voorspelt. De nieuwe indicator zal samengesteld worden op basis van (combinaties van) kenmerken die onderwijsachterstanden voorspellen. In deze rapportage geven we inzicht hoe goed verschillende (combinaties van) kenmerken dit doen. OCW zal op basis hiervan beslissen voor welke (combinaties van) kenmerken het in de volgende onderzoeksfase de verwachte onderwijsachterstanden per school(bestuur) wil laten schatten. Om tot een breed gedragen en kwalitatief hoogwaardige indicator te komen, worden CBS en OCW tijdens het gehele onderzoeksproces geadviseerd door een begeleidingscommissie van deskundigen uit de praktijk en wetenschap (zie bijlage 12 voor de deelnemerslijst).² Na afronding van het onderzoek beslist OCW over de invulling van het toekomstige onderwijsachterstandenbeleid.

Zoals gezegd, bespreken we in deze rapportage hoe goed verschillende (combinaties van) kenmerken uit CBS-registraties onderwijsachterstanden voorspellen. Terwijl de meeste CBS-registraties de hele bevolking beslaan, is de registratie van opleidingsniveaus onvolledig. Omdat het opleidingsniveau van ouders naar verwachting een belangrijk kenmerk bij het voorspellen

¹ Zie: https://www.duo.nl/zakelijk/images/controle_gewichtenregeling_staatssecretaris.pdf.

² Daarnaast heeft CBS in de loop van het onderzoek een aantal experts geraadpleegd naar aanleiding van specifieke vragen. We bedanken prof. dr. Wilma Reesing (Universiteit Leiden), prof. dr. Peter Dekker (Vrije Universiteit) en dr. Theo van Batenburg (GION) voor hun bijdrage aan de discussie over de rol van gemeten intelligentie in het schoolloopbaanonderzoek en prof. dr. Wout Ultee (emeritus-hoogleraar Radboud Universiteit) voor het beschikbaar stellen van materiaal over taalafstanden en zijn toelichting daarop.

van onderwijsachterstanden, gaan we in het eerste deel (hoofdstuk 2) van deze rapportage in op de mogelijkheden om ontbrekende waarden te voorspellen en in te vullen (imputeren) en in welke mate registergegevens te gebruiken zijn om het opleidingsniveau van ouders van basisschoolleerlingen te bepalen. We bespreken hoe we dit doen en wat de kwaliteit van de imputaties is. Als de beschikbare onderwijsregistratie na imputatie bruikbaar blijkt, heeft dit als voordeel dat de kwaliteit van de gegevens waarschijnlijk verbetert en de administratieve lasten voor scholen en OCW verminderen.

In het tweede deel (hoofdstuk 3 t/m 5) van deze rapportage gaan we in op de samenhang tussen onderwijsachterstand en (combinaties van) omgevingskenmerken. In dit onderzoek spreken we, in navolging van Kloprugge en De Wit (2015), van onderwijsachterstand wanneer leerlingen door ongunstige omgevingskenmerken slechter presteren op school dan op basis van hun potentie (intelligentie) kan worden verwacht. Om slecht presterende van onderpresterende kinderen te kunnen onderscheiden, houden we bij het schatten van de onderwijsprestaties van leerlingen rekening met een meting van hun intelligentie.

Om relevante omgevingskenmerken te identificeren, maken we eerst op grond van eerder empirisch onderzoek een brede inventarisatie van de omgevingskenmerken die mogelijk samenhangen met de schoolprestaties (hoofdstuk 3). Bij de omgevingskenmerken wordt een onderscheid gemaakt tussen kenmerken van het gezin (zoals opleidingsniveau van de ouders, inkomen en herkomst), van de buurt waar het kind woont (zoals stedelijkheid en armoede) en van de school (zoals een clustering van leerlingen uit zwakkere milieus). Deze brede selectie aan achtergrondkenmerken zal de basis vormen voor een te ontwikkelen analysemodel dat de schoolprestaties van leerlingen op schoolniveau zo goed mogelijk schat. De opbouw van dit model wordt in hoofdstuk 4 beschreven. De uitkomsten uit dit model zullen in hoofdstuk 5 vergeleken worden met een aantal door OCW voorgestelde modellen. Bepaald zal worden hoe goed de verschillende modellen bij de data passen en hoe goed schoolprestaties ermee worden verklaard. Tot slot volgt in hoofdstuk 6 de conclusie en gaan we kort in op de vervolgstappen van dit onderzoek.

Deel I: het gebruik van onderwijsregistraties

OCW heeft aan CBS gevraagd te onderzoeken welke kenmerken van invloed zijn op schoolprestaties om een betere indicator van onderwijsachterstanden samen te kunnen stellen. Van het opleidingsniveau van de ouders wordt een grote invloed verwacht. Voor de huidige gewichtenregeling wordt het opleidingsniveau van ouders geadmistreerd door de scholen zelf via ouderverklaringen. Uit eerdere analyses van OCW³ blijkt dat de informatie uit de ouderverklaringen relatief vaak onvolledig is ingevuld of niet correspondeert met het opleidingsniveau van de ouders zoals dat geregistreerd staat, wat ertoe leidt dat het beschikbare budget niet altijd bij de juiste scholen terecht komt. Ook zijn de controlewerkzaamheden tijd- en kostenintensief. Hierom heeft OCW gevraagd te onderzoeken of er in de toekomst gebruik gemaakt kan worden van de onderwijsregistraties zoals die in het Stelsel van Sociaal-statistische Bestanden (SSB) van CBS aanwezig zijn.⁴ Deze registratie is echter incompleet voor een substantieel en selectief deel van de Nederlandse bevolking. We hebben daarom een aanpak ontwikkeld om met deze ontbrekende opleidingsgegevens om te gaan. Hoe deze aanpak eruit ziet wordt in dit deel van het rapport uiteengezet. We zullen laten zien dat de resulterende schattingen een valide en betrouwbaar beeld geven van het aantal risicoleerlingen per school.

³ Zie: https://www.duo.nl/zakelijk/images/controle_gewichtenregeling_staatssecretaris.pdf.

⁴ Voor meer informatie over het SSB zie bijlage 2.

2. Onderwijsregistraties en imputatie

2.1 Achtergrond

De opleidingsinformatie in het SSB is afkomstig uit het Opleidingsniveaubestand van CBS.⁵ Dit bestand wordt jaarlijks afgeleid uit een longitudinale database van beschikbare informatie over opleidingen uit diverse opleidingsregisters en alle jaargangen van de Enquête Beroepsbevolking (EBB) sinds 1996, het zogenaamde Opleidingsarchief. Dit archief bevat opleidingsinformatie op een zeer gedetailleerd niveau in de vorm van opleidingsnummers. Voor publicatiedoeleinden kunnen hieruit verschillende indelingen van opleidingsniveau worden afgeleid.

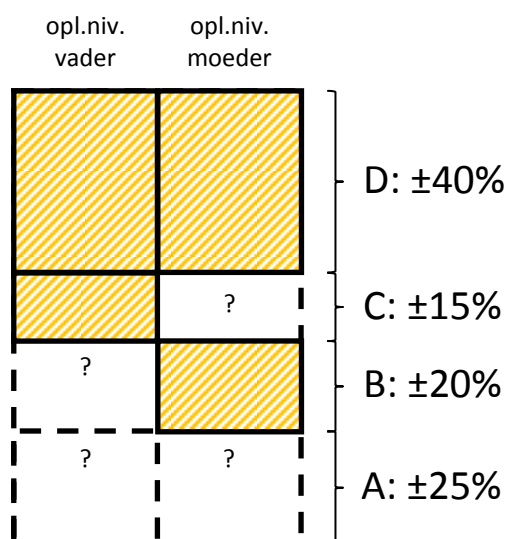
Het Opleidingsniveaubestand bevat het hoogst behaalde en hoogst gevolgde opleidingsniveau op een zeker peilmoment (per jaar de laatste vrijdag van september) voor alle personen uit de Nederlandse bevolking voor wie valide informatie aanwezig is in het Opleidingsarchief. Hierbij wordt alleen informatie meegenomen die, volgens bepaalde beslisregels, geldig is verklaard op het peilmoment. Voor personen van 15 jaar of ouder is het hoogste opleidingsniveau ofwel afkomstig uit een register, ofwel uit een EBB-jaargang. Voor personen die niet voorkomen in een register en die ook niet hebben meegedaan aan een EBB is geen opleidingsniveau bekend.

De aanwezige opleidingsniveaus zijn selectief naar diverse achtergrondkenmerken, zoals leeftijd en herkomst, en ook naar opleidingsniveau zelf. De selectiviteit naar leeftijd komt doordat centrale opleidingsregisters in Nederland pas recent worden bijgehouden (zie het einde van deze alinea). Jongeren zijn daardoor oververtegenwoordigd in het Opleidingsniveaubestand. Oudere personen die hun opleiding hadden afgerond voordat de relevante registers beschikbaar waren kunnen alleen voorkomen in het bestand als ze respondent zijn geweest in een EBB-jaargang sinds 1996. De gegevens zijn selectief naar herkomst omdat in het buitenland behaalde diploma's niet in Nederlandse opleidingsregisters zijn vastgelegd. Bovendien zijn mensen met buitenlandse herkomst ondervertegenwoordigd in de EBB-gegevens: enerzijds is de non-respons bij deze groep relatief hoog, en anderzijds hebben mensen die later dan 1996 naar Nederland zijn gekomen niet bij alle hier gebruikte EBB-jaren een kans gehad om in de steekproef te worden getrokken, in tegenstelling tot mensen die al vóór 1996 in Nederland woonachtig waren. Selectiviteit naar opleidingsniveau treedt op doordat het hoger onderwijs als eerste is begonnen met registreren (in de jaren '80). Registers over voortgezet onderwijs kwamen beschikbaar vanaf eind jaren '90, over het mbo vanaf schooljaar 2004/'05 en een centraal register van het basisonderwijs is er pas sinds 2010. Personen met een hoog opleidingsniveau zijn daardoor oververtegenwoordigd.

In 2011 was voor ongeveer 44 procent van de totale Nederlandse bevolking geen opleidingsniveau bekend in het SSB. Dit percentage zal in de toekomst geleidelijk kleiner worden, omdat de opleidingsregisters voor de nieuwe generaties vrijwel volledig dekkend zijn. Het percentage ontbrekende opleidingen daalt echter niet helemaal tot nul, omdat bijvoorbeeld opleidingen uit het buitenland ook in de toekomst zullen worden gemist.

Voor de huidige toepassing zijn we geïnteresseerd in het opleidingsniveau van ouders van leerlingen uit het basisonderwijs. Ter illustratie van het probleem toont figuur 2.1 de informatie die beschikbaar is over opleiding voor deze deelpopulatie voor het leerjaar 2013/'14, op basis van het Opleidingsniveaubestand met peildatum 27 september 2013. Voor ongeveer 40 procent van de leerlingen is het opleidingsniveau van beide ouders bekend in het SSB (deelpopulatie D). Voor ongeveer 35 procent is het opleidingsniveau van één ouder bekend (deelpopulaties C en B). Voor de resterende 25 procent is het voor beide ouders onbekend (deelpopulatie A).

⁵ Zie Linder, Van Roon en Bakker (2011) voor een uitgebreide beschrijving van de huidige werkwijze voor het maken van het Opleidingsniveaubestand.



Figuur 2.1: Ontbrekende opleidingsniveaus voor ouders van basisschoolleerlingen (leerjaar 2013/'14).

Om een goede weergave te krijgen van de verdeling van het opleidingsniveau voor alle ouders van basisschoolleerlingen, moet een bijschatting worden gemaakt voor de ontbrekende waarden in figuur 2.1. Hiervoor is vanuit het SSB veel hulpinformatie beschikbaar. Twee mogelijke aanpakken zijn: wegen en imputeren. Bij wegen wordt alleen gewerkt met de leerlingen voor wie het opleidingsniveau van beide ouders bekend is. Deze krijgen dan een gewicht zodat ze ook de overige leerlingen representeren. Bij imputatie worden geschatte waarden ingevuld voor de ontbrekende opleidingsniveaus. In dit project is gekozen voor imputatie, omdat dit beter past bij het beoogde gebruik van de data binnen een statistisch model. Bovendien kunnen we bij het imputeren van de records uit deelpopulaties B en C eenvoudig het opleidingsniveau van de ene ouder (dat wel bekend is) meenemen om een betere voorspelling te maken voor het opleidingsniveau van de andere ouder. Bij wegen is dat lastiger.

Uit eerder onderzoek naar de imputatie van ontbrekende waarden in het Opleidingsniveaubestand bleek dat de voorspelbaarheid van individuele opleidingsniveaus beperkt is (Scholtus en Pannekoek, 2015). Dat wil zeggen: op individueel niveau zal de geïmputeerde waarde vaak afwijken van het (onbekende) werkelijke opleidingsniveau. Een eerste analyse op de data van de ouders van basisschoolleerlingen bevestigde dit resultaat. Omdat onderwijsachterstandsgelden op schoolniveau verdeeld worden, is het echter voldoende als we erin slagen de opleidingsniveaus *per school* goed te voorspellen. De kans op juiste voorspellingen is groter op schoolniveau dan op individueel niveau. Als de imputatiemethode goed werkt, zullen de fouten die op individueel niveau gemaakt worden namelijk grotendeels tegen elkaar wegvallen wanneer we aggregeren naar schoolniveau. Bovendien hoeft het opleidingsniveau per school slechts voor een deel van de leerlingen te worden geïmputeerd. We verwachten daarom betere resultaten naarmate een school groter is en naarmate het opleidingsniveau voor meer leerlingen bekend is.

Tabel 2.1 toont voor scholen van verschillende grootte de percentages leerlingen waarvan het opleidingsniveau van beide ouders bekend is, op basis van de informatie uit het SSB voor leerjaar 2013/'14. Ter illustratie: uit de tabel blijkt bijvoorbeeld dat er volgens het SSB in 2013/'14 164 scholen waren met ten minste 20 en minder dan 40 leerlingen, en dat bij 12 procent van deze scholen het opleidingsniveau van beide ouders bekend was voor minder dan 20 procent van de leerlingen.

Het effect van de imputatie op schoolniveau zal groter zijn voor de scholen die linksboven in de tabel vallen en kleiner voor de scholen rechtsonder. Hierbij moet ook bedacht worden dat eventuele onzekerheid en/of vertekening in het geïmputeerde opleidingsniveau in dit

onderzoek alleen van belang is voor zover deze tot uitdrukking komen in de resultaten van het (in de komende onderzoeksfase te ontwikkelen) voorspel- en verdeelmodel. Het voorspelmodel, dat op basis van gezins- en omgevingskenmerken een schatting van de onderwijsachterstand van leerlingen maakt, zal naast opleiding waarschijnlijk ook nog andere variabelen bevatten. Het verdeelmodel zal gebruikt worden om uiteindelijk onderwijsachterstanden per school vast te stellen.

Tabel 2.1: Scholen in Nederland volgens het SSB (2013/'14) uitgesplitst naar grootte en percentages leerlingen met bekende waarden voor opleidingsniveau beide ouders. De percentages tellen per rij op tot 100 procent (op afrondingsverschillen na).

Aantal leerlingen	Aantal	% leerlingen met opleidingsniveau beide ouders bekend			
		[0,20%)	[20,30%)	[30%,40%)	[40,100%]
<20	18	22%	39%	6%	33%
20 – 39	164	12%	33%	29%	26%
40 – 99	1 278	10%	32%	36%	22%
100 – 199	2 422	4%	26%	43%	27%
≥200	3 059	2%	17%	38%	43%
Totaal	6 941	4%	23%	39%	33%

2.2 Imputatiemethoden

Om de ontbrekende opleidingsniveaus in figuur 2.1 te imputeren, kunnen we gebruikmaken van de verdeling van de waargenomen opleidingsniveaus in deelpopulatie D. In deze context wordt deelpopulatie D de *donorpool* genoemd. Records in deelpopulatie D zijn potentiële *donoren*. De overige records waar opleidingsniveaus moeten worden geïmputeerd heten *receptoren*. Verder is het belangrijk om op te merken dat opleidingsniveau een ordinale categoriale variabele is. In het vervolg noteren we opleidingsniveau als y en nummeren we de mogelijke categorieën van opleidingsniveau als $1, \dots, K$ (van laag naar hoog).

Een eenvoudige imputatiemethode voor categoriale variabelen is *random hot deck* donorimputatie. Hierbij wordt de populatie verdeeld in *strata* op basis van een of meer achtergrondvariabelen \mathbf{x} die zowel voor de donoren als de receptoren bekend zijn. Een stratum bestaat uit personen met dezelfde combinatie van scores op de achtergrondvariabelen. Voor elke receptor wordt een willekeurige donor uit hetzelfde stratum geselecteerd. De score van deze donor op opleidingsniveau wordt gebruikt als imputatie voor de receptor. Merk op: deze methode vereist dat de donor en receptor exact overeenkomen op alle achtergrondvariabelen. Er kan daarom slechts een beperkt aantal achtergrondkenmerken worden meegenomen, omdat elk stratum minimaal één donor (en bij voorkeur veel donoren) moet bevatten.

Iets abstracter beschouwd worden de imputaties bij *random hot deck* donorimputatie verkregen als een trekking uit een conditionele verdeling. Noteer de kans dat een willekeurig gekozen persoon met achtergrondkenmerken \mathbf{x} opleidingsniveau k heeft als $P(y = k | \mathbf{x})$. Deze conditionele kansen worden geschat voor elke mogelijke combinatie van scores op de achtergrondvariabelen \mathbf{x} , op basis van de empirische verdeling uit de donorpool. Vervolgens wordt voor een receptor met achtergrondkenmerken \mathbf{x}_r een opleidingsniveau geïmputeerd door één score te trekken uit $\{1, \dots, K\}$, waarbij de kans op categorie k gelijk is aan $P(y = k | \mathbf{x} = \mathbf{x}_r)$. Bij *random hot deck*-imputatie worden alle interacties tussen de achtergrondvariabelen automatisch meegenomen. Een interessant alternatief is om de

conditionele verdeling $P(y = k | \mathbf{x})$ te schatten via een zuiniger model, waarin bijvoorbeeld alleen de hoofdeffecten van de achtergrondvariabelen worden meegenomen. Op deze manier kan men meer hulpinformatie verwerken in de imputatiemethode, omdat de receptor niet meer exact wordt gekoppeld aan een bestaande donor met dezelfde scores op alle \mathbf{x} -variabelen.

In dit project is ervoor gekozen om $P(y = k | \mathbf{x})$ te modelleren via logistische regressie. Om rekening te houden met het ordinale karakter van opleidingsniveau is gebruikgemaakt van het zogenaamde *continuation-ratio-model* uit Agresti (1990). Dit model wordt kort toegelicht in bijlage 4. De praktische uitwerking is min of meer hetzelfde als bij de *random hot deck*-methode. De logistische regressiemodellen worden geschat op basis van de donorpool en toegepast om voor elke receptor de kansen $P(y = k | \mathbf{x} = \mathbf{x}_r)$ te voorspellen, voor $k = 1, \dots, K$. Vervolgens wordt voor elke receptor willekeurig één score uit $\{1, \dots, K\}$ getrokken uit een verdeling met deze kansen.

In deze toepassing bestaat de verzameling receptoren uit deelpopulaties A, B en C. Het imputatiemodel kan per deelpopulatie verschillen. Voor het imputeren van de ontbrekende opleidingsniveau van de moeder in deelpopulatie C kunnen we het bekende opleidingsniveau van de vader als hulpvariabele meenemen, en gezien de samenhang tussen deze twee variabelen is dat een goed idee (Scholtus en Pannekoek, 2015). Voor deelpopulatie B geldt, analoog, dat we het bekende opleidingsniveau van de moeder als hulpvariabele mee kunnen nemen bij het imputeren van het ontbrekende opleidingsniveau van de vader. In deelpopulatie A worden beide opleidingsniveaus geïmputeerd. Om de samenhang tussen de twee opleidingsniveaus te bewaren worden ze sequentieel geïmputeerd, waarbij de als eerste geïmputeerde variabele wordt meegenomen in het imputatiemodel voor de tweede variabele (Verboon, 1998).

2.3 Vergelijking imputatiemethoden

De keuze van een goed imputatiemodel voor opleidingsniveau hangt in belangrijke mate af van de toepassing waarvoor de imputaties gebruikt worden. In deze paragraaf gaan we in op de zoektocht naar een geschikt imputatiemodel voor de concrete toepassing: het schatten van het analysemodel voor onderwijsprestaties.

Hoewel het analysemodel wordt geschat op basis van gegevens uit het CohortOnderzoek OnderwijsLoopbanen van 5 tot 18 jaar (hierna: COOL⁵⁻¹⁸; zie Driessen et al., 2009, hoofdstuk 4 en bijlage 2), zijn de imputaties in eerste instantie afgeleid voor de gecombineerde register- en steekproefgegevens uit het SSB. Dit betreft een veel grotere populatie dan de COOL⁵⁻¹⁸-populatie: het gaat om 300 766 records van leerlingen die in 2007/08, 2010/11 en 2013/14 (de COOL⁵⁻¹⁸-cohortjaren) in groep 8 zaten en waarvan de Cito-scores Eindtoets Basisonderwijs (hierna: Cito-scores) bekend zijn. Door te werken met deze volledige verzameling leerlingen in de SSB-data zijn veel meer donoren beschikbaar voor het imputeren dan op basis van alleen de COOL⁵⁻¹⁸-data. De geïmputeerde opleidingsvariabelen worden vervolgens gekoppeld aan de COOL⁵⁻¹⁸-data.

We hebben verschillende imputatiemethoden en -modellen getest. Allereerst is een selectie gemaakt van (combinaties van) variabelen die het meest in aanmerking kwamen om in de imputatiemodellen op te nemen. In deze eerste selectie van variabelen hebben we gebruik gemaakt van variabelen uit het SSB en uit de COOL⁵⁻¹⁸-studie. Voor een groot aantal (combinaties van) hulpvariabelen is vervolgens nagegaan wat hun voorspellende waarde voor opleidingsniveau is. Hiertoe is gekeken naar de volgende associatiematen: Cramér's V, het percentage verwachte correcte imputaties bij stratificatie naar deze hulpvariabelen en het percentage imputaties dat naar verwachting maximaal één categorie naast het werkelijke opleidingsniveau van de ouder ligt.

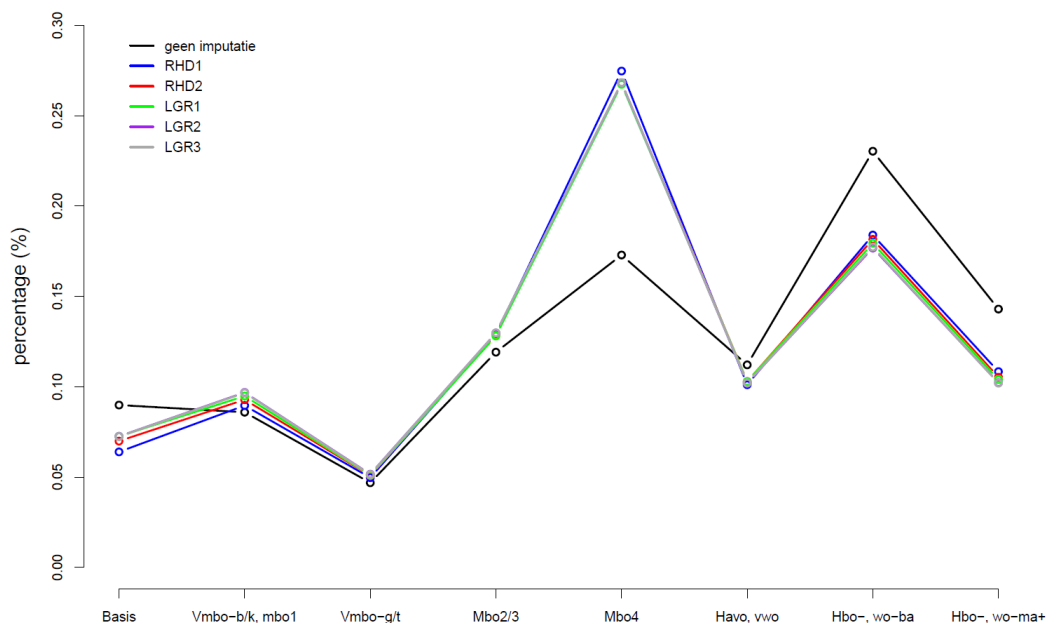
Uit de resultaten blijkt dat binnen de groep van modellen met één stratificatievariabele de variabele 'opleidingsniveau van de partner' naar verwachting de beste voorspeller voor het opleidingsniveau is: bij de moeders is Cramér's V gelijk aan 0,24 en het verwachte percentage

correcte imputaties is 23,8 procent, tegenover 19,7 procent bij imputatie zonder hulpvariabelen. Andere variabelen die het relatief goed doen zijn herkomst, persoonlijk bruto inkomen, leeftijd, stedelijkheid en burgerlijke staat. Een gezamenlijke stratificatie naar opleidingsniveau van de partner, herkomst, persoonlijk bruto inkomen, leeftijd en stedelijkheid geeft bij de moeders: Cramér's $V = 0,50$ en een verwacht percentage correcte imputaties van 37,4%. Bij de vaders werden vergelijkbare resultaten gevonden.

Op basis van deze tussenresultaten is gekozen een vijftal modellen verder uit te werken en onderling te vergelijken. Dit zijn de modellen die volgens Cramér's V en het verwachte percentage correcte imputaties binnen de donorpool het sterkst samenhangen met het opleidingsniveau. Voor de meest eenvoudige modellen met een of twee stratificatievariabelen in het model is random hot deck-imputatie (RHD) gebruikt. Voor de modellen met meer stratificatievariabelen is logistische regressie-imputatie (LGR) gebruikt:

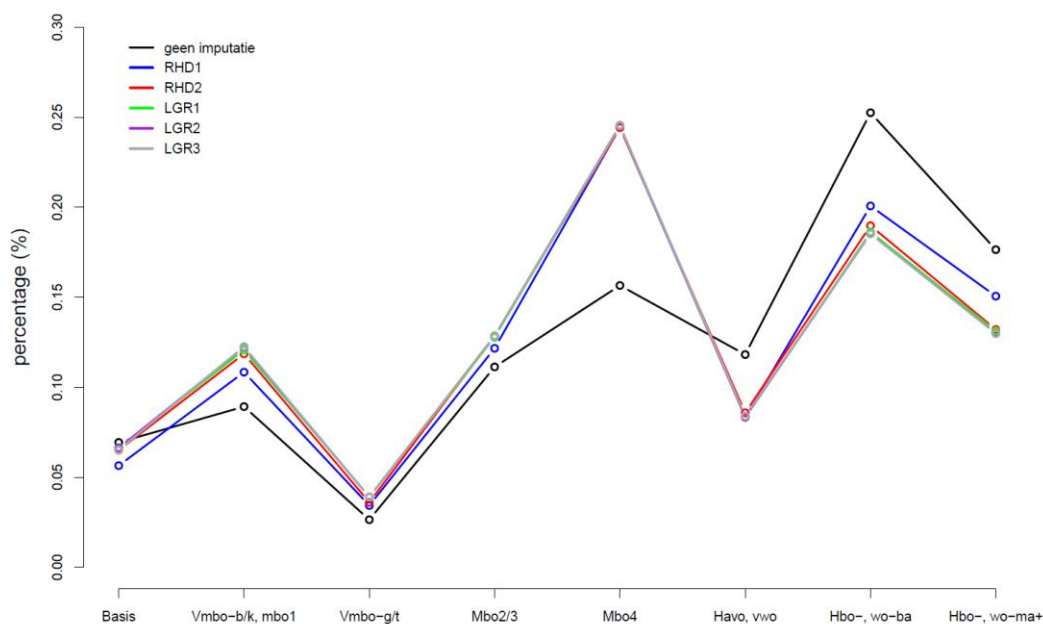
- random hot deck-imputatie met alleen het opleidingsniveau van de partner (indien bekend) als stratificatievariabele (RHD1);
- random hot deck-imputatie met persoonlijk bruto inkomen en het opleidingsniveau van de partner (indien bekend) als stratificatievariabelen (RHD2);
- logistische regressie-imputatie met stratificatie naar opleidingsniveau partner en verder als hulpvariabelen: persoonlijk bruto inkomen, herkomst, leeftijd, stedelijkheid, burgerlijke staat (LGR1);
- model LGR1 plus de CITO-groep 8-score van het kind (LGR2);
- model LGR2 plus thuistaal en inkomstenbron (LGR3).

