TIMSS 2019 | TRENDS IN INTERNATIONAL MATHEMATICS & SCIENCE STUDY

Universiteit Antwerpen

Research brief 5

Compendium

De samenhang tussen beleidsvoerend vermogen van scholen en prestaties wiskunde en wetenschappen van leerlingen in het 4de leerjaar

BELEIDSSAMENVATTING

Op vraag van het departement Onderwijs & Vorming van de Vlaamse Gemeenschap werd de samenhang geanalyseerd tussen het beleidsvoerend vermogen van scholen en de prestaties wiskunde en wetenschappen van leerlingen uit het vierde leerjaar. Hiervoor werden de resultaten van TIMSS 2019 (Trends in International Mathematics and Science Study) gecombineerd met een onderzoek in 77 scholen waarbij op basis van de Vragenlijst Beleidsvoerend Vermogen het beleidsvoerend vermogen in kaart werd gebracht. Voorafgaand aan het eigenlijke onderzoek werd de Vragenlijst Beleidsvoerend Vermogen gevalideerd. De samenhang tussen beleidsvoerend vermogen van scholen en de prestaties van leerlingen werd nagegaan door middel van multilevelanalyse. Er kon geen significante samenhang worden vastgesteld tussen het beleidsvoerend vermogen van scholen en de prestaties van leerlingen. Dit onderzoek legt grote verschillen bloot in het beleidsvoerend vermogen van scholen en in de percepties van schoolleiders en leerkrachten over het beleidsvoerend vermogen van hun school. Achteraan dit compendium worden adviezen geformuleerd.

EduBROn

Joris Van Elsen

Lies Appels

Marijn Gijsen

Sven De Maeyer

Peter Van Petegem

Februari 2022

INHOUDSOPGAVE

BELEIDSSAMENVATTING 1

INHOUDSOPGAVE 2

INLEIDING 4

1.1 Achtergrond 4

1.2 Ten geleide 5

2 TIMSS 6

2.1 Steekproef 6

2.2 Plausible values 6

IRT-analyse 7

Correlaties tussen TIMSS-parameters en Vlaamse parameters 7

Correlaties tussen plausible values en vaardigheidsscores 8

Conclusie 9

3 HET BELEIDSVOEREND VERMOGEN VAN SCHOLEN 10

3.1 Steekproef 10

3.2 De vragenlijst beleidsvoerend vermogen 10

3.3 Validering van de vragenlijst 11

Crossvalidering 11

Samenhang tussen de dragers van beleidsvoerend vermogen 12

Kwaliteit van het model 12

Dimensionaliteit, consistentie en betrouwbaarheid van de schalen 15

3.4 Van schaalscores naar schoolscores 15

Verschillen in percepties van beleidsvoerend vermogen 16

Verschillen in (gepercipieerd) beleidsvoerend vermogen van scholen 17

3.5 Conclusie 17

4 SAMENHANG BELEIDSVOEREND VERMOGEN EN LEERLINGENPRESTATIES 19

4.1 Steekproef 19

4.2 Multilevel analyse 19

4.3 Variabelen 20

Afhankelijke variabelen 20

Onafhankelijke variabelen 20

4.4 Gewichten 22

4.5 Beschrijvende statistieken 24

Vaardigheden van leerlingen 24

Dragers van beleidsvoerend vermogen 24

Controlevariabelen 24

Verschillen tussen en binnen scholen 25

4.6 Modellen 26

Modelopbouw 26

Modelvergelijking 27
Assumpties 28
4.7 Resultaten 28
5 CONCLUSIES 31
5.1 Duiding bij de resultaten 31
5.2 Aanbevelingen 32
REFERENTIES 34

Tabellen

Tabel 1 Correlaties tussen itemparameters
Tabel 2 Correlaties tussen plausible values en vaardigheidsscores
Tabel 3 Aantal deelnemers per functie
Tabel 4 Correlatiecoëfficiënten van de dragers van beleidsvoerend vermogen
Tabel 5 Kwaliteit van het oorspronkelijke en verfijnde model
Tabel 6 Factorladingen (cfa)
Tabel 7 Kwaliteit van het oorspronkelijke en verfijnde model
Tabel 8 Verschillen tussen functies
Tabel 9 Beschrijvende statistieken per drager van beleidsvoerend vermogen (geaggregeerde schoolscores)
Tabel 10 Aantal scholen per stratum
Tabel 11 Overzicht van de steekproef
Tabel 12 Overzicht van de variabelen
Tabel 13 Overzicht van het aantal leerlingen per categorie en de gemiddelde prestaties (gewogen)
Tabel 14 Overzicht van de modellen
Tabel 15 Modelvergelijking
Tabel 16a Parameters voor de modellen voor elke drager van beleidsvoerend vermogen (wiskunde)
Tabel 16b Parameters voor de modellen voor elke drager van beleidsvoerend vermogen (wetenschappen)

Figuren

Figuur 1a Correlaties itemparameters wiskunde
Figuur 1b Correlaties itemparameters wetenschappen
Figuur 2a Correlaties vaardigheidsscores wiskunde
Figuur 2b Correlaties vaardigheidsscores wetenschappen
Figuur 3 Spreiding van de schoolscores beleidsvoerend vermogen
Figuur 4 Verdeling van de vaardigheidsscores wiskunde en wetenschappen (pv1)
Figuur 5 Samenhang tussen schoolgrootte en vaardigheden van leerlingen (irt)
Figuur 6 Spreiding van de vaardigheden van leerlingen per school

INLEIDING

1.1 ACHTERGROND

Vlaamse scholen beschikken over een relatief grote autonomie om de kwaliteit van het onderwijs te bewaken (Valcke & Standaert, 2020; Vanhoof & Van Petegem, 2017). De overheid legt minimumdoelstellingen op en ziet toe op de kwaliteit van het onderwijs via inspecties. De besluitvaardigheid van scholen hangt af van hun beleidsvoerend vermogen. Vanhoof en Van Petegem (2017) omschrijven beleidsvormend vermogen als de mate waarin scholen hun autonomie succesvol inzetten in een continu proces van bijsturing van hun functioneren om de kwaliteit van het onderwijs te verbeteren en de door de overheid en henzelf gestelde doelen te bereiken.

Onderzoek naar beleidsvoerend vermogen van scholen past binnen het bredere schooleffectiviteitsonderzoek. Onderzoekers en beleidsmakers gaan ervan uit dat het beleidsvoerend vermogen van scholen invloed heeft op de prestaties van leerlingen. De precieze aard van deze relatie is moeilijk te ontrafelen. De Maeyer et al. (2007) beschouwen schoolklimaat als een intermediaire variabele tussen leiderschap en prestaties. Chang (2011) vond dat gedistribueerd (gedeeld) leiderschap invloed had op de prestaties van leerlingen via het academisch optimisme van leraren en Leithwood et al. (2020) identificeerden 11 factoren die leiderschap en leerlingenprestaties met elkaar verbinden, waaronder rationele, emotionele, organisatorische en familiale omstandigheden.

