

CODEBOEK DATA LVS-NCO-data

Mei 2022

Versie 1.2

Inhoud

Inleiding	3
1. Dataverzameling en - bewerking.....	5
2. Variabelen toelichting	10

Inleiding

Het Nationaal Regieorgaan Onderwijsonderzoek (NRO) heeft in een samenwerkingsverband met het Centraal Bureau voor de Statistiek (CBS) het Nederlands Cohortonderzoek Onderwijs (NCO) databestand mogelijk gemaakt. Dit databestand is mogelijk gemaakt door het slim combineren van bestaande registerdata uit het Stelsel van Sociaal-statistische Bestanden (SSB) van het CBS en bestaande onderwijsregistraties uit het Basisregister Onderwijs (BRON) van Dienst Uitvoerig Onderwijs (DUO).

Dit codeboek beschrijft data afkomstig uit de leerlingvolgsystemen (LVS) van basisschoolleerlingen. Dit databestand wordt beschikbaar gesteld vanaf mei 2021 waar toegang onder strikte voorwaarden voor kan worden verkregen.

Om onderzoek te doen en gebruik te maken van deze versie van het LVS-databestand, dient u gemachtigd te zijn om onderzoek te doen op de microdata van het CBS en al toegang te hebben verkregen tot de NCO-tool. Daarnaast dient u een extra onderzoeksvoorstel in te dienen bij het NRO en CBS voordat u in aanmerking kunt komen om met deze specifieke dataset aan de slag te gaan. Hierdoor wordt nagegaan of het onderzoeksvoorstel met de LVS-data te beantwoorden is en of het doel van uw onderzoek in lijn is met de doelen die door het NRO zijn gesteld.

Voor meer informatie kunt u de website van het NCO bezoeken of contact opnemen met info@nationaalcohortonderzoek.nl.

Aanvragen van de LVS-data

Om te werken met het NCO-databestand dient u eerst op de website van het NCO een aanvraag in te dienen om toegang te verkrijgen. Deze kunt u [hier](#) vinden. Als deze is goedgekeurd door het NRO wordt de toestemming bij het CBS aangevraagd om te werken met de microdata.

Let op: Het is niet mogelijk om deze dataset te koppelen aan bijzondere persoonsgegevens. Houd hier dus rekening met het aanvragen van de LVS-data.

Algemene Verordening Gegevensbescherming

Bij de ontwikkeling van het NCO staat de vertrouwelijkheid en anonimiteit van de gegevens van leerlingen voorop. Zo is voor de inrichting van het NCO juridisch advies over de verwerking van leerling gegevens ingewonnen bij bureau PrivacyCare. Daarvoor is bij aanvang rekening gehouden met de eisen van de Algemene Verordening Gegevensbescherming (AVG). Het belangrijkste punt voor het NCO is het uitsluiten dat persoonsgegevens te herleiden zijn tot individuen en scholen. Om deze zekerheid te garanderen, werkt het NCO samen met het CBS. Daarnaast, worden **geen bijzondere persoonsgegevens** van leerlingen opgevraagd. Bijzondere persoonsgegevens zijn bijvoorbeeld gegevens over godsdienst, seksuele voorkeur, lidmaatschap van een vakvereniging, gegevens omtrent de gezondheid, en strafrechtelijke of justitiële gegevens. Mocht voor aanvullend onderzoek in de toekomst bijzondere persoonsgegevens wenselijk zijn, dan moet daarvoor naar de mogelijkheden gekeken worden en moeten de betrokkenen geïnformeerd worden. Een schoolbestuur geeft alleen toestemming aan het CBS om niet-bijzondere gegevens uit het leerlingvolgsysteem te gebruiken en te koppelen aan het Stelsel van Sociaal-statistische Bestanden. Vervolgens haalt het CBS er in deze stap alles uit wat nog eventueel zou kunnen leiden tot herkenbaarheid, zoals de combinatie BRIN- en onderwijsnummer. Het CBS is gehouden aan de [CBS-wet](#) waarin geregeld is hoe het gegevens mag verwerken voor welke doelen. Geleverde gegevens worden door het CBS geanonimiseerd en zeer streng beveiligd bewaard. Meer informatie over hoe het CBS de privacyrichtlijnen naleeft, vindt u [hier](#). Wilt u het gehele rapport van PrivacyCare inzien, klik dan [hier](#).

