Centraal Planbureau

De voorspellende waarde van toetsen uit het leerlingvolgsysteem voor de eindtoets

Basisschoolleerlingen maken twee keer per jaar gestandaardiseerde toetsen voor rekenen/wiskunde en Nederlandse taal als onderdeel van het verplichte leerlingvolgsysteem (LVS). Aan het eind van de basisschool moeten de leerlingen ook een eindtoets maken (tegenwoordig doorstroomtoets), die mede wordt gebruikt om de plaatsing in het voortgezet onderwijs te bepalen.
Het CPB onderzoekt hoe goed de toetsen van het LVS de eindtoetsscore kunnen voorspellen.

Toetsen uit het LVS zijn een relatief goede voorspeller van de eindtoetsscores: ze verklaren ongeveer 75% van de variatie.
De voorspelkwaliteit neemt toe naarmate we meer toetsen toevoegen die dichter bij het einde van de basisschooltijd liggen.

CPB - mei 2024

Maria Zumbuehl, Lily Davies

Samenvatting

Toetsen uit het leerlingvolgsysteem (LVS) van de basisschool voorspellen de uitslag van de eindtoets in groep acht relatief goed. De voorspelkwaliteit neemt toe naarmate meer toetsen aan de analyse worden toegevoegd, die dichter bij het einde van de basisschooltijd liggen. Voorspellingen die gebaseerd zijn op een combinatie van alle beschikbare LVS-toetsen kunnen ongeveer 75% van de variatie in eindtoetsscores verklaren. Dit is aanzienlijk meer dan wat we kunnen voorspellen op basis van alleen de achtergrondkenmerken van de kinderen en scholen. Deze gegevens verklaren gezamenlijk slechts 16% van de variatie in eindtoetsscores. Bovendien zien we dat zodra we voldoende informatie hebben over de LVS-toetsen, het meenemen van achtergrondinformatie de voorspellingen niet meer significant verbetert.

Als alleen de drie meest recente LVS-toetsen gebruikt worden, is dat voldoende voor een goede voorspelling. Hoewel het gebruik van alle beschikbare LVS-gegevens leidt tot de beste voorspellingen, gaat er maar weinig precisie verloren als we de steekproef beperken tot de drie laatste toetsmomenten (midden en einde groep zeven en midden groep acht). Voor het hoofddoel van het LVS, namelijk het monitoren van leerlingen tijdens hun basisschooljaren om hen optimaal te ondersteunen, blijven alle LVS-toetsen nuttig. Ook voor andere onderzoeksvragen zijn de uitgebreidere datasets van waarde.

De meest relevante LVS-toets in onze steekproef is de tussentijdse toets in groep zeven. De scores op de tussentijdse toets van groep zeven zijn even voorspellend voor de eindtoetsscore als de scores op alle toetsen die zijn afgenomen in groep drie tot en met groep zes, met een verklarende kracht van 66%. Ook als latere toetsmomenten worden toegevoegd, blijven de toetsen uit midden groep zeven de belangrijkste informatiebron voor de voorspellingen. Veel scholen stellen eind groep zeven al voor het eerst een voorspelling van het schooladvies op, het zogenoemde 'pre-advies'. De tussentijdse toets in groep zeven is dus de laatste toets die leerlingen maken voor dat pre-advies. Deze toets kan door leerlingen als even belangrijk (*high stakes*) worden beschouwd als de eindtoetsen.

De voorspellingen zijn voor bepaalde groepen leerlingen minder nauwkeurig dan voor andere. Aan het begin van het basisonderwijs zijn de voorspellingen minder precies voor kinderen uit gezinnen met een lagere sociaal-economische status en voor kinderen met een migratieachtergrond, maar die verschillen verdwijnen tegen het einde van de basisschool. Het verschil naar sociaal-economische status komt overeen met bevindingen uit andere landen. Verder vinden we dat een verschil in voorspelkwaliteit tussen meisjes en jongens pas tegen het einde van de basisschool ontstaat, waarbij de eindtoetsscores van meisjes beter te voorspellen zijn.

De voorspellingen zijn in het algemeen nauwkeuriger voor gemiddelde leerlingen, die het op de eindtoets noch veel beter, noch slechter doen dan hun medeleerlingen. Deze bevinding is belangrijker voor de leerlingen met lagere scores: het spectrum waarin de voorspelkwaliteit afneemt, omvat relevante schooladviezen naar verschillende vmbo-niveaus. Voor leerlingen met hoge toetsscores maakt het minder uit: de kwaliteit van de voorspellingen verslechtert binnen het spectrum van scores waarvoor al een vwo-advies wordt gegeven.

De LVS-gegevens zijn uitermate geschikt om het onderwijssysteem en kansenongelijkheid te onderzoeken. Met dank aan het Nationaal Cohortonderzoek Onderwijs (NCO) zijn de LVS-gegevens beschikbaar om de efficiëntie en eerlijkheid van het Nederlandse basisonderwijs te onderzoeken. Zonder de toetslast voor leerlingen te verhogen, maken de LVS-gegevens het mogelijk om beleid te evalueren. Hierbij gaat het bijvoorbeeld om veranderingen van de rol van de eindtoets, en om inzicht te krijgen in gebeurtenissen die niet waren voorzien, zoals de sluiting van scholen tijdens de coronacrisis.

1 Inleiding

Basisschoolleerlingen worden regelmatig getest door middel van gestandaardiseerde toetsen. Vanaf groep drie op de basisschool moeten leerlingen twee keer per jaar deelnemen aan gestandaardiseerde toetsen (LVS – leerlingvolgsysteem). Het doel van deze verplichte toetsen is om de leerlingen en leerkrachten een goede inschatting te geven van de individuele leervorderingen, en ook om inzicht te geven in hoe goed een klas of school het doet ten opzichte van andere. De toetsscores van de verplichte LVS-onderdelen taal en rekenen worden op nationaal niveau verzameld door het NCO (Nationaal Cohortonderzoek Onderwijs) om onderzoek te doen naar het onderwijssysteem, zonder dat leerlingen extra toetsen hoeven af te leggen.

Aan het einde van de basisschool moeten leerlingen ook een gestandaardiseerde toets maken (voorheen eindtoets, tegenwoordig doorstroomtoets), die gebruikt wordt om te bepalen welke vorm van voortgezet onderwijs het beste is voor de betreffende leerling. Het tijdstip van deze toets en de specifieke rol ervan in het schooladvies is gewijzigd in 2016 en opnieuw in 2024, toen de eindtoets de doorstroomtoets werd. De doelstelling van deze toets bleef echter hetzelfde, namelijk ondersteuning bieden bij het bepalen van het schooladvies door het verstrekken van een tweede, objectief oordeel. Zowel de LVS-toetsen als de eindtoets meten de prestaties van een leerling in kernvakken (taal en rekenen), maar ze verschillen in de manier waarop ze worden afgenomen, wanneer ze worden afgenomen en de aandacht die eraan wordt besteed.

Deze CPB-studie onderzoekt op verzoek van het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) hoe voorspellend de toetsen uit het LVS zijn voor de scores op de eindtoets. Het doel van de eindtoets is om de opgebouwde kennis en vaardigheden van leerlingen aan het einde van de basisschool te meten, zodat de leerlingen een passend schooladvies kunnen krijgen. We onderzoeken in hoeverre de toetsscores van het LVS kunnen voorspellen hoe de leerlingen zullen presteren tijdens de eindtoets, en dus in hoeverre de specifieke vaardigheden die in de eindtoets worden gemeten al eerder bekend zijn.[1] Verder onderzoeken we vanaf welk leerjaar we goede voorspellingen kunnen doen, en of goede voorspellingen mogelijk zijn met minder datapunten.

Niet alle leerlingen hebben dezelfde gemeten leercurve, waardoor de scores op de eindtoets voor sommige groepen lastiger te voorspellen zijn. Uit een onderzoek onder leerlingen in het voortgezet onderwijs in het Verenigd Koninkrijk blijkt dat het moeilijker is om de eindexamenscores te voorspellen voor leerlingen uit families met een lagere sociaal-economische status (Wyness, 2022). Er zijn meerdere redenen waarom er ook in onze studie groepsspecifieke verschillen in voorspelkwaliteit kunnen zijn, zoals verschillen in motivatie voor de eindtoets, verschillende patronen in de leercurve of verschillen in ontbrekende uitslagen van LVS-toetsen. Dit zou kunnen leiden tot een lagere voorspelkwaliteit voor bepaalde groepen in het algemeen, of op verschillende momenten in de basisschool. Het zou bijvoorbeeld kunnen dat we voor één groep slechter voorspellen aan het begin van de basisschool, maar dat de kwaliteit van de voorspellingen naar elkaar toegroeien, maar ook het tegenovergestelde is mogelijk. Daarom onderzoeken we ook de groepsspecifieke voorspelkwaliteit gedurende de basisschool.

Deze publicatie is als volgt opgezet. Hoofdstuk twee geeft een toelichting op de gestandaardiseerde toetsen die we onderzoeken en de steekproef die we gebruiken. In hoofdstuk drie worden de voorspellingen

[1] De voorspellingen geven aan in welke mate de resultaten van de eindtoets al in een eerder stadium kunnen worden verwacht. De verschillen tussen de voorspelde en de uiteindelijke eindtoetsscore kunnen het gevolg zijn van verschillen in de gemeten vaardigheden tussen LVS en eindtoetsscores, verschillen in de geleverde inspanning in de verschillende toetstypen, of meetfouten. We kunnen deze verschillende redenen echter niet van elkaar onderscheiden.

geïntroduceerd. Hoofdstuk vier toont de belangrijkste resultaten voor de volledige steekproef, terwijl de resultaten die inzoomen op groepsverschillen worden besproken in hoofdstuk 5.

2 Data

Sinds het schooljaar 2014/2015 zijn scholen verplicht om de voortgang van alle leerlingen in groep drie tot en met acht bij te houden met een leerlingvolgsysteem, en om bij leerlingen aan het eind van hun basisschooltijd, in groep acht, een eindtoets af te nemen. Voor deze beleidswijziging deden veel scholen dit echter al vrijwillig. Leerlingen worden regelmatig getest met gestandaardiseerde Nederlandse taal- en rekentoetsen. Op leerlingniveau helpen deze toetsen bij het volgen van de leergroei (LVS) en de plaatsing in een passend traject in het voortgezet onderwijs. Op een meer geaggregeerd niveau kunnen deze gegevens helpen om specifieke groepen leerlingen of scholen te vergelijken. Hoewel de Dienst Uitvoering Onderwijs (DUO) de gegevens van de eindtoetsen, en nu de doorstroomtoets op nationaal niveau verzamelt en deze gegevens vaak zijn gebruikt in onderzoek, waren LVS-gegevens tot voor kort niet beschikbaar in een geaggregeerde databank.

Vanaf 2019 verzamelt het Nationaal Cohortonderzoek Onderwijs (NCO) de toetsresultaten van het LVS. Scholen zijn niet verplicht om deze gegevens te delen, maar er is een groeiend aantal scholen dat dit wel doet en gebruikmaakt van de door NCO verstrekte schoolrapporten waarin de eigen prestaties worden vergeleken met die van andere scholen. In tegenstelling tot vorige cohortonderzoeken, bijvoorbeeld COOL of PRIMA, hoeven leerlingen geen extra toetsen te maken voor het onderzoek, terwijl de steekproefomvang van het LVS aanzienlijk groter is. Het eerste cohort dat door het NCO is opgenomen, begon in 2013 aan groep drie en deed de eindtoets in 2019. Dit cohort is ook het laatste cohort dat in hun de basisschoolperiode niet werd geraakt door de coronacrisis. Aangezien de latere cohorten te maken kregen met schoolsluitingen tijdens de basisschoolperiode of geen eindtoets hebben gemaakt (het cohort dat in schooljaar 2019/20 in groep acht zat), beperken we onze studie tot het eerste cohort van het NCO.