Op vraag van het Departement Onderwijs van de Vlaamse Gemeenschap onderzochten we de samenhang tussen het beleidsvoerend vermogen van scholen en de prestaties wiskunde en wetenschappen van leerlingen. Hiervoor werden de gegevens van TIMSS 2019 gekoppeld aan de gegevens uit het onderzoek naar beleidsvoerend vermogen van scholen. De centrale vraag van dit onderzoek luidt:

> Is er een samenhang tussen beleidsvoerend vermogen van scholen en de prestaties wiskunde een wetenschappen van leerlingen?

Voor we deze vraag kunnen beantwoorden, dienen we de validiteit en betrouwbaarheid van de vragenlijst beleidsvoerend vermogen na te gaan. De vraag hierbij is of de vragenlijst meet wat hij beoogt te meten en of we vanuit de individuele percepties van respondenten uitspraken kunnen doen over het beleidsvoerend vermogen van scholen. Dit brengt ons bij de eerste deelvraag:

> Is de validiteit en betrouwbaarheid van de vragenlijst beleidsvoerend vermogen voldoende groot om uitspraken te doen over het beleidsvoerend vermogen van scholen?

De vragenlijst werd uitgedeeld aan het volledige pedagogische team van de deelnemende scholen. Dat betekent dat zowel de percepties van directeurs als die van (zorg)leerkrachten, beleidsondersteuners en (zorg)coördinatoren bevraagd. We formuleren de tweede deelvraag als volgt:

> Zijn er grote verschillen in de percepties van beleidsvoerend vermogen tussen personeelsleden van eenzelfde school?

Omdat de vragenlijst in zijn huidige vorm niet eerder op deze schaal werd ingezet, loont het de moeite om de data van naderbij te bekijken. Voor we de samenhang tussen beleidsvoerend vermogen van scholen en leerlingenprestaties verder onderzoeken, is het zinvol om deze verschillen tussen scholen in kaart te brengen. De derde deelvraag luidt:

> Zijn er grote verschillen in de percepties van beleidsvoerend vermogen tussen personeelsleden van verschillende scholen?

1.2 TEN GELEIDE

>> Dit compendium kan worden opgevat als een wetenschappelijke onderbouwing bij de research briefs. We hebben ernaar gestreefd de informatie op een zo begrijpelijk mogelijk manier te presenteren. Dit betekent dat we ons beperken tot die informatie die nodig is om de gezette stappen te volgen, zonder ons te verliezen in alle techniciteit waaraan de analyses onderhevig zijn. Om ervoor te zorgen dat de kritische lezer niet op zijn of haar honger blijft zitten, verwijzen we waar nodig door naar relevante wetenschappelijke literatuur.

Het compendium bestaat uit vier delen. In het eerste deel staan we stil bij de eigenheden van TIMSS en geven we een korte beschrijving van de Vlaamse TIMSS-data. Het tweede deel legt de focus op het onderzoek naar het beleidsvoerend vermogen van scholen. Hierbij worden de verschillende stappen in het valideringsproces toegelicht. Het derde deel omvat de analyses om te komen tot een antwoord op de hoofdvraag. We sluiten dit rapport af met conclusies en aanbevelingen.

2 TIMSS

Trends in International Mathematics and Science Study (TIMSS) is een internationaal vergelijkend onderzoek dat de prestaties voor wiskunde en wetenschappen meet van leerlingen uit het vierde leerjaar en het tweede middelbaar. Het cyclische karakter van TIMSS maakt het mogelijk om trends in de leerlingenprestaties te meten. Het internationale karakter van TIMSS zorgt ervoor dat leerlingenprestaties tussen landen vergeleken kunnen worden. De *International Association for the Evaluation of Educational Achievement* (IEA) coördineert het onderzoek. De IEA stelt alle data publiek ter beschikking via haar website en voorziet een uitgebreide handleiding voor onderzoekers (https://timss2019.org/international-database).

De dataverzameling en analyses van Internationaal vergelijkend onderzoek zoals TIMSS gebeurt volgens specifieke procedures. In dit deel lichten we de steekproeftrekking toe en de manier waarop TIMSS vaardigheden van leerlingen schat (plausible values).

2.1 STEEKPROEF

De steekproeftrekking voor TIMSS gebeurt in twee stappen. Eerst worden scholen geselecteerd en vervolgens klassen. De selectie van scholen gebeurt door een gestratificeerde steekproef (*Stratified Two-Stage Cluster Sample Design*). Voor Vlaanderen werden scholen ingedeeld in expliciete strata op basis van provincie, het aantal leerlingen dat aantikt op de Vlaamse SES-indicatoren (hoog, midden, laag) en het onderwijstype (officieel en vrij, enkel voor Antwerpen). Alle Vlaamse scholen uit het Brussels hoofdstedelijk gewest en het buitengewoon basisonderwijs (BuBaO) werden in één stratum ingedeeld. Binnen elk stratum werden vervolgens willekeurig acht scholen geselecteerd. Voor elke geselecteerde school werden ook twee vervangscholen geselecteerd voor het geval de eerste school niet kon of wenste deel te nemen aan het onderzoek. Binnen de geselecteerde scholen werden indien mogelijk opnieuw willekeurig twee klassen van het vierde leerjaar geselecteerd. Alle leerlingen uit deze klassen werden uitgenodigd deel te nemen aan het onderzoek. In totaal namen 4655 leerlingen uit 147 scholen deel aan TIMSS 2019. Meer uitgebreide informatie over de Vlaamse steekproef vind je terug in het Vlaams TIMSS-Rapport (Faddar et al., 2020) en het technisch rapport van TIMSS (Martin et al., 2020).

2.2 PLAUSIBLE VALUES

Door het design van *Large-Scale Assessments* zoals TIMSS is het schatten van prestatieniveaus of vaardigheden van leerlingen complex. Elke leerling krijgt slechts een fractie (ongeveer 15 %) van de opdrachten uit de totale itempool voorgelegd. Om een vaardigheidsscore te berekenen voor wiskunde en wetenschappen maakt TIMSS gebruik van zogenaamde plausible values. Plausible values zijn schattingen van de vaardigheden van leerlingen rekening houdend met de antwoorden van de leerling op de toetsitems, de achtergrondkenmerken van de leerling en de moeilijkheidsgraad, het discriminerend vermogen en indien van toepassing de gokfactor van de items. Plausible values kunnen niet geïnterpreteerd worden als individuele vaardigheidscores van leerlingen. Ze worden in essentie gebruikt om uitspraken te doen op systeemniveau. De IEA raadt het gebruik van de aangeleverde plausible values bij secundaire analyses sterk aan. (Martin et al., 2020).