1. Dataverzameling en bewerking

1.1 Dataverzameling

In oktober 2020 zijn alle schoolbesturen in het primair onderwijs (PO) benaderd met het verzoek om gegevens uit het LeerlingVolgSysteem (LVS) beschikbaar te stellen aan het NCO. Het gaat om toetsgegevens van de zogeheten Midden- en Eind-toetsen (M- en E-toetsen) van groep 3 t/m 8 voor de domeinen begrijpend lezen, spelling en rekenen-wiskunde. Het Doel is om zicht te krijgen op de effecten van de eerste schoolsluiting naar aanleiding van de COVID-19-crisis op de leergroei in het schooljaar 2019/2020. Gelet op de beperkte historische data van andere toetsaanbieders dan Cito, is besloten om voor deze dataverzameling alleen gegevens van Cito-toetsen mee te nemen. Daarbij zijn – voor zover mogelijk – de scores van oudere generaties van een toets omgezet naar de meest recente generatie.

De scholen die meedoen, krijgen tweejaarlijks een rapport over de leergroei van hun leerlingen, vooral met betrekking op de situatie omtrent de COVID-19-crisis. Schoolbesturen hadden in eerste instantie tot eind oktober de tijd om zich aan te melden en vervolgens vier weken om bezwaren van ouders te registreren. Die periode is later verlengd tot aanmelden tot eind november en bezwaren registreren tot midden januari.

De procedure was als volgt. De besturen hebben eerst een contract getekend met NRO. De besturen zijn immers juridisch verantwoordelijk en ‘eigenaar’ van de data. Omdat het om niet-bijzondere persoonsgegevens gaat is het niet nodig dat ouders hiervoor toestemming geven. Wel hebben de deelnemende scholen alle ouders geïnformeerd over het voornemen om de toetsgegevens te leveren aan CBS met het doel te koppelen aan NCO¹. Hiervoor zijn voorbeeldbrieven beschikbaar gesteld. Ouders zijn in de gelegenheid gesteld om daar bezwaar tegen te maken gedurende een bepaalde periode. De scholen registreren deze bezwaren in het LeerlingAdministratieSysteem (LAS). De softwareleveranciers van de LAS-en in het PO (Cito LOVS, ParnasSys en ESIS) hebben hiervoor een aparte ‘bezwaarknop’ ingebouwd. Voordat de feitelijke upload naar het CBS plaatsvindt, moeten de scholen op een ‘verzendknop’ duwen, om aan te geven dat men alle procedures doorlopen heeft en de softwareleveranciers tot levering over kunnen gaan. De softwareleveranciers leveren vervolgens de gegevens via een beveiligd kanaal aan het CBS. Het CBS ‘verrint’ deze gegevens (proces van pseudonimisering) en stelt de data beschikbaar aan het NCO-team die de data verder bewerkt.

¹ Voor meer informatie over de NCO-dataset, zie: Haelermans, C., Huijgen, T., Jacobs, M., Levels, M., van der Velden, R., van Vugt, L., van Wetten, S., (2020). Using Data to Advance Educational Research, Policy, and Practice: Design, Content, and Research Potential of the Netherlands Cohort Study on Education. *European Sociological Review* 36(4), p. 643–662. <https://doi.org/10.1093/esr/jcaa027>

1.2 Data-aanlevering

De data is verzameld via vier leveringen, de exporten van de data vonden plaats op: 30 november 2020, 18 januari 2021, 1 april 2021 en 1 augustus 2021. De gegevens van de leerlingen waarvoor de ouder(s) en/of het bevoegd gezag geen bezwaar hebben gemaakt voor het leveren van de data, zijn via de softwareleveranciers Cito, ParnasSys en ESIS verstrekt aan het Centraal Bureau voor de Statistiek (CBS). CBS pseudonimiseert vervolgens de identificerende informatie van de leerlingen en van de school. Op deze manier kunnen onderzoekers (in dit geval de onderzoekers van het NCO) niet achterhalen om welke leerling en om welke school het gaat.