Voor dit onderzoek zijn individuele LVS- en eindtoetsgegevens van leerlingen nodig. Omdat de toetsresultaten van LVS- én eindtoetsen niet goed vergelijkbaar zijn tussen verschillende toetsaanbieders (zie tekstkader), gebruiken we alleen de observaties waarbij de LVS-toetsen afkomstig zijn van Cito en waar de leerlingen de Centrale Eindtoets hebben gemaakt. Dat is de meest gebruikelijke eindtoets. Dit leidt tot een steekproef van 24.646 leerlingen, wat ongeveer 14% is van alle leerlingen die in 2019 hebben deelgenomen aan een eindtoets (en ongeveer 28% van de leerlingen die de Centrale Eindtoets hebben gemaakt). Met behulp van CBS-registergegevens kunnen we sociaaleconomische en demografische kenmerken koppelen aan leerlingen, en ook de kenmerken van de scholen die de leerlingen in hun laatste jaar van de basisschool bezochten. Meer informatie over de steekproef is te lezen in bijlage A.

Voor veel leerlingen observeren we een aantal, maar niet alle mogelijke LVS-toetsscores. Er zijn verschillende redenen waarom een bepaalde LVS-toetsscore niet in onze gegevens voorkomt. Bijvoorbeeld omdat de leerling de toets mist, omdat de klas de toets niet maakt of omdat er structurele administratieve problemen zijn. De belangrijkste oorzaak van ontbrekende toetsscores betreft het type toets dat is afgenomen. Alleen de latere generaties toetsen zijn opgenomen in de gegevens die door de NCO worden verzameld. Hoewel de scores van de oude generaties reken- en leestoetsen zijn omgezet naar scores van de laatste generatie, was dit niet mogelijk met scores van de oudere generaties spellingtoetsen. Daarom zien we veel meer ontbrekende spellingtoetsen in onze data. Ontbrekende LVS-toetsen worden uitgebreid besproken in bijlage B.

Toetsen in het basisonderwijs

Er zijn verschillende soorten toetsen in het basisonderwijs en elke toets kan bij verschillende aanbieders worden ingekocht. Dit kan leiden tot verwarring. Met name de term Cito-toets kan gebruikt worden voor zowel de LVS-toetsen als de eindtoets, en het is in ons onderzoek belangrijk om deze van elkaar te onderscheiden. In dit tekstkader specificeren we daarom de toetsen die we in ons onderzoek gebruiken en hoe we ernaar verwijzen.

In deze publicatie gebruiken we alleen informatie van de verplichte onderdelen van gestandaardiseerde toetsen, te weten taal en rekenen. Verder gebruiken scholen ook hun eigen toetsen die meten of de leerlingen de huidige leerdoelen behalen. Deze toetsen zijn niet beperkt tot kernvakken, maar kunnen in elk vak worden toegepast. Zulke toetsen maken geen deel uit van het NCO en kunnen daarom niet worden gebruikt voor onze analyses.

Leerlingvolgsysteem (LVS)

Vanaf groep drie doen de meeste basisschoolleerlingen twee keer per jaar mee aan gestandaardiseerde toetsen, in ten minste Nederlandse taal en rekenen/wiskunde. Deze toetsen maken deel uit van het leerlingvolgsysteem dat basisscholen moeten gebruiken en dat de voortgang van individuele leerlingen en groepen in kaart brengt. Vanaf groep drie van de basisschool maken leerlingen tussen december en maart een M-toets (midden) en tussen april en juli een E-toets (eind). Al deze toetsen worden LVS-toetsen genoemd.

Scholen in het basisonderwijs kunnen kiezen uit vijf verschillende geaccrediteerde aanbieders van LVS-toetsen. De toetsen verschillen per aanbieder in wat ze precies meten en hoe de toetsen worden afgenomen. Dit maakt het moeilijk om toetsresultaten van verschillende aanbieders met elkaar te vergelijken. Het NCO beperkt daarom op dit moment de beschikbare gegevens voor onderzoek tot toetsen van de grootste aanbieder, namelijk Cito.

Eindtoets

Aan het einde van de basisschool maken de meeste leerlingen wat tegenwoordig de doorstroomtoets heet (voorheen: de eindtoets). Het belangrijkste doel van die toets is om te helpen bij het bepalen van het schooladvies voor het voortgezet onderwijs. De toets omvat de vakken Nederlandse taal en rekenen. De gecombineerde scores van de eindtoetsen kunnen worden vertaald in een trajectadvies voor het voortgezet onderwijs. Scholen zijn vrij om een toetsaanbieder te kiezen uit een lijst van zes geaccrediteerde aanbieders (in 2019 waren het vijf aanbieders).

De aanbieder van de LVS-toets kan, maar hoeft niet dezelfde te zijn als de aanbieder van de toets aan het einde van het basisonderwijs. Hoewel verschillende aanbieders zowel eindtoetsen als LVS-toetsen aanbieden, staat het scholen vrij om verschillende aanbieders te kiezen voor de twee verschillende soorten toetsen. We beperken onze steekproef in dit onderzoek tot leerlingen die op scholen zaten die gebruik maakten van de Cito-versie van de LVS-toetsen en de Centrale Eindtoets van het College voor Toetsen en Examens (CvTE).

Onze voorspelmethode is geschikt voor analyses met ontbrekende datapunten, zodat ook leerlingen met ontbrekende toetsscores in het onderzoek kunnen worden meegenomen. Dit betekent dat de steekproef niet beperkt hoeft te worden tot alleen de leerlingen van wie elke toetsscore bekend is, maar dat alle leerlingen die een of meer LVS-toetsscores hebben en de eindtoets hebben gemaakt, in de steekproef kunnen worden meegenomen. De ontbrekende scores hebben invloed op de nauwkeurigheid van de voorspellingen, vooral in de eerste jaren van de basisschool omdat er dan meer ontbrekende scores zijn, maar ze hebben geen invloed op onze steekproefselectie. De voorspelmethode wordt in meer detail uitgelegd in bijlage C.

3 De waarde van LVS-toetsscores bij het voorspellen van de eindtoets

Voor elke leerling voorspellen we de eindtoets diverse keren, waarbij we bij elke nieuwe voorspelling steeds meer informatie over de LVS-toets toevoegen. De voorspellingen zijn gebaseerd op de LVS-toetsscores van rekenen/wiskunde, begrijpend lezen en spelling. Iedere voorspelling is gebaseerd op scores van toetsen die op bepaalde momenten tijdens de schoolperiode van een leerling zijn gemaakt (bijvoorbeeld alle toetsen van groep drie tot groep vijf). Daarnaast gebruiken we individuele achtergrondinformatie in elke voorspelling, zoals geslacht, opleiding en inkomen van de ouders, migratieachtergrond en de regio en onderwijsachterstandsscore van de school waarop de leerlingen zitten. Om deze voorspellingen te maken, gebruiken we een *machine learning*-algoritme, namelijk de *random forest*, zie bijlage C voor meer toelichting.[2] De *random forest*-methode baseert de voorspellingen niet op een vooraf gedefinieerd structureel model, dus we kunnen geen causaal verband afleiden tussen specifieke LVS-toetsscores en de eindtoetsscore.

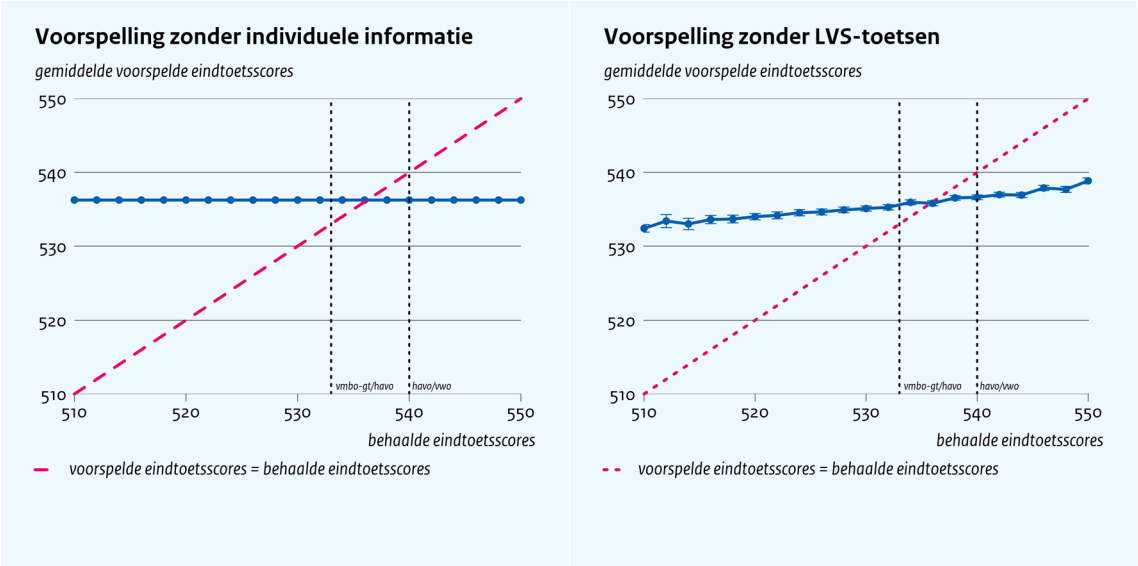
Uitgaande van een dataset zonder individuele informatie voorspellen we voor elke leerling dezelfde gemiddelde eindtoetsscore. In het linkerpaneel van figuur 1, waarin de gemiddelde voorspelde toetsscore voor elke werkelijk behaalde toetsscore wordt weergegeven, zien we dat deze voorspelling een horizontale lijn is bij de gemiddelde score van 536. De lijn van 45 graden staat voor de punten waarop de voorspelde toetsscores gelijk zijn aan de werkelijke toetsscores van de leerlingen. Hoe verder onze voorspellingen van de diagonale lijn afliggen, hoe slechter de kwaliteit van die voorspellingen is. De gemiddelde voorspelde toetsscore ligt op de meeste punten en voor bijna alle leerlingen ver van de werkelijke toetsscore. Naarmate scores dichter bij het gemiddelde komen, wordt deze afstand kleiner. Door het gemiddelde aan elke leerling toe te kennen, voorspellen we de juiste eindtoetsscore alleen voor die leerlingen die precies de gemiddelde score behalen. Tegelijkertijd overschatten we de scores voor leerlingen die lager dan het gemiddelde scoren en onderschatten we de scores voor leerlingen die boven het gemiddelde scoren.

Het toevoegen van individuele kenmerken (rechterpaneel van figuur 1) verbetert de voorspellingen enigszins. De afstand tussen de voorspelde en de behaalde scores wordt kleiner, en dus komen de voorspellingen dichter bij de 45 graden-lijn. De blauwe lijn, die de verhouding tussen de behaalde en voorspelde eindtoetsscores toont, blijft echter relatief vlak. Ook hier zien we dat scores rond het gemiddelde heel goed voorspeld worden, en dat de voorspellingen slechter worden richting de uitersten van de verdeling.