In figuur 2.2 representeert de zwarte lijn de verdeling van het opleidingsniveau van moeders op basis van de beschikbare gegevens in het Opleidingsniveaubestand. De overige lijnen geven de geschatte verdeling van het opleidingsniveau weer op basis van de verschillende imputatiemodellen. De imputatie heeft een duidelijk effect op de verdeling naar onderwijsniveau: het aandeel moeders met een mbo-diploma is aanzienlijk gestegen. Dit ging met name ten koste van het aandeel personen met een hoger onderwijsdiploma – de groep die oververtegenwoordigd is in het Opleidingsniveaubestand (zie paragraaf 2.1). Verder valt op dat er op dit geaggregeerd niveau weinig verschil te zien is tussen de verschillende imputatiemodellen.



Figuur 2.2: (Geschatte) verdeling onderwijsniveau van moeders in het onderzoeksbestand.

In figuur 2.3 staan dezelfde verdelingen voor de vaders. Ook hier is zichtbaar dat imputatie een effect heeft op de verdeling van onderwijsniveau. Wederom is het effect van de imputatie conform verwachting het grootst voor het aandeel personen met een mbo-diploma. Op model RHD1 na geven de imputatiemodellen op dit aggregatieniveau nagenoeg dezelfde uitkomsten. Terwijl bij de moeders geldt dat het toevoegen van het persoonlijk inkomen als hulpvariabele naast het opleidingsniveau van de partner (model RHD2 versus RHD1) tot vrijwel dezelfde geschatte opleidingsniveaus leidt, maakt dit bij de vaders wel enig verschil. Met oog op de uitkomsten van de andere modellen lijkt dit te betekenen dat het opleidingsniveau van de partner voor moeders een iets betere schatter is voor het eigen opleidingsniveau dan voor vaders.



Figuur 2.3: (geschatte) verdeling onderwijsniveau van vaders in het onderzoeksbestand.

We hebben de imputatiemodellen ook vergeleken voor deelpopulaties: moeders en vaders uitgesplitst naar herkomst- en inkomenscategorie (hier niet weergegeven). Ook uit deze modellen blijkt dat na toevoeging van geïmputeerde waarden de geschatte verdeling verandert. Hieruit blijkt nogmaals dat imputatie zinvol is. Uitgesplitst naar herkomst of inkomen zijn meer verschillen te zien tussen RHD en LGR, maar de LGR-methoden geven steeds ongeveer dezelfde verdeling. Dit geeft vertrouwen dat binnen de LGR-methoden de belangrijkste voorspellende variabelen voor opleidingsniveau opgenomen zijn.

Conclusie

Het model LGR3 geeft ons inziens het beste resultaat. In de figuren leveren LGR2 en LGR3 ongeveer dezelfde uitkomsten, maar het lijkt interessant om ook inkomstenbron op te nemen in het model. Dit vanuit de veronderstelling dat de relatie tussen inkomen en opleiding wellicht anders is voor werkenden dan voor niet-werkenden. Mutatis mutandis geldt dit ook voor thuistaal en het verband tussen herkomst en opleiding. De imputaties op basis van model LGR3 zijn gebruikt bij het schatten van de analysemodellen in hoofdstuk 5.

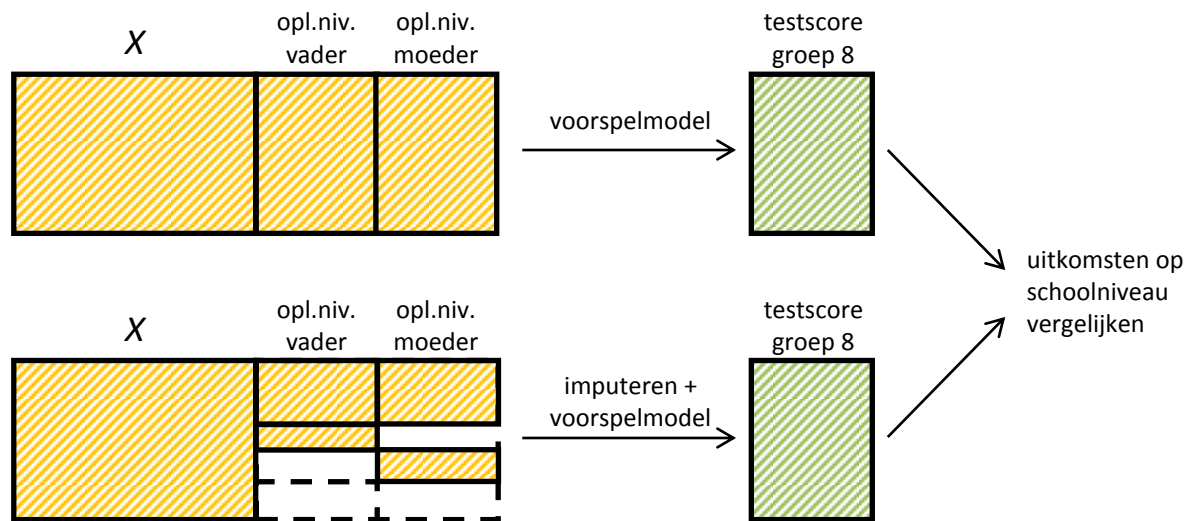
2.4 Simulatiestudie

Het is niet mogelijk om de kwaliteit van de geïmputeerde waarden rechtstreeks te evalueren, aangezien de bijbehorende werkelijke waarden onbekend zijn. Om toch iets te kunnen zeggen over de bruikbaarheid van de imputaties hebben we een simulatiestudie gedaan op een deel van de data uit COOL⁵⁻¹⁸, waarvoor de echte opleidingsniveaus bekend waren. We hebben de

imputatiemethode gevalideerd door ontbrekende waarden aan te brengen en deze te imputeren. De overeenkomsten en verschillen tussen geïmputeerde en werkelijke waarden geven een idee van de kwaliteit van de imputatie.

Het doel van de simulatiestudie is om dezelfde stappen te doorlopen als bij het imputeren van ontbrekende opleidingsniveaus ten behoeve van het schatten van het analysemodel (zoals beschreven in paragraaf 2.3), maar dan voor een bestand waarin alle werkelijke opleidingsniveaus bekend zijn. Verder zouden we idealiter willen onderzoeken wat het effect van de imputaties is op de uitkomsten van het verdeelmodel op schoolniveau. Dat laatste is echter mede afhankelijk van beslissingen die pas later zullen worden genomen, zoals de keuze van het definitieve voorspelmodel voor onderwijsachterstand en de keuze hoe de uitkomsten worden samengevat op schoolniveau in het definitieve verdeelmodel. Op dit moment kunnen we daarom alleen laten zien wat het effect van de imputaties is bij enkele mogelijke keuzes voor het voorspel- en verdeelmodel.

Figuur 2.4 geeft de opzet van de simulatiestudie schematisch weer. We hebben gebruikgemaakt van de enquêtedata van de COOL⁵⁻¹⁸-cohorten 2007/'08, 2010/'11 en 2013/'14. Uit deze cohorten hebben we alle leerlingen geselecteerd met een Cito-score in groep 8 en met een vanuit de ouderverklaringen bekend opleidingsniveau voor beide ouders. Dit zijn 16 212 records. De SSB-data zijn gekoppeld aan dit bestand. In het SSB zijn de opleidingsniveaus van sommige ouders onbekend, volgens het patroon uit figuur 2.1. Ten behoeve van de simulatiestudie hebben we dit patroon van ontbrekende waarden overgenomen in de COOL⁵⁻¹⁸-dataset. Dat wil zeggen: de COOL⁵⁻¹⁸-opleidingsniveaus van alle ouders die ontbreken in het Opleidingsniveaubestand van 2013 zijn vervangen door ontbrekende waarden (figuur 2.4 onderaan).



Figuur 2.4: Opzet validatiestudie.

Vervolgens zijn verschillende imputatiemodellen gebruikt om het gesimuleerde bestand te imputeren. We hebben hier gekozen voor dezelfde imputatiemodellen als in paragraaf 2.3. De donoren zijn hier, net als in paragraaf 2.3, afkomstig uit de SSB-data. Een belangrijk verschil is dat in de COOL⁵⁻¹⁸-data een andere indeling van opleidingsniveau wordt gebruikt dan in de SSB-data (zeven categorieën in plaats van acht). Ten behoeve van de simulatiestudie hebben we de opleidingsvariabelen uit het SSB daarom vertaald naar de indeling uit de COOL⁵⁻¹⁸-data.

We hebben gekozen voor de opzet uit figuur 2.4, omdat het op deze manier mogelijk is om de effecten van imputatie te onderzoeken in een realistische situatie, waarbij het gesimuleerde patroon van ontbrekende waarden zo goed mogelijk lijkt op het patroon dat in werkelijkheid optreedt in het SSB. Een nadeel van deze opzet is dat we moeten werken met de opleidingsvariabelen uit de COOL⁵⁻¹⁸-data die, zoals eerder vermeld, niet dezelfde indeling

hebben als de opleidingsvariabelen uit het SSB en waarschijnlijk ook minder betrouwbaar gemeten zijn. Bij wijze van alternatief hadden we een vergelijkbare simulatie uit kunnen voeren op het deel van de SSB-data waar beide opleidingsvariabelen bekend zijn (deelpopulatie D uit figuur 2.1). In dat geval hadden we zelf een realistisch patroon voor de ontbrekende waarden moeten bedenken. Ook hadden we dan moeten aannemen dat de resultaten van de simulatie niet gevoelig zijn voor selectiviteit van deelpopulatie D t.o.v. de rest van het bestand. Omdat met name deze laatste aanname ons niet geloofwaardig leek, hebben we gekozen voor een validatiestudie op COOL⁵⁻¹⁸-data.

Aangezien de werkelijke (door de ouders opgegeven) opleidingsniveaus in de COOL⁵⁻¹⁸-studie bekend zijn, kunnen we de kwaliteit van de geïmputeerde waarden rechtstreeks vaststellen. Zoals hiervoor is opgemerkt zijn we hier niet zo zeer geïnteresseerd in de kwaliteit van de afzonderlijke imputaties, maar meer in de wijze waarop de imputaties doorwerken in de uitkomsten van het voorspel- en verdeelmodel op schoolniveau. Ten behoeve van deze simulatiestudie zijn we uitgegaan van een van model dat later in hoofdstuk 4 en 5 wordt besproken: een structureel vergelijkingsmodel met alleen het opleidingsniveau van beide ouders, herkomst en intelligentie als verklarende variabelen voor de Cito-score in groep 8. Dit model is steeds geschat op basis van de data (inclusief imputaties) voor alle kinderen in het bestand uit figuur 2.4 voor wie een intelligentiemeting aanwezig is. Verder zijn in deze simulatiestudie alle operationalisering en correcties meegenomen die ook zijn gebruikt bij het schatten van de analysemodellen in het eigenlijke onderzoek (zie hoofdstuk 4, in het bijzonder paragraaf 4.2).⁶

Het geschatte analysemodel geeft in eerste instantie een voorspelling voor *onderwijsprestatie* in de vorm van de Cito-score in groep 8. Hieruit is een voorspelmodel voor de mate van *onderwijsachterstand* afgeleid door de bijdrage van intelligentie aan de voorspelde Cito-score te vervangen door de verwachte bijdrage voor een leerling met het gemiddelde intelligentieniveau.⁷ Op basis van dit voorspelmodel zijn, voor elke geïmputeerde dataset, voorspelde scores voor onderwijsachterstand afgeleid voor alle 16 212 leerlingen in deze simulatiestudie.

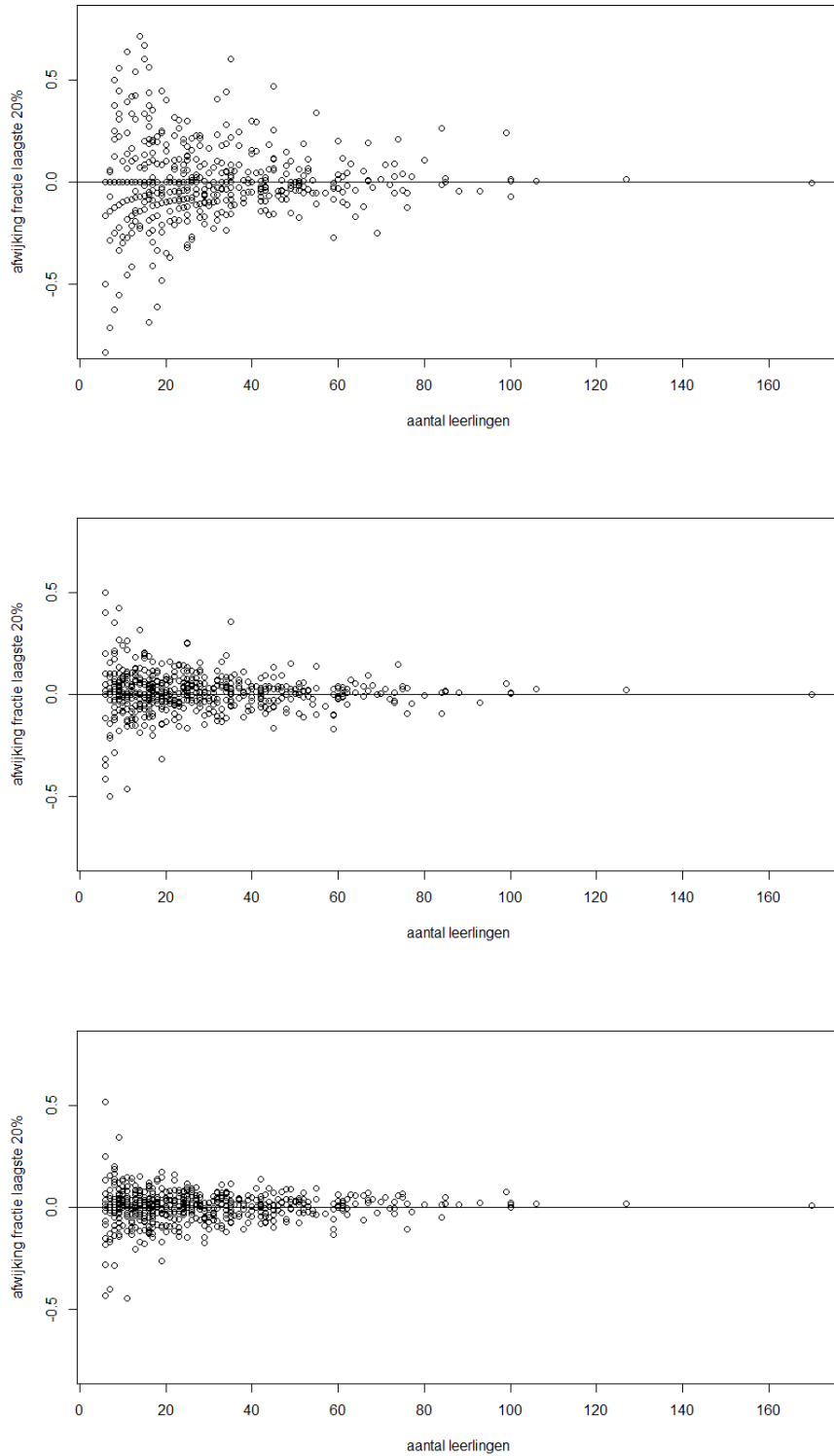
Deze scores kunnen op verschillende manieren worden samengevat op schoolniveau. Een mogelijke keuze is om een leerling mee te tellen als achterstandsleerling indien zijn voorspelde onderwijsachterstand behoort tot de laagste p% uit het bestand, met p gelijk aan 10, 20 of 30. Vervolgens kan dan per school het aantal achterstandsleerlingen worden geteld. Een andere mogelijkheid is om direct de gemiddelde voorspelde onderwijsachterstand-score per school te berekenen. De scholen met de laagste gemiddelde scores zijn dan de scholen waar de meeste onderwijsachterstanden voorkomen.⁸

Het analyse- en voorspelmodel is toegepast op de geïmputeerde versies van het bestand en op het oorspronkelijke, volledig gevulde bestand. De uitkomsten op basis van het volledig gevulde bestand vormen een benchmark voor de uitkomsten op basis van de geïmputeerde bestanden. Concreet hebben we de fractie achterstandsleerlingen per school geschat op basis van elke imputatiemethode (figuur 2.4 onderaan) en vergeleken met de bijbehorende schatting op basis van het oorspronkelijke, volledig gevulde bestand (figuur 2.4 bovenaan). Dat wil zeggen: per school is steeds het aantal voorspelde achterstandsleerlingen geteld en gedeeld door het totaal aantal leerlingen van die school in het bestand. Aangezien de geïmputeerde waarden stochastisch zijn, om stabielere uitkomsten te krijgen, bij elke imputatiemethode gemiddeld over 10 realisaties. Verder kijken we naar fracties in plaats van absolute aantallen zodat de uitkomsten vergelijkbaar zijn over scholen van verschillende grootte.

⁶ Voor de nscct-score (meting van intelligentie) is in deze simulatiestudie steeds de variant gebruikt die is gecorrigeerd voor 50% van het effect van opleiding en 75% van het effect van herkomst; zie paragraaf 4.2.4.

⁷ Dit is op dit moment hoe we van plan zijn voor de werkelijke data het analysemodel voor onderwijsprestatie om te zetten in een voorspelmodel voor onderwijsachterstand; zie ook hoofdstuk 3. De definitieve aanpak wordt echter pas uitgewerkt in de volgende fase van het project.

⁸ Zoals reeds is opgemerkt zal de definitieve aanpak voor het samenvatten van de uitkomsten op schoolniveau pas later in dit project worden uitgewerkt.



Figuur 2.5: Afwijkingen per school van de voorspelde fractie achterstandslerlingen (gedefinieerd als leerlingen met de laagste 20% voorspelde onderwijsachterstand-scores van Nederland) t.o.v. de benchmark in drie situaties: geen imputatie (boven), hot deck imputatie (midden) en logistische regressie-imputatie (onder). Elk punt stelt een school voor. Op de x-as staat het totaal aantal leerlingen per school. De resultaten in de onderste twee afbeeldingen zijn gemiddeld over tien imputatieronden.

Bij de hieronder te bespreken resultaten moet worden bedacht dat het COOL⁵⁻¹⁸-bestand alleen leerlingen bevat uit groep 2, 5 en 8; de scholen lijken in deze simulatiestudie daarom kleiner dan 'normaal' (vergelijk de x-as uit figuur 2.5 met tabel 2.1). Scholen met minder dan 5 leerlingen in het bestand zijn niet meegenomen in de onderstaande resultaten.

Figuur 2.5 toont spreidingsdiagrammen van de afwijkingen in de voorspelde fractie achterstandsleerlingen per school ten opzichte van de benchmark bij drie imputatiemethoden, wanneer het verdeelcriterium uitgaat van de leerlingen met de laagste 20 procent voorspelde testcores in het bestand.⁹

In het bovenste spreidingsdiagram in figuur 2.5 is niet geïmputeerd, maar zijn alleen de data van de leerlingen voor wie het opleidingsniveau van beide ouders bekend is meegenomen. Er is te zien dat dit op schoolniveau kan leiden tot relatief grote afwijkingen, met name bij de kleinste scholen. [N.B. Een afwijking van bijvoorbeeld +0.2 op de y-as betekent dat een werkelijke fractie achterstandsleerlingen van $f \times 100\%$ (met $0 \leq f \leq 1$) wordt geschat als $(f + 0.2) \times 100\%$.] Ook blijkt dat de fractie achterstandsleerlingen gemiddeld iets wordt onderschat. Dit laatste is conform verwachting, aangezien vooral ouders met lagere opleidingsniveaus ontbreken in het Opleidingsniveaubestand.

Het middelste spreidingsdiagram in figuur 2.5 toont de afwijkingen bij de meest uitgebreide *random hot deck*-imputatiemethode uit paragraaf 2.3 (RHD2): hierbij wordt alleen inkomen en waar mogelijk het opleidingsniveau van de partner gebruikt om strata te definiëren. Te zien is dat dit leidt tot een duidelijke verbetering. De afwijkingen ten opzichte van de benchmark zijn nu kleiner, maar met name voor de kleinere scholen komen nog steeds grote afwijkingen voor.

Het onderste spreidingsdiagram toont de afwijkingen bij het meest uitgebreide imputatiemodel uit paragraaf 2.3 (LGR3). De imputaties in dit model zijn gebaseerd op een logistisch regressiemodel. Deze aanpak leidt tot iets betere resultaten dan de *random hot deck*-methode. Voor scholen met meer dan 40 leerlingen zijn de uitkomsten zeer nauwkeurig en lijkt geen sprake van vertekening. Voor kleinere scholen is de nauwkeurigheid minder goed en bij scholen met 20 leerlingen of minder komen nog steeds grote afwijkingen voor. We vonden overigens vergelijkbare resultaten als de imputaties vaker dan 10 keer werden herhaald.

Een alternatieve aanpak is, zoals gezegd, om direct gemiddelde voorspelde scores voor onderwijsachterstand per school te berekenen. Ook hier kunnen we de uitkomsten bij verschillende imputatiemethoden vergelijken met de benchmark. Tabel 2.2 toont de correlatie tussen enerzijds de gemiddelde voorspelde score per school na imputatie (wederom gemiddeld over 10 imputatieronden) en anderzijds de gemiddelde voorspelde score per school op basis van het oorspronkelijke bestand met volledige data. Gezien de uitkomsten uit figuur 2.5 zijn alle scholen met 20 leerlingen of minder in het bestand buiten beschouwing gelaten. We hebben zowel de Pearson-correlatiecoëfficiënt berekend als de Spearman-rangcorrelatiecoëfficiënt. Te zien is dat imputatie een gunstig effect heeft op de correlaties. Voor de meest uitgebreide imputatiemethode (LGR3) zijn beide correlatiecoëfficiënten groter dan 0.97. In het bijzonder blijkt uit de hoge rangcorrelatiecoëfficiënt dat het ontbreken van onderwijsgegevens na imputatie met model LGR3 nog maar weinig effect heeft op een ordening van deze scholen naar gemiddelde voorspelde onderwijsachterstand-score (bijvoorbeeld van laag naar hoog).

⁹ Dit percentage is enigszins arbitrair. We hebben ook resultaten bij de laagste 10 procent en 30 procent berekend. Deze zijn vergelijkbaar. In een latere fase van het onderzoek wordt een definitief verdeelcriterium gekozen.

Tabel 2.2: Correlaties tussen de gemiddelde voorspelde onderwijsachterstand-score per school na imputatie en op basis van volledige data (benchmark). In de berekening zijn alleen scholen meegenomen met meer dan 20 leerlingen in het simulatiebestand; dit zijn 308 scholen.

Imputatiemethode	Correlatie tussen gemiddelde score per school	
	Pearson-correlatie	Spearman-rangcorrelatie
geen imputatie	0.899	0.864
RHD1	0.943	0.926
RHD2	0.968	0.957
LGR1	0.974	0.968
LGR2	0.983	0.966
LGR3	0.985	0.972

Conclusie

Op basis van deze simulatiestudie kan geconcludeerd worden dat imputatie van ontbrekende opleidingsniveaus op basis van een logistisch regressiemodel mogelijk is om voldoende nauwkeurige schattingen van onderwijsachterstanden te maken voor scholen met minimaal 40 leerlingen.¹⁰ Voor scholen met minder dan 20 leerlingen leidt imputatie tot onvoldoende nauwkeurige schattingen. Voor deze groep van kleinste scholen kunnen we uitwijken naar een analyse op schoolbestuurniveau in de gewichtenregeling om dit probleem te verhelpen. Zoals in tabel 2.1 is te zien betreft dit een zeer kleine minderheid van alle scholen in Nederland (ongeveer 20 van de 7 000 basisscholen in Nederland). Voor scholen met minimaal 20 maar minder dan 40 leerlingen gaan we er vooralsnog van uit dat voldoende nauwkeurige schattingen gemaakt kunnen worden.

Zoals is opgemerkt aan het eind van paragraaf 2.3 is in het vervolg van dit onderzoek gewerkt met geïmputeerde opleidingsniveaus op basis van imputatiemodel LGR3.

¹⁰ De aantallen leerlingen in deze alinea hebben betrekking op het totaal aantal leerlingen per school, dus niet alleen de leerlingen die zijn waargenomen in de COOL⁵⁻¹⁸-enquête.

Deel II: modelleren van onderwijsachterstanden

In dit tweede deel van de rapportage gaan we in op de samenhang van schoolprestatie en (combinaties van) achtergrondvariabelen. Eerst zal op grond van eerder empirisch onderzoek een brede inventarisatie gemaakt worden van achtergrondkenmerken van het kind en omgevingskenmerken die mogelijk samenhangen met de schoolprestatie (zie hoofdstuk 3). Bij de omgevingskenmerken wordt een onderscheid gemaakt tussen kenmerken van het gezin (zoals opleidingsniveau van de ouders, inkomen en herkomst), van de buurt waar het kind woont (zoals stedelijkheid en armoede) en van de school (zoals de clustering van leerlingen uit zwakkere milieus). Ook beschrijven we de rol van het achtergrondkenmerk intelligentie en de kanttekeningen die bij haar meting gezet moeten worden.

Deze brede selectie aan achtergrondkenmerken zal de basis vormen voor een te ontwikkelen analysemodel dat de schoolprestaties van leerlingen zo goed mogelijk schat (zie hoofdstuk 4). De resultaten uit dit model zullen in hoofdstuk 5 vergeleken worden met een aantal door OCW voorgestelde interessante, modellen. Bepaald zal worden hoe goed de verschillende modellen bij de data passen en hoe goed onderwijsachterstanden ermee worden geschat.

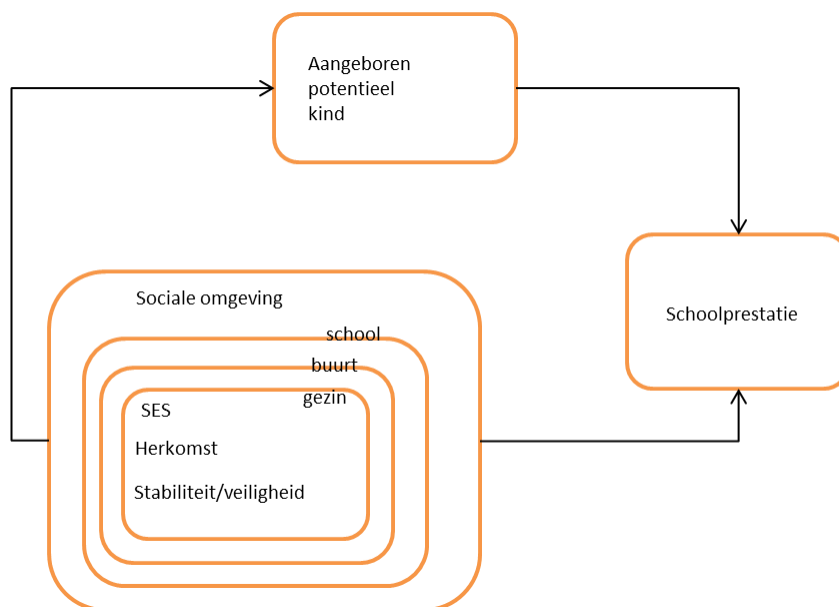
Op basis van deze rapportage kan OCW besluiten of en hoe zij de kenmerken die in de gewichtenregeling worden meegenomen wil aanpassen. Wanneer OCW haar voorkeur heeft uitgesproken, zal het CBS in de volgende fase van dit onderzoek op basis hiervan modellen ontwikkelen waarmee de onderwijsachterstanden per school vastgesteld kunnen worden. De uitkomsten van deze volgende fase van het onderzoek zullen naar verwachting begin 2017 gepubliceerd worden. Deze vervolgstappen worden besproken in het afsluitende hoofdstuk 6.