Om vaardigheden van leerlingen tussen landen te kunnen vergelijken vertrekt TIMSS vanuit de aanname dat de kans op een juist antwoord voor leerlingen met een gelijkaardig vaardigheidsniveau onafhankelijk is van de populatie waartoe de leerling behoort. De geschatte moeilijkheid van een opdracht hangt echter niet alleen af van de vaardigheid van leerlingen, maar ook van het aangeboden curriculum en de cultuur waarin een leerling leeft. Omdat we voor het huidige onderzoek niet de intentie hebben om vergelijkingen te maken met andere landen en er zeker van willen zijn dat de itemparameters en plausible values niet te sterk vertekend zijn door verschillen tussen het Vlaamse curriculum en de curricula uit andere landen, onderzochten we met een IRT-analyse de samenhang tussen itemparameters van TIMSS en itemparameters berekend op basis van de Vlaamse data. Vervolgens controleerden

we de correlatie tussen de plausible values van TIMSS en puntschattingen van de vaardigheden van leerlingen op basis van de Vlaamse itemparameters.

IRT-analyse

Het schatten van vaardigheden van leerlingen gebeurt op basis van de *Item Respons Theorie* (IRT). IRT is een set van statistische modellen die gebaseerd is op de relatie tussen individuele prestaties op een toetsvraag en het algemene prestatieniveau van een respondent. IRT-analyses maken het mogelijk om op basis van toetsgegevens iets te zeggen over de moeilijkheid van een item, de mate waarin een item erin slaagt een onderscheid te maken tussen goede en minder goede presteerders (discriminatievermogen) en de vaardigheid van deelnemers.

IRT-analyse vertrekt vanuit het idee dat vaardige leerlingen meer kans hebben een vraag juist te beantwoorden dan minder vaardige leerlingen. De kans dat een leerling het antwoord op een vraag juist heeft is hierbij afhankelijk van de moeilijkheid van het item, het discriminatievermogen van het item en de vaardigheid van de leerling. Afhankelijk van het type vraag wordt de kans op een juist antwoord anders berekend. Voor dit onderzoek worden drie IRT-modellen gebruikt: Het 2-parameter model (2PL), het 3 parameter model (3PL) en het generalized partial credit model (gpcm).

Voor het herberekenen van de vaardigheidsscores van Vlaamse leerlingen werd een nieuw IRT-model geschat op basis van de Vlaamse TIMSS-data. Eerst werden de toetsresultaten omgezet naar toetsscores met behulp van het door IEA aangeleverde hulpprogramma ASASCRM7 (SPSS, te downloaden van https://timss2019.org/international-database). Vervolgens werd de dataset gesplitst in items wiskunde en items wetenschappen. Op basis van de iteminformatie werd voor elk item het geschikte IRT-model bepaald. Voor dichotoom gescoorde open vragen werd een 2PL-model gebruikt, voor meerkeuzevragen 3PL en voor open vragen die een score 0, 1 of 2 kunnen opleveren een gpcm. Bij het schatten van de item-parameters werden de leerlingewichten (TOTWGT) gebruikt. Dit zorgt ervoor dat de invloed van de antwoorden van een leerling op de item-parameters steeds in verhouding staat tot het aantal leerlingen in de respectievelijke strata. Vervolgens worden op basis van deze parameterschattingen nieuwe vaardigheidsscores wiskunde en wetenschappen geschat voor alle leerlingen.

Correlaties tussen TIMSS-parameters en Vlaamse parameters

Tabel 1 geeft de correlatiecoëfficiënten weer tussen de item parameters van TIMSS en de geschatte parameters op basis van de Vlaamse data. Correlaties boven de 0.50 worden doorgaans als sterk beschouwd (Cohen, 1988). Voor wiskunde hangen alle itemparameters sterk samen, met uitzondering van de gokparameter. De discriminatieparameters voor de wetenschapsitems correleren slechts matig. Ook voor wetenschappen is de samenhang tussen de gokparameters van TIMSS en de Vlaamse data eerder aan de lage kant.

TABEL 1 CORRELATIES TUSSEN ITEMPARAMETERS

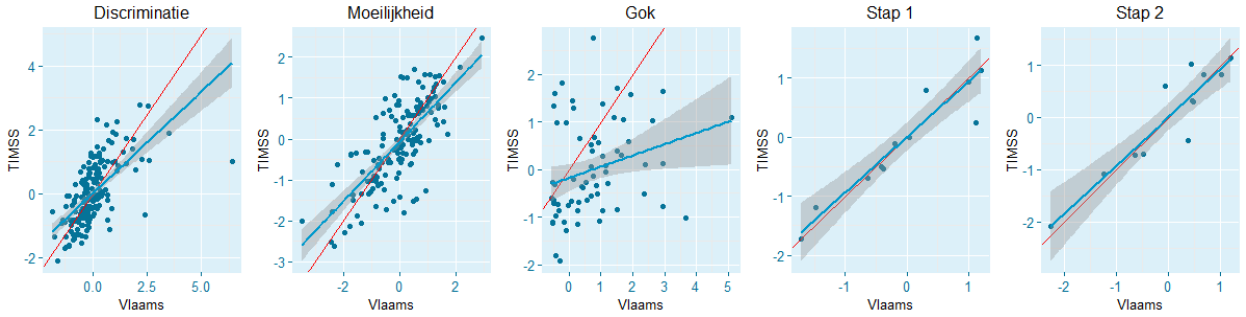
	Wiskunde	Wetenschappen
Discriminatie (slope/a)	0.640***	0.418***
Moeilijkheid (Location/b)	0.733***	0.662***
Gokparameters (Guessing/g)	0.274*	0.358***
Stap 1 (Step1/d1)	0.936***	0.961***
Stap 2 (Step2/d2)	0.923***	0.992***
Stap 2 (Step2/d2)	0.923***	0.992***