In de drie losse datasets van Cito, ParnasSys en ESIS zitten gegevens die de school in hun softwaresysteem over de leerling heeft geregistreerd zoals:

- Brinnummer (CBS pseudonimiseert deze);
- Vestigingsnummer;
- Postcode school;
- LeerlingID (CBS pseudonimiseert deze);
- Inschrijfdatum op school;
- Eventuele uitschrijfdatum van school;
- Jaargroep;
- Klasnaam;
- Vaardigheidsscore;
- OSO-toetscode;
- Jaargroep van afname toets;
- Afnamedatum toets;
- Geslacht;
- Geboortemaand en geboortjaar.

De volgende groepen zijn opgevraagd:

- Schooljaar 2013/2014: leerlingen groep 3;
- Schooljaar 2014/2015: leerlingen groep 3 t/m 4;
- Schooljaar 2015/2016: leerlingen groep 3 t/m 5;
- Schooljaar 2016/2017: leerlingen groep 3 t/m 6;
- Schooljaar 2017/2018: leerlingen groep 3 t/m 7;
- Schooljaar 2018/2019: leerlingen groep 3 t/m 8 – *vanaf hier volledig cohort*;
- Schooljaar 2019/2020: leerlingen groep 3 t/m 8;
- Schooljaar 2020/2021: leerlingen groep 3 t/m 8.

In **Tabel 1** ziet u een overzicht van de ontvangen data, uitgesplitst naar aantal scholen, aantal unieke leerlingen en aantal toetsrecords per softwareleverancier. De enige selectie die hiervoor heeft plaatsgevonden is dat er enkel gegevens van afnames in groep 3 t/m 8 meegenomen worden. Verder is te zien dat data zijn ontvangen van zo'n 1.900 scholen.

Tabel 1 Inhoud ruwe data

	Aantal unieke scholen	Aantal unieke leerlingen	Aantal toetsrecords
Cito LOVS	252	56.420	1.232.271
ParnasSys	1.409	366.203	10.048.088
ESIS	275	81.216	1.642.787
Totaal ⁱ	1.936	503.839	12.923.146

Noot: Data levering december 2020 + januari 2021 + april 2021 + augustus 2021

i. Leerlingen en scholen kunnen in de verschillende data van de softwareleveranciers voorkomen bijvoorbeeld doordat een leerling van school wisselt die een andere softwareleverancier heeft of omdat de school is overstapt van softwareleverancier. Het totaal is daarom geen exacte optelling van de aantallen van Cito LOVS, ParnasSys en ESIS.

1.2 Opschoning van ruwe data

Het NCO heeft de data opgeschoond zodat er zo weinig mogelijk gegevens met administratieve fouten en dubbele leerlingen (bijvoorbeeld door een schoolwissel) in de data voorkomen. Hieronder wordt uitgelegd welke stappen en beslissingen er zijn genomen om tot een opgeschoonde dataset te komen.

Stappen en selecties:

1. Niet alle leerlingen zijn door CBS gekoppeld. Enkel leerlingen die in de Gemeentelijke Basis Administratie (GBA) staan ingeschreven zijn gekoppeld en dus behouden in het bestand.
2. Onjuiste school. Een deel van de dubbele toetsrecords lijkt te zijn ontstaan door schoolwisselaars, waarbij de nieuwe school ook de gegevens van de oude school heeft overgenomen. Deze leerlingen hebben op twee verschillende scholen exact dezelfde toetsen gemaakt met dezelfde toetsresultaten. Om hiervoor te corrigeren wordt gekeken naar de in- en uitschrijfdatum van de leerling. Wanneer de toets is afgenomen in de periode tussen de

inschrijfdatum en eventuele uitschrijfdatum dan is het aannemelijk dat de toets op die desbetreffende school is afgenomen. Op deze manier zijn enkel de toetsrecords overgebleven die op de school ten tijde van de afnamedatum hebben plaatsgevonden.