3. Effecten op de schoolprestatie

3.1 Inleiding

In dit hoofdstuk inventariseren we welke kind-, gezins-, school- en buurtkenmerken volgens eerder onderzoek mogelijk van invloed zijn op schoolprestaties. Daarbij gaan we vrij uitgebreid in op de invloed van de individuele potentie van leerlingen. Dit doen we omdat we, aansluitend op Kloprugge en de Wit (2015), menen dat er sprake is van onderwijsachterstand als leerlingen door een ongunstige economische, sociale of culturele omgeving (met name de thuissituatie) op school slechter presteren dan zij bij een gunstiger situatie zouden kunnen.

In figuur 3.1 geven we schematisch weer welke relaties we in dit hoofdstuk aan de hand van eerdere literatuur bespreken. Veel van de kenmerken in het figuur zijn onderling aan elkaar gerelateerd. Bijvoorbeeld, het sociaaleconomisch milieu van een gezin (SES) zal in bepaalde mate gerelateerd zijn aan de stabiliteit van het gezin. Het doel van dit onderzoek is echter niet om onderlinge relaties tussen al deze kenmerken te onderzoeken, maar juist om de relatie tussen deze kenmerken en schoolprestatie van basisschoolleerlingen te bekijken. Hier houden we in de uiteindelijke analyses rekening mee, door de kenmerken zo op te nemen dat hun invloed op schoolprestaties wordt gecorrigeerd voor de invloed van de andere kenmerken.



Figuur 3.1: Conceptueel model.

In het vervolg van dit hoofdstuk zullen we op basis van eerder wetenschappelijk onderzoek een voorselectie van kenmerken maken, die naar verwachting de meest invloedrijke voorspellers zijn van schoolprestaties bij kinderen. We hebben daarbij niet de pretentie een volledig overzicht te geven. Hierna gaan we allereerst in op het concept schoolprestatie en de maat die we daarvoor gekozen hebben – de Cito-eindtoets. Dit is de afhankelijke variabele in het onderzoek. Vervolgens gaan we in op een aantal kindkenmerken die van invloed zijn op schoolprestatie en onderwijsachterstand. Hierbij gaan we ook in op het ‘gemeten potentieel’ van het kind en de omgevingsinvloeden die een effect hebben op dit specifieke kenmerk. Vanwege deze invloeden, zullen we de gemeten intelligentie later ‘corrigeren’ (zie hoofdstuk 4). Vervolgens wordt de invloed van omgevingskenmerken (gezin, buurt, school) op onderwijsachterstand en schoolprestatie besproken. Per type achtergrondkenmerk is steeds kort aangegeven welke variabelen daarover beschikbaar en bruikbaar zijn voor onze analyses.

Meer informatie over de bronnen waaruit deze variabelen komen, geven we in hoofdstuk 4 en bijlage 2. De operationalisering en lichten we uitgebreider toe in hoofdstuk 4 en bijlage 3.

3.2 Schoolprestatie

Schoolprestaties van kinderen worden vaak gemeten aan de hand van afgenomen tests. Deze bestaan uit een combinatie van taken die cognitieve capaciteiten en op school opgedane kennis meten (Rindermann, 2007). In het Nederlandse onderwijs worden verschillende toetsen afgenomen om de prestaties en vorderingen van leerlingen in kaart te brengen. Op de meeste basisscholen in Nederland wordt gewerkt met het leerlingvolgsysteem van het Cito. Dit volgsysteem van Cito bestaat uit de zogenaamde LVS-toetsen (leerlingvolgsysteem-toetsen), de entreetoetsen (groep 5 t/m 7) en de eindtoets (groep 8). De toetsen uit het volgsysteem worden op de basisschool ook wel de Cito-toetsen genoemd.

Vanwege invloeden van de school en het eventuele beleid op school op prestatie is onderzocht of het mogelijk is om een prestatie maat te hanteren die zo vroeg mogelijk bij het kind gemeten is, namelijk de Cito-toets die in groep 2 wordt afgenomen. De betrouwbaarheid van deze toets bleek echter onvoldoende. Zo bleek er een zeer sterke clustering van toetsscores op schoolniveau te bestaan. Dit duidt op een sterk testafname-effect: de gemeten scores zijn niet alleen het resultaat van de prestaties van een leerling, maar ook in belangrijke mate van degene die de test afneemt. Hierom is uiteindelijk gekozen om de veel betrouwbaarder eindtoets (Cito-toets groep 8) te gebruiken. Via Cito beschikt CBS over deze gegevens.

3.3 Kindkenmerken

Allereerst worden schoolprestaties beïnvloed door kenmerken van het kind zelf. Een belangrijk kindkenmerk dat van invloed is op schoolprestaties is (leeftijdsgelateerde) potentie (Van Grastek en Lems, 2014; Resing en Blok, 2002). Deze potentie, die meestal gemeten wordt met intelligentietesten, is in dit onderzoek een belangrijk kenmerk omdat onderwijsachterstand hiermee theoretisch gezien van schoolprestatie kan worden onderscheiden. In paragraaf 3.3.1 gaan we hier dan ook dieper op in.

Daarnaast speelt het sociaal-emotioneel functioneren van kinderen een rol. Hoewel er geen consensus bestaat over het verband tussen het sociaal-emotioneel functioneren van kinderen en hun schoolprestaties (Veen et al., 2013), wordt door verschillende onderzoekers (Van Hoorn et al., 2005) onderschreven dat voor een succesvolle schoolloopbaan niet alleen cognitieve capaciteiten, maar ook sociaalemotionele vaardigheden van belang zijn. Andersom kunnen schoolprestaties van invloed zijn op de sociaalemotionele gesteldheid. Voor het vaststellen van de sociaalemotionele toestand van een kind bestaan testen, zoals de Nederlandse Differentiatie Test Sociaal Emotioneel Functioneren (NDT-SEF). Helaas beschikt het CBS niet over deze testgegevens. Wel is onderzocht of het mogelijk en zinvol is om gerelateerde variabelen zoals het gebruik van eerste- en tweedelijns GGZ en het al dan niet aangemerkt zijn als zorgleerling als indicatoren voor de sociaalemotionele toestand van een kind te gebruiken.

3.3.1 Intelligentie en invloeden op de meting ervan

Resing en Drenth (2007) definiëren intelligentie als 'het geheel van cognitieve of verstandelijke vermogens dat nodig is om kennis te verwerven en daar op een goede wijze gebruik van te maken, teneinde problemen op te lossen die een vast omschreven doel en structuur hebben.' Intelligentie wordt gemeten met behulp van een intelligentietest en doorgaans uitgedrukt in intelligentiequotiënt (IQ).

Het IQ is een genormaliseerde testscore en het geeft aan hoe een score op een intelligentietest zich verhoudt tot de gemiddelde score van leeftijdsgenoten. Het gemeten IQ wordt gezien als een schatting van de ware, maar onbekende, intelligentie. Het gemiddelde is vastgesteld op 100 punten met een standaardafwijking van 15 punten. Dit betekent dat 68 procent van de bevolking tussen de 85 en de 115 scoort (binnen 1 standaardafwijking van het gemiddelde) en

95 procent van de bevolking tussen de 70 en de 130 (binnen 2 standaardafwijkingen van het gemiddelde).

Ondanks verschillen in onderzoeksopzet (bijvoorbeeld verschillende typen IQ-metingen) en context (verschillende landen) wordt in de meeste studies waargenomen dat intelligentie van groot belang is voor schoolprestatie (Neisser et al., 1996; Ogbu, 1994; Lynn en Vanhanen, 2002; Helms-Lorenz, 2003; Lynn, 1983).

In dit onderzoek nemen we intelligentie mee om te kunnen onderzoeken wat de invloed van omgevingskenmerken op schoolprestaties is wanneer voor de invloed van intelligentie wordt gecorrigeerd. Oftewel, om vast te stellen wat de invloed van omgevingskenmerken op onderwijsachterstanden is. Onderwijsachterstand definiëren we immers als de situatie waarin een leerling lager presteert dan mag worden verwacht op basis van zijn intelligentie (individuele potentie). Veel onderzoeken die onderwijsachterstanden proberen te verklaren, corrigeren echter niet voor de potentie van leerlingen. Een belangrijke reden hiervoor zijn de meetproblemen. Het is namelijk lastig om de intelligentie van een persoon vast te stellen, gecorrigeerd voor de externe effecten die medebepalend zijn geweest voor de ontwikkeling van de intelligentie. Zo ook is het complex te corrigeren voor de externe invloeden die iemand in meer of mindere mate in staat stellen een IQ-toets te maken. In het onderstaande worden deze meetproblemen afzonderlijk besproken.

Invloed van sociaaleconomische omgeving

Problematisch bij het gebruik van intelligentie als maat voor individuele potentie, is dat het IQ voor een deel wordt beïnvloed door de omgevingskenmerken van het kind en dus niet puur iets zegt over het individu. Over de manier waarop dit gebeurt, bestaan verschillende theorieën. Volgens de achterstands-omgevingshypothese kan de potentie van kinderen in families met lagere sociaaleconomische status (SES) getemperd worden door de beperkingen die armoede met zich meebrengt.¹¹ Volgens de bio-ecologische hypothese kan de potentie van kinderen bij hoger opgeleide en welgestelde families vergroot worden doordat deze gezinnen meer middelen hebben en mogelijkheden kunnen bieden aan de kinderen om zich te ontplooiën (Turkheimer et al., 2003; Hanscombe et al., 2012).

Aangezien het onze doelstelling is om de effecten van de omgeving op de schoolprestatie zo zuiver mogelijk te schatten, willen we in ons onderzoek enkel voor het ‘in potentie aanwezige’ of ‘aangeboren’ deel van intelligentie corrigeren. Een onderscheid dat in deze context weleens gemaakt wordt is dat tussen ‘vloeibare’ en ‘gekristalliseerde’ intelligentie. Vloeibare intelligentie betreft de mate van flexibiliteit in denken en het vermogen tot abstract redeneren. Gekristalliseerde intelligentie betreft de mate waarin een accumulatie van kennis en vaardigheden heeft plaatsgevonden gedurende de levensloop. Vloeibare intelligentie wordt geacht meer de ‘potentiële intelligentie’ te weerspiegelen. Voor het huidige onderzoek zou de vloeibare intelligentie het meest interessant zijn, omdat dit het meest in de buurt komt van de niet-door-de-omgeving-beïnvloede intelligentie. Het is echter moeilijk dit aspect van intelligentie vast te stellen. In dit verband is het interessant te noemen dat er veel studies zijn gedaan onder tweelingen waarbij de impact van erfelijke eigenschappen en de impact van omgevings- en opvoedingskenmerken onderzocht zijn. Het ‘aangeboren deel’ in intelligentie wordt op basis van verschillen in IQ-testscorecorrelaties bij eeneiige tweelingen en bij tweeeiige tweelingen vastgesteld (Hanscombe et al., 2012; Hart et al., 2007). Volgens dit soort studies¹² neemt het aangeboren deel de meeste variantie in intelligentiescores voor zijn rekening, maar speelt ook de invloed van de sociale omgeving een belangrijke rol (Hart et al., 2007).

¹¹ Turkheimer et al. (2003) concluderen dat bij arme gezinnen de sociaaleconomische status wel 60 procent van de variantie in IQ bij 7-jarige kinderen verklaart, terwijl dat effect bij welgestelde gezinnen bijna nihil is.

¹² Dit type studies onder mono-zygote en di-zygote tweelingen is in verschillende landen uitgevoerd.

Leeftijd en cognitieve ontwikkeling

De cognitieve ontwikkeling van de mens is leeftijdgerelateerd. Bepaalde facetten van de intelligentie veranderen sterker met de leeftijd dan andere. Er wordt geschat dat de intelligentie bij een vijfjarige al voor 50 procent vastligt en bij een achtjarige zelfs al voor 80 procent. Gedurende de eerste 18 maanden valt nog niet te voorspellen hoe intelligent het kind zal worden, maar de cognitieve ontwikkeling kan dan wel al gestimuleerd worden.

Bij tieners neemt de intelligentie steeds minder toe. Aan het einde van de puberteit is de cognitieve ontwikkeling gestabiliseerd. Rond de leeftijd van 17 jaar is een IQ-test betrouwbaar en zal het IQ in de loop van de tijd ook niet meer veel veranderen. Bij gezonde mensen blijft de intelligentie tot zeker na het 65ste levensjaar stabiel (Van Thiel en Engelen, 2016; Resing en Drenth, 2007).

Aangezien bij basisschoolkinderen de cognitieve capaciteiten nog in ontwikkeling zijn, zou idealiter de mate van ontwikkeling van de intelligentie van een kind in het analysemodel opgenomen worden. Hier bestaan op dit moment echter geen haalbare mogelijkheden voor. In dit onderzoek vindt de meting van cognitieve capaciteiten plaats bij kinderen van ongeveer 9 jaar (zie ook 'beschikbare gegevens over intelligentie'). Aangezien deze kinderen al kunnen lezen en schrijven, en de intelligentie hoogstwaarschijnlijk voor 80 procent vastligt, brengt het meten van IQ voor deze groep minder onzekerheden met zich mee.

Herkomst, opleiding en vertekening van toetsscores

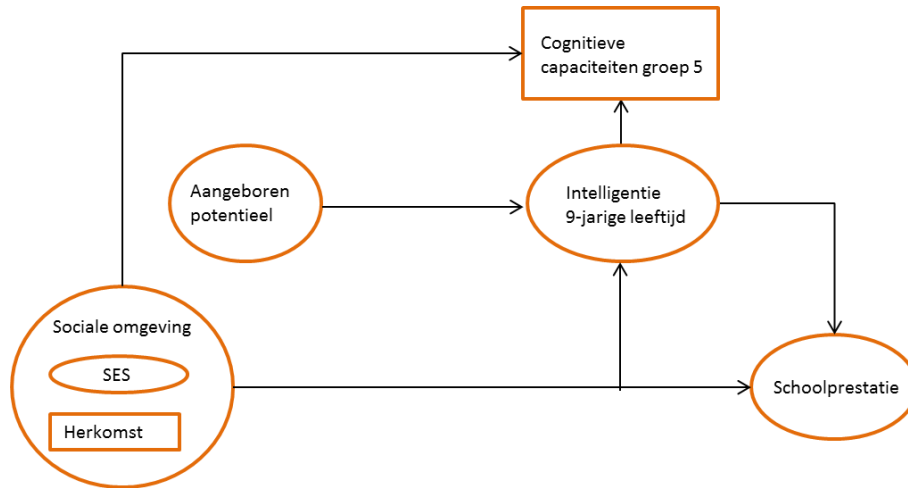
Een bijkomende moeilijkheid bij het gebruik van intelligentiescores is dat zij kunnen vertekenen voor minderheidsgroepen. Uit eerder onderzoek blijkt dat minderheidsgroepen, zoals niet-westerse allochtonen, gemiddeld lager scoren op gestandaardiseerde IQ-testen (de Jong & van Batenburg, 1984; Resing, Bleichrodt, & Drenth, 1986). Deze lagere IQ-scores onder allochtonen kunnen te maken hebben met een algeheel lager opleidingsniveau onder (niet-westerse) allochtonen (CBS Statline, 2016). Mensen die langer onderwijs hebben gevolgd en een hoger opleidingsniveau hebben scoren gemiddeld hoger op een IQ-test (Murray, 1998). De verschillen in scores tussen bevolkingsgroepen kunnen daarnaast te maken hebben met de gevoeligheid van de vraagstelling in de IQ-tests voor culturele verschillen (Resing en Drenth, 2007). Andere verklaringen voor lagere scores onder niet-westerse allochtonen kunnen gezocht worden in verschil van scholing (meer training op bepaalde facetten van de intelligentie) of aspecten als cultureel bepaalde ijver. Behalve door culturele vertekening, kunnen de intelligentiescores van minderheidsgroepen ook lager uitvallen als er sprake is geweest van selectieve migratie.

Voor dit onderzoek is het van belang dat we enkel corrigeren voor de 'aangeboren' intelligentie van minderheidsgroepen. Welk deel van de intelligentiescores van minderheidsgroepen hierdoor wordt bepaald en welk deel het resultaat is van vertekeningen, wordt uit eerder onderzoek echter niet duidelijk.

Beschikbare gegevens over intelligentie

In dit onderzoek beschikken wij over een maat die de intelligentie van leerlingen enigszins benadert: de niet-schoolse cognitieve capaciteiten test (nscct), die is afgenomen in groep 5 van scholen die hebben deelgenomen in de COOL⁵⁻¹⁸-cohortstudie. De nscct heeft dezelfde beperkingen als uitgebreidere intelligentietesten.

Bij de nscct-scores die in dit onderzoek gebruikt zullen worden is er dus sprake van mogelijke vertekening vanwege de invloed van de sociale omgeving, waaronder het opleidingsniveau en de culturele verschillen ten gevolge van verschil in herkomst. Daarnaast speelt de invloed van leeftijd een rol aangezien het onderzoek basisschoolleerlingen betreft. In figuur 3.2 worden deze invloeden grafisch weergegeven. De sociale omgeving heeft een invloed op de schoolprestatie van het kind, maar ook op de intelligentieontwikkeling van het kind. Dit geldt zowel voor de 'werkelijke' als voor de gemeten intelligentie.



Figuur 3.2: Conceptuele weergave van de invloeden van sociale omgeving en leeftijd op intelligentie. De rechthoeken geven aan dat het om bekende, meetbare variabelen gaat; de ovalen geven aan dat het om onbekende en/of onmeetbare variabelen gaat.

Om deze ‘culturele vertekening’ tegen te gaan, zullen we de nscct-scores later in dit onderzoek corrigeren op basis van informatie over het opleidingsniveau en herkomstgroepering van de ouders. Dit zijn de variabelen waarvan we verwachten dat zij de meetfout van de nscct, oftewel de afwijking tussen de gemeten en het werkelijke potentieel, het beste verklaren. Door de correctie die wij uitvoeren komen de gemiddelde nscct-scores van de kinderen met hoger en lager opgeleide ouders en autochtone en allochtone kinderen minder ver uit elkaar te liggen. De correctiemethode wordt in detail behandeld in paragraaf 4.2.4.

3.4 Gezinsinvloeden

Naast de individuele kenmerken, zoals de intelligentie van het kind, zijn de kenmerken van het gezin medebepalend voor schoolprestaties (Mulder et al., 2014). Hierna bespreken we de gezinskenmerken die in eerder onderzoek het vaakst gerelateerd worden aan schoolprestaties.

3.4.1 SES

Eerder onderzoek laat zien dat de sociaaleconomische status (SES) van de ouders van een kind, van invloed is op zijn/haar schoolprestaties (Sirin, 2005). De drie meest voorkomende componenten die in dit verband in de literatuur worden genoemd zijn: de opleiding van de ouders, het beroep of de beroepsstatus van de ouders en het inkomen van het gezin (Reynders et al., 2005). Deze kenmerken worden in eerder onderzoek gezamenlijk maar ook afzonderlijk gehanteerd om de sociaaleconomische achtergrond te meten. Van deze drie blijken opleiding en beroep de meeste voorspellende waarde te hebben op schoolprestatie (Reynders et al., 2005).

Beschikbare gegevens over SES

In het huidige onderzoek maken we gebruik van informatie over de opleiding van beide ouders en verschillende variabelen die de inkomenspositie meten om de sociaaleconomische positie van het gezin in kaart te brengen. Daarnaast wordt er gebruik gemaakt van een variabele die aangeeft of de ouders al dan niet inkomsten uit werk hebben. Informatie over beroepen is slechts voor een zeer klein deel van de ouders beschikbaar. Deze informatie nemen we daarom niet mee.

3.4.2 Cultureel kapitaal en herkomst

In westerse landen zijn de vaak lagere schoolprestaties van etnische minderheden een onderwerp van zorg. Sinds de jaren '70 bestaat het idee dat het onderwijssysteem de sociale ongelijkheid tussen autochtonen en allochtonen in de hand werkt (Driessen, 2001). Daarbij

speelt de culturele overdracht van sociale ongelijkheid een rol. Cultureel kapitaal, bestaande uit facetten als omgangsvormen, taalgebruik, interesses, kunst en cultuur voorkeuren, sociale en culturele opvattingen, wordt van generatie op generatie overgebracht. Kinderen verkrijgen hun cultureel kapitaal in eerste instantie tijdens hun socialisatie binnen het gezin. Aangezien de dominante cultuur ten grondslag ligt aan het curriculum dat de scholen aanbieden, wordt wel verondersteld dat hoe dichterbij het gezin zich bij de dominante cultuur bevindt, hoe makkelijker het wordt voor het kind om zich comfortabel te voelen en mee te komen op school (Driessen, 2001). Dit zou er dan toe leiden dat de verschillen in de schoolprestaties tussen allochtonen en autochtonen worden vergroot (Driessen, 2001).

Beschikbare gegevens over cultureel kapitaal en herkomst

In dit onderzoek wordt in de eerste plaats de herkomstgroepering op een zodanige manier geoperationaliseerd (zie 4.2.4) dat zij een zo goed mogelijke maat vormt voor culturele afstand. Daarnaast worden ook de generatie van de ouders, de verblijfsduur van de ouders in Nederland en thuistaal als afzonderlijke kenmerken meegenomen om te onderzoeken of een afwijkende culturele achtergrond gerelateerd is aan schoolprestaties.

3.4.3 Stabiliteit van de gezinsomstandigheden

De stabiliteit in een gezin wordt gerelateerd aan allerlei kenmerken. Gezinnen worden bijvoorbeeld als instabiel getypeerd wanneer het gaat om één-oudergezinnen, gezinnen waar de ouders kampen met psychische of huwelijksproblemen, wanneer een autoritaire opvoedingsstijl wordt gehanteerd, of als er in een gezin ingrijpende gebeurtenissen hebben plaatsgevonden zoals ontslag en ziekte van gezinsleden (Leseman, 2007). De stabiliteit binnen een gezin is net als herkomst een kenmerk dat gerelateerd is aan socio-economische status, maar nog een aanvullende invloed heeft. De stabiliteit in het gezin is indicatief voor de capaciteit van de ouders om een stimulerende omgeving aan hun kind te bieden. Een stimulerende omgeving houdt bijvoorbeeld in dat ouders een opvoedprincipe zoals 'informeel leren' toepassen. Daar horen bijvoorbeeld taalspelletjes of klokkijken bij, die als stimulerend kunnen worden beschouwd voor het lerend denkvermogen van het kind (Mesman, 2011). De veronderstelling is dat er in gezinnen met instabiele kenmerken minder nadruk ligt en minder ruimte bestaat voor stimulering en zo iets als 'informeel leren', omdat andere zaken de aandacht vragen. In dit onderzoek kijken we naar stabiliteit als indicator van de omstandigheden in het gezin, maar het is goed denkbaar dat instabiliteit in een gezin (gezinsniveau) ook een effect heeft op het sociaal-emotioneel functioneren van het kind (individueel niveau).

Beschikbare gegevens over de stabiliteit van de gezinsomstandigheden

Er zijn verschillende manieren waarop de stabiliteit van een gezin met de bij CBS beschikbare data kan worden geoperationaliseerd. In dit onderzoek zien we de informatie over gezinssamenstelling, scheidingen en overlijden van ouders als indicatie van (in)stabiliteit. Daarnaast nemen we de leeftijd van de ouders mee als kenmerken vanuit de veronderstelling dat zeer jonge ouders mogelijk minder goed in staat zijn om de nodige stabiliteit te bieden. We nemen ook gegevens over verdacht zijn van een misdrijf en schuldsanering mee. Bij gezinnen waarvan bekend is dat de vader en/of moeder verdacht zijn geweest van een misdrijf of in de schuldsanering hebben gezeten wordt er vanuit gegaan dat de instabiele situatie minder ruimte biedt voor het stimuleren van de cognitieve ontwikkeling van het kind. Tot slot wordt het gegeven of de ouders ooit asielen hebben aangevraagd meegenomen als afzonderlijke indicator van stabiliteit in de gezinsomgeving. Indien men asielen heeft aangevraagd dan kan er vanuit worden gegaan dat men een onzekere en turbulente periode achter de rug heeft, niet alleen in het thuisland, maar ook door de verplaatsing van thuisland naar Nederland en de asielenprocedure in Nederland.

3.5 Buurtkenmerken

De buurt waarin kinderen opgroeien, kan van invloed zijn op de schoolprestaties. Kinderen spelen in de buurt en krijgen (onbewust) de situatie in een bepaalde buurt mee. Een woonomgeving die gekenmerkt wordt door een sterke aanwezigheid van laagopgeleiden, werkloosheid, taalarmoede of een belemmerende dorps/straatcultuur kan als niet-stimulerend voor het kind beschouwd worden (Mulder et al., 2014). Als een kind in een taalarm gebied opgroeit, is de kans groot dat het kind niet gestimuleerd wordt, wat nadelig is voor hun schoolprestaties. Buurten waar werkloosheid overheerst kunnen de instelling, werkhouding of opvattingen over de samenleving van kinderen beïnvloeden.

Een soortgelijke invloed geldt voor buurtveiligheid (Leseman, 2007). Criminaliteit en onveiligheid in de buurt worden, net als armoede en werkloosheid, beschouwd als negatieve kenmerken, die niet stimulerend zijn voor de ontwikkeling van een kind. Verondersteld wordt dat kinderen omringd door meer criminaliteit en onveiligheid een ander beeld van de samenleving (rechten, plichten, autoriteit) kunnen ontwikkelen, maar ook bijvoorbeeld over discipline, dat uiteindelijk door kan werken in de manier waarop ze zich op school gedragen, inzetten en presteren. Eerder onderzoek wijst inderdaad uit dat schoolprestaties gemiddeld lager zijn in stedelijke gebieden, waar dit soort buurtkenmerken vaker voorkomen (Mulder et al., 2014). Om deze effecten tegen te gaan, hanteert OCW de impulsregeling waarbij scholen die in armoedeprobleemcumulatiegebieden liggen extra financiering krijgen voor gewichtenleerlingen.

Beschikbare gegevens over buurtkenmerken

In het huidige onderzoek verkennen we de invloed van de buurt op schoolprestaties. Verschillende kenmerken zullen worden gebruikt om een buurt te typeren. De welvaart of armoede in een buurt zal bepaald worden aan de hand van inkomensgegevens. Daarnaast nemen we de gemiddelde woningwaarde (WOZ) van de buurt mee. Ook onderzoeken we of variabelen als het aandeel uitkeringen in het kader van de Wet Werk en Bijstand (WWB), het aandeel mensen met een laag inkomen en het aandeel huurhuizen in een buurt, opgenomen zouden moeten worden in het verdeelmodel. Als grove proxy voor de aanwezige culturele middelen wordt gebruik gemaakt van informatie over het aandeel niet-westerse allochtonen. Verder nemen we de stedelijkheidsgraad mee als buurtkenmerk. Onder de noemer van 'leefbaarheid' vallen aspecten van veiligheid, voorzieningen en kwaliteit van de fysieke leefomgeving. Als indicator voor leefbaarheid hebben we de leefbaarheidsscore uit de Leefbaarometer¹³ opgenomen als buurtkenmerk.

3.6 Schoolkenmerken

Scholen kunnen de bestaande verschillen in schoolse cognitieve vaardigheden tussen kinderen vergroten of verkleinen (Van Rooijen en Zoon, 2012). Onderzoek onder autochtone leerlingen met lage schoolprestaties laat zien dat deze leerlingen vaker op een school met ongunstige kenmerken zitten (Mulder et al., 2014). Dit zijn vaak scholen met een concentratie van leerlingen met lage schoolprestaties (Mulder et al., 2014). Op scholen met veel laag opgeleide leerlingen ontbreekt de stimulans van goede leerlingen ofwel leerlingen die het goede voorbeeld geven. Daarnaast stellen de leraren de doelstellingen naar beneden bij in een dergelijke omgeving, waardoor ook de druk om te presteren wordt afgezwakt. Volgens het onderzoek van Mulder et al. (2014) heeft dit een sterke negatieve invloed op de schoolprestaties van kinderen. Ook is het zelden het geval dat de beste leerkrachten worden ingezet voor de zwakste niveaugroep (Van Rooijen en Zoon, 2012; Houtveen et al., 2011).