[2] De *random forest*-methode doet voorspellingen voor individuen door hun een gemiddelde toe te wijzen dat is gebaseerd op de gemiddelde uitkomst van vergelijkbare groepen. Omdat deze methode voorspellingen berekent op basis van gemiddelden, zal die nooit voorspellingen maken buiten of aan de rand van het spectrum van mogelijke eindtoetsscores. Dit leidt ertoe dat er zeer weinig voorspellingen worden gemaakt voor punten aan het uiterste van de verdeling.

Scores tussen gemengde vmbo-gt- en havo-adviezen en gemengde havo- en vwo-adviezen worden goed voorspeld, omdat ze rond de gemiddelde behaalde score liggen. De consistente overschatting van scores onder het gemiddelde en onderschatting van scores boven het gemiddelde duidt op een systematische afwijking in onze voorspellingen. Alle voorspellingen liggen binnen een bereik van 528-543 punten (y-as), terwijl de werkelijke scores die we observeren binnen een bereik van 501-550 punten liggen (x-as). Dit betekent dat we op jonge leeftijd niet goed kunnen voorspellen wat voor schooladvies kinderen zullen krijgen.

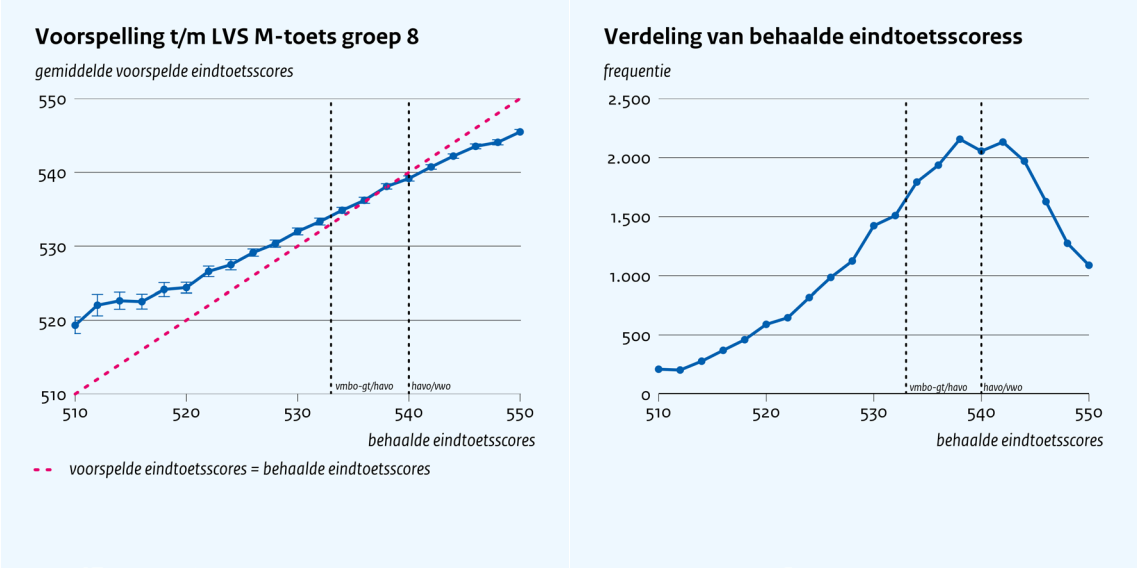
Figuur 1 Gemiddelde van voorspellingen per eindtoetsscore indien geen individuele informatie beschikbaar is (links) of er alleen achtergrondinformatie wordt gebruikt voor de voorspellingen (rechts)



Noot: De figuur links geeft de beste voorspelling van de eindtoetsscores weer die gedaan kan worden zonder extra informatie: namelijk het toekennen van de gemiddelde eindtoetsscore aan iedereen. De figuur rechts toont de voorspellingen van de eindtoets als alleen achtergrondkenmerken worden gebruikt voor de voorspelling. De gemiddelde voorspelde scores zijn gegroepeerd naar de resultaten van de werkelijk ontvangen score op de eindtoets. Om kleine aantallen leerlingen per groep te vermijden, omvat elke groep twee toetsscores en worden alle toetsscores onder de 510 gegroepeerd in de eerste groep.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

De voorspelling is relatief goed als alle LVS-toetsscores aan het model worden toegevoegd. De gemiddelde voorspelde toetsscores liggen dicht bij de werkelijke toetsscores, maar dit geldt wederom meer voor scores rond het gemiddelde en het neemt af richting de uitersten (figuur 2, links). De afwijking die wordt geïntroduceerd door de overschatting van lage scores en onderschatting van hoge scores vermindert als we alle scores gebruiken om voorspellingen te doen. De afwijking blijft echter aanwezig en onze voorspellingen vallen nog steeds in een beperkt bereik, namelijk van 514- 549 punten. Als we kijken naar de spreiding van toetsscores zien we dat de voorspelkwaliteit samenhangt met het aantal leerlingen dat een bepaalde score op de eindtoets behaalt. De frequentie van toetsscores piekt tussen de adviesdrempels vmbo-gt/havo en havo/vwo (figuur 2, rechts), waar de toetsscores het best voorspeld worden. Hier zijn meer datapunten beschikbaar voor het doen van de voorspellingen, wat kan bijdragen aan nauwkeurigere voorspellingen. Verder weg van de gemiddelde toetsscore neemt de frequentie van toetsscores ook af en verslechtert de voorspelkwaliteit. Scores aan de onderkant van de verdeling worden slechter voorspeld dan scores aan de bovenkant. Dit komt overeen met het feit dat er aan de onderkant minder informatie is om het model te trainen dan aan de bovenkant.

Figuur 2 Gemiddelde van voorspellingen per eindtoetsscore als voorspellingen zijn gebaseerd op alle LVS-toetsscores en achtergrondkenmerken

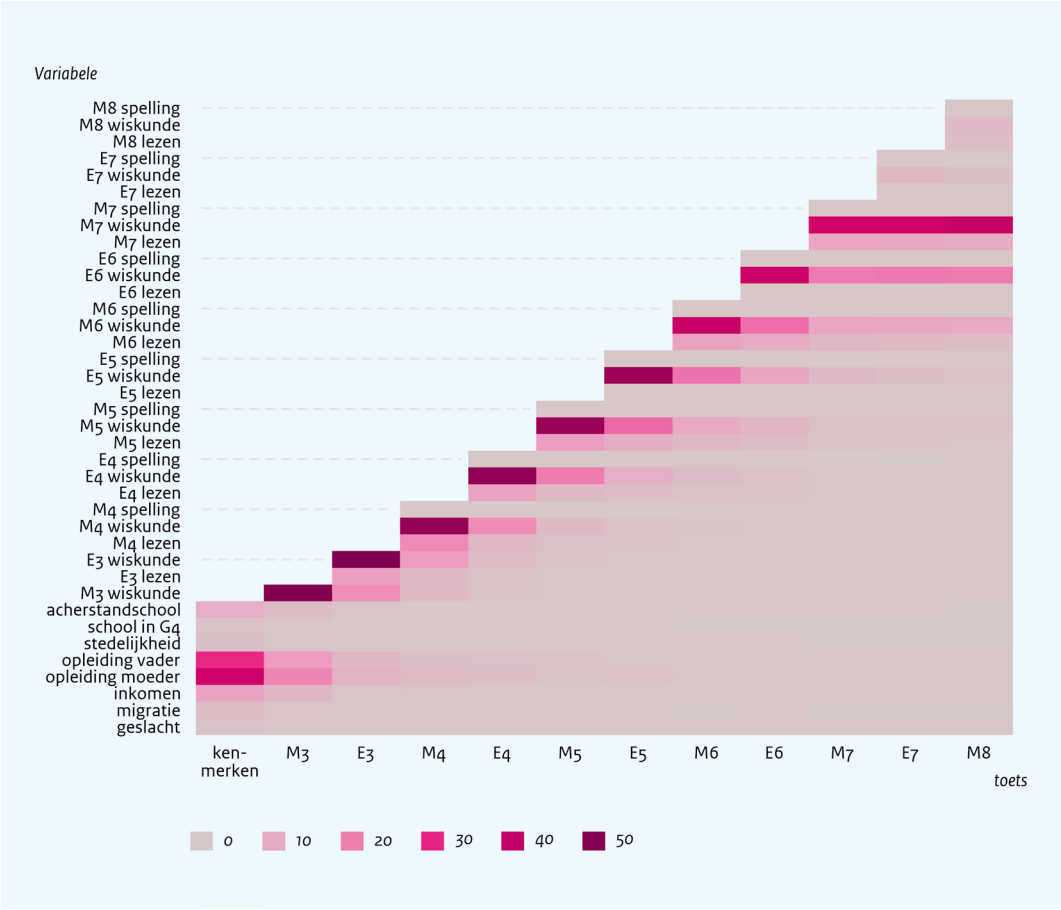


Noot: Het linker paneel toont de gemiddelde voorspellingen van de eindtoets gebaseerd op alle LVS-toetsscores. Het rechter paneel toont hoeveel leerlingen in de volledige steekproef (test en training) een specifieke eindtoetsscore hebben gekregen. Om kleine aantallen leerlingen per groep te voorkomen, omvat elke groep twee toetsscores en zijn alle toetsscores onder de 510 samengevoegd in de eerste groep. Deze aanpassingen leiden tot een relatief groter aantal observaties in de eerste groep (510 omvat nu alle observaties van leerlingen die tussen 501-510 punten scoren). Ook verdwijnt aan de rechterkant van de scoreverdeling de karakteristieke piek bij de laatste score (topcodering door de toets) omdat deze is opgenomen in een grotere groep.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

Voorspellingen van eindtoetsscores zijn voornamelijk gebaseerd op de meest recente rekentoets, terwijl de kenmerken van leerlingen minder belangrijk worden naarmate we meer LVS-toetsen gebruiken om voorspellingen te doen. Wanneer voorspellingen alleen gebaseerd zijn op leerlingkenmerken, draagt de opleiding van de ouders van de leerling het meest bij aan de voorspelling van de eindtoetsscore. Naarmate er meer toetsscores worden toegevoegd, zien we dat de opleiding van de ouders steeds minder bijdraagt aan de voorspelling van de eindtoetsscore. Naarmate nieuwere toetsscores worden meegenomen in de voorspelling van de eindtoets, worden de rekentoetsscores van de laatst afgenomen toets de belangrijkste voorspeller, terwijl de leesscores enigszins belangrijk zijn en de spellingscores onbelangrijk blijven. Dit komt waarschijnlijk doordat de rekentoetsen de best gevulde toetsen in onze dataset zijn, met 86% van de LVS-rekentoetsen aanwezig, terwijl gemiddeld twee derde van de leestoetsen aanwezig is en slechts een derde van de spellingtoetsen.[3] We kunnen niet met zekerheid zeggen of de rekentoets de belangrijkste bron van informatie zou blijven als alle toetsscores even goed gevuld zouden zijn. Met name de M-toets rekenen/wiskunde van groep zeven – de toets die werd afgenomen voordat leerlingen hun voorlopig schooladvies kregen – blijft de belangrijkste variabele voor de voorspellingen, zelfs als de toetsen aan het einde van de groep zeven en de M-toets van groep acht worden toegevoegd. Merk op dat, hoewel figuur 3 variabelen toont die vaak bijdragen aan de voorspelling, er geen causaal verband kan worden afgeleid tussen belangrijke variabelen en de eindtoetsscore.

[3] Dit patroon van ontbrekende datapunten is specifiek voor het eerste cohort van de LVS-gegevens in het NCO. In de toekomst verwachten we dat ook begrijpend lezen en spelling beter gevuld zullen zijn, waardoor de voorspellende kracht van deze variabelen zou kunnen toenemen.