3. Toetsrecords waar de vaardigheidsscore, OSO-toetscode en afnamedatum van ontbreekt worden verwijderd.
4. Identieke dubbele leerlingen zijn verwijderd. Hierbij is gekeken naar identieke waardes op: brinnummer, vestigingsnummer, geslacht, leerlingID, inschrijfdatum, uitschrijfdatum, jaargroep, klasnaam, vaardigheidsscore, OSO-toetscode, afnamedatum, bronbestand, postcode school en geboortedatum.
5. Van toetsrecords waarbij enkel de jaargroep niet identiek is, maar de overige variabelen wel, wordt de hoogste jaargroep behouden.
6. Onder één rinpersoon kunnen meerdere leerlingID's bestaan en onder één leerlingID blijken soms verschillende personen te zijn gekoppeld. Er is bekeken in hoeverre dit kwam door een eventuele schoolwisseling of door een incorrecte combinatie van rinpersoon en leerlingID. Indien dat laatste het geval bleek hebben wij via de NCO-data gekeken of achterhaald kon worden welke combinatie de juiste was (op basis van geboortedatum). De toetsrecords waar rinpersoon en leerlingID met de geboortedatum overeenkwamen zijn behouden in de data. De overige toetsrecords binnen de dubbelingen zijn verwijderd.
7. Toetsrecords van toetsen met een andere toetsaanbieder dan Cito zijn verwijderd.
8. Indien een behaalde vaardigheidsscore niet binnen de juiste minimale en maximale score valt die hoort bij de betreffende OSO-toetscode, dan wordt deze op missing gezet.
9. Enkel toetsrecords waarvan de jaargroep afname bekend is zijn behouden.
10. Enkel toetsrecords van Midden-toetsen (M-toetsen) en E-toetsen zijn behouden.
11. Enkel reguliere toetsen zijn behouden, S(B)O toetsen zijn verwijderd. Digi-toetsen blijven wel in de data.
12. Begrijpend lezen en rekenen-wiskunde generatie 2 toetsen worden omgezet naar generatie 3 toetsen door middel van een formule die is geleverd door Cito.
13. Spelling generatie 2 vaardigheidsscores worden verwijderd aangezien deze niet omgezet kunnen worden naar generatie 3 vaardigheidsscores.
14. Enkel de opgevraagde cohorten zoals genoemd op pagina 7 worden behouden.
15. De verwijzing van een M-toets en E-toets is gebaseerd op de afnamemaand. Toetsen die tussen september en maart zijn afgenomen noemen we M-toets. Toetsen die tussen april en augustus zijn afgenomen noemen we E-toetsen. Bij minder dan 2% van de toetsen blijkt dat de toets niet is afgenomen in het juiste tijdsframe.

16. Tot slot blijken sommige leerlingen meerdere keren in een schooljaar getoetst. In eerste instantie wordt de toets meegenomen die op het juiste afnamemoment is afgenomen. Dat wil zeggen: een M-toets in de maanden september t/m maart of een E-toets in de maanden april t/m augustus. Indien dit geen uitsluitsel geeft over welke toets de juiste is wordt de laatst afgenomen toets behouden zodat de data uiteindelijk van iedere leerling per schooljaar maximaal één M-toets en één E-toets bevat.

Na deze stappen zijn de toetsrecords gekoppeld aan de desbetreffende leerling waardoor de data op iedere rij één leerling bevat met daarbij de desbetreffende toetsresultaten over de hele basisschoolloopbaan.

De variabele die (per toets) in het bestand blijven staan zijn:

- Brinnummer + vestigingsnummer school;
- Jaargroep ten tijde van de toets;
- Afnamedatum toets;
- Vaardigheidsscore;
- Soort toets (bijvoorbeeld M7 Digi-toets).

Alle variabelen zijn gemaakt voor ieder schooljaar 2013/2014 t/m 2020/2021 en voor ieder domein: begrijpend lezen, spelling (niet werkwoorden), rekenen-wiskunde.