De omgeving van een school is vaak bepalend voor de samenstelling van de school en de klas waarin het kind zich bevindt. Schoolsegregatie kan dus gemakkelijk ontstaan als er eveneens sprake is van een gesegregeerde woonomgeving (Van Rooijen en Zoon, 2012). Vaak gaat het om concentratie van allochtone leerlingen waardoor ook wel de zogenoemde 'zwarte scholen'

¹³ Zie bijlage 2 voor meer informatie over de Leefbaarometer.

ontstaan. De concentratie van niet-westerse allochtone leerlingen op 'zwarte scholen' versterkt volgens eerder onderzoek het negatieve effect van herkomst op leerprestatie (Mulder et al., 2014; Ledoux et al., 2015).

Concentratie van leerlingen wier ouders een hoge opleiding hebben afgerond kan zorgen voor een positief effect op de schoolprestaties: andere kinderen kunnen 'gemobiliseerd'/gemotiveerd worden om ook beter te presteren of te leren van de aanpak, inzet en gedrag van de hoger presterende leerling.

Beschikbare gegevens over schoolkenmerken

De invloed van scholen op schoolprestaties van kinderen nemen wij allereerst mee in het analysemodel door rekening te houden met het opleidingsniveau van de moeders van leerlingen. Dit doen we op drie manieren: door te kijken naar het aandeel moeders met een lage opleiding, hoge opleiding en hun gemiddelde opleiding. Daarnaast kijken we naar de concentratie van allochtone leerlingen op scholen, omdat verondersteld wordt dat schoolprestaties gemiddeld lager liggen op scholen waar een clustering hiervan voorkomt. Ten slotte nemen we ook het aandeel ouders op school met een laag inkomen mee als mogelijke verklarende variabele.

4. Onderzoeksmethode

4.1 Inleiding

OCW heeft CBS gevraagd een aantal modellen door te rekenen die zij op voorhand interessant vinden. Het gaat om negen modellen met de volgende voorspellers:

1. Opleidingsniveau
2. Opleidingsniveau en inkomen
3. Opleidingsniveau en herkomst
4. Opleidingsniveau, inkomen en herkomst
5. Opleidingsniveau, inkomen, herkomst en buurtgegevens
6. Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur
7. Inkomen
8. Inkomen, herkomst, buurtgegevens
9. Inkomen, herkomst, buurtgegevens en verblijfsduur.

Daarnaast heeft het ministerie gevraagd om, op basis van beschikbare gegevens over de theoretisch relevante kenmerken die in hoofdstuk 3 zijn besproken, een model te ontwikkelen dat onderwijsachterstanden zo goed mogelijk voorspelt. In dit hoofdstuk bespreken we de data en methode waarmee we een model hebben ontwikkeld om onderwijsachterstanden zo goed mogelijk te voorspellen.

Het hoofdstuk is als volgt opgezet. In paragraaf 4.2 bespreken we de gebruikte gegevens. Voor enkele kernvariabelen (opleiding, herkomst en intelligentie) zijn bij het operationaliseren ten behoeve van het modelleren specifieke keuzes gemaakt die enige toelichting behoeven; deze toelichting wordt eveneens gegeven in paragraaf 4.2. In paragraaf 4.3 gaan we in op de methode. We bespreken waarom in dit onderzoek is gekozen om onderwijsachterstand te modelleren met behulp van een structureel vergelijkingsmodel. Ook gaan we in op de fitmaten die gebruikt zijn om de modellen te beoordelen en de procedure die gevolgd is om het best passende analysemodel te vinden. De resultaten volgen in hoofdstuk 5.

4.2 Data

4.2.1 Beschrijving gebruikte gegevens

We hebben onze analysemodellen ontwikkeld op basis van een dataset die bestaat uit gegevens die zijn verzameld tijdens het cohortonderzoek COOL⁵⁻¹⁸ en gegevens uit het SSB. De belangrijkste reden om ook gebruik te maken van de COOL⁵⁻¹⁸-gegevens is de beschikbaarheid van een intelligentiemaat per leerling. In tegenstelling tot het SSB, bevatten COOL⁵⁻¹⁸-bestanden nscct-testgegevens die in groep 5 af zijn genomen. Deze test is speciaal ontwikkeld om leerlingen op te kunnen sporen die onder hun niveau presteren.

Binnen de COOL⁵⁻¹⁸-cohortstudie werden leerlingen uit groep 2, 5 en 8 door de jaren heen gevolgd.¹⁴ De metingen werden verricht in de schooljaren 2007/2008, 2010/2011 en 2013/2014 in groep 2, 5 en 8. Omdat wij in ons model voor iedere leerling zowel de Cito-score in groep 8 als de score op de nscct in groep 5 willen gebruiken, kunnen we slechts twee cohorten in onze studie meenemen, namelijk de leerlingen die in schooljaar 2007/'08 of 2010/'11 in groep 5 zaten. Van deze leerlingen hebben we immers ook de Cito-scores uit groep 8.

Niet alle leerlingen bij wie een nscct is afgenomen komen in de uiteindelijke analyse terecht. Dit heeft verschillende oorzaken. Zo kunnen we alleen leerlingen meenemen voor wie de nscct is afgenomen en dit is niet voor alle leerlingen in groep 5 het geval. Ook als leerlingen ontbrekende waarden hebben op andere variabelen die voorkomen in de modellen die we

¹⁴ Voor meer achtergrondinformatie over COOL⁵⁻¹⁸, zie bijlage 2.

onderzoeken, nemen we ze niet mee. Daarnaast is er uitval in het COOL⁵⁻¹⁸-onderzoek door verwezen leerlingen (zie Paas et al., 2013). De analyses zijn uiteindelijk gebaseerd op gegevens van 13 466 leerlingen. Bijlage 10 bevat een overzicht van het aantal leerlingen dat afvalt.

De gegevens over leerlingen en hun nsct-scores uit de COOL⁵⁻¹⁸-cohortstudie zijn verrijkt met gegevens uit het SSB. Het SSB is door CBS opgezet en bevat een groot aantal microdatabestanden over sociaaleconomische en ruimtelijke statistieken.¹⁵ Door gegevens uit het SSB aan de COOL⁵⁻¹⁸-data te koppelen hebben we een onderzoeksbestand samengesteld dat alle in hoofdstuk 3 genoemde variabelen bevat.

Voor de interpretatie van de resultaten die in het volgende hoofdstuk besproken worden is het belangrijk om te weten hoe de variabelen exact geoperationaliseerd zijn. In bijlage 3 is daarom een overzicht van alle variabelen en hun operationalisering opgenomen. De variabelen opleidingsniveau van de ouders, herkomst (als maat voor culturele afstand) en intelligentie behoeven een uitgebreidere beschrijving. Daarom gaan we in paragraaf 4.2.2 t/m 4.2.4 hieronder uitvoeriger in op de operationalisering van deze concepten.

4.2.2 Operationalisering opleidingsniveau van de ouders

Ter vereenvoudiging van de analysemodellen hebben we ervoor gekozen om de ordinale variabele opleidingsniveau van de ouders (zie de acht categorieën in bijlage 3) te transformeren naar een continue variabele. Op die manier wordt het effect van opleiding opgenomen in het model door middel van slechts twee variabelen (opleidingsniveau vader en opleidingsniveau moeder), in plaats van dummy's voor alle afzonderlijke categorieën.

De continue variabelen zijn afgeleid door de categorieën van opleidingsniveau af te beelden op een continue schaal, op zodanige wijze dat de correlatie tussen het afgeleide continue opleidingsniveau en de schoolprestatie wordt gemaximaliseerd, rekening houdend met eventuele andere verklarende variabelen in het te schatten analysemodel. In de praktijk is dit gedaan door eerst een afzonderlijk regressiemodel¹⁶ te schatten waarin schoolprestatie wordt verklaard door alle categorieën van het opleidingsniveau van zowel de vader als de moeder als dummyvariabelen. Daarnaast zijn in dit regressiemodel steeds alle andere verklarende variabelen meegenomen die voorkomen in het analysemodel. In het regressiemodel heeft elke opleidingscategorie een geschatte coëfficiënt. De afgeleide continue opleidingsvariabele bestaat nu uit de waarden van deze coëfficiënten. Bijvoorbeeld: als de coëfficiënt van een Mbo4-opleiding voor de moeder 0.175 is, dan heeft de afgeleide continue variabele 'opleidingsniveau moeder' de waarde 0.175 voor alle leerlingen met een moeder met een Mbo4-opleiding. Deze procedure wordt herhaald voor elk te schatten analysemodel, en de schaalwaarden van de afgeleide continue opleidingsvariabelen kunnen dan ook per model verschillen.

4.2.3 Operationalisering herkomst

In dit onderzoek wordt de herkomst van de ouders gebruikt om te onderzoeken of een afwijkende culturele achtergrond van invloed is op schoolprestaties. We hebben herkomstlanden van de ouders zo ingedeeld dat zij relatief homogeen zijn naar verschillende indicatoren voor culturele afstand. Relatief homogeen, want binnen de groepen is zeker ook een flinke variatie in culturele kenmerken. Hetzelfde geldt natuurlijk binnen ieder land afzonderlijk. De gebruikte indicatoren zijn de volgende:

- De mate van ontwikkeling van de economie. Naarmate de economie van een herkomstland meer ontwikkeld is, is de culturele afstand tot Nederland kleiner. Daarbij is uitgegaan van de indeling die de Verenigde Naties hanteert (UN, 2015).

¹⁵ Zie bijlage 2 voor meer informatie over de bestanden uit het SSB die zijn gebruikt.

¹⁶ Hiervoor zijn, omwille van de eenvoud, Ordinary Least Squares (OLS) regressiemodellen zonder intelligentie gebruikt in plaats van structurele vergelijkingsmodellen met intelligentie. De invloed hiervan is zeer gering: de continue variabelen die volgen uit de OLS-procedure hebben een correlatie tussen de 0.95 en 0.99 met de continue variabelen zoals deze zouden volgen uit een afleidingsprocedure met een structureel vergelijkingsmodel.

- De taalafstand (zie voor een overzicht van enkele indelingen daarvoor Van Tubbergen & Kalmijn, 2005; Kooistra et al., 2008). Naarmate de taal van het herkomstland meer verschilt van het Nederlands, is de culturele afstand groter.
- Of het herkomstland een voormalige kolonie is van Nederland. Als dat zo is, dan is de culturele afstand tot Nederland *ceteris paribus* kleiner.
- De verschillen in geloof en de daarmee samenhangende omgangsvormen, waarden en normen, tussen het land van herkomst en Nederland.
- De regionale nabijheid van de herkomstlanden onderling.

Naast deze inhoudelijke indicatoren hebben we bij de indeling ook rekening gehouden met de omvang van de categorieën. Het heeft geen zin om categorieën te onderscheiden die vrijwel geen leerlingen tellen.

Hierna benoemen we per categorie de bijbehorende landen op volgorde van frequentie. Daarbij benoemen we alleen de landen met acht of meer waarnemingen. Voor de laatste categorie benoemen we de landen, gezien het grote aantal, niet allemaal afzonderlijk.

1. Nederland
2. Suriname en (voormalige) Nederlandse Antillen
3. Turkije
4. Noord-Afrika (Marokko, Egypte, Tunesië, Algerije, Soedan)
5. EU-15, andere ontwikkelde economieën: Duitsland, België, Verenigd Koninkrijk, Frankrijk, Spanje, Italië, Portugal, Zwitserland, Griekenland, Oostenrijk, Denemarken, Finland, Zweden, Noorwegen, IJsland, Canada, Verenigde Staten, Australië, Nieuw-Zeeland, Japan, 2e generatie Indonesië
6. Nieuwe EU-landen en economieën in transitie: Polen, Tsjechoslowakije, Hongarije, Bulgarije, Roemenië, Litouwen, Joegoslavië, (Voormalige) Sovjet Unie
7. Oost Azië: Indonesië (1e generatie), China, Filipijnen, Thailand, Vietnam, Hongkong, Zuid-Korea, Maleisië, Burma, Cambodja, Singapore, Straits Settlements
8. Rest Afrika, rest Azië, rest Latijns Amerika: Somalië, Ghana, Irak, Syrië, Pakistan, Afghanistan, Dominicaanse republiek, India, Iran, Colombia, Brazilië, Kaapverdië, Sri Lanka, Zuid-Afrika, etc.

Omdat het te verwachten is dat de culturele afstand van kinderen met één ouder uit Nederland beperkt zal zijn, worden kinderen ook in categorie 1 ingedeeld als de herkomst van maar één van de ouders Nederland is¹⁷. Als beide ouders uit een ander land komen, wordt het kind ingedeeld in de categorie waartoe het herkomstland van de moeder behoort. De herkomstlanden van de vader en moeder zijn niet allebei meegenomen omdat er sprake is van multicollineariteit. Uit de literatuur blijkt dat een zeer belangrijke rol weggelegd is voor de moeder als het aankomt op ontwikkeling en schoolprestatie van het kind (Morrison et al., 2003). Omdat de invloed van de moeder groot is en omdat de resultaten van de verkennende analyses laten zien dat het herkomstland van de moeder de grootste invloed heeft op schoolprestaties, hebben we ervoor gekozen om kinderen met ouders die geen van beiden een Nederlandse herkomst hebben, in de categorie van het herkomstland van de moeder in te delen.

Herkomst wordt hier gebruikt als maat voor culturele afstand, wat gezien kan worden als een continue begrip. Ter vereenvoudiging van het analysemodel hebben we daarom uit de bovengedefinieerde categorieën van herkomst een continue variabele afgeleid, op dezelfde manier als bij opleidingsniveau (zie paragraaf 4.2.2). In feite is steeds een gezamenlijk regressiemodel geschat waarin zowel opleiding als herkomst voorkomt, waaruit voor al deze variabelen tegelijk de schaalwaarden voor de bijbehorende continue variabelen zijn afgeleid.

¹⁷ Bij het analyseren van de data blijkt dat kinderen met één Nederlandse ouder en één niet-Nederlandse ouder vergelijkbaar met autochtone kinderen scoren op de Cito toets.

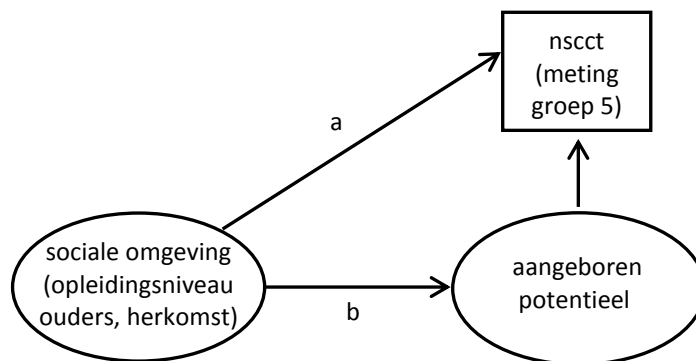
4.2.4 Operationalisering intelligentie

De scores op de nscct-test die bij leerlingen in groep 5 is afgenomen gebruiken we als meting voor intelligentie. Op deze manier kan een inschatting gemaakt worden van de op basis van intelligentie te verwachten schoolprestatie. Deze kan dan worden vergeleken met de werkelijke schoolprestatie om tot een maat van onderpresteren (dan wel excelleren) te komen. Een dergelijke analyse sluit aan bij het beoogde doel van het onderwijsachterstandenbeleid: het beschikbaar stellen van extra geld voor kinderen die, ongeacht hun prestatieniveau, onderpresteren.

In eerdere onderzoeken naar het effect van gezins- en omgevingskenmerken op onderwijsachterstand/schoolprestatie, is deze aanpak echter ongebruikelijk. Een belangrijke reden hiervoor is dat gegevens over intelligentie in lang niet alle onderzoeken beschikbaar zijn. Daarnaast bestaat het risico dat intelligentietoetsen zoals de nscct niet slechts aangeboren, niet-schoolese, capaciteiten meten, maar onbedoeld ook 'schoolse vaardigheden' zoals leesvaardigheid. Wanneer een intelligent kind bijvoorbeeld niet in staat is de testvragen te begrijpen, omdat de in de vraag voorkomende woorden bij hem/haar niet bekend zijn, kan ten onrechte geconcludeerd worden dat dit kind niet over de juiste capaciteiten beschikt en dus geen extra aandacht (en geld) nodig heeft.

Dit risico moet afgewogen worden tegen het nadeel van het niet meenemen van deze variabele. In dat geval is de enige overgebleven relevante uitkomstmaat immers de 'schoolprestatie'. Het achterstandenbeleid zal dan in de eerste plaats gericht zijn op het extra financieel ondersteunen van slecht presterende kinderen, ongeacht of deze kinderen de potentie hebben om beter te presteren. Om een model te kunnen maken dat daadwerkelijk aansluit op de achterliggende gedachte van het onderwijsachterstandenbeleid, kiezen we er daarom voor intelligentie, zoals gemeten door de nscct, in de analyse op te nemen. Daarbij zijn we ons bewust van, en houden we rekening met, verschillende haken en ogen van de nscct. Ter vergelijking voeren we ook analyses uit waarbij we niet voor intelligentie corrigeren.

De nscct is, zoals in principe alle intelligentietesten, niet vrij van invloeden van de sociale omgeving. De sociale omgeving is intrinsiek verbonden met de ontwikkeling van intelligentie. Figuur 4.1 geeft het probleem schematisch weer in de vorm van een pad-diagram. De figuur laat zien dat de score op onze intelligentiemaat, de nscct in groep 5, op twee manieren wordt beïnvloed door de sociale omgeving: enerzijds doordat de sociale omgeving invloed heeft op het werkelijke aangeboren potentieel van een leerling (pijl b) en anderzijds doordat de sociale omgeving invloed heeft op de meting van intelligentie, los van het werkelijke aangeboren potentieel van een leerling (pijl a). In de context van ons onderzoek zien we pijl a als een bron van vertekening.



Figuur 4.1: Schematische weergave van verschillende invloeden van de sociale omgeving op gemeten intelligentie.

De 'sociale omgeving' is een breed concept. Zoals in paragraaf 3.3.1 is uitgelegd, zijn er verschillende kenmerken die een vertekende weergave van intelligentie kunnen opleveren. Hier beschouwen we alleen de invloed van het opleidingsniveau en de herkomst van de ouders. Opleidingsniveau van de ouders, als een van de belangrijke indicatoren van SES, kan van invloed

zijn op de gemeten intelligentie omdat de potentie die kinderen hebben zich anders kan ontwikkelen in een milieu met hoger dan wel lager opgeleide ouders. Over de mate van invloed van het opleidingsniveau van de ouders op deze ontwikkeling ontbreekt eenduidigheid, maar uit verschillende studies is af te leiden dat de invloed ergens tussen 25 procent en 80 procent ligt (Turkheimer et al., 2003; Hanscombe et al., 2012; Hart et al., 2007).

Een ander kenmerk dat besproken is in paragraaf 3.3.1 is de vertekening die ontstaat door herkomst, bijvoorbeeld door gevoeligheid van de vraagstelling in een intelligentietest voor culturele verschillen. Over de grootte van het effect van herkomst op de meting van intelligentie zijn geen bruikbare cijfers te vinden in de literatuur. We verwachten dat een groot deel van de invloed van herkomst op waargenomen nscct-scores toe te schrijven is aan pure meeteffecten (culturele vertekening). Daarnaast zou een kleiner deel van de invloed kunnen bestaan uit werkelijke verschillen in intelligentie, als gevolg van een selectie-effect, omdat mensen met een niet-Nederlandse herkomst in ons bestand een selectieve deelgroep zijn van de volledige populaties uit de landen van herkomst.

De vraag is nu: hoe krijgen we een intelligentie- of liever nscct-score die vrij is van de invloed van opleidingsniveau en herkomst op de meting? Idealiter zouden we de nscct willen corrigeren voor de invloed van pijl a in figuur 4.1, maar niet voor pijl b. Echter, aangezien we maar één nscct-meting per leerling hebben, zijn deze twee effecten niet van elkaar te scheiden. Bij wijze van alternatieve oplossing corrigeren we de nscct-scores voor bepaalde percentages van de invloed van opleiding en herkomst, waarbij de gekozen percentages overeenkomen met onze verwachtingen over de relatieve verhouding tussen pijl a en pijl b in figuur 4.1. Om de gevoeligheid van de uitkomsten voor de mate van correctie te bepalen hebben we bovendien verschillende varianten onderzocht.

Voor opleiding hebben we, gezien de bovengenoemde percentages uit de literatuur, drie varianten uitgevoerd waarbij de nscct-score is gecorrigeerd voor respectievelijk 25%, 50% en 75% van de invloed van het opleidingsniveau van de ouders. Voor herkomst hebben we, bij gebrek aan meer informatie, ervoor gekozen om twee varianten te onderzoeken, waarbij de nscct-score gecorrigeerd is voor 75% en 100% van de invloed van herkomstgroepering van de ouders. In totaal geeft dit $3 \times 2 = 6$ varianten. Daarnaast hebben we een variant doorgerekend waarin de nscct-score ongecorrigeerd blijft.

De precieze uitwerking van de correctiemethode voor nscct-scores wordt gegeven in bijlage 5.

4.2.5 Clustering gegevens op school- en buurtniveau

De COOL⁵⁻¹⁸-gegevens hebben betrekking op leerlingen, maar zijn afkomstig van een steekproef van scholen. Tussen de waargenomen schoolprestaties van verschillende leerlingen in het bestand kunnen afhankelijkheden bestaan (clustereffecten). In het bijzonder verwachten we dat de schoolprestaties van twee kinderen van dezelfde school en/of uit dezelfde buurt meer overeenkomsten vertonen dan die van twee willekeurig gekozen kinderen. De aanname van onafhankelijke waarnemingen die gemaakt wordt in standaard regressiemodellen en structurele vergelijkingsmodellen (zie paragraaf 4.3.1) gaat hier dus mogelijk niet op. Dit betekent dat varianties foutief geschat kunnen worden en bepaalde effecten te gemakkelijk significant verklaard kunnen worden. Bovendien kunnen de geschatte effecten zelf vertekend zijn wanneer we bij het schatten geen rekening houden met het feit dat het COOL⁵⁻¹⁸-onderzoek een clustersteekproef van scholen betreft.

Om de omvang van de clustereffecten op school- en buurtniveau in onze gegevens vast te stellen hebben we een aantal zogenaamde multilevel-modellen geschat. Een multilevel-model is een regressiemodel waarin rekening wordt gehouden met afhankelijkheden die kunnen ontstaan doordat de waarnemingen geclusterd zijn in groepen (Snijders en Bosker, 2012). Hier hebben we gekeken naar multilevel-modellen met clustering op school- en/of buurtniveau. De omvang van een clustereffect kan in een dergelijk model worden weergegeven door de zogenaamde intraklasse-correlatiecoëfficiënt. Deze drukt uit hoeveel procent van de totale onverklaarde variantie in het model bestaat uit samenhang tussen waarnemingen uit dezelfde school, respectievelijk buurt.

Omwillen van de eenvoud is intelligentie niet meegenomen in de multilevel-modellen. Dit betekent dat we deze modellen kunnen schatten op basis van alle beschikbare SSB-data over leerlingen in het basisonderwijs, en niet alleen het deel van het bestand dat gekoppeld kan worden aan de COOL⁵⁻¹⁸-gegevens uit paragraaf 4.2.1. Een belangrijk voordeel van de SSB-gegevens is in dit geval dat de school- en buurteffecten hierin goed van elkaar kunnen worden gescheiden. Met de COOL⁵⁻¹⁸-gegevens is dit minder goed mogelijk: vanwege het steekproefontwerp van het COOL⁵⁻¹⁸-onderzoek komt het zelden voor dat meer dan één school per buurt is waargenomen, met als gevolg dat de waargenomen school- en buurteffecten in deze gegevens sterk overlappen.

De multilevel-analyses zijn uitgevoerd op het bestand met SSB-data dat eerder is geïntroduceerd in paragraaf 2.3 bij het imputeren. Voor de 75 017 leerlingen in dit bestand van wie de opleidingsniveaus van beide ouders bekend waren, werden in een model met alleen opleiding en herkomst als verklarende variabelen intraklasse-correlatiecoëfficiënten gevonden van respectievelijk 4.0 procent op schoolniveau en 1.0 procent op buurniveau. Na toevoeging van meer verklarende variabelen, waaronder kenmerken op school- en buurniveau, daalden deze coëfficiënten licht (3.9 procent op schoolniveau en 0.9 procent op buurniveau). Beide effecten zijn klein, maar wel significant. Hierbij moet wel worden bedacht dat het bestand een groot aantal waarnemingen bevat waardoor de kans op significante resultaten wordt vergroot. Een herhaling van de analyse op hetzelfde bestand inclusief geïmputeerde opleidingsniveaus leidde tot vergelijkbare uitkomsten.

Het gevonden buurteffect is dusdanig klein dat wij hebben besloten hier in de vervolganalyses geen rekening mee te houden. Het schooleffect daarentegen kan niet worden verwaarloosd. We hebben daarom in de verdere analyses rekening gehouden met samenhang op schoolniveau; zie paragraaf 4.3.1.

4.3 Methode

4.3.1 Structurele vergelijkingsmodellen

Om onderwijsprestaties zo goed mogelijk te kunnen verklaren is het belangrijk dat een geschikt analysemodel gekozen wordt. In dit onderzoek ontwikkelen we modellen waarbij we corrigeren voor intelligentie. Ter vergelijking zullen we eveneens modellen zonder intelligentie tonen.

Voor de modellen waarbij we corrigeren voor intelligentie hebben we in dit onderzoek gebruik gemaakt van structurele vergelijkingsmodellen (ook wel SEM genoemd). Het is in een structureel vergelijkingsmodel namelijk mogelijk om onderscheid te maken tussen enerzijds directe effecten van achtergrondvariabelen op schoolprestatie en anderzijds indirecte effecten die via intelligentie lopen (zie figuur 3.2). Dit is van belang omdat we uiteindelijk onderwijsachterstanden, oftewel voor intelligentie gecorrigeerde schoolprestaties, willen verklaren. Het gaat ons dus om de directe effecten van de overige achtergrondvariabelen. Om deze effecten goed te kunnen schatten moeten ook de indirecte effecten via intelligentie worden gemodelleerd. Een structureel vergelijkingsmodel biedt deze mogelijkheid.

Daarnaast gebruiken wij structurele vergelijkingsmodellen omdat zij de mogelijkheid bieden om te corrigeren voor toevallige meetfouten in nsct-scores.¹⁸ In ons model wordt de nsct-score gebruikt als proxy voor het ongemeten concept intelligentie. Uit eerdere tests in Van Batenburg (2015) blijkt dat de nsct in groep 5 een betrouwbaarheid van 0.88 heeft. Dit houdt in dat 12 procent van de totale variantie in de nsct-scores bestaat uit ruis die niet samenhangt met het onderliggende concept intelligentie. Door dit gegeven op te nemen in het structurele

¹⁸ In deze alinea verwijst de term 'meetfout' niet naar een systematische vertekening voor bepaalde groepen, zoals in paragraaf 4.2.4, maar naar toevallige onnauwkeurigheid in gemeten intelligentie. Deze onnauwkeurigheid blijkt bijvoorbeeld uit het feit dat wanneer dezelfde intelligentietoets tweemaal kort na elkaar wordt afgenomen bij dezelfde persoon, dit niet precies dezelfde intelligentiescore oplevert.

vergelijkingsmodel houden we er rekening mee dat de nscct-score niet precies samenvalt met intelligentie. We doen dit door restricties op te leggen aan de parameters van het model.¹⁹

Structurele vergelijkingsmodellen worden vaak geschat via *maximum likelihood* (zie bijvoorbeeld Bollen, 1989). Hierbij wordt aangenomen dat alle variabelen in het model een gezamenlijke normale verdeling volgen. Ook wordt aangenomen dat alle waarnemingen onafhankelijk van elkaar zijn. In onze toepassing worden beide aannames tot op zekere hoogte geschonden. Ten eerste zijn sommige achtergrondvariabelen niet-normaal verdeeld. Ten tweede schatten we de modellen op COOL⁵⁻¹⁸-gegevens van een clustersteekproef van scholen waarbij bovendien afhankelijkheid bestaat tussen waarnemingen van leerlingen binnen dezelfde school (zoals besproken in paragraaf 4.2.5). Niet-normaliteit van de data heeft doorgaans geen effect op de zuiverheid van de geschatte parameters, maar wel op de geschatte standaardfouten en fitmaten (Bollen, 1989). Afhankelijkheid tussen waarnemingen op schoolniveau kan zowel effect hebben op parameterschattingen als op standaardfouten en fitmaten.