*** p < .001 ** p < .01 * p < .05

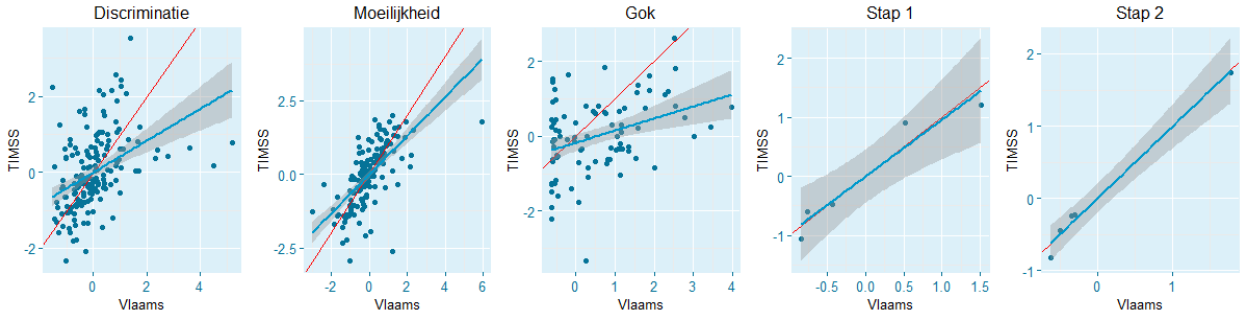
Figuur 1a en 1b geven de correlaties tussen de itemparameters visueel weer. Elk puntje op de grafiek staat voor een toetsitem. Om de figuur beter te kunnen interpreteren werden de parameters gestandaardiseerd. Een oefening met een gemiddelde moeilijkheidsgraad krijgt zo de waarde 0 voor de moeilijkheidsparameter. Een oefening die 1 *SD* moeilijker is dan gemiddeld krijgt de waarde 1 enzovoort. De rode lijn op de figuur geeft een indicatie van de perfecte correlatie. Hoe verder het punt verwijderd is van de rode lijn, hoe minder groot de overeenkomst tussen de TIMSS-

parameter en de Vlaamse parameter. De blauwe lijn is de regressielijn die het best bij de correlaties past. Hoe minder de beide lijnen van elkaar afwijken, hoe beter de correlatie.

FIGUUR 1A CORRELATIES ITEMPARAMETERS WISKUNDE



FIGUUR 1B CORRELATIES ITEMPARAMETERS WETENSCHAPPEN



Correlaties tussen plausible values en vaardigheidsscores

Vervolgens werd de correlatie nagegaan tussen de vaardigheidsscores die werden geschat op basis van de Vlaamse item-parameters en de vijf plausible values van TIMSS. Wanneer de geschatte vaardigheidsscores van individuele leerlingen sterk afwijken van de plausible values van TIMSS, kan dit mogelijk de resultaten van dit onderzoek vertekenen.

Tabel 2 geeft de correlatiecoëfficiënten weer tussen de vaardigheidsscores op basis van de Vlaamse data en de plausible values van TIMSS. Alle correlaties voor wiskunde en wetenschappen liggen boven de 0.80 en zijn sterk (Cohen, 1988). Alle correlaties zijn bovendien significant.

TABEL 2 CORRELATIES TUSSEN PLAUSIBLE VALUES EN VAARDIGHEIDSSCORES

	Vaarigheidsscores	
Plausible value	Wiskunde	Wetenschappen
PV 1	0.894***	0.824***
PV 2	0.898***	0.827***
PV 3	0.894***	0.822***
PV 4	0.892***	0.822***
PV 5	0.895***	0.821***

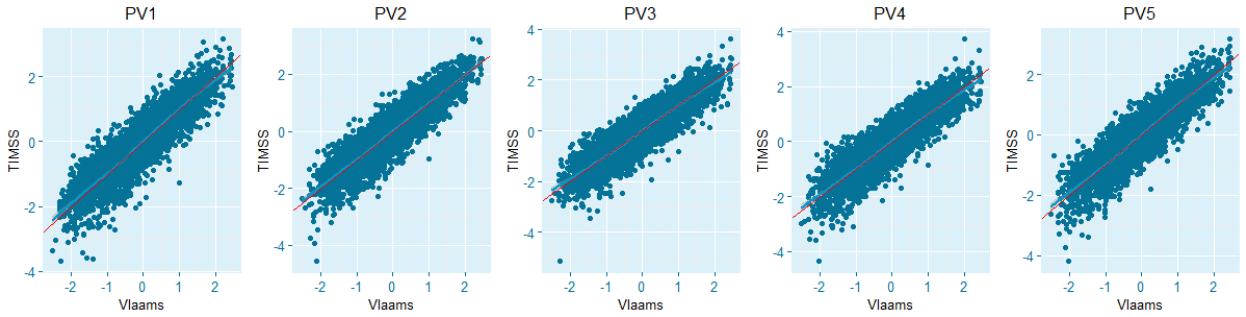
*** p < .001 ** p < .01 * p < .05

Figuur 2a en 2b geven de correlaties tussen de vaardigheidsscores en de plausible values visueel weer. Elk puntje op de grafiek staat voor een Vlaamse leerling die deelnam aan TIMSS 2019. Om de figuur beter te kunnen interpreteren, werden de vaardigheidsscores en plausible values gestandaardiseerd. De rode lijn op de figuur geeft een indicatie van de perfecte correlatie. De blauwe lijn is de regressielijn die het best bij de correlaties past.

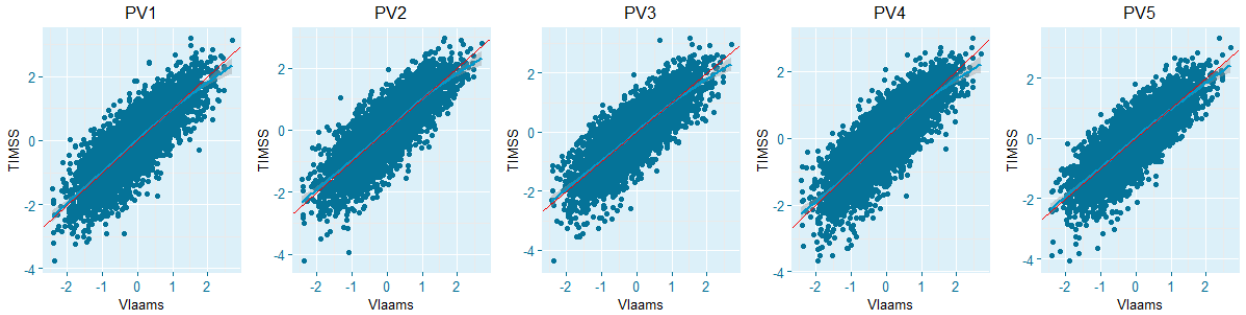
Hierbij kan worden opgemerkt dat de correlaties voor wiskunde net iets sterker zijn dan voor wetenschappen. De puntenwolken met betrekking tot wetenschappen in Figuur 2b zijn meer uitgespreid dan die voor wiskunde in

Figuur 2a. Dit is niet onlogisch aangezien de wetenschapsvragen meer appelleren aan de leefomgeving van de leerling dan de wiskundevragen. De moeilijkheid van de wetenschapsvragen is mogelijk meer curriculum- of contextgebonden dan de meer algemene en abstracte wiskundevragen.