Er zijn nog wel een paar dingen waar rekening mee gehouden moet worden:

- De generatie 3 toets voor spelling bestaat sinds schooljaar 2014/2015 dus voor schooljaar 2013/2014 is geen informatie over spelling bekend.
- De M-toets in begrijpend lezen wordt niet afgenomen in jaargroep 3. Vandaar dat deze gegevens dus ook niet aanwezig zijn.

Tabel 2. Inhoud LVS-NCO-data

	Aantal leerlingen
Totaal	438.355

Noot: Data levering dec. 2020 + jan. 2021+ april 2021 + augustus 2021

2. Variabelen toelichting

2.1 Variabelenlijst

RINPERSOONS	Soort rinpersoon;
RINPERSOON	Samen met “RINSPERSOONS” is dit de persoon;
JRGRPAFNAME_	Groep ten tijde van afname toets;
AFNAMEDATUM_	Datum ten tijde van afname toets;
TOETSSOORT_	Extra gegevens over de toets;
VASC_	Vaardigheidsscore.

2.1 Toelichting variabelen

RINPERSOONS - Soort rinpersoon

Format: string

R Rinpersoon wel in GBA

Samen met het rinpersoons-nummer identificeert deze code een natuurlijk en uniek geanonimiseerd persoon. De codes van deze variabele geven aan uit welke bron het rinpersoons-nummer is afgeleid. In de meeste gevallen is de bron van de rinpersoons-nummer de basisregistratie personen (BRP), categorie ‘*R = Rinpersoon wel in gemeentelijke basisadministratie*’. Voor personen die geen geldig administratienummer hebben, is de koppeling helaas niet gelukt en personen met een andere waarde ‘R’ hebben wij daarom moeten verwijderen uit de analyse.

Afkomstig uit: [GBAPERSONSTAB](#) en [NIETGBAPERSON](#) van het Stelsel van Sociaal-statistische bestanden van het CBS.

RINPERSOON - Samen met RINPERSOONS identificeert dit nummer met de unieke persoon

Format: string

Dit nummer identificeert een natuurlijk persoon. Het is een betekenis- en dimensieloos 9-cijferig nummer. RIN staat voor Record Identification Number en is de versleutelde aanpassing van het A-nummer uit de BRP. Omdat het A-nummer zeer identificerend is, wordt het vervangen door RINPERSOON.

Afkomstig uit: [GBAPERSONSTAB](#) en [NIETGBAPERSON](#) van het Stelsel van Sociaal-statistische bestanden van het CBS.

AFNAMEDATUM - Afnamedatum toets

Format: string

Deze variabele geeft aan op welke datum de toets is afgenomen. Het datumformat is JJJJMMDD.

Afhankelijk van het domein van de toets en de soort toets die afgenomen is wordt deze variabelennaam aangevuld met het domein (BG = begrijpend lezen; S = spelling; RW = rekenen-wiskunde) en de soort toets (M = Midden-toets; E = Eind-toets). Er wordt een uitzondering gemaakt voor de E-toetsen die in augustus, september of oktober 2020 zijn afgenomen, deze vindt u terug via de toevoeging van '_aug_sep_okt' aan de variabelennaam.

Bijbehorende variabelen zijn:

AFNAMEDATUM_BG_M	AFNAMEDATUM_S_M	AFNAMEDATUM_RW_M
AFNAMEDATUM_BG_E	AFNAMEDATUM_S_E	AFNAMEDATUM_RW_E

Afkomstig uit: Leerlingvolgsystemen via Cito LOVS, ParnasSys en ESIS.

JRGRPAFNAME - Groep ten tijde van afname toets

Format: numeriek

Deze variabele geeft aan in welke groep de leerlingen zaten ten tijde van de afgenomen toets. Let op: we hebben gemerkt dat kan afwijken van de variabele WPOLEERJAAR die u terug kunt vinden in de NCO-tool.

Afhankelijk van het domein van de toets, het schooljaar en de soort toets die afgenomen is wordt deze variabelennaam aangevuld met het domein (BG = begrijpend lezen; S = spelling; RW = rekenen-wiskunde), het schooljaar (2013 t/m 2019) en de soort toets (M = Midden-toets; E = Eind-toets). Er wordt een uitzondering gemaakt voor de E-toetsen die in augustus, september of oktober 2020 zijn afgenomen, deze vindt u terug via de toevoeging van '_aug_sep_okt' aan de variabelennaam.