Figuur 3 Bijdrage van variabelen aan het voorspellen van eindtoetsscores



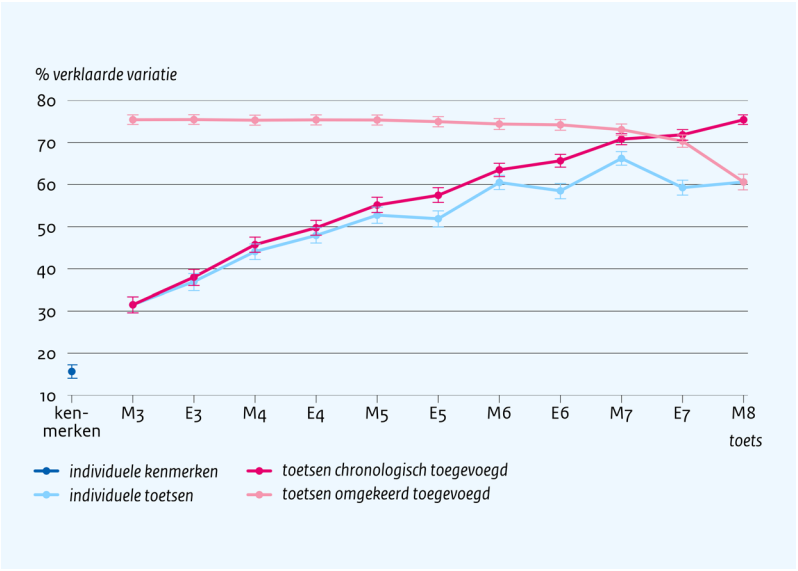
Noot: De figuur illustreert de bijdrage van elke inputvariabele, weergegeven op de y-as, tot de totale nauwkeurigheid van de voorspellingen. Deze bijdragen worden uitgedrukt in percentages en geven het relatieve belang van elke variabele weer. De hoogste waarde voor een unieke input bedraagt 50,8%, en behoort tot een model met weinig LVS-inputs. Het belang van een variabele is gebaseerd op het aantal keer dat de variabele gekozen is om uitsplitsingen te maken, waarbij uitsplitsingen aan het begin meer gewicht krijgen dan splitsingen aan het eind (Tibshirani et al., 2022). Elke voorspelling wordt weergegeven op de x-as. De eerste voorspelling gebruikt alleen kenmerken, en dus hebben daar alleen deze inputs bijbehorende belangrijkheidspercentages. Bij elke volgende voorspelling worden de scores van het volgende LVS-toetsmoment toegevoegd aan het model.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

4 Voorspelkwaliteit op basis van verschillende testmomenten

De voorspelling van de eindtoetsscores verbetert continu naarmate leerlingen de basisschool doorlopen en verklaart uiteindelijk ongeveer 75% van de variatie in de eindtoetsscores. Voorspellingen op basis van toetsscores in groep drie met achtergrondvariabelen verklaren slechts 32% van de variatie in de eindtoetsscore. Dit is echter al een grote verbetering ten opzichte van de voorspellingen op basis van alleen individuele kenmerken, die slechts 16% van de variatie verklaren. De donkerroze lijn in figuur 4 laat zien dat de voorspellingen continu verbeteren, naarmate er meer toetsen aan het model worden toegevoegd. Het beste

model bevat scores van elke LVS-toets die op de basisschool is afgenomen, samen met individuele kenmerken, en is in staat om meer dan 75% van de variatie in eindtoetsscores te verklaren.[4]

Figuur 4 Kwaliteit van voorspelling van eindtoetsscores afhankelijk van meegenomen inputvariabelen



Noot: Deze figuur laat zien hoeveel van de variatie in eindtoetsscores kan worden verklaard door voorspellingen die gebaseerd zijn op verschillende sets van inputvariabelen. De blauwe punt aan de linkerkant is de voorspelling die alleen is gebaseerd op kenmerken van de leerlingen en hun scholen. De donkerroze lijn toont de voorspellingen die gebaseerd zijn op achtergrondkenmerken en alle LVS-toetsen tot en met het toetsmoment dat op de x-as staat. De lichtroze lijn laat de voorspellingen zien die de leerlingkenmerken en de meest recente toets omvatten, tot en met het relevante toetsmoment. De lichtblauwe lijn toont de voorspellingen die alleen gebaseerd zijn op één specifiek toetsmoment (en achtergrondkenmerken). De verticale lijnen (*whiskers*) geven het 95%-betrouwbaarheidsinterval aan.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

Met alleen de scores van de laatste drie LVS-toetsen kunnen we de eindtoetsscores al goed voorspellen. Het gebruik van de laatste drie toetsen in de voorspelling levert zelfs bijna dezelfde voorspelkwaliteit op als het meenemen van alle toetsscores. Als we de eindtoets goed willen voorspellen, hoeven we alleen te kijken naar de laatste twee jaar van de LVS-toetsen. De complete dataset van toetsscores is dus niet nodig voor het specifieke doel van het voorspellen van de eindtoets. Tegelijkertijd kunnen deze toetsscores wel van belang zijn voor mogelijk ander onderzoek.

De meest relevante toets in onze steekproef is de M-toets in groep zeven. De scores op de M-toets in groep zeven zijn zelfs even voorspellend voor de eindtoetsscore als de scores op alle toetsen die zijn afgenomen in groep drie tot en met groep zes. De M-toets in groep zeven is de laatste toets die leerlingen maken voordat ze een pre-advies krijgen: een eerste inschatting van de leerkracht over welk schooladvies de leerlingen gaan krijgen. Deze toets kan door leerlingen als even belangrijk (*high stakes*) worden beschouwd als de eindtoetsen. Daarom kunnen leerlingen hun studie-inzet voor deze toets verhogen, op een vergelijkbare manier als voor de eindtoets (Bach & Fischer, 2020). Vanaf schooljaar 2023/2024 wordt de doorstroomtoets (voorheen: eindtoets) afgenomen vóór de M-toets in groep acht. Het hebben van één LVS-toets minder om de eindtoetsscores te voorspellen, lijkt de voorspelkwaliteit echter niet veel te beïnvloeden.

[4] In deze publicatie gebruiken we de term 'variatie' om over het verklaarde deel van de eindtoetsscores te spreken. In de statistiek is het echter gebruikelijker om te spreken over 'variantie'.

5 Verschillen in voorspelkwaliteit

Verschillen in voorspelkwaliteit tussen leerlingen

Voorspellingen gaan gepaard met voorspelfouten, wat deels willekeurig is, maar ook deels gerelateerd kan zijn aan individuele kenmerken. Er zijn veel redenen waarom het moeilijker kan zijn om de eindtoetsscores van sommige leerlingen te voorspellen dan die van anderen. De leercurve kan tussen leerlingen verschillen, waarbij sommige leerlingen hun potentieel eerder bereiken dan anderen, wat leidt tot verschillen in voorspelkwaliteit gedurende de basisschoolperiode. Ook de motivatie om goed te presteren, varieert tussen leerlingen in de loop van de tijd. Met name de motivatie om goed te presteren op de eindtoets in vergelijking met de LVS-toetsen kan sterk variëren tussen leerlingen. Er zijn ook verschillen tussen scholen, zoals in de kwaliteit en middelen, die van invloed kunnen zijn op het ontwikkelingstraject van de scores.

Bovenop individuele verschillen in leerontwikkeling kan ook het LVS zelf de leercurve beïnvloeden. Terwijl dit onderzoek het LVS gebruikt als een instrument om de vaardigheden van leerlingen op een bepaald punt in hun onderwijstraject te meten, is het hoofddoel van het LVS van diagnostische aard. Leerkrachten leren meer over de actuele sterktes en zwaktes van hun leerlingen en kunnen bepaalde delen van de lessen afstemmen op de behoeften van de leerlingen. Leerlingen met een achterstand kunnen specifieke ondersteuning krijgen, terwijl hoogbegaafde leerlingen extra lesopdrachten erbij kunnen krijgen. Dergelijke reacties op de LVS-resultaten kunnen het academische traject van een leerling veranderen, waardoor de voorspelbaarheid van hun prestaties afneemt.

Toetsresultaten van leerlingen die lager scoren, zijn moeilijker te voorspellen. Omdat elke voorspelling gebaseerd is op gemiddelde scores van kleine subgroepen van onze steekproef, zijn leerlingen met hele lage of hoge scores in het algemeen moeilijker te voorspellen. Zeker als er voor die leerlingen niet veel beschikbare datapunten zijn. Hoewel dat aan beide kanten van de verdeling speelt, is dit relevanter voor de leerlingen met lagere scores. Het spectrum van scores waar de voorspelkwaliteit afneemt omvat relevante schooladviezen binnen het vmbo, terwijl aan de bovenkant de voorspelkwaliteit verslechtert binnen het spectrum van scores die met een vwo-advies al het hoogste schooladvies krijgen.

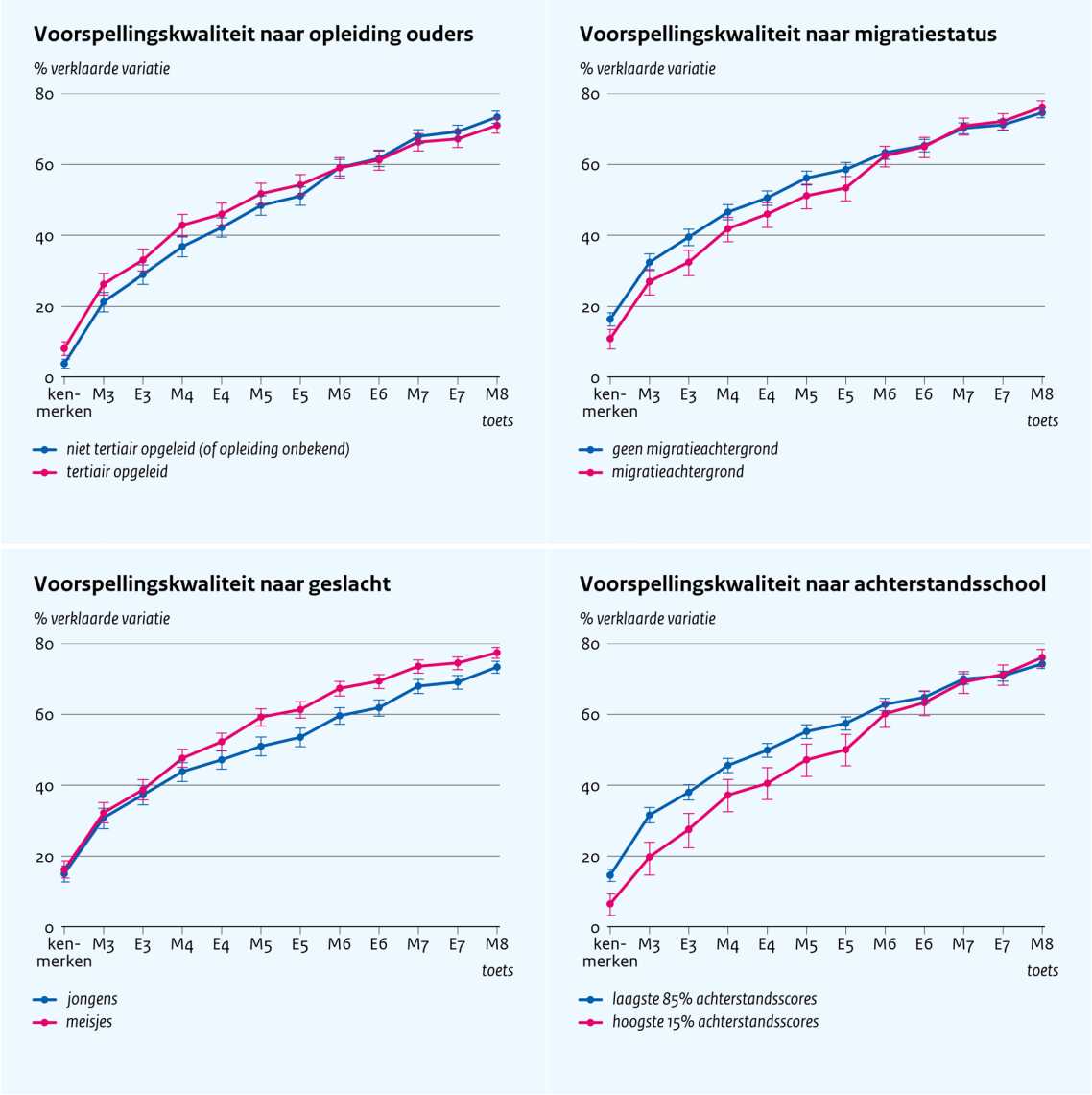
Verschillen in voorspelkwaliteit tussen leerlingen kunnen ook optreden als we niet voor iedereen evenveel toetsen observeren. We kunnen eindtoetsscores beter voorspellen naarmate we meer LVS-toetsen voor een leerling waarnemen. Er is echter enige variatie in het aantal LVS-toetsscores dat we per leerling observeren. Verder zien we ook variatie in het aantal geregistreerde toetsen tussen groepen van leerlingen op basis van hun achtergrondkenmerken. Met name leerlingen met een migratieachtergrond, die in een G4-stad wonen of in een niet-stedelijk gebied, en leerlingen van wie de ouders in het laagste inkomenskwintiel vallen, hebben significant meer ontbrekende toetsscores dan de gemiddelde leerling. Dit wordt in detail beschreven in bijlage B.