FIGUUR 2A CORRELATIES VAARDIGHEIDSSCORES WISKUNDE



FIGUUR 2B CORRELATIES VAARDIGHEIDSSCORES WETENSCHAPPEN



Conclusie

Om na te gaan of de plausible values een goede weerspiegeling zijn van de vaardigheden van leerlingen in de specifieke, Vlaamse context, werden nieuwe itemparameters en vaardigheden van leerlingen geschat door middel van IRT-analyse. Zowel wat betreft de itemparameters als wat betreft de geschatte vaardigheden van leerlingen (puntschatting of plausible values) is er een overwegend sterke samenhang tussen de TIMSS-data en de vaardigheidsschattingen op basis van de Vlaamse data. Ondanks deze sterke samenhang is enige voorzichtigheid geboden bij het gebruik van TIMSS-data voor secundaire analyses op Vlaams niveau. Plausible values zijn uitermate geschikt om uitspraken te doen op systeemniveau ("Vlaamse leerlingen …"). Maar Wanneer we geïnteresseerd zijn in de gemiddelde vaardigheid van leerlingen op schoolniveau, kan het gebruik van plausible values mogelijk vertekeningen opleveren. Dit geldt zeker voor kleine scholen met een beperkt aantal deelnemende leerlingen. Daarom voerden we de uiteindelijke analyses parallel uit met de plausible values van TIMSS en de geschatte vaardigheidsscores op basis van de Vlaamse data. Om dit compendium overzichtelijk te houden wordt enkel over de analyses met de plausible values gerapporteerd. Wanneer deze resultaten substantieel en significant afwijken van de analyses met de Vlaamse vaardigheidsscores, wordt dit steeds vermeld.

3 HET BELEIDSVOEREND VERMOGEN VAN SCHOLEN

3.1 STEEKPROEF

Aan de directeurs van de scholen die werden geselecteerd voor TIMSS 2019 werd gevraagd om ook deel te nemen aan het onderzoek rond beleidsvoerend vermogen. Alle leden van het pedagogische team van de school (leerkrachten, zorgleerkrachten, (zorg)coördinatoren, beleidsondersteuners en directieleden) werden uitgenodigd de vragenlijst in te vullen. In totaal vulden 1696 personeelsleden uit 77 scholen de vragenlijst in. Het gemiddelde aantal respondenten per school bedraagt 21.95 (*SD* = 10.10, min. = 6, max = 56). De overgrote meerderheid van de respondenten is vrouwelijk (n = 1343, 86.65 %).

Tabel 3 toont het aantal respondenten dat de vragenlijst invulde per functie. De meeste vragenlijsten werden ingevuld door leerkrachten (70.64 %). Voor 59 scholen vulden één of meerdere directieleden de vragenlijst in. Voor 13 andere scholen vulde geen enkel directielid de vragenlijst in. Per school vulden gemiddeld 22.49 respondenten de vragenlijst in (*SD* = 9.75). De gemiddelde responsgraad binnen de deelnemende scholen bedraagt 76.11 % (*SD* = 17.70).

TABEL 3 AANTAL DEELNEMERS PER FUNCTIE

Functie	Aantal	Percentage
Leerkracht	1206	71.15
Zorgleerkracht	143	8.44
Beleidsondersteuner/zorgcoördinator	108	6.37
Directie	68	4.01
Andere	170	10.03
NA	*1*	
Totaal	1696	100

3.2 DE VRAGENLIJST BELEIDSVOEREND VERMOGEN

De Vragenlijst Beleidsvoerend Vermogen die voor dit onderzoek werd gebruikt, is ontwikkeld door Van Petegem et al. (2004) in het kader van een onderzoek naar het zelfevaluerend en beleidsvoerend vermogen van scholen. De vragenlijst is opgebouwd rond acht pijlers. Vanhoof et al. (2011) omschrijven de acht pijlers als volgt.

Effectieve communicatie. Effectieve communicatie verwijst naar de mate waarin formele en informele communicatie-initiatieven tegemoetkomen aan de informatiebehoeftes en aan de wensen ot wederzijdse beïnvloeding en afstemmen.

Gedeeld leiderschap. Gedeeld leiderschap verwijst naar de participatie van het schoolteam in de besluitvorming op school.

Geïntegreerd beleid. Geïntegreerd beleid verwijst naar de onderlinge afstemming van de verschillende afzonderlijke beleidsdomeinen binnen het ruimere schoolbeleid.

Gezamenlijke doelgerichtheid. Gezamenlijke doelgerichtheid omvat het werken vanuit een richtinggevende visie die door verschillende betrokkenen gedragen wordt.

Innovatief vermogen. Innovatief vermogen verwijst naar de mate waarin de school open staat voor vernieuwingen, naar de wijze waarop ze omgaat met veranderingen en erin slaagt om vernieuwingen op een succesvolle manier te implementeren.

Ondersteunende relaties. Ondersteunende relaties verwijst naar de mate waarin professionele en persoonlijke relaties tussen leden van het schoolteam als ondersteunend ervaren worden.

Reflectief vermogen. Reflectief vermogen verwijst naar de mate waarin de school initiatieven neemt om in het eigen functioneren sterke en zwakke punten te identificeren.

Responsief vermogen. Responsief vermogen verwijst naar de mate waarin de school open staat voor, en in staat is een antwoord te bieden op externe vragen en verwachtingen.

Verondersteld wordt dat deze acht dragers samenhangen met beleidseffectiviteit van scholen. Een beleidseffectieve school is:

> een school die haar beschikbare beleidsruimte ten volle aanwendt om te komen tot een voortdurend proces van autonoom en bewust veranderen van haar organisatorische en onderwijskundige functioneren met als doel het verbeteren van haar onderwijskwaliteit en het bereiken van de haar opgelegde en eigen doelen. (Van Petegem et al., 2004, p. 32)

De vragenlijst kan gebruikt worden als zelfevaluatieinstrument voor scholen of als onderzoeksinstrument om de relatie na te gaan tussen beleidsvoerend vermogen van scholen en andere variabelen. De vragenlijst is een dynamisch instrument dat kan worden aangepast aan de noden van scholen of onderzoekers. De voor het huidige onderzoek gebruikte vragenlijst is gebaseerd op de vragen uit het *Praktijkboek Beleidsvoerend Vermogen in Scholen* (Vanhoof & Van Petegem, 2017). De vragenlijst bestaat uit 78 vragen[1] verspreid over de acht dragers van beleidsvoerend vermogen. De vragen werden beoordeeld op een 5-punts Likertschaal van 1 (*helemaal oneens*) tot 5 (*helemaal eens*) met een neutrale antwoordoptie 3 (*noch oneens/noch eens*).