Om rekening te houden met niet-normaliteit en clustering op schoolniveau hebben we de structurele vergelijkingsmodellen geschat via *robust maximum likelihood*. Dit houdt in dat rekening wordt gehouden met het steekproefontwerp van het COOL⁵⁻¹⁸-onderzoek tijdens het schatten van de parameters van het model. Verder worden de standaardfouten en fitmaten achteraf gecorrigeerd voor niet-normaliteit en afhankelijkheid op schoolniveau. In het geval van de fitmaten is hier de zogenaamde Satorra-Bentler-correctiefactor gebruikt.²⁰

Voor de varianten van het analysemodel waarin intelligentie niet wordt meegenomen als controlevariabele heeft het gebruik van structurele vergelijkingsmodellen hier geen toegevoegde waarde. Deze varianten zouden ook kunnen worden geschat als regressiemodel of eventueel als multilevel-model. Echter, om de vergelijkbaarheid van de uitkomsten en fitmaten tussen de verschillende analysemodellen te bevorderen, hebben we ervoor gekozen om alle modellen hier te schatten als structurele vergelijkingsmodellen. Om dezelfde reden hebben we de modellen zonder intelligentie geschat op hetzelfde onderzoeksbestand (de COOL⁵⁻¹⁸-gegevens) als de modellen met intelligentie, ook al zouden de modellen zonder intelligentie in principe kunnen worden geschat op basis van alle beschikbare SSB-data.

4.3.2 Stepwise-aanpak

In hoofdstuk 5 berekenen we de resultaten uit de modellen die OCW bij voorbaat interessant vond. Daarnaast hebben we gezocht naar het analysemodel dat het beste past bij de data in het onderzoeksbestand: welke combinatie van variabelen leidt tot de beste voorspelling van schoolprestatie, zoals gemeten door de Cito-score in groep 8? Hierbij is een stepwise-procedure gevolgd die we hier zullen toelichten.

Ten eerste hebben we gezocht naar de variabelen die een significante bijdrage leveren aan het voorspellen van de Cito-score in groep 8. Daarbij worden alle in hoofdstuk 3 benoemde en in bijlage 3 geoperationaliseerde variabelen een voor een getest in losse regressiemodellen. Soms zijn variabelen zowel discreet en continu getest: bijvoorbeeld inkomensklassen (discreet) en inkomen (continu). Hieruit volgt een voorselectie van significante variabelen die in aanmerking komen voor de eigenlijke stepwise-procedure.

¹⁹ In dit geval is het opleggen van restricties de enige manier om rekening te houden met de (eerder geschatte) betrouwbaarheid van de nscct-scores, omdat we slechts beschikken over één meting van intelligentie per kind. Bij meerdere metingen van dezelfde conceptuele variabele zou men de betrouwbaarheid van deze variabelen daadwerkelijk kunnen schatten via een structureel vergelijkingsmodel.

²⁰ Een alternatieve aanpak om rekening te houden met de clustereffecten op schoolniveau is door het structurele vergelijkingsmodel uit te breiden met een multilevel-component. Dit leidt tot een tamelijk complex model. Voor het doel van het huidige onderzoek – het voorspellen van de mate van onderprestatie van leerlingen per school – heeft dit geavanceerdere model echter geen toegevoegde waarde. (De multilevel-toevoeging zou hooguit leiden tot betere voorspellingen voor de scholen die meedoen aan het COOL⁵⁻¹⁸-onderzoek, maar niet voor de andere scholen in Nederland.) Hier is daarom gekozen voor de eenvoudigere *robust maximum likelihood*-aanpak.

We hebben een startmodel opgesteld waarin twee variabelen worden opgenomen die uit de eerste analyses van het grootste belang bleken: opleiding en herkomst. Vervolgens hebben we onderzocht welke van de variabelen die eerder bivariaat significant waren gebleken in een regressiemodel, ook multivariaat significant zijn in een structureel vergelijkingsmodel en leiden tot een betere fit van het model bij de data. Daartoe hebben we programmatuur ontwikkeld die de variabele die het huidige model het meest verbetert als eerste toevoegt. Deze stap wordt net zolang herhaald totdat er geen verbetering meer mogelijk is. Om te bepalen of het toevoegen van een variabele het model verbetert hebben we gekeken naar verschillende fitmaten: AIC, BIC en aangepaste R-kwadraat (zie het kader 'Fitmaten' voor een toelichting). In de automatische procedure is in de eerste plaats gekeken naar de AIC; de andere maten zijn vooral gebruikt bij het achteraf vergelijken van de uitkomsten voor verschillende modellen. De precieze stepwise-procedure wordt hieronder toegelicht.

Om in aanmerking te komen voor toevoeging aan het huidige model moet een variabele een significant direct effect hebben op schoolprestatie. Hierbij is een variabele beoordeeld als significant indien de p-waarde van het betreffende effect kleiner is dan 0.10. Mocht een variabele in eerste instantie niet direct significant zijn, dan wordt geprobeerd of dit wel zo is wanneer hij daarnaast ook een indirect effect via intelligentie heeft. Ook wordt gekeken of het verwijderen van variabelen die door de toevoeging van de nieuwe variabele zelf niet-significant zijn geworden, ertoe leidt dat de toegevoegde variabele alsnog significant wordt. Wanneer één van deze zaken het geval is wordt getest of het resulterende model beter is dan het huidige model door de AIC-waarden te vergelijken. Gezien de definitie van de AIC (zie 'Fitmaten') wordt het kandidaat-model beschouwd als beter dan het huidige model indien de AIC-waarde afneemt met ten minste 2 per extra geschatte parameter. Nadat de best scorende variabele aan het model is toegevoegd wordt de procedure herhaald, net zo lang totdat het model niet meer verbeterd kan worden.

Naast de AIC is ook bekeken wat er met de aangepaste R-kwadraat gebeurt wanneer er variabelen aan het model toegevoegd worden. Hierbij moeten we echter voorzichtig zijn. Omdat we een grote dataset hebben kan het zo zijn dat wanneer er kenmerken worden toegevoegd die alleen gelden voor kleine groepen (bijvoorbeeld of ouders van een kind asielzoekers zijn of niet) dit niet zorgt voor verbetering van de aangepaste R-kwadraat; de verklaarde variantie neemt slechts een klein beetje toe. Dit betekent niet dat dit geen belangrijke toevoeging is. Voor scholen met een clustering van kinderen met deze kenmerken kan de variabele immers wel degelijk relevant zijn. AIC is een maat die hier beter mee om kan gaan. Voor de variabelen die gelden voor een kleine groep kijken we daarom eerst of de AIC afneemt.

Categoriale verklarende variabelen worden in het SEM-model meegenomen via dummyvariabelen. Bij variabelen met meer dan twee categorieën hebben we ervoor gekozen om steeds alle dummyvariabelen tegelijk wel of niet op te nemen in het model, ook als niet alle categorieën afzonderlijk een significant effect hadden op schoolprestatie. Op deze manier voorkomen we dat eventuele niet-significante categorieën impliciet worden samengevoegd met de referentiecategorie. Dit laatste zou een correcte interpretatie van het model lastiger maken.

Fitmaten

Om de fit van verschillende SEM-modellen te vergelijken gebruiken we Akaikes Informatie-Criterium (AIC) en het Bayesiaanse Informatie-Criterium (BIC):

$$\begin{aligned} \text{AIC} &= 2k - 2 \ln \hat{L}, \\ \text{BIC} &= k \ln n - 2 \ln \hat{L}. \end{aligned}$$

Hierbij is n het aantal waarnemingen in het onderzoeksbestand, k het aantal te schatten parameters en \hat{L} de waarde van de likelihood-functie voor het geschatte model.

Een lagere AIC- of BIC-waarde wijst op een betere modelfit. Doorgaans neemt de waarde van de term $-2 \ln \hat{L}$ af naarmate meer verklarende variabelen worden opgenomen in een model. Aan de andere kant neemt de kans op overfitten in dat geval ook toe: het model reproduceert dan voornamelijk toevallige eigenschappen van de steekproefdata. De AIC en BIC houden rekening met de kans op overfitten via een positieve term die toeneemt met het aantal parameters in het model (respectievelijk $2k$ en $k \ln n$), zodat complexere modellen in het nadeel zijn. Deze 'penalty-term' is groter bij de BIC dan bij de AIC.

Naast de AIC en BIC hebben we ook gekeken naar de verklaarde variantie in schoolprestaties. De aangepaste R-kwadraat meet de fractie verklaarde variantie, rekening houdend met de complexiteit van het model:

$$R^2 = 1 - \frac{n}{n - p - 1} \frac{\sigma_e^2}{\sigma_y^2}.$$

Hierbij zijn σ_e^2 en σ_y^2 respectievelijk de onverklaarde en totale variantie van schoolprestatie. Verder is n opnieuw het aantal waarnemingen en p het aantal verklarende variabelen dat is opgenomen in het model met een direct effect op schoolprestatie. In tegenstelling tot de gewone R-kwadraat kan de aangepaste R^2 afnemen bij toevoeging van meer variabelen aan het model, als de toegevoegde variabelen te weinig verklarende kracht hebben met betrekking tot schoolprestatie, gegeven de variabelen die al in het model zaten.

Ten slotte vermelden we bij de modellen inclusief intelligentie ook de partiële R-kwadraat van elk model, in vergelijking met een model dat alleen intelligentie bevat. Dit is de fractie van de onverklaarde variantie van schoolprestatie in een model met alleen intelligentie die alsnog verklaard wordt door het toevoegen van de overige kenmerken:

$$R_{part}^2 = \frac{R^2 - R_0^2}{1 - R_0^2},$$

met R^2 als boven en R_0^2 de aangepaste R-kwadraat van een model waarin alleen intelligentie is meegenomen. Aangezien intelligentie in het vervolg niet wordt gebruikt bij het bepalen van verwachte onderwijsachterstanden is de partiële R-kwadraat een interessante maat om de toegevoegde waarde van de achtergrondkenmerken in verschillende modellen te vergelijken.

5. Resultaten

5.1 Inleiding

In dit hoofdstuk presenteren we de resultaten van de verschillende modellen: de door OCW vooraf opgestelde modellen en het best passende model dat we hebben geschat volgens de onderzoeksmethode zoals beschreven in hoofdstuk 4. We beginnen met de uitkomsten van de modellen waarin voor de invloed van intelligentie is gecorrigeerd (paragraaf 5.2). Vervolgens bespreken we de uitkomsten van de modellen zonder intelligentie (paragraaf 5.3).

5.2 Modellen met intelligentie

Tabel 5.1 laat het best passende model zien volgens de stepwise-SEM-procedure.²¹ In dit model is de intelligentiemaat (nscct-score) opgenomen waarbij het effect van opleidingsniveau voor 50% is gecorrigeerd en het effect van herkomst voor 75% is gecorrigeerd (voor uitleg hierover, zie paragraaf 4.2.4). We hebben ervoor gekozen om het best passende model bij deze correctievariant hier te presenteren, omdat uit de literatuur blijkt dat de erfelijke component van intelligentie ongeveer tussen de 25 en 80% ligt en het in de lijn der verwachting ligt dat er sprake is van selectieve migratie naar Nederland. De resultaten voor de andere correctievarianten hebben we opgenomen in bijlage 8. Voor de volledigheid is in bijlage 11 de onderliggende correlatiematrix opgenomen van de variabelen in het model uit tabel 5.1.

Tabel 5.1: Resultaten best passende model met intelligentie, nscct-correctie 50%-75% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.819	0.202	0.030	27.069	<0.001
Cito groep 8	~ Opleiding vader	0.826	0.181	0.033	25.080	<0.001
Cito groep 8	~ Herkomst	-0.704	-0.071	0.109	-6.437	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.156	0.037	0.028	5.498	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8	~ Intelligentie	0.542	0.479	0.010	54.356	<0.001
Intelligentie	~ Opleiding moeder	0.331	0.092	0.037	8.842	<0.001
Intelligentie	~ Opleiding vader	0.325	0.081	0.040	8.123	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 181	-30 813	0.418	15 640

Ook als we de nscct in andere mate voor herkomst en opleidingsniveau corrigeren, worden dezelfde variabelen geselecteerd voor het best passende model. Verder blijft de modelfit nagenoeg onveranderd. We zien bij de uiteenlopende correcties vooral een verschuiving tussen de directe en indirecte effecten van de opleidingsniveaus op de Cito-score. Wanneer er een grotere correctie voor opleiding wordt gedaan op de nscct-score is het directe effect van opleidingsniveau op Cito-score groter en het indirecte effect (via intelligentie) kleiner, zoals men

²¹ Er is één model dat volgens de stepwise-procedure nog beter bij de data past: het model uit tabel 5.1 met de variabele 'vluchtelingenkind' in plaats van 'verblijfsduur moeder' (AIC = -40 281; BIC = -39 981; Adjusted R-squared = 0.417). De variabele 'vluchtelingenkind' is afgeleid uit informatie over het migratiemotief van de ouders. Deze informatie zal in de toekomst waarschijnlijk niet meer beschikbaar zijn in het SSB. Omwille van de toekomstbestendigheid van het analysemodel – en aangezien beide modellen elkaar weinig ontlopen wat modelfit betreft – hebben we ervoor gekozen om in dit rapport het model met verblijfsduur te presenteren als best passend model.

ook zou verwachten. In alle gevallen blijft het effect dat opleiding heeft op intelligentie significant: een deel van het effect van opleiding op Cito-score loopt dus via intelligentie.

Het gekozen model verklaart ongeveer 42 procent van de variantie in de Cito-scores in groep 8. De intelligentie van het kind (zoals gemeten door de nscct-score) blijkt van grote verklarende waarde te zijn voor zijn/haar schoolprestatie. Van de gezinskenmerken zijn er, naast de variabelen opleidingsniveau en herkomst die in het startmodel zitten, nog twee andere van meerwaarde voor het model: verblijfsduur van de moeder en het wel/niet hebben van een ouder in de schuldsanering. Daarnaast bevat het best passende model één schoolkenmerk: het gemiddelde opleidingsniveau van de moeders per school.

Van de variabelen naast intelligentie heeft het opleidingsniveau van de ouders duidelijk de sterkste invloed op Cito-scores in groep 8. Om het gevonden effect van opleidingsniveau in het model uit tabel 5.1 te duiden, zijn in tabel 5.2 de schaalwaarden voor de afzonderlijke opleidingscategorieën opgenomen van de afgeleide continue opleidingsvariabelen bij dit model. De positieve directe effecten van de twee opleidingsvariabelen op Cito-score (respectievelijk 0.819 en 0.826 voor moeders en vaders) laten zien dat lagere opleidingsniveaus samenhangen met lagere Cito-scores, en hogere opleidingsniveaus met hogere Cito-scores. Daarnaast hebben de opleidingsniveaus een indirect effect op Cito-score via intelligentie, dat eveneens positief is (respectievelijk 0.331 en 0.325 voor moeders en vaders). Dat wil zeggen: ouders met hogere opleidingsniveaus hebben kinderen met gemiddeld hogere intelligentiescores in groep 5, en kinderen met hogere intelligentiescores in groep 5 behalen op hun beurt gemiddeld hogere Cito-scores in groep 8.

Tabel 5.2: Coëfficiënten voor continue opleidingsvariabelen bij het model in tabel 5.1.

Categorie	Opleiding vader	Opleiding moeder
Basisonderwijs	-0.208	-0.101
vmbo-b/k, mbo1	-0.137	-0.179
Vmbo-g/t, avo onderbouw	0.071	0.141
Mbo2 en mbo3	0.000	0.000
Mbo4	0.181	0.175
Havo, vwo	0.355	0.379
Hbo-, wo-bachelor	0.348	0.492
Hbo-, wo-master, doctor	0.479	0.616

Herkomst (culturele afstand) heeft een negatief effect op Cito-score (-0.704). Tabel 5.3 toont de schaalwaarden van de afgeleide herkomstvariabele die in het model is opgenomen. Hieruit blijkt dat de Cito-scores van leerlingen negatief beïnvloed worden als de ouders een niet-Nederlandse herkomst hebben, tenzij het gaat om ouders uit de categorie Oost-Azië: kinderen uit deze laatste groep presteren gemiddeld juist iets beter dan kinderen van ouders met een Nederlandse herkomst.

Tabel 5.3: Coëfficiënten voor continue herkomstvariabele bij het model in tabel 5.1.

Categorie	Herkomst
Nederland	0.000
Suriname en Nederlandse Antillen	0.320
Turkije	0.322
Noord Afrika	0.154
EU-15, andere ontwikkelde economieën en eerste generatie Indonesiërs	0.181
Nieuwe EU-landen en economieën in transitie	0.090
Oost-Azië	-0.073
Rest Afrika, rest Azië, rest Latijns Amerika	0.146

Het hebben van (ten minste) één ouder die in de schuldsanering zit heeft een negatief effect op de Cito-score van een leerling (-0.270). Dit is conform onze verwachting. De relatie tussen verblijfsduur en Cito-score is daarentegen op het eerste gezicht minder logisch: het hebben van een moeder die korter dan 5 jaar of tussen 5 en 15 jaar in Nederland verblijft heeft een positief effect op de Cito-score in vergelijking met de referentiecategorie (het hebben van een moeder die 15 jaar of langer in Nederland woont). Dit effect is overigens alleen significant voor de categorie moeders met een verblijfsduur tussen 5 en 15 jaar (0.156), en niet voor de categorie met een kortere verblijfsduur (0.019).

Om dit tegen-intuïtieve effect te duiden is het nuttig om de onderliggende correlatiematrix in bijlage 11 te bekijken. Hier is te zien dat de één-op-één-relatie tussen verblijfsduur en Cito-score voor beide categorieën negatief is: leerlingen met ouders die hier korter dan 15 jaar wonen, hebben gemiddeld lagere Cito-scores, zoals verwacht. De positieve coëfficiënten uit tabel 5.1 ontstaan doordat in het SEM-model rekening wordt gehouden met andere kenmerken. De interpretatie van deze positieve coëfficiënten is dan dat leerlingen van ouders met een kortere verblijfsduur relatief hoge Cito-scores hebben, gegeven de overige kenmerken van hun ouders. Uit de correlatiematrix in bijlage 11 blijkt in het bijzonder dat ouders met een kortere verblijfsduur gemiddeld lagere opleidingsniveaus hebben dan ouders die langer dan 15 jaar in Nederland verblijven. Blijkbaar zijn de Cito-scores van kinderen in deze categorie relatief beter dan die van kinderen van Nederlandse ouders met een vergelijkbaar laag opleidingsniveau.

Tenslotte blijkt het gemiddelde opleidingsniveau van de moeders per school een positief effect te hebben op de Cito-scores van de leerlingen. Leerlingen op een school met relatief veel (weinig) hoogopgeleide ouders scoren dus gemiddeld beter (slechter) op de Cito-toets, los van de invloed van het opleidingsniveau van hun eigen ouders. Dit is het enige gezinsoverstijgende kenmerk dat door de stepwise-procedure is geselecteerd in het best passende model.²²

Om de robuustheid van het gevonden model te toetsen is gecontroleerd of het model consistent is over verschillende sub-samples. Daarbij wordt het onderzoeksbestand op basis van een random selectie in tweeën gedeeld. Vervolgens wordt de stepwise-SEM-procedure op beide halve datasets herhaald. In beide gevallen werden dezelfde variabelen geselecteerd in het best passende model. Hieruit blijkt dat de gevonden verbanden robuust zijn, en niet toevallig goed passen bij de steekproef in ons onderzoeksbestand.

Naast het model dat volgt uit een stepwise-SEM-aanpak zijn ook de in paragraaf 4.1 genoemde negen modellen geschat waarin OCW op voorhand geïnteresseerd was. De fitmaten hiervan zijn in tabel 5.4 uiteengezet; de geschatte coëfficiënten zijn te vinden in bijlage 6. Ter vergelijking is,

²² We hebben ook onderzocht of een stapeling van achtergrondkenmerken de schoolprestaties significant beïnvloedt, door interactie-effecten op te nemen in het model (bijvoorbeeld een interactie van herkomst en opleidingsniveau). Helaas blijken er te weinig kinderen met een stapeling van achtergrondkenmerken in onze data te zitten om deze effecten betrouwbaar te kunnen schatten.

om de partiële R-kwadraat van deze modellen te bepalen, ook een model geschat met alleen intelligentie. Dit model heeft een R-kwadraat van 0.288. Met andere woorden: intelligentie verklaart ongeveer 29 procent van de variantie in schoolprestatie. Het stepwise-SEM model uit tabel 5.1 heeft een partiële R-kwadraat van 0.181. De omgevingskenmerken die in dit model zijn meegenomen zijn dus in staat om ongeveer 18 procent te verklaren van het deel van de variantie in schoolprestatie dat niet verklaard wordt door intelligentie.

Uit tabel 5.4 blijkt dat naast het stepwise-SEM, ook het derde model de schoolprestatie (Cito-score groep 8) goed voorspelt. Dit model bevat alleen opleidingsniveau, herkomst en intelligentie als verklarende variabelen. Modellen met inkomen en buurtgegevens doen het over het algemeen een stuk slechter, deze modellen laten een lagere verklaarde variantie zien.

Tabel 5.4: Resultaten modellen met intelligentie, nscct correctie 50%-75% (n = 13 466).²³

Model	AIC	BIC	Adjusted R ²	Partial R ²	Log-likelihood
1 Opleiding	63 076	63 182	0.420	0.184	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.418	0.182	-41 734
3 Opleidingsniveau en herkomst	35 905	36 048	0.417	0.181	-17 934
4 Opleidingsniveau, inkomen en herkomst	51 406	51 594	0.417	0.181	-25 678
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 142	453 906	0.419	0.183	-225 836
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 127	469 057	0.414	0.177	-233 307
7 Inkomen (kwartielen)	109 878	110 028	0.381	0.130	-54 919
7a Inkomen (decielen)	104 508	105 086	0.381	0.130	-52 177
8 Inkomen, herkomst en buurtgegevens	480 840	482 289	0.373	0.118	-240 227
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 632	463 389	0.373	0.119	-230 582
10 Stepwise SEM (model uit tabel 5.1)	-31 181	-30 813	0.418	0.181	15 640

Wat verder opvalt is dat voor model 1 (alleen opleiding en intelligentie) de aangepaste én partiële R-kwadragen iets hoger zijn dan bij het model dat uiteindelijk is gekozen bij een stepwise-SEM-aanpak. Dit heeft te maken met de aanpassing die is gedaan aan de R-kwadraat statistiek; er is gecorrigeerd voor het aantal variabelen in het model. Zoals in paragraaf 4.3.2 is beredeneerd kan een variabele die een goede verklaringskracht heeft voor een kleine groep kinderen slechts een klein beetje verklaarde variantie toevoegen aan het totaal. De correctie voor het aantal toegevoegde variabelen is in dat geval groter dan de toename van de verklaarde variantie, dit resulteert in een lagere aangepaste R-kwadraat, terwijl het model wel beter bij de data past (volgens AIC en BIC). Dit betekent niet per se dat een model met minder variabelen ook inhoudelijk beter is.

5.3 Modellen zonder intelligentie

Alle modellen hebben we ook zonder intelligentie (nscct-score) als verklarende variabele geschat. Om deze modellen te kunnen vergelijken met die waarin intelligentie is opgenomen, wordt in deze modellen op dezelfde manier voor het steekproefontwerp gecorrigeerd als in de SEM-modellen. Deze correctie was nodig omdat uit eerdere analyse is gebleken dat er significante clustering op schoolniveau te vinden is in de data; zie paragraaf 4.2.5.

Tabel 5.5 laat de resultaten zien van het best passende model wanneer intelligentie wordt weggelaten. Er is te zien dat de verklaarde variantie ongeveer gehalveerd is in vergelijking met het SEM-model, maar nog steeds bijna 20 procent bedraagt.

²³ Voor fitmaten bij andere nscct-correcties, zie bijlage 7. Bij model 7 met alleen inkomen (en intelligentie) is inkomen steeds in kwartielen meegenomen. Voor de variant met een 50%-75% correctie op nscct die in de hoofdttekst wordt gepresenteerd hebben we ook een model geschat met inkomen in decielen (model 7a).

Ook voor de overige modellen zijn de analyses zonder intelligentie uitgevoerd. De fitmaten voor deze modellen staan in tabel 5.6; de geschatte coëfficiënten zijn te vinden in bijlage 9. De modellen die herkomst en opleiding als verklarende variabelen bevatten (model 3, 4, 5 en 10) presteren beter dan de overige modellen. De modellen met inkomen in plaats van opleiding laten een daling zien van de verklaarde variantie en een stijging van de AIC. De verklaarde variantie bedraagt 19 à 20 procent voor de beste modellen zonder intelligentie. Een model met alleen inkomen verklaart ongeveer 8 procent van de variantie in schoolprestaties.

Tabel 5.5: Resultaten best passende model zonder intelligentie (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	0.952	0.234	0.037	25.547	<0.001
Cito groep 8 ~	Opleiding vader	0.969	0.212	0.039	25.010	<0.001
Cito groep 8 ~	Herkomst	-0.805	-0.082	0.117	-6.909	<0.001
Cito groep 8 ~	Gemiddelde opleiding school	0.479	0.050	0.126	3.787	<0.001
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	-0.031	-0.004	0.061	-0.514	0.608
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.201	0.047	0.033	6.046	<0.001
Cito groep 8 ~	Ouder(s) in schuldsanering	-0.357	-0.040	0.071	-5.028	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			35 340	35 400	0.198	-17 662

Tabel 5.6: Resultaten modellen zonder intelligentie (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	35 528	35 551	0.187	-17 761
2 Opleidingsniveau en inkomen	35 467	35 497	0.190	-17 730
3 Opleidingsniveau en herkomst	35 414	35 444	0.194	-17 703
4 Opleidingsniveau, inkomen en herkomst	35 372	35 409	0.196	-17 681
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	35 365	35 515	0.198	-17 663
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	35 421	35 579	0.194	-17 689
7 Inkomen (kwartielen)	37 131	37 161	0.084	-18 562
7a Inkomen (decielen)	37 303	37 378	0.088	-18 641
8 Inkomen, herkomst en buurtgegevens	36 905	37 040	0.100	-18 435
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	36 874	37 024	0.102	-18 417
10 Stepwise-procedure (model uit tabel 5.5)	35 340	35 400	0.198	-17 662

6. Conclusies en vervolg

6.1 Conclusies

In het kader van een mogelijke herziening van het onderwijsachterstandenbeleid, heeft het ministerie van OCW CBS gevraagd een betere indicator te ontwikkelen voor verwachte onderwijsachterstanden. Om aan dit verzoek te voldoen, is onderzocht welke (combinatie van) kenmerken – die kunnen worden afgeleid uit registraties die aanwezig zijn bij CBS – onderwijsachterstanden in het primair onderwijs het beste voorspellen.