Verschillen in voorspelkwaliteit tussen groepen

Hoewel de voorspelkwaliteit voor iedereen toeneemt gedurende het basisonderwijs, verschilt de voorspelkwaliteit en de verbetering van de voorspelkwaliteit tussen groepen. We kunnen de voorspelkwaliteit op elk meetmoment afzonderlijk in kaart brengen voor verschillende leerlingkenmerken. Op basis hiervan kunnen we beoordelen voor welke groep leerlingen al vroeg in de basisschool de

eindtoetsscores relatief goed voorspeld kunnen worden en voor wie we zouden moeten wachten tot er meer toetsen zijn afgenomen. Hierbij moet worden opgemerkt dat een betere voorspelkwaliteit niet noodzakelijkerwijs betekent dat de toetsscores van een groep leerlingen beter zijn dan van anderen, maar dat de voorspelde score dichter bij de eindtoetsscore ligt die ze uiteindelijk behalen. Met behulp van de 95% betrouwbaarheidsintervallen kunnen we de voorspelkwaliteit tussen de verschillende groepen vergelijken.

Figuur 5 Heterogeniteit in voorspelling



Noot: Deze cijfers tonen voor verschillende groepen hoeveel van de variatie in eindtoetsscores kan worden verklaard door voorspellingen op basis van achtergrondkenmerken en alle beschikbare LVS-toetsscores op verschillende momenten van de basisschool.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

De voorspellingen zijn aan het begin van het basisonderwijs relatief beter voor leerlingen uit gezinnen met een hogere sociaal-economische status dan voor leerlingen uit gezinnen met een lage sociaal-economische status, maar de kloof wordt kleiner in latere jaren. Leerlingen uit huishoudens met een hoger inkomen of met tertiair opgeleide ouders (hoger beroepsonderwijs en wetenschappelijk onderwijs) hebben mogelijk al op jongere leeftijd toegang tot meer onderwijsmiddelen (Dickson et al., 2016). Met deze middelen kan een leerling al op jongere leeftijd zijn potentieel bereiken, wat leidt tot een hogere voorspelkwaliteit van de eindtoetsscores op jongere leeftijd. Omgekeerd kunnen leerlingen uit lagere sociaal-

economische milieus te maken krijgen met uitdagingen zoals een minder stabiele thuisomgeving, wat leidt tot wisselvallige schoolprestaties, en daarmee tot een slechtere voorspelkwaliteit (Wyness, 2022). Bovendien, zoals figuur D.1 in bijlage D laat zien, voorspellen we de eindtoetsscores beter voor leerlingen die in de top van de scoreverdeling zitten. Leerlingen uit gezinnen met een hogere sociaaleconomische status zitten vaker in deze groep. Over het geheel genomen is het verschil in voorspellingen echter niet erg groot en de vroege kloof is nauwelijks statistisch significant op 95%-niveau.

Aan het begin van de basisschool is er een groter verschil in voorspelkwaliteit tussen leerlingen met en zonder migratieachtergrond dan aan het eind. De variatie in taalvaardigheid van de migrantengroep kan bijdragen aan de slechtere voorspellingen op jongere leeftijden. De groep van leerlingen met migratieachtergrond bevat eerste- en tweedegeneratiemigranten, die buiten schooltijd wisselend in aanraking komen met de Nederlandse taal, wat leidt tot een uiteenlopend taalvaardigheidsniveau in de eerste jaren van de basisschool. De grote variatie in scores binnen deze groep maakt het moeilijk om eindtoetsscores voor deze groep te voorspellen. Naarmate leerlingen met een migrantenachtergrond ouder worden, worden de verschillen in het taalniveau kleiner, waardoor de groep homogener wordt (Zumbuehl & Dillingh, 2020).

Meisjes en jongens beginnen met een vergelijkbare voorspelkwaliteit, maar dit verbetert gaandeweg de basisschool meer voor meisjes dan voor jongens. Meisjes vertonen meer gedragsregulatie en betere studiegewoonten, wat zou kunnen leiden tot meer consistente verbeteringen in toetsscores (Weis et al., 2013). Daarnaast verschilt de voorbereiding op LVS-toetsen en de eindtoets tussen jongens en meisjes. Terwijl meisjes zich op een meer vergelijkbare manier voorbereiden op de eindtoets als op andere toetsen, presteren jongens veel beter op toetsen met een hoog belang, zoals de eindtoets (Azmat et al., 2016).

Het patroon voor leerlingen op scholen met een hoge onderwijsachterstandsscore is vergelijkbaar met leerlingen uit gezinnen met een lagere sociaal-economische status. Aan het begin van de basisschool is de voorspelkwaliteit significant slechter voor scholen met een hoge onderwijsachterstandsscore. Dit verschil wordt kleiner in groep zes. Het verschil in voorspellingsnauwkeurigheid is in lijn met het verschil in het percentage gemiste toetsscores tussen scholen met een lagere en hogere achterstandsscore. Minder geobserveerde vroege toetsscores voor de leerlingen die naar achterstandsscholen gaan, maakt het moeilijker om al op jonge leeftijd de eindtoets te voorspellen. Dit zou voor een deel ook kunnen verklaren waarom we verschillen zien tussen leerlingen met verschillende opleidingsniveaus van de ouders en migratieachtergrond. De scores van leerlingen in de grote steden (G4) lijken iets makkelijker te voorspellen dan de scores van andere leerlingen, hoewel dit verschil statistisch niet significant is.

Referenties

Athey, S., Tibshirani, J., & Wager, W. (2019). Generalized random forests. *Ann. Statist.* *47*(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender Differences in Response to Big Stakes. *Journal of the European Economic Association*, *14*(6), 1372–1400. https://doi.org/10.1111/jeea.12180

Bach, M., & Fischer, M. (2020). Understanding the Response to High-Stakes Incentives in Primary Education. *ZEW – Centre for European Economic Research Discussion Paper No. 20-066*. https://doi.org/10.2139/ssrn.3736769

Dickson, M., Gregg, P., & Robinson, H. (2016). Early, Late or Never? When Does Parental Education Impact Child Outcomes? *The Economic Journal*, *126*(596), F184–F231. https://doi.org/10.1111/ecoj.12356

Fitzek, F. H., Granelli, F., & Seeling, P. (2020). *Computing in Communication Networks: From Theory to Practice*. Academic Press.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Nationaal Regieorgaan Onderwijsonderzoek (NRO) & Centraal Bureau voor de Statistiek (CBS). (2022). *Codeboek Data LVS-NCO-data (Versie 1.2)*.

Tibshirani, J., S. Athey, E. Sverdrup, and S. Wager (2022). The grf algorithm. https://grf-labs.github.io/grf/REFERENCE.html, Accessed 08-04-2024.

Weis, M., Heikamp, T., & Trommsdorff, G. (2013). Gender differences in school achievement: The role of self-regulation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00442

Wyness, G., Macmillan, L., Anders, J., & Dilnot, C. (2022). Grade expectations: How well can past performance predict future grades? *Education Economics*, *1*–22. https://doi.org/10.1080/09645292.2022.2113861

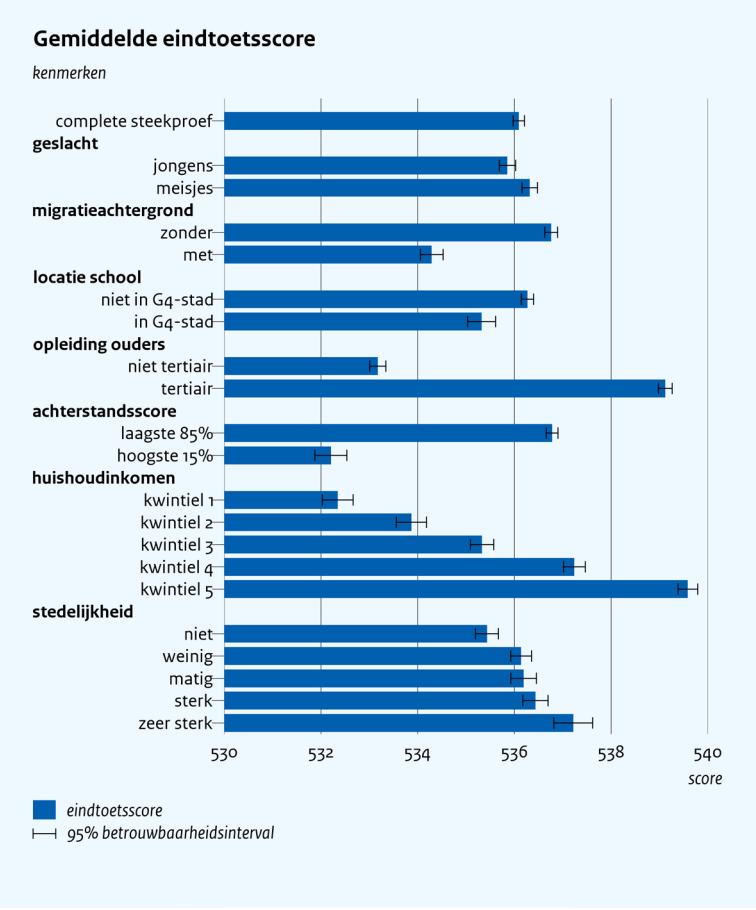
Twala, B. E. T. H., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956. https://doi.org/10.1016/j.patrec.2008.01.010

Bijlage A Steekproef

De steekproef van deze studie bestaat uit basisschoolleerlingen die in april 2019 (schooljaar 2018/19) een Centrale Eindtoets hebben gemaakt en van wie de LVS-scores, eveneens afkomstig van Cito, zijn verzameld door het Nationaal Cohortonderzoek Onderwijs (NCO). De NCO-steekproef die voor onderzoek beschikbaar is, bevat alleen leerlingen die de Cito-LVS toetsen hebben gemaakt. In het eerste cohort zijn er van 49.817 leerlingen toetsuitslagen bekend bij NCO. Als we de steekproef van leerlingen beperken tot alleen de leerlingen die de Centrale Eindtoets (van CvTE) hebben gemaakt, daalt de steekproefomvang naar 24.646. We nemen alleen leerlingen op die in groep drie tot en met acht dezelfde toetsen in dezelfde groep in hetzelfde jaar hebben gemaakt. Leerlingen die een klas hebben gedoubleerd , worden niet meegenomen.