3.3 VALIDERING VAN DE VRAGENLIJST

Crossvalidering

De validiteit van de vragenlijst beleidsvoerend vermogen werd nagegaan door crossvalidering. Crossvalidering is een methode om de validiteit van onder andere vragenlijsten na te gaan door een dataset willekeurig in twee te splitsen. Het ene deel van de dataset (traindata) wordt gebruikt om een model op te bouwen of te verbeteren, het andere deel (testdata) om de validiteit van dat model na te gaan. Bij een analyse op basis van één dataset blijft de vraag in welke mate het resulterende model past bij andere steekproeven. Door de dataset willekeurig te splitsen, creëer je twee onafhankelijke steekproeven waardoor het opgestelde model meteen getoetst kan worden op een tweede dataset.

Voor de validering van de vragenlijst en het verbeteren van het model werd de dataset willekeurig in twee gelijke delen gesplitst (N_{Train}=848, N_{Test}=848). Met een *exploratieve factoranalyse* (EFA) op de ene helft (traindata) werd gezocht naar het model dat het best bij de data past. Vervolgens werd dit model getoetst met een *confirmatieve factoranalyse* (CFA) op de andere helft van de data (testdata).

Een EFA is een techniek om de factorstructuur van een dataset na te gaan op basis van de gegevens zelf. Hierbij wordt gezocht naar de factorstructuur die het best bij de data past of anders gezegd: de factorstructuur waarbij de variantie over alle manifeste variabelen heen (de verschillende vragen uit de vragenlijst) zoveel mogelijk toe te schrijven is aan de verschillen tussen de latente variabelen (de dragers van beleidsvoerend vermogen).

Op basis van de resultaten van de EFA werden verschillende items verwijderd omdat ze niet goed passen bij de andere items van dezelfde drager. De drager geïntegreerd beleid viel hierbij volledig weg. Geïntegreerd of integraal beleid heeft betrekking op de mate waarin scholen erin slagen verschillende beleidsdomeinen op elkaar af te

[1] Twee vragen van de pijler effectieve communicatie werden niet weerhouden voor het huidige onderzoek.

stemmen, vertrekkende vanuit een overkoepelende visie (Vanhoof & Van Petegem, 2017). Geïntegreerd beleid kan als dusdanig worden beschouwd als een overkoepelende drager die de andere dragers met elkaar verbindt. Zo beschouwd lijkt geïntegreerd beleid moeilijk te meten met een vragenlijst. De drager responsief vermogen valt uiteen in twee factoren. Inhoudelijke analyse van de vragen bevestigt deze tweedeling. De eerste vijf vragen die peilen naar het responsief vermogen van scholen verwijzen expliciet naar responsiviteit ten aanzien van ouders, de laatste vijf vragen hebben betrekking op responsief vermogen ten aanzien van de bredere samenleving. Deze tweedeling is niet verwonderlijk en kon worden verwacht. In het eerste deel van het oorspronkelijke onderzoek waarbij de vragenlijst beleidsvoerend vermogen werd ontwikkeld, maakten Van Petegem et al. (2004) een onderscheid tussen responsiviteit ten aanzien van ouders en responsiviteit ten aanzien van de omgeving. In het tweede deel van hun onderzoek werden beide schalen samengevoegd.

Het resulterende model uit de EFA werd vervolgens getoetst op de andere helft van de data (testdata) met een CFA. Hiermee werd nagegaan of het model ook goed blijkt te zijn wanneer een andere steekproef wordt gebruikt. Voor elke schaal werden vervolgens enkele controles uitgevoerd. Het uiteindelijke model bestaat uit 49 items, verspreid over acht dragers van beleidsvoerend vermogen.

Samenhang tussen de dragers van beleidsvoerend vermogen

Omdat de verschillende dragers een onderdeel zijn van beleidsvoerend vermogen, gaan we ervan uit dat de dragers onderling samenhangen. Tabel 4 toont de correlatiecoëfficiënten tussen de verschillende dragers van beleidsvoerend vermogen. Correlaties groter dan .50 worden in de onderwijswetenschappen als sterk beschouwd (Cohen, 1988). Effectieve communicatie hangt sterk samen met gedeeld leiderschap (r = .64) en ondersteunende relaties (r = 51). Gedeeld leiderschap hangt sterk samen met gezamenlijke doelgerichtheid (r = .52) en ondersteunende relaties hangen sterk samen met reflectief vermogen.

TABEL 4 CORRELATIECOËFFICIËNTEN VAN DE DRAGERS VAN BELEIDSVOEREND VERMOGEN

	Doeltreffende communicatie	Gedeeld leiderschap	Gezamenlijke doelgerichtheid	Innovatief vermogen	Ondersteunende relaties	Reflectief vermogen	Responsief vermogen t.a.v. ouders	Responsief vermogen t.a.v. samenleving
Doeltreffende communicatie	1							
Gedeeld leiderschap	.64	1						
Gezamenlijke doelgerichtheid	.42	.52	1					
Innovatief vermogen	.32	.33	.45	1				
Ondersteunende relaties	.51	.43	.32	.34	1			
Reflectief vermogen	.36	.32	.42	.47	.51	1		
Responsief vermogen t.a.v. ouders	.22	.31	.25	.21	.23	.34	1	
Responsief vermogen t.a.v. samenleving	.20	.27	.31	.38	.31	.35	.43	1

Kwaliteit van het model

Om de kwaliteit van het verfijnde model te controleren, werd een CFA uitgevoerd met de testgegevens. De fit van het verfijnde model met 49 items is aanvaardbaar en beter dan die van het oorspronkelijke model met 78 items. Tabel 5 toont de fit-indices voor het oorspronkelijke en het verfijnde model. De RMSR (.048) en RMSEA (.048) van het verfijnde model liggen ruim onder de. De CFI en TLI benaderen .95. De p-waarde van de chi-kwadraat is < .05.

Dit is grotendeels het gevolg van de steekproefomvang. De p-waarde van de chi-kwadraat neemt immers af naarmate de steekproefgrootte toeneemt.

TABEL 5 KWALITEIT VAN HET OORSPRONKELIJKE EN VERFIJNDE MODEL

		Oorspronkelijke model	Verfijnde model	Drempelwaarde
Chi-kwadraat		9564.730	3113.448	
	Vrijheidsgraden	2897	1099	
	p-waarde	.000	.000	> .05
CFI		.828	.923	> .95
TLI		.822	.918	> .95
RMSR		.056	.046	< .08
RMSEA		.056	.048	< .06
RMSEA		.056	.048	< .06

Tabel 6 toont de factorladingen van de items. Bijna alle gestandaardiseerde factorladingen liggen ruim boven .70. Slechts zes items hebben een factorlading tussen .62 en .70.