Het opleidingsniveau van ouders is naar verwachting een belangrijke voorspeller van onderwijsachterstanden. De registratie van opleidingsniveaus is, anders dan de overige gebruikte registraties, echter incompleet en selectief onbekend naar leeftijd, herkomst en opleidingsniveau. Deze selectiviteit zou ertoe kunnen leiden dat resultaten vertekenen zodra we opleidingsniveau als één van de voorspellers van onderwijsprestaties opnemen. We hebben daarom allereerst onderzocht of we hiervoor kunnen corrigeren door te imputeren. Omdat bij de gewichtenregeling financiële middelen op schoolniveau verdeeld worden, is het bij de correctie niet zozeer van belang dat we voor iedere ouder het juiste opleidingsniveau weten, maar dat we per school de samenstelling van de opleidingsniveaus van ouders juist meten. Na imputatie van onbekende opleidingsniveaus met behulp van een continuation-ratio-model blijkt het inderdaad mogelijk om op schoolniveau (uitgezonderd de allerkleinste scholen) opleidingsniveaus voldoende betrouwbaar en zonder noemenswaardige vertekening te meten. Hierdoor is het niet langer nodig om dit door scholen zelf te laten registreren en vervolgens te laten controleren.

Verder blijkt uit het onderzoek dat er naast het opleidingsniveau van de ouders andere kenmerken zijn die invloed hebben op schoolprestaties. Uit onze analyses blijkt dat schoolprestaties, gemeten als Cito-scores in groep 8, het best verklaard kunnen worden wanneer naast het opleidingsniveau van ouders ook rekening wordt gehouden met hun herkomst, verblijfsduur, of zij in de schuldsanering zitten en het gemiddelde opleidingsniveau op school. Als we verder rekening houden met de intelligentie van leerlingen, waardoor we voor hun leerpotentie corrigeren, kunnen we 42 procent van de verschillen in de Cito-scores in groep 8 verklaren. Dat is in vergelijking met andere sociaalwetenschappelijke studies hoog.²⁴

Omdat we door intelligentie mee te nemen de invloed van omgevingskenmerken op leerprestaties zo zuiver mogelijk schatten, geven we de voorkeur aan het beste passende model dat voor intelligentie corrigeert. Daarbij houden we er rekening mee dat intelligentie niet zonder de invloed van de sociale omgeving kan worden gemeten. Namelijk door de helft van de invloed van opleidingsniveau en driekwart van de invloed van herkomst op intelligentie te verwijderen (zie ook paragraaf 4.2.4 en bijlage 5).

In het best passende model houden we niet alleen rekening met intelligentie, maar ook met het gegeven dat kinderen van eenzelfde school meer overeenkomsten vertonen dan twee willekeurig gekozen kinderen. Daarnaast is onderzocht of kenmerken van buurten en scholen in aanvulling op de kenmerken van het gezin of het kind van invloed zijn op leerprestaties. Alleen het gemiddelde opleidingsniveau op school blijkt van invloed.

Naast het beste passende model, zijn er ook nog andere modellen met behoorlijke verklarende kracht. Van de vooraf door het ministerie van OCW gesuggereerde modellen, blijkt het model met daarin herkomst en opleiding het meest geschikt om onderwijsprestaties te verklaren.

²⁴ Het is ook ongeveer vier keer zo hoog als wanneer de schoolprestaties alleen door de kenmerken die in de huidige gewichtenregeling worden gebruikt worden verklaard (Driessen, 2015). Dit verschil komt met name doordat Driessen niet corrigeert voor de nscct-score. De verklaarde variantie in schoolprestatie door intelligentie is in ons onderzoek bijna 29 procent. Van de resterende onverklaarde variantie wordt ongeveer 18 procent verklaard door de overige kenmerken in ons model.

Bij alle uitkomsten moet worden opgemerkt dat zij betrekking hebben op de Cito-scores in groep 8 en dat het huidige onderwijsachterstandenbeleid hier dus op van invloed kan zijn geweest. Deze potentiële invloed was minder groot geweest wanneer we gebruik hadden kunnen maken van een vroegere en kwalitatief hoogwaardige meting van schoolprestaties. Dit bleek in dit onderzoek echter niet mogelijk.

6.2 Vervolg

Op basis van de uitkomsten die in dit rapport zijn opgetekend en de adviezen van de begeleidingscommissie, beslist het ministerie van OCW op basis van welke modellen zij de gewichtenregeling mogelijk wil herzien. Die analysemodellen zal CBS in de volgende onderzoeksfase vertalen naar voorspel- en verdeelmodellen die de verwachte onderwijsachterstanden per school schatten. Hieruit zal blijken in welke mate het gebruik van verschillende analysemodellen leidt tot uiteenlopende schattingen van de onderwijsachterstanden op schoolniveau. Er zijn verschillende manieren om de achterstanden op schoolniveau te schatten. In het vervolgonderzoek zullen maximaal drie schattingswijzen worden uitgewerkt en onderzocht. Anders dan het analysemodel, worden de schattingen uitgevoerd voor alle scholen en leerlingen en niet alleen voor de scholen en leerlingen die voorkomen in het onderzoeksbestand dat is gebruikt voor het analysemodel. Ook in deze onderzoeksfase worden CBS en OCW weer geadviseerd door de begeleidingscommissie. Na afronding van deze fase kan het ministerie van OCW een beredeneerde keuze maken voor de herziening van het beleid. Dit is een politieke keuze en CBS en de begeleidingscommissie hebben hier dan ook geen stem in.

Bijlage 1. Gebruikte literatuur

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons: New York.

van Batenburg, T.A. (2015). Niet Schoolse Cognitieve Capaciteiten Test. Verantwoording, normering en handleiding. Gronings Instituut voor Onderzoek van Onderwijs, Rijksuniversiteit Groningen.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons: New York.

CBS Statline. (2016). Bevolking; hoogst behaald onderwijsniveau; geslacht, leeftijd en herkomst. <http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=82275NED&D1=0&D2=0&D3=0&D4=1,4-5&D5=0,4-5,8,12-13&D6=44,49,54,59,I&HDR=T,G1,G3,G5&STB=G2,G4&VW=T>.

Driessen, G. (2001). Ethnicity, forms of capital and educational achievement. *International Review of Education*, vol. 47, 513–538.

Driessen, G. (2015). De wankle empirische basis van het onderwijsachterstandenbeleid. *Mens & Maatschappij*, vol. 90, 221–243.

Driessen, G., Mulder, L., Ledoux, G., Roeleveld, J. en van der Veen, I. (2009). Cohortonderzoek COOL⁵⁻¹⁸. Technisch rapport basisonderwijs, eerste meting 2007/08. Stichting Instituut voor Toegepaste Sociologie (ITS), Radboud Universiteit Nijmegen.

Driessen, G. en van Langen, A. (2014). De onderwijsachterstand van jongens. Omvang, oorzaken en interventies. Stichting Instituut voor Toegepaste Sociologie (ITS), Radboud Universiteit Nijmegen en Kohnstamm Instituut Amsterdam.

van Grastek, S. en Lems, M. (2014). Cognitieve ontwikkeling kleuter en schoolkind. <http://www.cognitieve-ontwikkeling.nl>

Hanscombe, K.B., Trzaskowski, M., Haworth, C.M.A., Davis, O.S.P., Dale, P.S. en Plomin, R. (2012). Socioeconomic status (SES) and children's intelligence (IQ): in a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. *PLoS ONE*, vol. 7, 1-16.

Hart, S., Petrill, S., Deater-Deckard, K.D. en Thompson, L. (2007). SES and CHAOS as environmental mediators of cognitive ability: a longitudinal analysis. *Florida State University Libraries. Faculty Publications*. 1-38.

van Hoorn, W., van der Kamp, L., den Brinker, W., ter Laak, J. en de Groot, M. (2005). Nederlandse differentiatie testserie, vragenlijst sociaal emotioneel-functioneren. Handleiding. Amsterdam: Harcourt Test Publishers.

Houtveen, A.A.M., van der Velde, V. en van Niekerk, J.C.A. (2011). Opbrengst van taalbeleid: eindrapportage pilots taalbeleid onderwijsachterstanden. Utrecht: Hogeschool Utrecht/Zuidam en Uithof Utrecht BV.

Kloprogge, J. en de Wit, W. (2015). Het onderwijsachterstandenbeleid na 2015. Literatuurstudie t.b.v. expertbijeenkomst OAB september 2015. Nationaal Regieorgaan Onderwijsonderzoek.

Kooistra, J.P., Ultee, W. en Pelzer, B. (2008). Het effect van taalafstand en koloniaal verleden op leerprestaties. *Mens & Maatschappij*, vol. 93, 258-278.

Ledoux, G., Roeleveld, J., Veen, A., Karssen, M., van Daalen, M., Blok, H., Kuiper, E. en Dijkers, L. (2015). Het onderwijsachterstandenbeleid onderzocht. Werkt het zoals bedoeld? Stichting Instituut voor Toegepaste Sociologie (ITS), Radboud Universiteit Nijmegen en Kohnstamm Instituut Amsterdam.

Leseman, P.P.M. (2007). Achterstandenbeleid: voorbij de voor- en voerschoolse periode. In P.A.H. van Lieshout, M.S.S. van der Meij & J.C.I. de Pree (Red.), *Bouwstenen voor betrokken jeugdbeleid*. Amsterdam: Amsterdam University Press.

Linder, F., van Roon, D. en Bakker, B.F.M. (2011). Combining data from administrative sources and sample surveys; the single variable case. In: *ESSnet Data Integration. WP4 Case Studies*. Luxembourg: Eurostat. 39-97.

Lynn, R. (1983). IQ in Japan and the United States. *Nature*, 306, 291–292.

Lynn, R. en Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.

Helms-Lorenz, M., van de Vijver, F. J. R. en Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c? *Intelligence*, 31, 9–29.

Meijnen, G.W. (2003). *Onderwijsachterstanden in basisscholen*. Leuven: Garant.

Mesman, J. (2011). Oud geleerd, jong gedaan. *Het Jonge Kind*, 11, 9-12.

Morrison, E.F., Rimm-Kauffman, S. en Pianta, R.C. (2003). A longitudinal study of mother–child interactions at school entry and social and academic outcomes in middle school. *Journal of School Psychology*, vol 41, 185–200.

Mulder, L., Fettelaar, D., Schouwenaars, I., Ledoux, G., Dijkers, L. en Kuiper, E. (2014). De achterstand van autochtone doelgroepopleerlingen. Oorzaken en aanpak. Stichting Instituut voor Toegepaste Sociologie (ITS), Radboud Universiteit Nijmegen.

Murray, C. (1998). Income inequality and IQ. ED.GOV Institute of Education Sciences.

Neisser, U., Boodoo, G., Bouchard, T. J. J., Boykin, A. W., Brody, N., Ceci Diane, S. J., Halpern, E., Loehlin, J.C., Perloff, R., Sternberg, J. en Sternberg, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51, 277–101.

Ogbu, J.U. (1994). Culture and intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* New York: Macmillan. 328–338.

Paas, T., Mulder, L. en Roeleveld, J. (2013). Zittenblijvers en verwezen leerlingen in het cohortonderzoek COOL⁵⁻¹⁸. Stichting Instituut voor Toegepaste Sociologie (ITS), Radboud Universiteit Nijmegen.

Resing, W.C.M., Bleichrodt, N. en Drenth, P.J.D. (1986). Het gebruik van de RAKIT bij allochtoon etnische groepen. *Nederlands Tijdschrift voor de Psychologie*, vol. 41, 179–188.

Resing, W.C.M. en Blok, J.B. (2002). De classificatie van intelligentiescores. Voorstel voor een eenduidig systeem. *De Psycholoog*, vol. 37, 244-249.

Resing, W. en Drenth, P. (2007). *Intelligentie: weten en meten*. Amsterdam: Uitgeverij Nieuwezijds B.V.

Reynders, T., Nicaise, I. en Van Damme, J. (2005). Longitudinaal onderzoek in het basisonderwijs. De constructie van een SES-variabele voor het SiBO-onderzoek. *Schoolloopbanen in het BasisOnderwijs (SIBO)*, Leuven.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, vol. 21, 667–706.

van Rooijen, K. en Zoon, M. (2012). Risicofactoren voor onderwijsachterstanden. Nederlands Jeugd Instituut:
http://www.nji.nl/nji/dossierDownloads/Risicofactoren_Onderwijsachterstanden.pdf.

Scholtus, S. en Pannekoek, J. (2015). Massa-imputatie van opleidingsniveaus. Centraal Bureau voor de Statistiek, Den Haag.

Sirin, S.R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, vol. 75, 417–453.

Snijders, T.A.B. en Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*. London: Sage Publishers.

Thiel, Van, E. en Engelen, S. (2016). IQ scores en verschillen tussen groepen.
<https://www.123test.nl>.

van Tubergen, F. en Kalmijn, M. (2005). Destination-language proficiency in cross-national perspective: a study of immigrant groups in nine Western countries. *American Journal of Sociology*, vol. 110, 1412-1457.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B. en Gottesman, I.I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, vol. 14, 623-628.

UN (2015). World Economic Situation and Prospects 2015.
http://www.un.org/en/development/desa/policy/wesp/wesp_archive/2015wesp_full_en.pdf

Veen, A., Karssen, A.M., van Daalen, M.M., Roeleveld, J., Triesscheijn, J. en Elshof, D. (2013). De aansluiting tussen voor- en voerschoolse educatie en tussen voerschoolse educatie en groep 3. Kohnstamm Instituut Amsterdam.

Verboon, P. (1998). Massa-imputatie via lineaire modellen. Centraal Bureau voor de Statistiek, Den Haag.

Bijlage 2. Databronnen

In dit onderzoek is gebruik gemaakt van drie bronnen: COOL⁵⁻¹⁸, het Stelsel van Sociaal-Statistische Bestanden (SSB) en de Leefbarometer. De inhoud van deze bestanden wordt in deze bijlage kort toegelicht.

2.1 COOL5-18

Het COOL⁵⁻¹⁸ onderzoek wordt in dit onderzoek gebruikt als bron voor gegevens over niet schoolse cognitieve capaciteiten, thuistaal en zorgleerlingen.

COOL⁵⁻¹⁸ staat voor Cohort Onderzoek OnderwijsLoopbanen onder leerlingen van 5 tot 18 jaar. Dit onderzoek volgt leerlingen van 5 tot 18 jaar in hun schoolloopbaan door het primair en voortgezet onderwijs en het mbo. In het kader van dit onderzoek worden op diverse momenten toetsen en vragenlijsten afgenomen. Daarnaast wordt de gehele schoolloopbaan van de leerlingen in kaart gebracht.

COOL⁵⁻¹⁸ wordt uitgevoerd door een consortium. Het ITS en Kohnstamm Instituut zijn verantwoordelijk voor het primair onderwijs gedeelte en het Cito en GION zijn verantwoordelijk voor het voortgezet onderwijs en MBO gedeelte. CBS is verantwoordelijk voor de koppeling van de gegevens aan de onderwijsnummerbestanden. Het onderzoek wordt gesubsidieerd door NRO/PROO.

Er zijn drie ronden van gegevensverzameling geweest. De eerste ronde (2007-2008) richtte zich op leerlingen uit groep 2, 5 en 8 van het primair onderwijs en klas 3 van het voortgezet onderwijs. In de tweede en derde ronde (2010-2011 en 2013-2014) zijn deze groepen opnieuw onderzocht. Aanvullend zijn ook de leerlingen in de bovenbouw van het Havo en Vwo en in het MBO onderzocht. In elke ronde nemen er ongeveer 50 000 leerlingen in het primair onderwijs deel en ongeveer 20 000 leerlingen in het derde leerjaar van het voortgezet onderwijs.

Een van de testen die onder leerlingen in groep 5 is uitgevoerd in het kader van de COOL⁵⁻¹⁸ is de niet schoolse cognitieve capaciteiten test (nscct). De populatie van dit onderzoek bestaat uit de leerlingen die zowel de nscct als de cito-toets in groep 8 hebben gemaakt (zie ook bijlage 10). Omdat COOL⁵⁻¹⁸ de grootste verzameling van nscct scores bevat, betekent dit dat de populatie van dit onderzoek zich beperkt tot een deel van de COOL⁵⁻¹⁸ deelnemers. We maken niet alleen gebruik van de COOL⁵⁻¹⁸ voor informatie over de nscct, maar ook voor informatie over de thuistaal van leerlingen en of zij een zorgleerling zijn. Alle overige informatie komt uit de bronnen die we hierna bespreken.

Uitgebreidere informatie over het de COOL⁵⁻¹⁸ onderzoek is te vinden op haar eigen [website](#). Zie ook de technische rapporten²⁵ voor meer informatie.

²⁵ Zie hier de links naar de technische rapporten: [rapport onderzoek 1](#), [rapport onderzoek 2](#), [rapport onderzoek 3](#).

2.2 SSB

Het SSB is door CBS ontwikkeld en bevat geanonimiseerde microdata over sociaaleconomische en ruimtelijke statistieken. De microdatabestanden die voor dit onderzoek uit het SSB gebruikt zijn, worden hieronder kort toegelicht.

Kenmerken van deelnemers aan de Eindtoets Basisonderwijs van Cito

Het Citobestand bevat kenmerken van deelnemers aan de Eindtoets Basisonderwijs van Cito, zoals het soort toets waaraan ze hebben deelgenomen en hun thuistaal. Verder bevat het bestand verschillende soorten scores van de deelnemers op de vier onderdelen van de Citotoets.

Gemeentelijke Basisadministratie (GBA)

Dit bestand bevat van alle personen die vanaf 1 januari 1995 in de gemeentelijke bevolkingsregisters voorkomen de demografische achtergrondgegevens (bijvoorbeeld geslacht, geboortjaar, burgerlijke staat, herkomstgroepering, verblijfsduur, overlijden) die niet of nauwelijks wijzigen.

Door het CIZ afgegeven indicatie voor AWBZ-zorg

CBS beschikt over een bestand met gegevens over indicaties die zijn afgegeven door het Centrum Indicatiestelling Zorg (CIZ). Een indicatie van het CIZ geeft toegang tot zorg die wordt gefinancierd uit de Algemene Wet Bijzondere Ziektekosten(AWBZ). Deze gegevens worden in dit onderzoek gebruikt als indicatie voor het gebruik van langdurige zorg.

Zorgkosten van Nederlandse ingezetenen die een basisverzekering hebben

CBS beschikt via Vektis over een bestand dat per Nederlandse ingezetene, die verzekerd is via de basisverzekering, diens kosten per jaar voor zorg die verzekerd is via de basisverzekering bevat. De basisverzekering is wettelijk verplicht via de Zorgverzekeringswet (Zvw) voor vrijwel alle Nederlandse ingezetenen. De kosten zijn die kosten die daadwerkelijk vergoed zijn door de zorgverzekeraars. De kosten zijn onderverdeeld in zorgvormen, zoals huisartsenzorg, ziekenhuiszorg, fysiotherapie, etc. in dit onderzoek wordt in het bijzonder gekeken naar de aanwezigheid van kosten voor eerste- en/of tweedelijns GGZ. De zorgkosten zijn inclusief de kosten die vanwege het verplicht of vrijwillig eigen risico uiteindelijk door de verzekerden zelf zijn betaald. Alleen indien de verzekerde zelf een rekening heeft ontvangen en deze niet heeft ingediend bij de verzekering, bijvoorbeeld omdat het eigen risico niet is bereikt, zijn de kosten niet in de cijfers opgenomen.

Migratiemotieven van in de bevolkingsregisters van Nederlandse gemeenten ingeschreven immigranten met een buitenlandse nationaliteit

CBS beschikt over een bestand van de Immigratie- en Naturalisatiedienst (IND) met het migratiemotief en het afgeleid migratiedoel, afgeleid op basis van het feitelijk geregistreerde gedrag van de immigrant uit informatie in het Stelsel van Sociaal Statistische Bestanden. Bij de IND wordt per migratiebeweging maar één motief vastgelegd. Er worden 13 verschillende motieven onderscheiden. Voor de afgeleide migratiedoelen geldt dat er meerdere doelen tegelijkertijd kunnen voorkomen. De volgende afgeleide doelen worden onderscheiden: - gezinsmigratie; - arbeidsmigratie; - asielmigratie; - studiemigratie.

Inkomen van huishoudens en personen

De inkomensbestanden zijn voornamelijk gebaseerd op registers afkomstig van de Belastingdienst (fiscaal basisregister) en de bevolkingsregisters van gemeenten (GBA). Daarnaast worden gegevens verzameld uit de registraties van de studiefinanciering van de Dienst Uitvoering Onderwijs (DUO).

Opleidingsniveaubestand

Deze bron bevat alleen records van personen waarvan de hoogst behaalde opleiding bekend is. Deze informatie komt voor een deel van de personen uit registraties en voor een andere deel uit enquêteonderzoek dat is gebaseerd op steekproeven. Doordat onderwijsregistraties nog niet zo heel lang bestaan, ontbreekt registerinformatie over veel oudere Nederlanders. Ook particuliere opleidingen vallen buiten het bereik van de onderwijsregisters. Daarnaast is van veel allochtonen die hun opleiding in het buitenland hebben gevolgd op dit moment (nog) onvoldoende betrouwbare registerinformatie beschikbaar. Dit betekent dat er voor het samenstellen van het Opleidingsniveaubestand een aanvullend beroep moet worden gedaan op gegevens uit enquêteonderzoek, waarvan de Enquête Beroepsbevolking (EBB) de belangrijkste is. Anders dan bij de onderwijsregistraties zijn de gegevens van de EBB alleen op steekproefbasis beschikbaar.

Personen sociaaleconomische categorie

Dit bestand wordt door CBS samengesteld op basis van bestanden over verschillende inkomensbronnen. In het bestand wordt aangegeven tot welke sociaaleconomische categorie iemand behoort. De onderscheiden categorieën zijn: werknemer, directeur-groootaandeelhouder, zelfstandige, overig actief, ontvanger werkloosheidsuitkering, ontvanger bijstandsuitkering, ontvanger uitkering sociale voorziening overig, ontvanger ziektewet/AO, ontvanger pensioenuitkering, nog niet schoolgaand/schoolgaand/studerend met inkomen, nog niet schoolgaand/schoolgaand/studerend zonder inkomen, overig zonder inkomen. Om de sociaaleconomische categorie te bepalen worden alle inkomsten in de verslagmaand uit de verschillende inkomensbronnen die iemand heeft, met elkaar vergeleken. Het hoogste bedrag is in principe bepalend voor de sociaaleconomische categorie. Daarnaast wordt meegenomen of een persoon ingeschreven staat bij een onderwijsinstelling.

Delictkenmerken van personen tegen wie een proces-verbaal van misdrijf is opgemaakt

Dit bestand bevat een selectie van variabelen uit het herkenningssysteem (HKS) van de Koninklijke Landelijke Politie Dienst (KLPD), waarin alle verdachten van misdrijven worden geregistreerd tegen wie een proces-verbaal van misdrijf is opgemaakt. Van alle verdachten is opgenomen hoe vaak er proces verbaal van misdrijf is opgemaakt en voor welke delictgroep.

Personen die een schuldsanering hebben via de Wet schuldsanering natuurlijke personen

Dit bestand bevat alle natuurlijke personen die in de schuldsanering zitten of hebben gezeten. Op grond van de Wet schuldsanering natuurlijke personen (Wsnp) krijgen natuurlijke personen die in ernstige betalingsmoeilijkheden verkeren een saneringsregeling. Met de Wsnp wordt beoogd te voorkomen dat natuurlijke personen tot in lengte van dagen door schuldeisers achtervolgd worden wanneer zij in een problematisch financiële situatie zijn terechtgekomen. De Raad voor rechtsbijstand, Bureau Wsnp levert de gegevens uit de Centrale Database Schuldsanering (CDS; deze database bevat eveneens gegevens van de 285-verklaringen).

Overzicht van alle postcodes met bijbehorende stedelijkheidscategorie en omgevingsadressendichtheid

Het Geografisch Basisregister bevat alle adressen van Nederland die zijn voorzien van een postcode, gemeentecode en wijk- en buurtcode. Dit wordt gebruikt om de omgevingsadressendichtheid (OAD) van buurten, wijken en gemeenten te bepalen. Dit wordt gedaan door het gemiddeld aantal adressen per vierkante kilometer binnen een cirkel met een straal van één kilometer op 1 januari. De OAD beoogt de mate van concentratie van menselijke activiteiten (wonen, werken, schoolgaan, winkelen, uitgaan etc.) weer te geven. Het CBS gebruikt de OAD om de stedelijkheid van een bepaald gebied te bepalen. Voor de berekening hiervan wordt eerst voor ieder adres de OAD vastgesteld. Daarna is het gemiddelde berekend van de omgevingsadressendichtheden van alle afzonderlijke adressen binnen het beschouwde gebied.

Personen met een bijstandsuitkering

Bestand dat alle personen bevat die in een bepaald jaar een uitkering op grond van de (geldende) Bijstandswet (BIJSTAND (incl. WIJ in 2009/2010/2011), BBZ, IOAW, IOAZ, WWIK en bijstand aan eldersverzorgden en adreslozen) hebben ontvangen (exclusief nabetalingsaan personen die gedurende het jaar geen recht op uitkering hadden). Van deze personen is door middel van een tijdsleutel aangegeven op welke dagen in het jaar de persoon een uitkering had en welk(e) soort(en) bijstandsuitkering het dan betreft met hun aanvangsdatum. BBZ staat voor Besluit bijstandverlening zelfstandigen, IOAW staat voor Wet inkomensvoorziening oudere en gedeeltelijk arbeidsongeschikte werkloze werknemers, IOAZ staat voor Wet inkomensvoorziening oudere en gedeeltelijk arbeidsongeschikte gewezen zelfstandigen, WWIK staat voor Wet werk en inkomen kunstenaars.

Waarde van onroerende zaken

Dit bestand bevat de WOZ-waarden van alle WOZ-objecten in Nederland, met uitzondering van de WOZ-objecten die getypeerd zijn als "uitgezonderd gebouwd object" (gebruikscategorie 31). WOZ-objecten die voor de onroerend zaakbelasting (ozb) verplicht zijn vrijgesteld (onder meer infrastructurele werken en kerken) zijn in het algemeen niet meegenomen, terwijl WOZ-objecten die vanwege een facultatieve gemeentelijke belastingverordening voor de OZB zijn vrijgesteld (zoals gemeentelijke gebouwen en kassen) wel zijn meegenomen. Het WOZ-bestand bevat ook informatie over de eigendom van woningen.

2.3 Leefbaarometer

De Leefbaarometer is een monitoringsinstrument van het Ministerie van Binnenlandse Zaken om de leefbaarheidssituatie in buurten en wijken in kaart te brengen. De leefbaarheidssituatie is gebaseerd op de beleving van de leefbaarheid van bewoners en woongedrag. Op basis van een groot aantal kenmerken op diverse terreinen (woningvoorraad, publieke ruimte, voorzieningen, bevolkingssamenstelling, levensopbouw en sociale samenhang van de bevolking, overlast en onveiligheid) wordt per postcode 6 gebied een leefbaarheidsscore berekend. De berekende scores voor 2008, 2012 en 2014 worden in dit onderzoek gebruikt als voorspellers van de schoolprestatie van leerlingen. Voor meer informatie over de Leefbaarometer zie de [website](#).

Bijlage 3. Variabelen en operationalisering

Variabelen

Voor alle variabelen is het peilmoment 1 september van het jaar dat de COOL steekproef is afgenomen bij de leerlingen (2007, 2011, 2013).