De gemiddelde leerling haalt een score van 536 op de eindtoets. De mediane toetsscore (537) is hoger dan het gemiddelde, wat betekent dat de verdeling van de toetsscores scheef is in de richting van hogere scores. Leerlingen zonder migratieachtergrond, leerlingen die buiten de G4 wonen, leerlingen met ouders die tertiair onderwijs hebben gevolgd en in de hoogste inkomenskwintielen vallen, behalen gemiddeld hogere scores op de eindtoets. Zie ook figuur A.1.

Figuur A.1 Toetsscores per leerling



Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

Onze steekproef is niet volledig representatief voor de bredere cohortpopulatie, omdat het geen willekeurig getrokken steekproef is. De selectie is het resultaat van twee beslissingen van scholen. Ten eerste beslissen de scholen welke toetsaanbieders ze kiezen voor hun LVS en hun eindtoets (nu: doorstroomtoets). Alleen scholen die kozen voor LVS-toetsen van Cito en de Centrale Eindtoets, komen in onze steekproef terecht. Ten tweede beslissen de scholen of ze wel of niet deelnemen aan het NCO. Tabel A.1 toont het verschil in waarneembare kenmerken tussen onze steekproef en de bredere leerlingenpopulatie. Hiermee kunnen we beoordelen hoe representatief onze steekproef is. Er is een significant verschil in de gemiddelden van waarneembare kenmerken, wat aangeeft dat onze steekproef niet volledig representatief is voor de bredere populatie. Onze steekproef telt gemiddeld iets meer meisjes en iets meer leerlingen met een migratieachtergrond. Het inkomensniveau van de ouders is gemiddeld hoger in onze steekproef en ouders hebben vaker een tertiaire opleiding. Leerlingen in onze steekproef wonen vaker in een G4-stad, maar ook gemiddeld vaker in minder sterk verstedelijkte gebieden, en bezoeken vaker scholen met een hogere achterstandsscore.

Tabel A.1 Representativiteit van onze steekproef

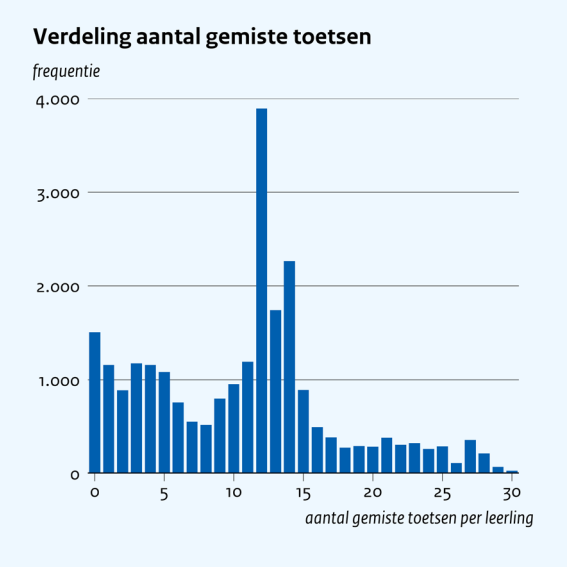
	steekproef			bredere leerlingpopulatie				
	N	gemid.	SD	N	gemid.	SD	verschil	T-stat
vrouw	24.646	0,51	0,50	178.445	0,50	0,50	0,01	3,36
met migratieachtergrond	24.646	0,27	0,44	178.447	0,26	0,44	0,02	6,44
inkomenskwintiel	24.477	3,26	1,37	176.825	3,17	1,36	0,10	11,04
stedelijkheid	24.645	2,51	1,26	178.360	2,63	1,26	-0,14	-16,08
ouders tertiair opgeleid	24.646	0,49	0,50	178.447	0,46	0,50	0,03	9,60
onderwijsachterstands-score van school	24.499	0,42	0,88	171.243	0,39	0,80	0,04	6,70
school in G4 stad	24.645	0,19	0,39	178.360	0,13	0,33	0,07	32,47

Noot: Verschillen in kenmerken tussen de steekproef van deze studie en het volledige cohort dat in 2019 de eindtoets deed. N en SD staan respectievelijk voor het aantal leerlingen voor wie het gemiddelde is berekend en voor de standaarddeviatie. T-stat staat voor *T-statistiek.*
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

Bijlage B Gemiste toetsen

Er ontbreekt een aanzienlijk aantal toetsscores in de LVS-gegevens. De LVS-gegevens bestaan uit dertig verschillende toetsen die gedurende de hele basisschool zijn afgenomen. Gemiddeld ontbreken er elf van de dertig toetsen per leerling. Van slechts 1500 leerlingen is elke LVS-toets geregistreerd, terwijl er 31 leerlingen zijn bij wie elke toetsscore ontbreekt. In figuur B.1 kunnen we zien dat veel leerlingen in onze dataset behoorlijk wat toetsscores missen. De meeste leerlingen bevinden zich in het lage tot middelste deel van de verdeling van ontbrekende toetsscores, waarbij de mediane leerling twaalf toetsen mist. In veel gevallen waarin de toets van een enkele leerling ontbreekt, ontbreken ook de toetsscores van de andere leerlingen in de klas. 91% van de scholen heeft ten minste één toetsmoment waarop voor elke leerling een toets ontbreekt. Als we de spellingtoets buiten beschouwing laten, daalt dit naar 85%.

Figuur B.1 Aantal gemiste toetsen per leerling

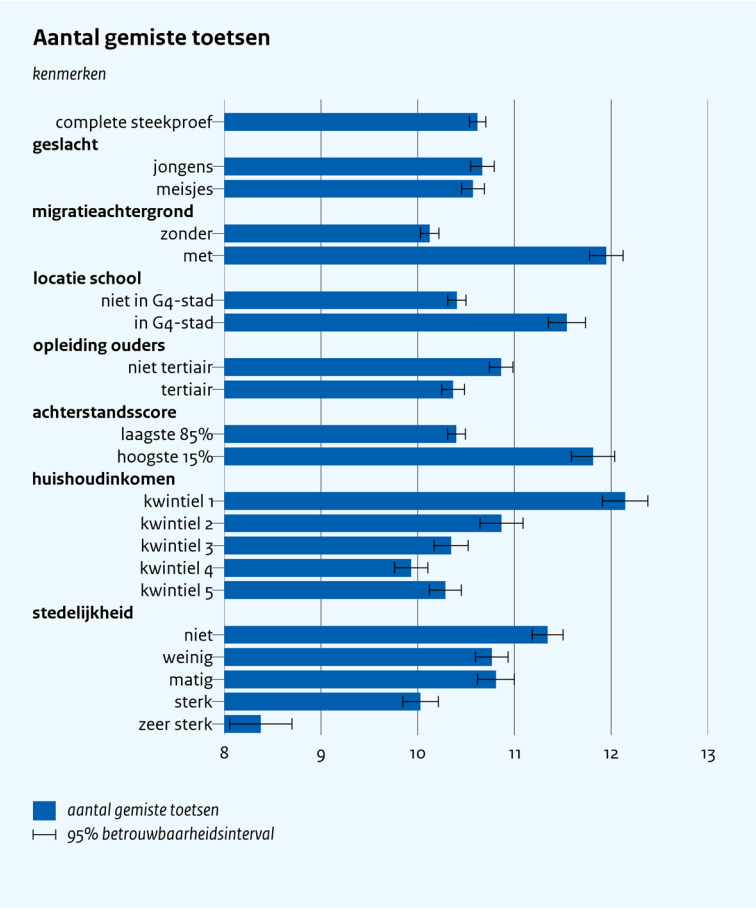


CPB-bewerking op basis van gegevens van CBS en NCO.

Toetsscores ontbreken in de dataset vaak om structurele of administratieve redenen. Als een leerling bijvoorbeeld wisselt van of naar een school die niet deelneemt aan het NCO, zien we alleen hun toetsscores vanaf het moment dat ze naar de basisschool gaan die wel deelneemt aan het NCO. Er zijn ook structurele redenen waarom sommige toetsen vaker ontbreken, die te maken hebben met de generatie van de afgenomen toets. Oudere generaties LVS-toetsen zijn niet opgenomen in de data, maar sommige zijn wel omgezet naar de nieuwste generatie, terwijl andere, namelijk de spellingtoets niet omgezet kunnen worden (Nationaal Regieorgaan Onderwijsonderzoek [NRO] & Centraal Bureau voor de Statistiek [CBS], 2022). Verder, als scholen van LVS-aanbieder wisselen van of naar Cito, kunnen we het deel van de toetsscores dat is geregistreerd door de niet-Cito-aanbieder niet waarnemen.

Bepaalde groepen leerlingen missen relatief veel toetsscores. Met name leerlingen met een migratieachtergrond, die in een G4-stad wonen of in een niet-stedelijk gebied, en leerlingen van wie de ouders in het laagste inkomenskwintiel vallen, missen significant meer toetsscores dan de gemiddelde leerling. Het aantal geobserveerde toetsscores draagt bij aan de kwaliteit van de voorspellingen, maar is niet de enige factor van belang.

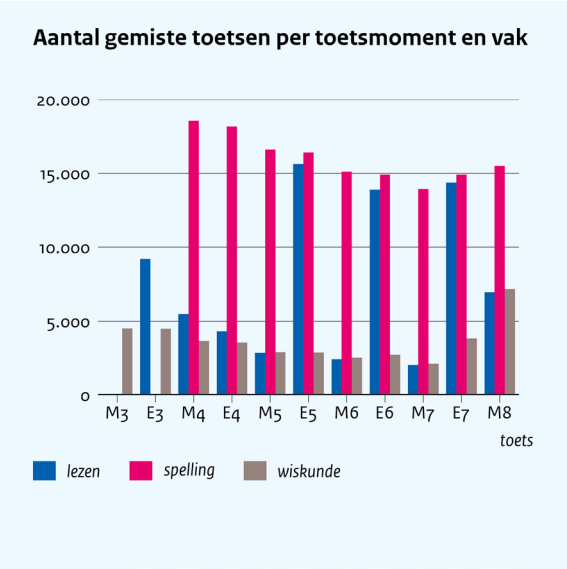
Figuur B.2 Gemiddeld aantal gemiste toetsscores, naar leerlingkenmerken



CPB-bewerking op basis van gegevens van CBS en NCO.

Er is veel variatie in het aantal gemiste toetsscores per vak. De meeste ontbrekende datapunten zijn gemiste spellingtoetsen, zie ook figuur B.3. Op elk meetmoment is de spellingtoets het slechtst gevuld. De spellingtoets is amper gevuld, omdat veel scholen ten tijde van de registratie van deze gegevens nog gebruikmaakten van de oudere generatie van de spelling LVS-toetsen. Deze oudere generatie toetsen kunnen niet in de data worden meegenomen (NRO & CBS, 2022). De oudere reken- en leestoetsen zijn wel herberekend naar de nieuwe generatie, maar de spellingtoets niet. Daarom is alleen de laatste generatie van de spellingtoets opgenomen in de data. Dit administratieprobleem zal niet blijven bestaan voor nieuwere LVS-cohorten, aangezien de toets van generatie 2 niet meer wordt gebruikt en geleidelijk uit deze dataset verdwijnt. In totaal ontbreekt 65% van de spellingtoetsen in de gegevens, 31% van de leestoetsen en slechts 14% van de rekentoetsen. We zien dat de leestoetsen vaker ontbreken bij de E-toets, maar zeer goed gevuld zijn bij de M-toets. Het aantal ontbrekende rekentoetsen is consistent laag op elk toetsmoment.