TABEL 6 FACTORLADINGEN (CFA)

	Factorlading	Gestandaardiseerde factorlading
Effectieve communicatie		
Q13_1	1	0.871
Q13_2	0.999	0.85
Q13_3	0.796	0.72
Q13_4	1.012	0.825
Q13_6	0.86	0.657
Gedeeld leiderschap		
Q14_1	1	0.804
Q14_2	0.986	0.804
Q14_3	0.924	0.779
Q14_4	1.135	0.809
Q14_5	1.014	0.799
Q14_6	0.833	0.725
Q14_7	0.953	0.733
Q14_9	1	0.798
Q14_10	0.99	0.815
Q15_6	0.784	0.686
Gezamenlijke doelgerichtheid		
Q16_1	1	0.779
Q16_2	1.088	0.833
Q16_3	0.951	0.76
Q16_9	1.03	0.764
Q16_10	0.924	0.769
Innovatief vermogen		
Q17_1	1	0.692
Q17_2	0.907	0.711
Q17_4	0.988	0.8
Q17_7	0.938	0.755
Q17_8	0.897	0.721
Q17_9	0.858	0.784
Ondersteunende relaties		
Q18_1	1	0.813
Q18_3	1.053	0.841
Q18_4	1.101	0.81
Q18_5	0.861	0.703
Q18_6	1.144	0.84
Q18_10	0.899	0.756
Q16_4	1.079	0.778
Reflectief vermogen		
Q19_3	1	0.775
Q19_4	0.944	0.779
Q19_5	0.894	0.754
Q19_6	0.901	0.628
Q19_8	0.912	0.694
Q19_9	0.93	0.719
Q19_10	0.993	0.803
Responsief vermogen t.a.v. ouders		
Q20_1	1	0.704
Q20_2	1.023	0.753
Q20_3	0.791	0.702
Q20_4	1.098	0.816
Q20_5	1.009	0.792
Responsief vermogen t.a.v. samenleving		
Q20_6	1	0.722
Q20_7	0.821	0.686
Q20_8	1.137	0.814
Q20_9	1.041	0.809

Dimensionaliteit, consistentie en betrouwbaarheid van de schalen

Naast de kwaliteit van het model is ook de dimensionaliteit en de interne consistentie van de schalen van belang. Een goede schaal meet één aspect van beleidsvoerend vermogen (eendimensionaliteit) en de items van de schaal passen goed bij elkaar en onderscheiden zich van de items van andere schalen (intern consistent). De interne consistentie wordt meestal gemeten door middel van Cronbach's alpha. Een Cronbach's alpha > .70 wordt doorgaans als goed beschouwd. Uit de analyse blijkt dat alle schalen ééndimensionaal zijn en intern consistent. Tabel 7 toont de Cronbach alpha's voor de verschillende schalen.

Voor we vanuit dit model gemiddelde schoolscores berekenen, is het belangrijk na te gaan of op basis van individuele antwoorden van leerkrachten uitspraken gedaan kunnen worden over de school. In het geval dat de percepties van respondenten binnen scholen sterk van elkaar verschillen, is het mogelijk problematisch om de individuele scores te aggregeren op schoolniveau. Om na te gaan of het berekenen van schoolscores zinvol is, bekeken we de overeenstemming en betrouwbaarheid van elke schaal en gingen we na of er significante verschillen zijn tussen respondenten met een verschillende functie in de school.

Uit Tabel 7 blijkt dat zowel de overeenstemming tussen de respondenten uit de verschillende scholen (*Within-Group Agreemant* (r_{WG})) als de betrouwbaarheid van de beoordelingen tussen de deelnemers (interbeoordelaars-betrouwbaarheid *ICC(1)* en *ICC(2)*) voldoende hoog is om de individuele scores te aggregeren op schoolniveau. Within group agreement geeft een indicatie van de mate waarin antwoorden van verschillende respondenten binnen een school uitwisselbaar zijn (Bliese, 2000). Een r_{WG} > .70 wordt doorgaans als acceptabel beschouwd (Lance et al., 2006). De ICC(1) is de proportie van de variantie die verklaard kan worden door het behoren tot de groep (i.c. school). Een hoge ICC(1) betekent dat een beoordeling van één deelnemer uit een school een relatief goede weerspiegeling is van het groepsgemiddelde. De ICC(2) is een schatting van de betrouwbaarheid van de groepsgemiddelden en geeft aan hoe betrouwbaar het geaggregeerde groepsgemiddelde (over de groepsleden heen) een onderscheid maakt tussen de verschillende groepen (Bliese, 2000). Hierbij baseren we ons op de meest gangbare criteria: r_{WG} ≥ , ICC(1) ≥ .10 en ICC(2) ≥ .70 (Biemann et al., 2012).

TABEL 7 KWALITEIT VAN HET OORSPRONKELIJKE EN VERFIJNDE MODEL

		r_{WG}		*ICC*	
Drager van Beleidsvoerend Vermogen	*α*	*M*	*SD*	*ICC(1)*	*ICC(2)*
Doeltreffende communicatie	.88	.90	.06	.28	.90
Gedeeld leiderschap	.94	.96	.02	.24	.88
Gezamenlijke doelgerichtheid	.89	.94	.04	.23	.87
Innovatief vermogen	.88	.95	.03	.23	.87
Ondersteunende relaties	.92	.93	.08	.20	.85
Reflectief vermogen	.89	.94	.03	.18	.82
Responsief vermogen t.a.v. ouders	.86	.94	.03	.19	.84
Responsief vermogen t.a.v. samenleving	.84	.94	.04	.11	.73

3.4 VAN SCHAALSCORES NAAR SCHOOLSCORES

Op basis van het aangepaste model werd voor elke school een score berekend per drager van beleidsvoerend vermogen. Dit gebeurde in drie stappen. Eerst werd voor elke respondent afzonderlijk een score berekend op elke schaal door de gemiddelde score over de items van de schaal te berekenen. In een tweede stap werden deze scores per school geaggregeerd door per schaal de gemiddelde score van de deelnemers te nemen. Deze scores werden ten slotte over de scholen heen gestandaardiseerd.

Het standaardiseren van schoolscores vergemakkelijkt de interpretatie. De score van scholen voor elke drager van beleidsvoerend vermogen is uitgedrukt in een betekenisloze schaal. We kunnen immers niet stellen dat een score 3 (*noch oneens/noch eens*) evenveel verschilt van score 4 (*eens*) als een score 4 van een score 5 (*helemaal eens*)

Standaardisatie is een lineaire transformatie van de oorspronkelijke schaal en heeft geen invloed op de verdeling van de scores. Een school met een gestandaardiseerde score 0, scoort gemiddeld voor de betrokken drager. Een school met score 1 scoort 1 standaardafwijking hoger dan gemiddeld en een school met score -1 scoort 1 standaardafwijking lager dan gemiddeld. Het resultaat van deze analyses is een databestand met voor elke school een gestandaardiseerde score voor de acht dragers van beleidsvoerend vermogen.