Bijbehorende variabelen zijn:

JRGRPAFNAME_BG_M	JRGRPAFNAME_S_M	JRGRPAFNAME_RW_M
JRGRPAFNAME_BG_E	JRGRPAFNAME_S_E	JRGRPAFNAME_RW_E

Afkomstig uit: Leerlingvolgsystemen via Cito LOVS, ParnasSys en ESIS.

AFNAMEDATUM - Datum ten tijde van afname toets

Format: string

Deze variabele geeft aan op welke datum leerling de toets heeft gemaakt (JJJMMDD).

Afhankelijk van het domein van de toets, het schooljaar en de soort toets die afgenomen is wordt deze variabelennaam aangevuld met het domein (BG = begrijpend lezen; S = spelling; RW = rekenen-wiskunde), het schooljaar (2013 t/m 2019) en de soort toets (M = Midden-toets; E = Eind-toets). Er

wordt een uitzondering gemaakt voor de E-toetsen die in augustus, september of oktober 2020 zijn afgenomen, deze vindt u terug via de toevoeging van ‘_aug_sep_okt’ aan de variabelennaam.

Bijbehorende variabelen zijn:

AFNAMEDATUM_BG_M	AFNAMEDATUM_S_M	AFNAMEDATUM_RW_M
AFNAMEDATUM_BG_E	AFNAMEDATUM_S_E	AFNAMEDATUM_RWEE

Afkomstig uit: Leerlingvolgsystemen via Cito LOVS, ParnasSys en ESIS.

TOETSSOORT - Extra gegevens over de toets

Format: string

Deze variabele geeft aan of de benoeming van de M-toets of E-toets overeenkomt met de oorspronkelijk toets. Zoals vermeld eerder in dit document blijkt dat bij minder dan 2% van de toetsen toets niet is afgenomen in het juiste tijdsframe (zie punt 15 onder **1.2 Opschoning van ruwe data**). Daarnaast vindt u in deze variabele ook of het om een starttoets (start / S) gaat en het niveau daarvan (+1 of +2), of om een digitale toets (digi).

Afhankelijk van het domein van de toets, het schooljaar en de soort toets die afgenomen is wordt deze variabelennaam aangevuld met het domein (BG = begrijpend lezen; S = spelling; RW = rekenen-wiskunde), het schooljaar (2013 t/m 2019) en de soort toets (M = Midden-toets; E = Eind-toets). Er wordt een uitzondering gemaakt voor de E-toetsen die in augustus, september of oktober 2020 zijn afgenomen, deze vindt u terug via de toevoeging van ‘_aug_sep_okt’ aan de variabelennaam.

Bijbehorende variabelen zijn:

TOETSSOORT_BG_M	TOETSSOORT_S_M	TOETSSOORT_RW_M
TOETSSOORT_BG_E	TOETSSOORT_S_E	TOETSSOORT_RW_E

Afkomstig uit: Leerlingvolgsystemen via Cito LOVS, ParnasSys en ESIS.

VASC - Vaardigheidsscore

Format: numeriek

Deze variabele geeft de vaardigheidsscore behorende bij de afgenomen toets aan.

Afhankelijk van het domein van de toets, het schooljaar en de soort toets die afgenomen is wordt deze variabelennaam aangevuld met het domein (BG = begrijpend lezen; S = spelling; RW = rekenen-wiskunde), het schooljaar (2013 t/m 2019) en de soort toets (M = Midden-toets; E = Eind-toets). Er wordt een uitzondering gemaakt voor de E-toetsen die in augustus, september of oktober 2020 zijn afgenomen, deze vindt u terug via de toevoeging van ‘_aug_sep_okt’ aan de variabelennaam.

Bijbehorende variabelen zijn:

VASC_BG_M	VASC_S_M	VASC_RW_M
VASC_BG_E	VASC_S_E	VASC_RW_E

Afkomstig uit: Leerlingvolgsystemen via Cito LOVS, ParnasSys en ESIS.