Figuur B.3 Aantal gemiste toetsscores, per toets en vak



CPB-bewerking op basis van gegevens van CBS en NCO.

Voor het eerste cohort ontbreken meer datapunten dan voor latere cohorten in de LVS-gegevens. Deze studie gebruikt gegevens van het eerste LVS-cohort dat volledig is gevolgd, namelijk het cohort dat in 2018 aan het laatste jaar van de basisschool begint. In dit cohort missen leerlingen gemiddeld 17,3 toetsscores. Dit daalt naar 13,3 voor het cohort van 2019 en daalt verder naar 12,8 in 2020, maar stijgt naar 14 in 2021.[5] Het significante verschil in gemiddelde gemiste toetsscores tussen het cohort dat in deze studie is onderzocht en latere cohorten komt waarschijnlijk doordat dit het eerste cohort is waarvan gegevens van alle groepen beschikbaar zijn en doordat de oude generatie toetsen meer wordt gebruikt. Studies die gegevens van latere cohorten gebruiken, zullen minder problemen hebben met ontbrekende gegevens, maar zullen wel te maken krijgen met ontbrekende toeten als gevolg van de coronacrisis.

Tabel B.1 Gemiddeld aantal gemiste toetsscores per leerling in elk LVS-cohort

jaar	N	gemiddeld	SD
2018/2019	49.817	17,35	6,93
2019/2020	56.252	13,31	8,29
2020/2021	56.546	12,81	8,14
2021/2022	55.975	14,00	7,20

Noot: N en SD staan respectievelijk voor het aantal leerlingen per eindtoets-cohort in de NCO steekproef en voor de standaarddeviatie.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

[5] Om het gemiddelde aantal gemiste toetsscores per cohort te berekenen, gebruiken we de volledige LVS-steekproef en beperken we deze niet tot alleen degenen die de Centrale Eindtoets hebben gemaakt. Dit komt doordat er in 2020 geen eindtoets is afgenomen, en we de gegevens over de eindtoets in 2021 nog niet hebben. Deze steekproefbeperking is voor deze cohorten dus niet mogelijk.

Bijlage C Methode

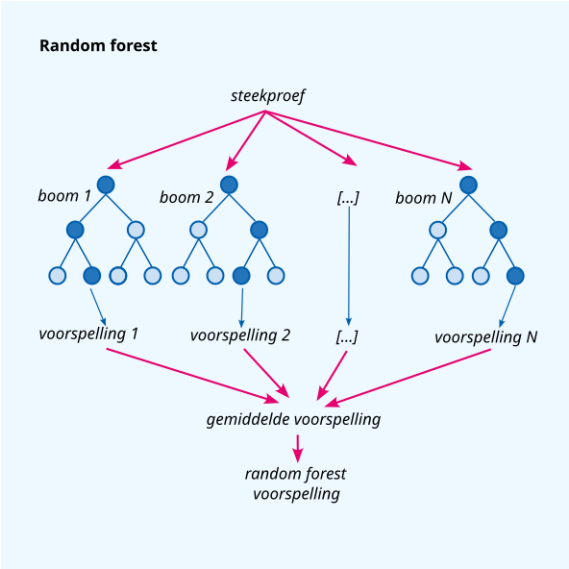
In dit onderzoek gebruiken we een *machine learning* model om de eindtoetsscores te voorspellen op basis van LVS-toetsscores en leerlingkenmerken. *Machine learning*-algoritmen kunnen ingewikkelde patronen en interacties in gegevens ontdekken, waardoor ze nauwkeurige voorspellingen kunnen doen voor nieuwe gevallen (Fitzek et al., 2020). In tegenstelling tot conventionele voorspelmethoden zoals de kleinste kwadratenmethoden of logistische regressie, wordt er geen expliciete relatie tussen kenmerken en te voorspellen uitkomst opgelegd. Hierdoor kunnen *machine learning*-methoden, mits voldoende data voorhanden zijn, een complexe structuur ontdekken die uniek is voor de gegevens en de taak in kwestie (Mullainathan & Spiess, 2017). Een ander voordeel is dat veel *machine learning*-methoden kunnen omgaan met ontbrekende datapunten en bovendien de informatie dat een datapunt ontbreekt kunnen gebruiken om de voorspelling te verbeteren (Twala et al., 2008).

We gebruiken een *random forest* om eindtoetsscore te voorspellen. Zoals figuur C.1 laat zien, bestaan *random forests* uit vele regressiebomen. Deze bomen verdelen de data herhaaldelijk in twee groepen. Aan het begin zitten alle datapunten boven in de regressieboom. Deze datapunten worden in twee groepen gesplitst op basis van of de waarde van een variabele groter of kleiner is dan een bepaalde grenswaarde.[6] Stel dat het huishoudinkomen gekozen wordt met als grenswaarde de waarde van het derde kwintiel. Dan worden alle leerlingen uit huishoudens met een inkomen beneden of binnen het derde kwintiel in de linkertak geplaatst, en alle leerlingen uit huishoudens met een inkomen in het vierde of vijfde kwintiel in de rechtertak. Eén of beide takken worden dan gesplitst in nog twee takken en het proces wordt herhaald tot een door de onderzoeker gekozen stopregel van toepassing is. Ieder datapunt zit nu in een groep met enkele andere datapunten, de zogeheten bladeren van de regressieboom. Voor ieder punt is de voorspelling gelijk aan de gemiddelde waarde van de eindtoetsscore in het blad. *Random forests* nemen het gemiddelde van de voorspellingen over vele (in ons geval 2000) van deze regressiebomen. Hier verschilt elke regressieboom in de variabelen en splitsingspunten, omdat elke boom is gegroeid op een aparte set 'gebootstrapte' gegevens binnen de trainingssteekproef.[7] Om deze reden heet het algoritme een *random forest*. *Random* verwijst naar het feit dat het algoritme, als er veel verklarende variabelen zijn, in ieder punt een willekeurige subset van variabelen gebruikt om uitsplitsingen te maken. Het combineren van vele zwak gecorreleerde regressiebomen verbetert de voorspelkwaliteit op nieuwe data (Hastie et al., 2007).

[6] De geselecteerde variabele om splitsingen van te maken en het splitsingspunt worden gekozen met behulp van een *greedy*-algoritme om de beste fit te bereiken.

[7] Strikt genomen gebruiken we het *regression forest*-algoritme uit het GRF-pakket voor R, een specifiek een consistente *random forest* (Athey et al., 2019). Sommige instellingen van het model (hyperparameters) kunnen door de onderzoeker worden gekozen om de prestaties van het model en de rekensnelheid te optimaliseren.
Het algoritme dat we hebben gebruikt, implementeert automatisch kruisvalidatie om de optimale hyperparameters voor het model te selecteren. Aangezien we 36 modellen hebben gebruikt, hebben ze elk iets andere hyperparameters. De minimale "*node size*", die bepaalt hoe diep elke boom is, of hoeveel keer de boom aftakt, ligt voor de meeste modellen rond de vijf of zes. De "*honesty*" van de bomen heeft betrekking op het aantal vertakkingen. De *honesty* van de bomen verwijst naar het feit of de boom wel of niet is opgebouwd en geëvalueerd op verschillende subsets van gegevens. Voor de meeste modellen is *honesty* ingeschakeld. Tot slot groeit het aantal variabelen, dat willekeurig wordt gekozen als kandidaten voor een splitsing bij elke tak, naarmate er meer tests worden toegevoegd aan de invoer van het model. Het varieert van vier, wanneer alleen de achtergrondvariabelen als invoer worden gebruikt, tot negentien wanneer alle variabelen worden meegenomen.

Figuur C1 Illustratie *random forest*-methode

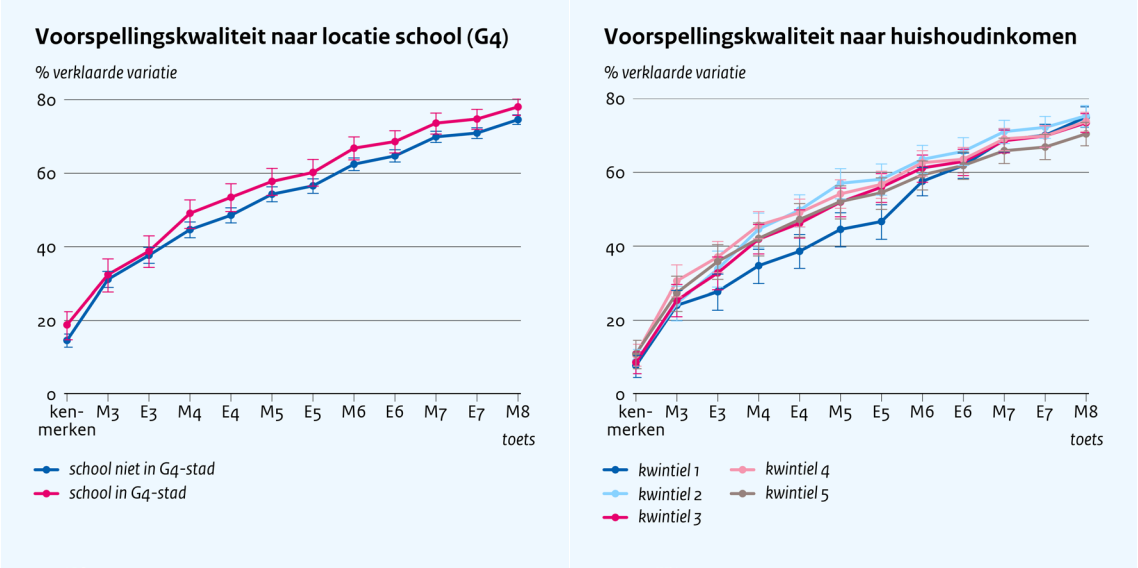


We hebben onze steekproef van 24.646 leerlingen willekeurig verdeeld in een train-dataset met 19.717 leerlingen (80%) en een test-dataset met 4.929 leerlingen (20%). De *random forests* worden gebouwd met behulp van de train-dataset, waar ze de patronen van de inputs leren en hoe deze zich verhouden tot de eindtoetsscores. Het model doet vervolgens voorspellingen op de ongeziene test-dataset, die alleen inputs bevat maar geen eindtoetsscores. Nadat de voorspellingen zijn gedaan, voegen we de eindtoetsscores toe en evalueren we de voorspellingen die zijn gedaan op de testdataset. Als we een traditionele methode hadden gebruikt om de voorspellingen te doen, zouden we het vooraf gespecificeerde model alleen kunnen toepassen op waarnemingen met volledige informatie, zonder ontbrekende datapunten. Dit zou betekenen dat we slechts 1509 leerlingen in onze data zouden hebben, van wie er 1207 in de train-dataset terecht zouden komen en slechts 302 in de test-dataset beschikbaar zouden zijn om voorspellingen te doen en te evalueren.