Identificatoren:

1. rinpersoons
2. rinpersoon
3. schoolnummer
4. buurt

Schoolprestaties:

1. Cito-scores Eindtoets Basisonderwijs
2. nscct-score

Kenmerken van kind/ouders/huishouden:

3. herkomstgroepering (semi-continu)
4. geslacht
5. huishoudenstype
6. generatie
7. vluchtelingenkind
8. leerling ontvangt zorg
9. leerling is een zorgleerling
10. hoogst behaalde opleiding moeder (semi-continu)
11. hoogst behaalde opleiding vader (semi-continu)
12. generatie moeder
13. generatie vader
14. ontbrekende moeder
15. ontbrekende vader
16. moeder overleden
17. vader overleden
18. inkomstenbron moeder
19. inkomstenbron vader
20. bruto huishoudinkomen
21. persoonlijk bruto inkomen moeder (continu en discreet)
22. persoonlijk bruto inkomen vader (continu en discreet)
23. inkomen juridische ouders (continu en discreet)
24. verblijfsduur moeder
25. verblijfsduur vader
26. burgerlijke staat (gescheiden) moeder
27. burgerlijke staat (gescheiden) vader
28. ouders zijn niet-samenwonend
29. leeftijd moeder
30. leeftijd vader
31. moeder is ooit verdachte geweest
32. vader is ooit verdachte geweest
33. moeder ontvanger schuldhulpverlening (in de zin van de WSNP) op peilmoment
34. vader ontvanger schuldhulpverlening (in de zin van de WSNP) op peilmoment
35. thuistaal

Schoolkenmerken (op basis van populatie in het onderzoeksbestand):

36. aandeel niet-westerse leerlingen

- 37. aandeel leerlingen met ouders met een lage opleiding
- 38. aandeel leerlingen met ouders met een hoge opleiding
- 39. gemiddelde opleidingsniveau (op basis van opleiding moeder)
- 40. aandeel leerlingen met ouders met een laag inkomen

Buurtkenmerken:

- 41. stedelijkheid
- 42. aandeel niet-westerse alloctonen
- 43. aandeel minima
- 44. aandeel huishoudens met lage inkomens
- 45. aandeel personen met bijstand
- 46. gemiddelde WOZ-waarde
- 47. aandeel huurwoningen
- 48. leefbaarheid

Operationalisering

Schoolprestaties

Cito-score Eindtoets Basisonderwijs

1: Continu (getransformeerd tot z-scores)

Niet-schoolse cognitieve capaciteiten toets

1: Continu (getransformeerd tot z-scores)

Kenmerken van kind/ouders/huishouden

Herkomstgroepering (juridische ouders; semi-continue variabele)

1: Autochtoon

2: Suriname en Nederlandse Antillen

3: Turkije

4: Noord-Afrika

5: EU-15, andere ontwikkelde economieën en tweede generatie Indonesiërs

6: Nieuwe EU-landen en economieën in transitie
--

7: Oost-Azië

8: Rest Afrika, rest Azië, rest Latijns Amerika

Generatie (juridische ouder(s))

1: Autochtoon

2: Eerste generatie allochtoon

3: Tweede generatie allochtoon

Geslacht (kind)

1: Man

2: Vrouw

Huishoudenstype (huishouden van kind)
1: Eenpersoonshuishouden 2: Niet-gehuwd paar zonder kinderen 3: Gehuwd paar zonder kinderen 4: Niet-gehuwd paar met kinderen 5: Gehuwd paar met kinderen 6: Eenouderhuishouden 7: Overig huishouden 8: Institutioneel huishouden

Vluchteling (kind) (migratie motief juridische ouder(s): asiel)
1: Ja 2: Nee

Zorgleerling (kind, uit COOL)
1: Ja 2: Nee

Ontvangt zorg (kind ontvangt eerste- en/of tweedelijns GGZ)
1: Ja 2: Nee

Opleidingsniveau (juridische ouders; semi-continue variabele)
1: Basisonderwijs 2: Vmbo-b/k, mbo1 3: Vmbo-g/t, avo onderbouw 4: Mbo2 en mbo3 5: Mbo 4 6: Havo, vwo 7: Hbo-, wo-bachelor 8: Hbo-, wo-master, doctor

Ontbrekende juridische ouder(s)
1: Ja 2: Nee

Overleden juridische ouder(s)
1: Ja 2: Nee

Inkomstenbron (SECM) (juridische ouder(s))
1: Inkomen uit werk 2: Zonder inkomen of met uitkering

Bruto huishoudinkomen, continu
Inkomen van het huishouden

Persoonlijk bruto inkomen (juridische ouder(s)) , kwartielen
1: [0,25%) 2: [25,50%) 3: [50,75%) 4: [75,100%]

Persoonlijk bruto inkomen (juridische ouder(s)) , decielen
1: [0,10%) 2: [10,20%) 3: [20,30%) 4: [30,40%) 5: [40,50%) 6: [50,60%) 7: [60,70%) 8: [70,80%) 9: [80,90%) 10: [90,100%]

Persoonlijk bruto inkomen (juridische ouder(s)) , continu en als logaritme van continu
1: continu

Juridisch gezinsinkomen, kwartielen
1: [0,25%) 2: [25,50%) 3: [50,75%) 4: [75,100%]

Juridisch gezinsinkomen, continu en als logaritme van continu
1: Continu

Verblijfsduur (juridische ouder(s))
1: <5 jaar 2: 5-14 jaar 3: ≥15 jaar

Burgerlijke staat: gescheiden (juridische ouder(s))
1: Ja 2: Nee

Juridische ouder(s) zijn niet samenwonend
1: Ja 2: Nee

Leeftijd (juridische ouder(s))
1: Continu

Persoon is ooit verdachte geweest van een misdrijf (juridische ouder(s))
--

- | |
|-----------------|
| 1: Ja
2: Nee |
|-----------------|

Ontvanger schuldhulpverlening in de zin van de WSNP op peilmoment (juridische ouder(s))

- | |
|-----------------|
| 1: Ja
2: Nee |
|-----------------|

Thuistaal (gezin)

- | |
|-------------------------------------|
| 1: Nederlands
2: Niet Nederlands |
|-------------------------------------|

Schoolkenmerken

Alle schoolkenmerken zijn afgeleid uit de populatie van het onderzoeksbestand. Dit zijn niet alle leerlingen van elke school, maar alleen de leerlingen die mee hebben gedaan aan de COOL cohort studie.

Niet-westerse leerlingen

- | |
|--|
| 1: [0,25%]
2: [25,50%]
3: [50,75%]
4: [75,100%] |
|--|

Leerlingen van wie de juridische ouder(s) een hoge opleiding hebben

- | |
|--|
| 1: [0,25%]
2: [25,50%]
3: [50,75%]
4: [75,100%] |
|--|

Leerlingen van wie de juridische ouder(s) een lage opleiding hebben

- | |
|--|
| 1: [0,25%]
2: [25,50%]
3: [50,75%]
4: [75,100%] |
|--|

Gemiddelde opleidingsniveau school (gebaseerd op de opleiding van de moeder)
--

- | |
|------------|
| 1: Continu |
|------------|

Leerlingen van wie de juridische ouder(s) een laag inkomen hebben

- | |
|--|
| 1: [0,25%]
2: [25,50%]
3: [50,75%]
4: [75,100%] |
|--|

Buurtkenmerken

Stedelijkheid
1: Zeer stedelijk
2: Sterk stedelijk
3: Matig stedelijk
4: Weinig stedelijk
5: Niet-stedelijk

Percentage niet-westerse allochtonen
1: [0,25%)
2: [25,75%)
3: [75,100%]

Minima (Inkomen onder sociaal minimum)
1: [0,10%) onder minimum inkomen
2: [10,20%) onder minimum inkomen
3: [20,30%) onder minimum inkomen
4: [30,100%] onder minimum inkomen

Percentage inwoners met laag inkomen (gebaseerd op kwartielen van de populatie)
1: [0,10%) laag inkomen
2: [10,30%) laag inkomen
3: [30,50%) laag inkomen
4: [50,100%] laag inkomen

Percentage inwoners met bijstand
1: [0,10%) bijstand
2: [10,30%) bijstand
3: [30,50%) bijstand
4: [50,100%] bijstand

Gemiddelde woningwaarde (WOZ)
1: Continue variabele

Percentage huurwoningen
1: Percentage huurwoningen ten opzichte van het totaal aantal woningen

Leefbaarheid (Leefbaarometer)
1: Continue variabele

Bijlage 4. Logistische regressie-imputatie

Het continuation-ratio-model is een uitbreiding van het standaard (binaire) logistische regressiemodel die bedoeld is voor het modelleren van ordinale variabelen met $K \geq 3$ categorieën, rekening houdend met de natuurlijke ordening in de categorieën. Hieronder volgt een korte beschrijving van dit model; zie Agresti (1990, hoofdstuk 9) voor meer details.

Gegeven een ordinale variabele y met categorieën $1, \dots, K$, definieer:

$$\pi_{ki} = P(y = k \mid y \geq k, \mathbf{x} = \mathbf{x}_i), \quad (k = 1, \dots, K - 1).$$

Dit is de conditionele kans op een score in categorie k , gegeven de achtergrondkenmerken én gegeven dat alle categorieën lager dan k zijn uitgesloten. Elk van deze kansen kunnen we apart modelleren via een standaard logistisch regressiemodel:

$$\ln \left(\frac{\pi_{ki}}{1 - \pi_{ki}} \right) = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{x}_i, \quad (k = 1, \dots, K - 1),$$

met α_k en $\boldsymbol{\beta}_k$ te schatten regressieparameters die per categorie k mogen verschillen. Het model voor π_{ki} kan worden geschat op basis van alle records in de donorpool die voldoen aan de conditie $y \geq k$, via maximum-likelihood.

Gegeven schattingen voor α_k en $\boldsymbol{\beta}_k$ worden de conditionele kansen voor een record met achtergrondkenmerken \mathbf{x}_i dan geschat door:

$$\hat{\pi}_{ki} = \frac{\exp(\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \mathbf{x}_i)}{1 + \exp(\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \mathbf{x}_i)}, \quad (k = 1, \dots, K - 1).$$

In de context van een imputatiemethode kunnen we deze uitdrukkingen in het bijzonder toepassen op de receptoren. Vervolgens kunnen deze conditionele kansen worden terugvertaald naar een (geschatte) marginale kansverdeling over de categorieën, $p_{ki} = P(y = k \mid \mathbf{x} = \mathbf{x}_i)$, door gebruik te maken van het volgende recursieve verband tussen p_{1i}, \dots, p_{Ki} en $\pi_{1i}, \dots, \pi_{(K-1)i}$:

$$\begin{aligned} p_{1i} &= \pi_{1i} \\ p_{ki} &= \pi_{ki} (1 - p_{1i} - \dots - p_{(k-1)i}), \quad (k = 2, \dots, K - 1), \\ p_{Ki} &= (1 - \pi_{(K-1)i}) (1 - p_{1i} - \dots - p_{(K-2)i}), \end{aligned}$$

met $\sum_{k=1}^K p_{ki} = 1$. Deze (geschatte) marginale kansverdeling kan dan worden gebruikt om receptoren te imputeren, zoals beschreven in de hoofdstekst.

Bijlage 5. Correctie van nscct-scores voor invloed sociale omgeving

In deze technische bijlage bij paragraaf 4.2.4 beschrijven we een methode voor het corrigeren van de nscct-scores van leerlingen voor een deel van de invloed van het opleidingsniveau en de herkomst van de ouders.

Als uitgangspunt nemen we de volgende methode om de invloed van herkomst en opleidingsniveau van de ouders op gemeten intelligentie volledig te verwijderen (pijl a én pijl b in figuur 4.1). Stel dat voor het opleidingsniveau van beide ouders en voor herkomst afgeleide continue variabelen zijn bepaald zoals beschreven in paragraaf 4.2.2 en paragraaf 4.2.3. Definieer het volgende regressiemodel:

$$\text{nscct} = \overline{\text{nscct}} + \alpha^{OM} (O_M - \overline{O_M}) + \alpha^{OV} (O_V - \overline{O_V}) + \alpha^H (H - \overline{H}) + \varepsilon.$$

Hierbij is O_M de afgeleide continue variabele voor het opleidingsniveau van de moeder, O_V de afgeleide continue variabele voor het opleidingsniveau van de vader en H de afgeleide continue variabele voor herkomst. Met de notatie $\overline{O_M}$, $\overline{O_V}$, etc. wordt steeds het gemiddelde van een variabele bedoeld. De verwachte nscct-score voor een leerling met ouders met opleidingsniveaus O_M en O_V en herkomst H is volgens dit regressiemodel gelijk aan $\overline{\text{nscct}} + \alpha^{OM} (O_M - \overline{O_M}) + \alpha^{OV} (O_V - \overline{O_V}) + \alpha^H (H - \overline{H})$.

Nadat het regressiemodel is gefit kan een volledig voor opleiding en herkomst gecorrigeerde nscct-score als volgt worden geschat:

$$\text{nscct}' = \text{nscct} - \alpha^{OM} (O_M - \overline{O_M}) - \alpha^{OV} (O_V - \overline{O_V}) - \alpha^H (H - \overline{H}).$$

De scores nscct' komen (op een constante na) overeen met de residuen van het geschatte regressiemodel. Deze zijn ongecorrleerd met herkomst en opleiding. We kunnen daarom stellen dat uit de gecorrigeerde scores nscct' de volledige invloed van herkomst en opleiding is verwijderd.

Om te voorkomen dat er te sterk gecorrigeerd wordt kunnen we dempingsfactoren ($0 \leq \gamma^O, \gamma^H \leq 1$) toevoegen aan bovenstaande correctieformule:

$$\text{nscct}''(\gamma^O, \gamma^H) = \text{nscct} - \gamma^O \alpha^{OM} (O_M - \overline{O_M}) - \gamma^O \alpha^{OV} (O_V - \overline{O_V}) - \gamma^H \alpha^H (H - \overline{H}).$$

Hiermee blijft een deel van de invloed van opleidingsniveau en/of herkomst op de gemeten intelligentie behouden. We hebben opleiding en herkomst in het bovenstaande regressiemodel opgenomen in afwijking van hun gemiddelden, zodat het gemiddelde van de nscct-scores behouden blijft onder deze correctie.

Zoals in de hoofdttekst is opgemerkt hebben we ervoor gekozen om varianten van $\text{nscct}''(\gamma^O, \gamma^H)$ uit te proberen met $\gamma^O = 0.75$, $\gamma^O = 0.50$ en $\gamma^O = 0.25$ voor opleidingsniveau en met $\gamma^H = 0.75$ en $\gamma^H = 1.00$ voor herkomstgroepering. Dit geeft in totaal zes varianten van het model met herkomst en opleidingsniveau ouders.

Bijlage 6. Geschatte SEM-modellen met 50%-75%-correctie nsct

Alle modellen die hierna worden beschreven, zijn geschat met een nsct die is gecorrigeerd zoals in hoofdstuk 4.2.4 is beschreven. De nsct-score is voor 50% gecorrigeerd voor opleiding en voor 75% voor herkomst. Dit betekent dat de effecten van opleiding en herkomst voor respectievelijk 50% en 75% uit de nsct-score zijn gehaald.

Tabel A6.1: Opleidingsniveau (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.876	0.217	0.031	28.155	<0.001
Cito groep 8	~ Opleiding vader	0.868	0.191	0.033	25.918	<0.001
Cito groep 8	~ Intelligentie	0.548	0.490	0.010	54.561	<0.001
Intelligentie	~ Opleiding moeder	0.364	0.101	0.039	9.402	<0.001
Intelligentie	~ Opleiding vader	0.353	0.087	0.041	8.525	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelijkheid
			63 076	63 182	0.420	-31 524

Tabel A6.2: Opleidingsniveau en inkomen (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.839	0.211	0.030	27.621	<0.001
Cito groep 8	~ Opleiding vader	0.857	0.179	0.037	23.407	<0.001
Cito groep 8	~ Log(Inkomen)	0.053	0.033	0.014	3.702	<0.001
Cito groep 8	~ Intelligentie	0.544	0.487	0.010	54.275	<0.001
Intelligentie	~ Opleiding moeder	0.368	0.103	0.039	9.560	<0.001
Intelligentie	~ Opleiding vader	0.371	0.087	0.044	8.472	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelijkheid
			83 506	83 648	0.418	-41 734

Tabel A6.3: Opleidingsniveau en herkomst (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.835	0.207	0.031	27.075	<0.001
Cito groep 8	~ Opleiding vader	0.835	0.184	0.033	24.999	<0.001
Cito groep 8	~ Herkomst	-0.793	-0.074	0.112	-7.050	<0.001
Cito groep 8	~ Intelligentie	0.544	0.481	0.010	54.662	<0.001
Intelligentie	~ Opleiding moeder	0.331	0.093	0.037	8.884	<0.001
Intelligentie	~ Opleiding vader	0.327	0.082	0.040	8.231	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelijkheid
			35 905	36 048	0.417	-17 934

Tabel A6.4: Opleidingsniveau, inkomen en herkomst (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.850	0.202	0.032	26.573	<0.001
Cito groep 8	~ Opleiding vader	0.860	0.176	0.037	23.173	<0.001
Cito groep 8	~ Herkomst	-0.844	-0.070	0.122	-6.935	<0.001
Cito groep 8	~ Log(Inkomen)	0.052	0.032	0.014	3.621	<0.001
Cito groep 8	~ Intelligentie	0.542	0.480	0.010	54.373	<0.001
Intelligentie	~ Opleiding moeder	0.353	0.094	0.039	9.047	<0.001
Intelligentie	~ Opleiding vader	0.361	0.083	0.043	8.408	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelijkheid
			51 406	51 594	0.417	-25 678

Tabel A6.5: Opleidingsniveau, inkomen, herkomst en buurtgegevens (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	0.858	0.205	0.032	26.628	<0.001
Cito groep 8 ~	Opleiding vader	0.864	0.177	0.036	23.826	<0.001
Cito groep 8 ~	Herkomst	-1.028	-0.076	0.129	-7.962	<0.001
Cito groep 8 ~	Log(Inkomen)	0.056	0.035	0.014	3.923	<0.001
Cito groep 8 ~	Leefbaarheid	0.005	0.014	0.004	1.303	0.193
Cito groep 8 ~	25-75% hoog opgeleiden buurt	-0.014	-0.007	0.020	-0.724	0.469
Cito groep 8 ~	75-100% hoog opgeleiden buurt	-0.073	-0.012	0.038	-1.917	0.055
Cito groep 8 ~	Zeer stedelijk	0.048	0.018	0.037	1.326	0.185
Cito groep 8 ~	Sterk stedelijk	0.020	0.008	0.032	0.606	0.544
Cito groep 8 ~	Niet stedelijk	-0.008	-0.003	0.038	-0.219	0.826
Cito groep 8 ~	Weinig stedelijk	0.000	0.000	0.034	-0.015	0.988
Cito groep 8 ~	Woningwaarde buurt	-0.000	-0.003	0.000	-0.285	0.776
Cito groep 8 ~	Percentage huurwoningen buurt	-0.001	-0.023	0.001	-1.680	0.093
Cito groep 8 ~	25-75% niet westers buurt	0.041	0.019	0.021	1.919	0.055
Cito groep 8 ~	75-100% niet westers buurt	0.014	0.002	0.061	0.225	0.822
Cito groep 8 ~	<10% laag inkomen buurt	0.019	0.004	0.043	0.440	0.660
Cito groep 8 ~	10-30% laag inkomen buurt	-0.014	-0.007	0.025	-0.585	0.559
Cito groep 8 ~	>50% laag inkomen buurt	0.045	0.019	0.026	1.734	0.083
Cito groep 8 ~	<10% bijstand buurt	0.037	0.017	0.025	1.461	0.144
Cito groep 8 ~	Intelligentie	0.542	0.481	0.010	54.853	<0.001
Intelligentie ~	Opleiding moeder	0.354	0.095	0.039	9.190	<0.001
Intelligentie ~	Opleiding vader	0.366	0.084	0.043	8.575	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			452 142	453 906	0.419	-225 836

Tabel A6.6: Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	0.888	0.213	0.032	27.818	<0.001
Cito groep 8 ~	Opleiding vader	0.882	0.179	0.037	24.069	<0.001
Cito groep 8 ~	Log(Inkomen)	0.071	0.044	0.015	4.871	<0.001
Cito groep 8 ~	Leefbaarheid	0.006	0.016	0.004	1.455	0.146
Cito groep 8 ~	25-75% hoog opgeleiden buurt	-0.006	-0.003	0.020	-0.315	0.753
Cito groep 8 ~	75-100% hoog opgeleiden buurt	-0.075	-0.012	0.038	-1.985	0.047
Cito groep 8 ~	Zeer stedelijk	0.008	0.003	0.039	0.207	0.836
Cito groep 8 ~	Sterk stedelijk	0.011	0.004	0.033	0.325	0.745
Cito groep 8 ~	Niet stedelijk	0.007	0.003	0.038	0.171	0.864
Cito groep 8 ~	Weinig stedelijk	0.012	0.005	0.035	0.333	0.739
Cito groep 8 ~	Woningwaarde buurt	0.000	0.001	0.000	0.072	0.943
Cito groep 8 ~	Percentage huurwoningen buurt	-0.001	-0.029	0.001	-2.011	0.044
Cito groep 8 ~	25-75% niet westers buurt	0.026	0.012	0.022	1.183	0.237
Cito groep 8 ~	75-100% niet westers buurt	0.025	0.003	0.062	0.402	0.687
Cito groep 8 ~	<10% laag inkomen buurt	0.014	0.003	0.044	0.314	0.754
Cito groep 8 ~	10-30% laag inkomen buurt	-0.013	-0.006	0.025	-0.510	0.610
Cito groep 8 ~	>50% laag inkomen buurt	0.026	0.011	0.027	0.968	0.333
Cito groep 8 ~	<10% bijstand buurt	0.034	0.016	0.026	1.315	0.189
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	0.026	0.003	0.058	0.445	0.656
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.135	0.032	0.029	4.685	<0.001
Cito groep 8 ~	Intelligentie	0.542	0.480	0.010	55.006	<0.001
Intelligentie ~	Opleiding moeder	0.349	0.094	0.038	9.129	<0.001
Intelligentie ~	Opleiding vader	0.361	0.083	0.043	8.438	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			467 127	469 057	0.414	-233 307

Tabel A6.7: Inkomen (kwartielen) (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	25-50% inkomen	0.055	0.024	0.021	2.649	0.008
Cito groep 8 ~	50-75% inkomen	0.245	0.105	0.021	11.942	<0.001
Cito groep 8 ~	75-100% inkomen	0.419	0.177	0.025	16.430	<0.001
Cito groep 8 ~	Intelligentie	0.606	0.560	0.010	57.855	<0.001
Intelligentie ~	25-50% inkomen	0.251	0.119	0.024	10.311	<0.001
Intelligentie ~	50-75% inkomen	0.398	0.185	0.025	16.096	<0.001
Intelligentie ~	75-100% inkomen	0.598	0.274	0.028	21.391	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			109 878	110 028	0.381	-54 919

Tabel A6.7a: Inkomen (decielen) (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	0-10% inkomen	-0.083	-0.025	0.033	-2.487	0.013
Cito groep 8 ~	10-20% inkomen	-0.076	-0.023	0.033	-2.320	0.020
Cito groep 8 ~	20-30% inkomen	-0.079	-0.024	0.033	-2.411	0.016
Cito groep 8 ~	30-40% inkomen	-0.032	-0.010	0.032	-0.987	0.324
Cito groep 8 ~	50-60% inkomen	0.114	0.034	0.031	3.619	<0.001
Cito groep 8 ~	60-70% inkomen	0.134	0.040	0.031	4.288	<0.001
Cito groep 8 ~	70-80% inkomen	0.245	0.073	0.031	7.995	<0.001
Cito groep 8 ~	80-90% inkomen	0.286	0.086	0.030	9.399	<0.001
Cito groep 8 ~	90-100% inkomen	0.421	0.126	0.031	13.776	<0.001
Cito groep 8 ~	Intelligentie	0.603	0.558	0.009	64.125	<0.001
Intelligentie ~	0-10% inkomen	-0.418	-0.135	0.038	-11.071	<0.001
Intelligentie ~	10-20% inkomen	-0.307	-0.099	0.037	-8.190	<0.001
Intelligentie ~	20-30% inkomen	-0.170	-0.055	0.037	-4.600	<0.001
Intelligentie ~	30-40% inkomen	-0.150	-0.049	0.037	-4.073	<0.001
Intelligentie ~	50-60% inkomen	-0.007	-0.002	0.036	-0.194	0.846
Intelligentie ~	60-70% inkomen	0.065	0.021	0.036	1.831	0.067
Intelligentie ~	70-80% inkomen	0.165	0.053	0.035	4.693	<0.001
Intelligentie ~	80-90% inkomen	0.216	0.070	0.035	6.137	<0.001
Intelligentie ~	90-100% inkomen	0.339	0.110	0.034	9.931	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			104 508	105 086	0.381	-52 177

Tabel A6.8: Inkomen, herkomst en buurtgegevens (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Log(Inkomen)	0.184	0.114	0.015	12.007	<0.001
Cito groep 8 ~	Herkomst	-1.054	-0.145	0.080	-13.107	<0.001
Cito groep 8 ~	Leefbaarheid	0.004	0.011	0.004	0.954	0.340
Cito groep 8 ~	25-75% hoog opgeleiden buurt	0.099	0.046	0.020	4.836	<0.001
Cito groep 8 ~	75-100% hoog opgeleiden buurt	0.235	0.038	0.040	5.894	<0.001
Cito groep 8 ~	Zeer stedelijk	0.073	0.027	0.039	1.856	0.063
Cito groep 8 ~	Sterk stedelijk	0.030	0.012	0.034	0.896	0.370
Cito groep 8 ~	Niet stedelijk	-0.031	-0.013	0.039	-0.782	0.434
Cito groep 8 ~	Weinig stedelijk	0.005	0.002	0.036	0.137	0.891
Cito groep 8 ~	Woningwaarde buurt	0.000	0.001	0.000	0.091	0.927
Cito groep 8 ~	Percentage huurwoningen buurt	-0.001	-0.030	0.001	-2.154	0.031
Cito groep 8 ~	25-75% niet westers buurt	0.043	0.020	0.023	1.883	0.060
Cito groep 8 ~	75-100% niet westers buurt	0.030	0.004	0.067	0.444	0.657
Cito groep 8 ~	<10% laag inkomen buurt	0.035	0.007	0.046	0.762	0.446
Cito groep 8 ~	10-30% laag inkomen buurt	-0.013	-0.006	0.026	-0.514	0.607
Cito groep 8 ~	>50% laag inkomen buurt	0.032	0.014	0.027	1.191	0.234
Cito groep 8 ~	<10% bijstand buurt	0.043	0.020	0.026	1.611	0.107
Cito groep 8 ~	Intelligentie	0.595	0.536	0.010	57.919	<0.001
Intelligentie ~	Log(Inkomen)	0.233	0.160	0.015	15.081	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			480 840	482 289	0.373	-240 227

Tabel A6.9: Inkomen, herkomst, buurtgegevens en verblijfsduur (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Log(Inkomen)	0.193	0.119	0.015	12.533	<0.001
Cito groep 8 ~	Herkomst	-1.089	-0.149	0.081	-13.441	<0.001
Cito groep 8 ~	Leefbaarheid	0.004	0.011	0.004	0.999	0.318
Cito groep 8 ~	25-75% hoog opgeleiden buurt	0.099	0.046	0.020	4.836	<0.001
Cito groep 8 ~	75-100% hoog opgeleiden buurt	0.230	0.037	0.040	5.770	<0.001
Cito groep 8 ~	Zeer stedelijk	0.071	0.027	0.039	1.808	0.071
Cito groep 8 ~	Sterk stedelijk	0.031	0.013	0.034	0.921	0.357
Cito groep 8 ~	Niet stedelijk	-0.027	-0.011	0.039	-0.681	0.496
Cito groep 8 ~	Weinig stedelijk	0.008	0.003	0.036	0.231	0.818
Cito groep 8 ~	Woningwaarde buurt	0.000	0.002	0.000	0.180	0.857
Cito groep 8 ~	Percentage huurwoningen buurt	-0.001	-0.031	0.001	-2.202	0.028
Cito groep 8 ~	25-75% niet westers buurt	0.041	0.019	0.023	1.830	0.067
Cito groep 8 ~	75-100% niet westers buurt	0.030	0.004	0.067	0.454	0.650
Cito groep 8 ~	<10% laag inkomen buurt	0.031	0.007	0.046	0.682	0.495
Cito groep 8 ~	10-30% laag inkomen buurt	-0.012	-0.006	0.026	-0.476	0.634
Cito groep 8 ~	>50% laag inkomen buurt	0.030	0.013	0.027	1.092	0.275
Cito groep 8 ~	<10% bijstand buurt	0.042	0.020	0.027	1.567	0.117
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	0.064	0.008	0.060	1.059	0.289
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.167	0.039	0.028	5.889	<0.001
Cito groep 8 ~	Intelligentie	0.594	0.535	0.010	57.786	<0.001
Intelligentie ~	Log(Inkomen)	0.233	0.160	0.015	15.081	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			461 632	463 389	0.373	-230 582

Tabel A6.10: Stepwise SEM met verblijfsduur (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	0.819	0.202	0.030	27.069	<0.001
Cito groep 8 ~	Opleiding vader	0.826	0.181	0.033	25.080	<0.001
Cito groep 8 ~	Herkomst	-0.704	-0.071	0.109	-6.437	<0.001
Cito groep 8 ~	Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.156	0.037	0.028	5.498	<0.001
Cito groep 8 ~	Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8 ~	Intelligentie	0.542	0.479	0.010	54.356	<0.001
Intelligentie ~	Opleiding moeder	0.331	0.092	0.037	8.842	<0.001
Intelligentie ~	Opleiding vader	0.325	0.081	0.040	8.123	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			-31 181	-30 813	0.418	15 640

Bijlage 7. Fitmaten voor SEM-modellen en nscct-varianten

Alle deelmodellen met SEM zijn met de verschillende nscct-correcties die in hoofdstuk 4.2.4 worden beschreven uitgevoerd. In deze bijlage worden de fitmaten voor deze verschillende modellen gepresenteerd. Bij elke tabel staat telkens aangegeven in welke mate de nscct-score gecorrigeerd is voor respectievelijk opleiding en herkomst. Zo geeft 75%-100% aan dat 75% van het effect van opleiding en 100% van het effect van herkomst op de nscct-score uit deze score gehaald is.