Hoewel iedere stap in de individuele regressiebomen verklaarbaar is, zorgt het grote aantal bomen ervoor dat de methode in de praktijk een 'black box' is. We observeren dus wel de inputs en output van het model, maar niet de interne werking. Dit betekent dat we hier geen verband afleiden tussen de inputs en de uitkomst. We zijn echter wel in staat om het belang van elke input-variabele in het model te beoordelen, wat we hebben gedaan in figuur 3. Het belang van een variabele is gebaseerd op het aantal keer dat de variabele gekozen is om uitsplitsingen te maken, waarbij uitsplitsingen aan het begin meer gewicht krijgen dan splitsingen aan het eind (Tibshirani et al., 2022).

Bijlage D Extra figuren

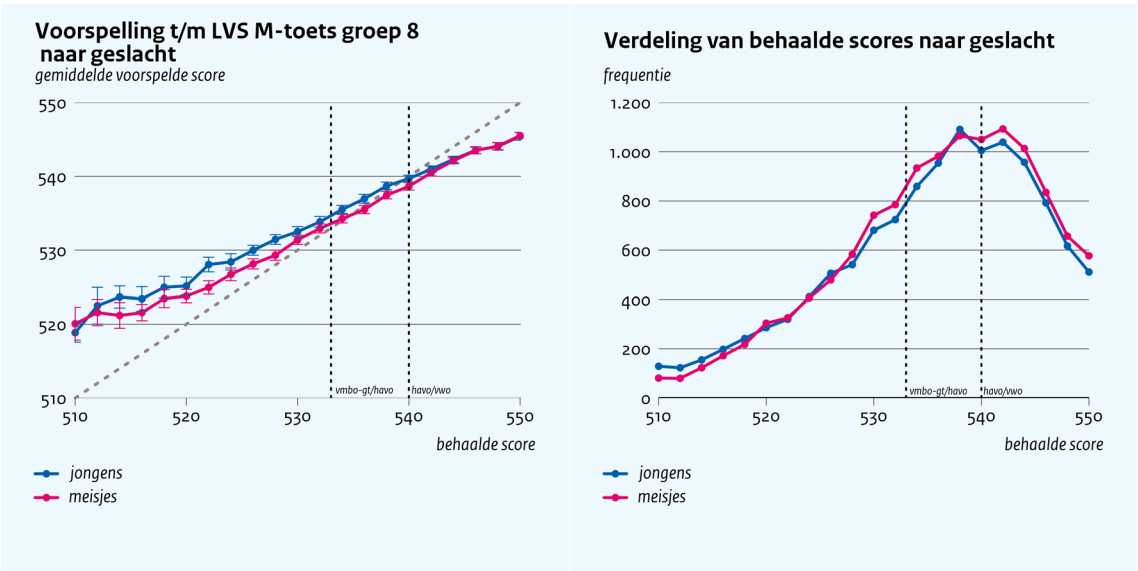
Figuur D.1 Heterogeniteit in voorspelling



Noot: Deze cijfers tonen voor verschillende groepen hoeveel van de variatie in eindtoetsscores kan worden verklaard door voorspellingen op basis van achtergrondkenmerken en alle beschikbare LVS-toetsscores op verschillende momenten van de basisschool.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.

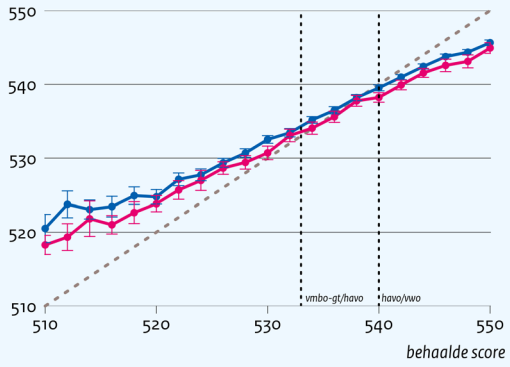
Verschillen in voorspelkwaliteit zijn niet altijd stabiel over de spreiding van toetsscores. De volgende figuren zoomen in op de voorspellingen van het beste model, waarbij alle LVS-toetsscores zijn gebruikt. Links worden de gemiddelde voorspellingen per eindtoetsscore per groep weergegeven en rechts de verdeling van de eindtoetsscores voor deze groepen. Deze figuren geven inzicht in de complexe relaties tussen de verdeling van scores en de nauwkeurigheid van voorspellingen.

Figuur D.2 Gemiddelde van voorspellingen voor elke eindtoetsscore en verdeling van eindtoetsscores per groep



Voorspelling t/m LVS M-toets groep 8 naar migratiestatus

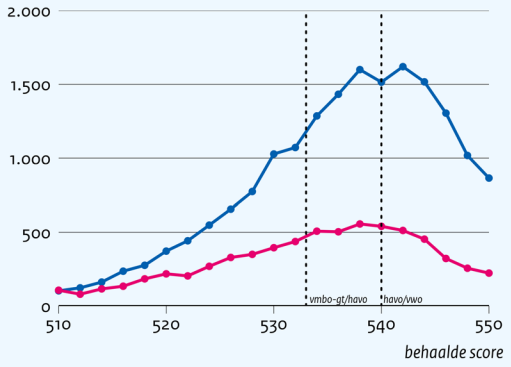
gemiddelde voorspelde score



- *geen migratieachtergrond*
- *migratieachtergrond*

Verdeling van behaalde scores naar migratiestatus

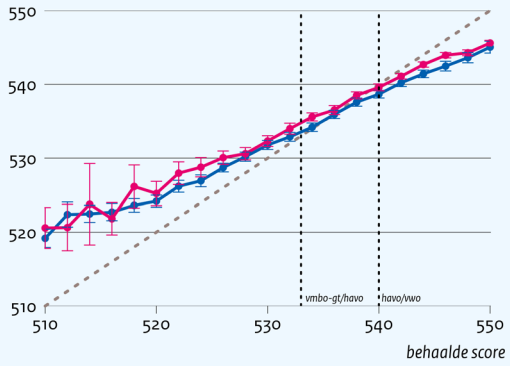
frequentie



- *geen migratieachtergrond*
- *migratieachtergrond*

Voorspelling t/m LVS M-toets groep 8 naar opleiding ouders

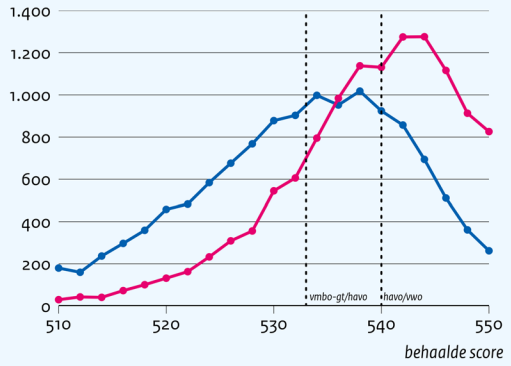
gemiddelde voorspelde score



- *niet tertiair opgeleid (of opleiding onbekend)*
- *tertiair opgeleid*

Verdeling van behaalde scores naar opleiding ouders

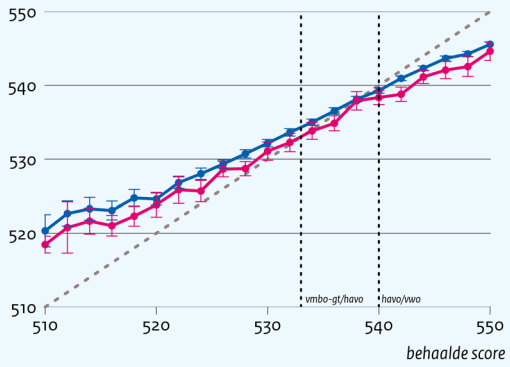
frequentie



- *niet tertiair opgeleid (of opleiding onbekend)*
- *tertiair opgeleid*

Voorspelling t/m LVS M-toets groep 8 naar achterstandsschool

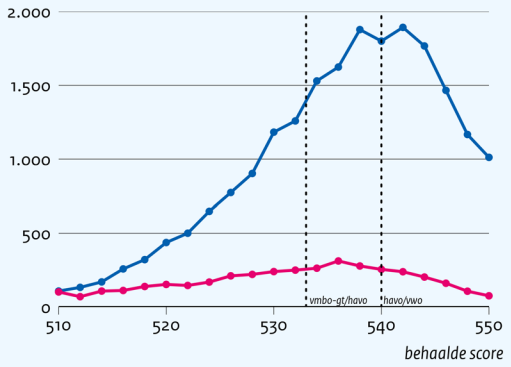
gemiddelde voorspelde score



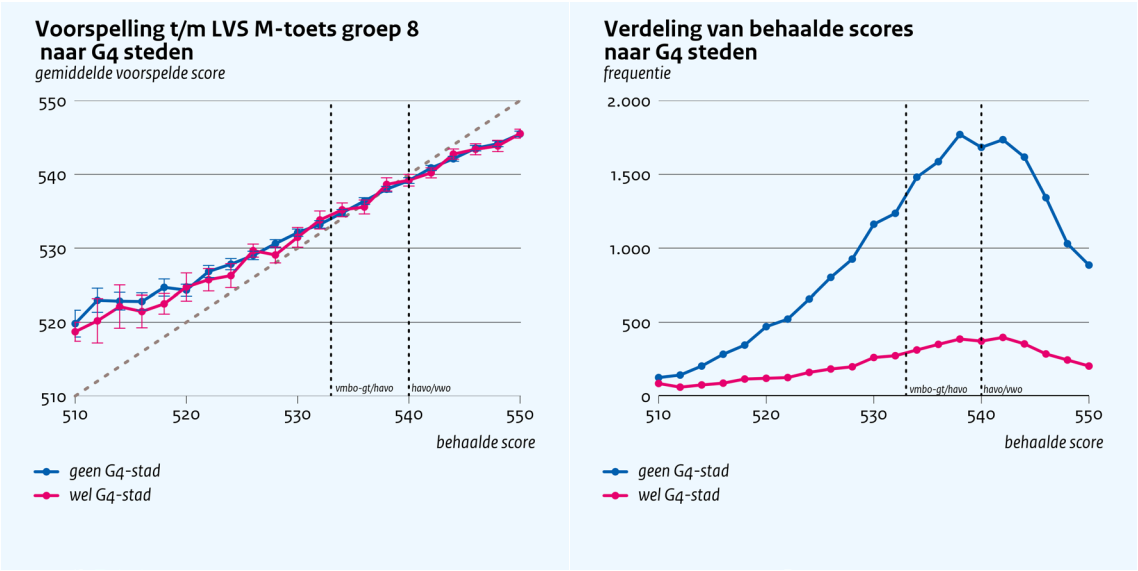
- *laagste 85% achterstandsscore*
- *hoogste 15% achterstandsscore*

Verdeling van behaalde scores naar achterstandsschool

frequentie



- *laagste 85% achterstandsscore*
- *hoogste 15% achterstandsscore*



Noot: De linker panelen tonen de gemiddelde voorspellingen van de eindtoets per groep gebaseerd op alle LVS-toetsscores. De rechter panelen tonen hoeveel leerlingen per groep in de volledige steekproef (test en training) een specifieke eindtoetsscore hebben gekregen. Om kleine aantallen leerling per groep te voorkomen, omvat elke groep twee toetsscores en zijn alle toetsscores onder de 510 samengevoegd in de eerste groep. Deze aanpassingen leiden tot een relatief groter aantal observaties in de eerste groep (510 omvat nu alle observaties van leerlingen die tussen 501-510 punten scoren). Ook verdwijnt aan de rechterkant van de scoreverdeling de karakteristieke piek bij de laatste score (topcodering door de toets) omdat deze is opgenomen in een grotere groep.
Bron: CPB-bewerking op basis van gegevens van CBS en NCO.