Tabel A7.1: Fitmaten voor nscct-correctie 75%-100% (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	63 076	63 182	0.419	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.418	-41 734
3 Opleidingsniveau en herkomst	35 882	36 025	0.418	-17 922
4 Opleidingsniveau, inkomen en herkomst	51 388	51 575	0.418	-25 669
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 127	453 891	0.419	-225 828
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 160	469 090	0.411	-233 323
7 Inkomen	109 878	110 028	0.381	-54 919
8 Inkomen, herkomst en buurtgegevens	480 865	482 314	0.378	-240 240
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 657	463 414	0.379	-230 595
10 Stepwise SEM met verblijfsduur (model uit tabel A8.1)	-31 207	-30 839	0.419	15 652
11 Stepwise SEM met asielmotief	-40 305	-40 005	0.418	20 193

Tabel A7.2: Fitmaten voor nscct-correctie 50%-100% (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	63 076	63 182	0.420	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.418	-41 734
3 Opleidingsniveau en herkomst	35 882	36 025	0.418	-17 922
4 Opleidingsniveau, inkomen en herkomst	51 388	51 575	0.418	-25 669
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 127	453 891	0.420	-225 828
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 160	469 090	0.411	-233 323
7 Inkomen	109 878	110 028	0.381	-54 919
8 Inkomen, herkomst en buurtgegevens	480 865	482 314	0.378	-240 240
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 657	463 414	0.379	-230 595
10 Stepwise SEM met verblijfsduur (model uit tabel A8.2)	-31 207	-30 839	0.419	15 652
11 Stepwise SEM met asielmotief	-40 305	-40 005	0.419	20 193

Tabel A7.3: Fitmaten voor nscct-correctie 25%-100% (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	63 076	63 182	0.420	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.419	-41 734
3 Opleidingsniveau en herkomst	35 882	36 025	0.419	-17 922
4 Opleidingsniveau, inkomen en herkomst	51 388	51 575	0.419	-25 669
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 127	453 891	0.421	-225 828
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 160	469 090	0.412	-233 323
7 Inkomen	109 878	110 028	0.381	-54 919
8 Inkomen, herkomst en buurtgegevens	480 865	482 314	0.378	-240 240
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 657	463 414	0.379	-230 595
10 Stepwise SEM met verblijfsduur (model uit tabel A8.3)	-31 207	-30 839	0.420	15 652
11 Stepwise SEM met asielmotief	-40 305	-40 005	0.419	20 193

Tabel A7.4: Fitmaten voor nscct-correctie 75%-75% (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	63 076	63 182	0.419	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.418	-41 734
3 Opleidingsniveau en herkomst	35 905	36 048	0.417	-17 934
4 Opleidingsniveau, inkomen en herkomst	51 406	51 594	0.417	-25 678
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 142	453 906	0.418	-225 836
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 127	469 057	0.414	-233 307
7 Inkomen	109 878	110 028	0.381	-54 919
8 Inkomen, herkomst en buurtgegevens	480 840	482 289	0.373	-240 227
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 632	463 389	0.373	-230 582
10 Stepwise SEM met verblijfsduur (model uit tabel A8.4)	-31 181	-30 813	0.417	15 640
11 Stepwise SEM met asielmotief	-40 281	-39 981	0.417	20 181

Tabel A7.5: Fitmaten voor nscct-correctie 50%-75% (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	63 076	63 182	0.420	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.418	-41 734
3 Opleidingsniveau en herkomst	35 905	36 048	0.417	-17 934
4 Opleidingsniveau, inkomen en herkomst	51 406	51 594	0.417	-25 678
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 142	453 906	0.419	-225 836
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 127	469 057	0.414	-233 307
7 Inkomen	109 878	110 028	0.381	-54 919
8 Inkomen, herkomst en buurtgegevens	480 840	482 289	0.373	-240 227
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 632	463 389	0.373	-230 582
10 Stepwise SEM met verblijfsduur (model uit tabel 5.1/A8.5)	-31 181	-30 813	0.418	15 640
11 Stepwise SEM met asielmotief	-40 281	-39 981	0.417	20 181

Tabel A7.6: Fitmaten voor nscct-correctie 25%-75% (n = 13 466).

Model	AIC	BIC	Adjusted R-squared	Log-likelihood
1 Opleiding	63 076	63 182	0.420	-31 524
2 Opleidingsniveau en inkomen	83 506	83 648	0.419	-41 734
3 Opleidingsniveau en herkomst	35 905	36 048	0.418	-17 934
4 Opleidingsniveau, inkomen en herkomst	51 406	51 594	0.418	-25 678
5 Opleidingsniveau, inkomen, herkomst en buurtgegevens	452 142	453 906	0.419	-225 836
6 Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur	467 127	469 057	0.415	-233 307
7 Inkomen	109 878	110 028	0.381	-54 919
8 Inkomen, herkomst en buurtgegevens	480 840	482 289	0.373	-240 227
9 Inkomen, herkomst, buurtgegevens en verblijfsduur	461 632	463 389	0.373	-230 582
10 Stepwise SEM met verblijfsduur (model uit tabel A8.6)	-31 181	-30 813	0.418	15 640
11 Stepwise SEM met asielmotief	-40 281	-39 981	0.418	20 181

Bijlage 8. Stepwise SEM-modellen voor nscct-varianten

Voor het best passende SEM-model presenteren we niet alleen de fitmaten (bijlage 7), maar eveneens de coëfficiënten wanneer we gebruik maken van de verschillende nscct-correcties. In de titels verwijst het eerste getal naar de correctie voor opleiding en het tweede getal naar de correctie voor herkomst. Zo geeft 75%-100% aan dat 75% van het effect van opleiding en 100% van het effect van herkomst op de nscct-score uit deze score gehaald is.

Tabel A8.1: Stepwise SEM NSCCT 75%-100% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.901	0.222	0.030	29.992	<0.001
Cito groep 8	~ Opleiding vader	0.908	0.199	0.033	27.602	<0.001
Cito groep 8	~ Herkomst	-0.904	-0.092	0.109	-8.270	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.156	0.036	0.028	5.498	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8	~ Intelligentie	0.541	0.473	0.010	54.666	<0.001
Intelligentie	~ Opleiding moeder	0.147	0.041	0.037	3.954	<0.001
Intelligentie	~ Opleiding vader	0.148	0.037	0.040	3.719	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 207	-30 839	0.419	15 652

Tabel A8.2: Stepwise SEM NSCCT 50%-100% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.821	0.202	0.030	27.217	<0.001
Cito groep 8	~ Opleiding vader	0.828	0.181	0.033	25.169	<0.001
Cito groep 8	~ Herkomst	-0.904	-0.092	0.109	-8.270	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.156	0.036	0.028	5.498	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8	~ Intelligentie	0.542	0.476	0.010	54.644	<0.001
Intelligentie	~ Opleiding moeder	0.294	0.082	0.037	7.908	<0.001
Intelligentie	~ Opleiding vader	0.295	0.074	0.040	7.437	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 207	-30 839	0.419	15 652

Tabel A8.3: Stepwise SEM NSCCT 25%-100% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.741	0.182	0.030	24.391	<0.001
Cito groep 8	~ Opleiding vader	0.747	0.164	0.033	22.673	<0.001
Cito groep 8	~ Herkomst	-0.904	-0.092	0.109	-8.270	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.156	0.036	0.028	5.498	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8	~ Intelligentie	0.544	0.483	0.010	54.607	<0.001
Intelligentie	~ Opleiding moeder	0.441	0.122	0.037	11.862	<0.001
Intelligentie	~ Opleiding vader	0.443	0.109	0.040	11.156	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 207	-30 839	0.420	15 652

Tabel A8.4: Stepwise SEM NSCCT 75%-75% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.899	0.222	0.030	29.870	<0.001
Cito groep 8	~ Opleiding vader	0.906	0.199	0.033	27.533	<0.001
Cito groep 8	~ Herkomst	-0.704	-0.071	0.109	-6.437	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.156	0.037	0.028	5.498	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8	~ Intelligentie	0.541	0.475	0.010	54.382	<0.001
Intelligentie	~ Opleiding moeder	0.184	0.052	0.037	4.910	<0.001
Intelligentie	~ Opleiding vader	0.177	0.044	0.040	4.434	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 181	-30 813	0.417	15 640

Tabel A8.5: Stepwise SEM NSCCT 50%-75% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.819	0.202	0.030	27.069	<0.001
Cito groep 8	~ Opleiding vader	0.826	0.181	0.033	25.080	<0.001
Cito groep 8	~ Herkomst	-0.704	-0.071	0.109	-6.437	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.156	0.037	0.028	5.498	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8	~ Intelligentie	0.542	0.479	0.010	54.356	<0.001
Intelligentie	~ Opleiding moeder	0.331	0.092	0.037	8.842	<0.001
Intelligentie	~ Opleiding vader	0.325	0.081	0.040	8.123	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 181	-30 813	0.418	15 640

Tabel A8.6: Stepwise SEM NSCCT 25%-75% (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	0.738	0.182	0.030	24.220	<0.001
Cito groep 8 ~	Opleiding vader	0.745	0.163	0.033	22.564	<0.001
Cito groep 8 ~	Herkomst	-0.704	-0.071	0.109	-6.437	<0.001
Cito groep 8 ~	Gemiddelde opleiding school	0.205	0.021	0.122	1.682	0.093
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	0.019	0.002	0.057	0.340	0.734
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.156	0.037	0.028	5.498	<0.001
Cito groep 8 ~	Ouder(s) in schuldsanering	-0.270	-0.030	0.065	-4.148	<0.001
Cito groep 8 ~	Intelligentie	0.544	0.485	0.010	54.316	<0.001
Intelligentie ~	Opleiding moeder	0.478	0.132	0.037	12.774	<0.001
Intelligentie ~	Opleiding vader	0.473	0.116	0.040	11.812	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			-31 181	-30 813	0.418	15 640

Bijlage 9. Modellen zonder intelligentie

Alle modellen die in deze bijlage worden beschreven, zijn geschat zonder nsct.

Tabel A9.1: Opleidingsniveau (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	1.000	0.269	0.035	28.321	<0.001
Cito groep 8	~ Opleiding vader	1.000	0.230	0.037	26.782	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			35 528	35 551	0.187	-17 761

Tabel A9.2: Opleidingsniveau en inkomen (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	1.000	0.251	0.037	26.870	<0.001
Cito groep 8	~ Opleiding vader	1.000	0.209	0.043	23.277	<0.001
Cito groep 8	~ Log(Inkomen)	0.117	0.072	0.017	7.002	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			35 467	35 497	0.190	-17 730

Tabel A9.3: Opleidingsniveau en herkomst (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	1.000	0.247	0.037	26.816	<0.001
Cito groep 8	~ Opleiding vader	1.000	0.221	0.038	26.248	<0.001
Cito groep 8	~ Herkomst	-1.000	-0.093	0.121	-8.296	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			35 414	35 444	0.194	-17 703

Tabel A9.4: Opleidingsniveau, inkomen en herkomst (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	1.000	0.237	0.039	25.721	<0.001
Cito groep 8	~ Opleiding vader	1.000	0.204	0.043	23.214	<0.001
Cito groep 8	~ Herkomst	-1.000	-0.083	0.131	-7.637	<0.001
Cito groep 8	~ Log(Inkomen)	0.102	0.063	0.016	6.212	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			35 372	35 409	0.196	-17 681

Tabel A9.5: Opleidingsniveau, inkomen, herkomst en buurtgegevens (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	1.000	0.238	0.040	24.985	<0.001
Cito groep 8 ~	Opleiding vader	1.000	0.204	0.043	23.223	<0.001
Cito groep 8 ~	Herkomst	-1.000	-0.078	0.131	-7.612	<0.001
Cito groep 8 ~	Log(Inkomen)	0.096	0.059	0.017	5.697	<0.001
Cito groep 8 ~	Leefbaarheid	0.009	0.024	0.004	2.164	0.031
Cito groep 8 ~	25-75% hoog opgeleiden buurt	0.003	0.001	0.021	0.152	0.879
Cito groep 8 ~	75-100% hoog opgeleiden buurt	-0.106	-0.017	0.043	-2.474	0.014
Cito groep 8 ~	Zeer stedelijk	0.048	0.018	0.039	1.241	0.215
Cito groep 8 ~	Sterk stedelijk	0.032	0.013	0.034	0.922	0.357
Cito groep 8 ~	Niet stedelijk	0.004	0.002	0.041	0.106	0.916
Cito groep 8 ~	Weinig stedelijk	0.002	0.001	0.036	0.060	0.952
Cito groep 8 ~	Woningwaarde buurt	0.000	0.007	0.000	0.538	0.591
Cito groep 8 ~	Percentage huurwoningen buurt	-0.001	-0.034	0.001	-2.319	0.021
Cito groep 8 ~	25-75% niet westers buurt	0.016	0.008	0.023	0.718	0.473
Cito groep 8 ~	75-100% niet westers buurt	0.022	0.003	0.068	0.322	0.747
Cito groep 8 ~	<10% laag inkomen buurt	-0.003	-0.001	0.047	-0.061	0.951
Cito groep 8 ~	10-30% laag inkomen buurt	-0.002	-0.001	0.026	-0.079	0.937
Cito groep 8 ~	>50% laag inkomen buurt	0.042	0.018	0.028	1.495	0.136
Cito groep 8 ~	<10% bijstand buurt	0.035	0.017	0.028	1.250	0.212
			AIC	BIC	Adjusted R-squared	Log- likelihood
			35 365	35 515	0.198	-17 663

Tabel A9.6: Opleidingsniveau, inkomen, buurtgegevens en verblijfsduur (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Opleiding moeder	1.000	0.249	0.038	26.097	<0.001
Cito groep 8 ~	Opleiding vader	1.000	0.206	0.043	23.222	<0.001
Cito groep 8 ~	Log(Inkomen)	0.106	0.066	0.017	6.241	<0.001
Cito groep 8 ~	Leefbaarheid	0.009	0.025	0.004	2.277	0.023
Cito groep 8 ~	25-75% hoog opgeleiden buurt	0.011	0.005	0.021	0.539	0.590
Cito groep 8 ~	75-100% hoog opgeleiden buurt	-0.104	-0.017	0.043	-2.425	0.016
Cito groep 8 ~	Zeer stedelijk	0.016	0.006	0.041	0.388	0.698
Cito groep 8 ~	Sterk stedelijk	0.025	0.010	0.035	0.706	0.480
Cito groep 8 ~	Niet stedelijk	0.015	0.006	0.041	0.357	0.721
Cito groep 8 ~	Weinig stedelijk	0.012	0.005	0.037	0.317	0.751
Cito groep 8 ~	Woningwaarde buurt	0.000	0.011	0.000	0.865	0.387
Cito groep 8 ~	Percentage huurwoningen buurt	-0.002	-0.039	0.001	-2.462	0.014
Cito groep 8 ~	25-75% niet westers buurt	0.003	0.001	0.023	0.135	0.893
Cito groep 8 ~	75-100% niet westers buurt	0.033	0.004	0.068	0.484	0.629
Cito groep 8 ~	<10% laag inkomen buurt	-0.006	-0.001	0.047	-0.132	0.895
Cito groep 8 ~	10-30% laag inkomen buurt	0.000	0.000	0.027	0.006	0.996
Cito groep 8 ~	>50% laag inkomen buurt	0.025	0.011	0.029	0.862	0.389
Cito groep 8 ~	<10% bijstand buurt	0.032	0.015	0.028	1.116	0.265
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	-0.016	-0.002	0.062	-0.264	0.792
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.176	0.041	0.033	5.296	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			35 421	35 579	0.194	-17 689

Tabel A9.7: Inkomen (kwartielen) (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	25-50% inkomen	0.207	0.091	0.025	8.283	<0.001
Cito groep 8 ~	50-75% inkomen	0.486	0.209	0.025	19.557	<0.001
Cito groep 8 ~	75-100% inkomen	0.781	0.331	0.030	26.350	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			37 131	37 161	0.084	-18 562

Tabel A9.7a: Inkomen (decielen) (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	0-10% inkomen	-0.335	-0.100	0.039	-8.610	<0.001
Cito groep 8 ~	10-20% inkomen	-0.261	-0.078	0.039	-6.726	<0.001
Cito groep 8 ~	20-30% inkomen	-0.181	-0.054	0.037	-4.964	<0.001
Cito groep 8 ~	30-40% inkomen	-0.122	-0.037	0.036	-3.368	<0.001
Cito groep 8 ~	50-60% inkomen	0.109	0.033	0.035	3.131	0.002
Cito groep 8 ~	60-70% inkomen	0.174	0.052	0.035	4.945	<0.001
Cito groep 8 ~	70-80% inkomen	0.344	0.103	0.034	10.029	<0.001
Cito groep 8 ~	80-90% inkomen	0.416	0.125	0.036	11.465	<0.001
Cito groep 8 ~	90-100% inkomen	0.625	0.187	0.042	14.837	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelihood
			37 303	37 378	0.088	-18 641

Tabel A9.8: Inkomen, herkomst en buurtgegevens (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Log(Inkomen)	0.301	0.186	0.020	15.284	<0.001
Cito groep 8 ~	Herkomst	-1.000	-0.138	0.086	-11.613	<0.001
Cito groep 8 ~	Leefbaarheid	0.007	0.020	0.004	1.649	0.100
Cito groep 8 ~	25-75% hoog opgeleiden buurt	0.163	0.075	0.023	7.235	<0.001
Cito groep 8 ~	75-100% hoog opgeleiden buurt	0.324	0.053	0.044	7.398	<0.001
Cito groep 8 ~	Zeer stedelijk	0.099	0.037	0.043	2.274	0.023
Cito groep 8 ~	Sterk stedelijk	0.053	0.022	0.036	1.469	0.142
Cito groep 8 ~	Niet stedelijk	-0.030	-0.012	0.044	-0.679	0.497
Cito groep 8 ~	Weinig stedelijk	0.007	0.003	0.039	0.182	0.855
Cito groep 8 ~	Woningwaarde buurt	0.000	0.014	0.000	1.008	0.314
Cito groep 8 ~	Percentage huurwoningen buurt	-0.002	-0.043	0.001	-2.729	0.007
Cito groep 8 ~	25-75% niet westers buurt	0.021	0.010	0.024	0.859	0.391
Cito groep 8 ~	75-100% niet westers buurt	0.038	0.005	0.075	0.514	0.608
Cito groep 8 ~	<10% laag inkomen buurt	0.017	0.004	0.050	0.343	0.732
Cito groep 8 ~	10-30% laag inkomen buurt	0.001	0.001	0.028	0.048	0.962
Cito groep 8 ~	>50% laag inkomen buurt	0.029	0.012	0.029	0.981	0.327
Cito groep 8 ~	<10% bijstand buurt	0.045	0.021	0.030	1.499	0.134
			AIC	BIC	Adjusted R-squared	Log- likelhood
			36 905	37 040	0.100	-18 435

Tabel A9.9: Inkomen, herkomst, buurtgegevens en verblijfsduur (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaard- diseerd	Standaard- fout	Z-score	P-waarde
Cito groep 8 ~	Log(Inkomen)	0.305	0.188	0.020	15.509	<0.001
Cito groep 8 ~	Herkomst	-1.000	-0.147	0.084	-11.843	<0.001
Cito groep 8 ~	Leefbaarheid	0.007	0.020	0.004	1.683	0.093
Cito groep 8 ~	25-75% hoog opgeleiden buurt	0.162	0.074	0.022	7.200	<0.001
Cito groep 8 ~	75-100% hoog opgeleiden buurt	0.318	0.052	0.044	7.263	<0.001
Cito groep 8 ~	Zeer stedelijk	0.100	0.037	0.044	2.299	0.022
Cito groep 8 ~	Sterk stedelijk	0.055	0.023	0.036	1.514	0.130
Cito groep 8 ~	Niet stedelijk	-0.028	-0.011	0.043	-0.633	0.527
Cito groep 8 ~	Weinig stedelijk	0.009	0.004	0.039	0.235	0.814
Cito groep 8 ~	Woningwaarde buurt	0.000	0.015	0.000	1.079	0.281
Cito groep 8 ~	Percentage huurwoningen buurt	-0.002	-0.043	0.001	-2.707	0.007
Cito groep 8 ~	25-75% niet westers buurt	0.021	0.010	0.024	0.841	0.400
Cito groep 8 ~	75-100% niet westers buurt	0.040	0.005	0.074	0.538	0.591
Cito groep 8 ~	<10% laag inkomen buurt	0.014	0.003	0.050	0.291	0.771
Cito groep 8 ~	10-30% laag inkomen buurt	0.003	0.001	0.028	0.096	0.923
Cito groep 8 ~	>50% laag inkomen buurt	0.028	0.012	0.029	0.939	0.348
Cito groep 8 ~	<10% bijstand buurt	0.043	0.020	0.030	1.433	0.152
Cito groep 8 ~	Verblijfsduur moeder 0-5 jaar	0.023	0.003	0.067	0.342	0.732
Cito groep 8 ~	Verblijfsduur moeder 5-15 jaar	0.224	0.052	0.033	6.833	<0.001
			AIC	BIC	Adjusted R-squared	Log- likelhood
			36 874	37 024	0.102	-18 417

Tabel A9.10: Stepwise-procedure met verblijfsduur (n = 13 466).

Afhankelijke	Onafhankelijke	Coëfficiënt	Gestandaardiseerd	Standaardfout	Z-score	P-waarde
Cito groep 8	~ Opleiding moeder	0.952	0.234	0.037	25.547	<0.001
Cito groep 8	~ Opleiding vader	0.969	0.212	0.039	25.010	<0.001
Cito groep 8	~ Herkomst	-0.805	-0.082	0.117	-6.909	<0.001
Cito groep 8	~ Gemiddelde opleiding school	0.479	0.050	0.126	3.787	<0.001
Cito groep 8	~ Verblijfsduur moeder 0-5 jaar	-0.031	-0.004	0.061	-0.514	0.608
Cito groep 8	~ Verblijfsduur moeder 5-15 jaar	0.201	0.047	0.033	6.046	<0.001
Cito groep 8	~ Ouder(s) in schuldsanering	-0.357	-0.040	0.071	-5.028	<0.001
			AIC	BIC	Adjusted R-squared	Log-likelihood
			35 340	35 400	0.198	-17 662

Bijlage 10. Afgevalen records

Tabel A10.1: Afgevalen records.

Leerling komt voor in cohort 07-08	Leerling komt voor in cohort 11-12	Leerling komt voor in cohort 13-14	<i>Aantal</i>
Nee	Nee	Nee	0
		Ja	17 672
	Ja	Nee	18 240
		Ja	8 110
Ja	Nee	Nee	26 240
		Ja	371
	Ja	Nee	9 073
		Ja	2 376
Totaal			82 082

	<i>Aantal</i>
Totaal	82 082
w.v.	
Met Cito groep 8 score	36 548
w.v.	
Met nscct score	14 509
w.v.	
Met achtergrondkenmerken	13 466
Restant	13 466

Bijlage 11. Correlatiematrix variabelen in stepwise SEM-model

Tabel A11.1: Correlatiematrix modelvariabelen stepwise SEM, nscct-correctie 50%-75% (n = 13 466).

	nscct	Cito groep 8	Opleiding moeder	Opleiding vader	Herkomst	Gemidd. opleiding school	Verbl.dr. moeder 0-5 jaar	Verbl.dr. moeder 5-15 jaar	In schuld-sanering	Intelligentie
nscct	1***									
Cito groep 8	0.500***	1***								
Opleiding moeder	0.123***	0.381***	1***							
Opleiding vader	0.118***	0.364***	0.485***	1***						
Herkomst	-0.044***	-0.202***	-0.286***	-0.256***	1***					
Gemiddelde opleiding school	0.064***	0.236***	0.424***	0.357***	-0.511***	1***				
Verblijfsduur moeder 0-5 jaar	-0.004	-0.015*	-0.026***	-0.020**	0.069***	-0.044***	1***			
Verblijfsduur moeder 5-15 jaar	-0.014*	-0.027***	-0.099***	-0.078***	0.239***	-0.179***	-0.032***	1***		
In schuld-sanering	-0.011	-0.066***	-0.077***	-0.060***	0.057***	-0.087***	-0.009	0.051***	1***	
Intelligentie	0.938***	0.533***	0.132***	0.126***	-0.047***	0.068***	-0.004	-0.015*	-0.012	1***

Notitie: *p<0.1; **p<0.05; ***p<0.01

Bijlage 12. Begeleidingscommissie

Arnold Jonk – Inspectie van het Onderwijs
Elly Dekker – Vereniging van Nederlandse gemeenten
Fons Dingelstad – Ministerie van Onderwijs Cultuur en Wetenschap
Herman van de Werfhorst – Universiteit van Amsterdam
Hilde Huizenga – Universiteit van Amsterdam
Karin van der Wiel – Centraal Planbureau
Kees Zandvliet – SEOR
Lex Herweijer – Sociaal en Cultureel Planbureau
Paul Jungbluth – Universiteit Maastricht
Paul Leseman – Universiteit Utrecht
Simone Walvisch – PO-Raad

Verklaring van tekens

Niets (blanco)	Een cijfer kan op logische gronden niet voorkomen
.	Het cijfer is onbekend, onvoldoende betrouwbaar of geheim
*	Voorlopige cijfers
**	Nader voorlopige cijfers
2015–2016	2015 tot en met 2016
2015/2016	Het gemiddelde over de jaren 2015 tot en met 2016
2015/'16	Oogstjaar, boekjaar, schooljaar enz., beginnend in 2015 en eindigend in 2016
2013/'14–2015/'16	Oogstjaar, boekjaar, enz., 2013/'14 tot en met 2015/'16

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
CCN Creatie, Den Haag

Ontwerp
Edenspiekermann

Inlichtingen
Tel. 088 570 7070
Via contactformulier: www.cbs.nl/infoservice

© Centraal Bureau voor de Statistiek, Den Haag/Heerlen/Bonaire, 2016.
Verveelvoudigen is toegestaan, mits CBS als bron wordt vermeld.