Verschillen in percepties van beleidsvoerend vermogen

De vragenlijst beleidsvoerend vermogen werd bezorgd aan alle leden van het pedagogische team van de deelnemende scholen. Dat betekent dat de vragen werden ingevuld door zowel leerkrachten als zorgleerkrachten, coördinatoren, beleidsondersteuners en directeurs. Om na te gaan of er verschillen zijn in de antwoorden tussen respondenten met een verschillende functie, werden multilevelanalyses uitgevoerd voor elke drager. Multilevelanalyse stelt ons in staat om de significantie van de verschillen tussen respondenten met een verschillende functie vast te stellen, rekening houdend met de school. Tabel 8 toont de verschillen tussen de functies per drager van beleidsvoerend vermogen. Omdat de meeste respondenten leerkrachten zijn, werden leerkrachten als referentiecategorie genomen. Om de resultaten beter interpreteerbaar te maken, werden de scores per drager gestandaardiseerd op het niveau van de individuele respondent. De waarden in de kolom 'effect' geven aan hoeveel standaarddeviaties hoger of lager respondenten met de betrokken functie scoorden. Enkel significante effecten werden opgenomen in de tabel.

TABEL 8 VERSCHILLEN TUSSEN FUNCTIES (GESTANDAARDISEERD)

	Com		Ged		Gez		Inn	
Random effects	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
School	.33	.58	.30	.54	.25	.50	.28	.53
Resterend	.71	.84	.74	.86	.76	.87	.76	.87
Fixed effects	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Intercept (leerkracht)								
Zorgleerkracht								
Beleidsondersteuner/ZorgCo.			.27**	.09			-.26**	.09
Directie	.52***	.11	.74***	.11	.31**	.11		
Andere								
	Rel		Ref		ResPar		ResSoc	
Random effects	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
School	.21	.46	.19	.44	.19	.44	.11	.33
Resterend	.80	.89	.81	.90	.81	.90	.89	.94
Fixed effects	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Intercept (leerkracht)								
Zorgleerkracht								
Beleidsondersteuner/ZorgCo.	-.20*	.09	-.49***	.09				
Directie			-.40***	.11			.24*	.12
Andere							-.20*	.08

*** p < .001 ** p < .01 * p < .05 ° p < .1

Deze analyse leert dat de percepties van respondenten met beleids(ondersteunende)functies voor enkele dragers significant afwijken van de percepties van leerkrachten. Directeurs schatten doeltreffende communicatie, gedeeld leiderschap en gezamenlijke doelgerichtheid van hun school significant hoger in dan leerkrachten. Voor de drager reflectief vermogen liggen de percepties van directeurs significant lager dan die van leerkrachten. Ook de percepties van beleidsondersteuners en zorgcoördinatoren wijken voor sommige dragers significant af van de percepties van leerkrachten. We kunnen dus stellen dat de functie die de respondenten bekleden binnen een school ertoe doet.

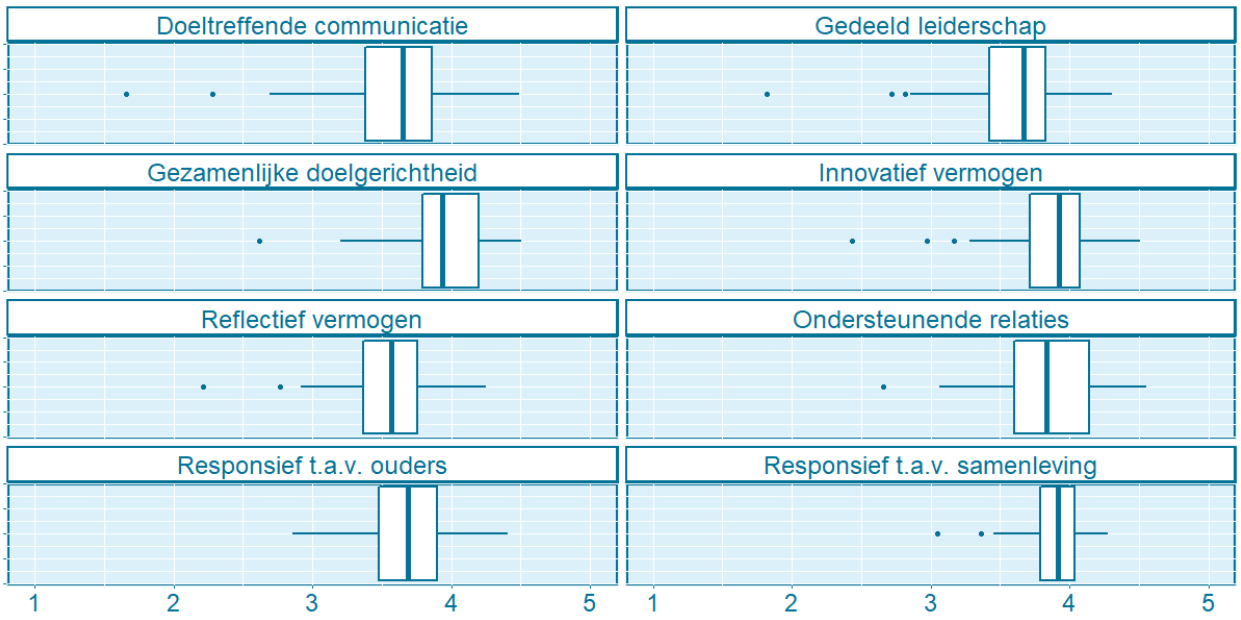
Verschillen in gepercipieerd beleidsvoerend vermogen van scholen

Uit de analyse van de resultaten blijkt dat er ook verschillen zijn in (gepercipieerd) beleidsvoerend vermogen tussen scholen. Tabel 9 toont de spreiding (minimum en maximum schoolscore), het gemiddelde en de standaarddeviatie voor de geaggregeerde schoolscores voor elke drager van beleidsvoerend vermogen. Een kleine 10 % van de scholen scoort minder dan 3 op doeltreffende communicatie. Dit betekent dat leerkrachten aangeven dat ze het overwegend oneens zijn met de (positief geformuleerde) stellingen over doeltreffende communicatie. Figuur 3 toont de spreiding van de schoolscores. Hier merken we enkele uitbijters (outliers) op. Dat zijn scholen die lager scoren dan het eerste kwartiel min anderhalve keer de interkwartielafstand. Twee scholen nemen het gros van de outliers voor hun rekening. Eén school is een outlier voor zes van de acht dragers van beleidsvoerend vermogen, een andere voor vijf van de acht. Drie andere scholen zijn een outlier voor telkens één van de dragers.

TABEL 9 BESCHRIJVENDE STATISTIEKEN PER DRAGER VAN BELEIDSVOEREND VERMOGEN (GEAGGREGEERDE SCHOOLSCORES)

Drager van beleidsvoerend vermogen	*Min*	*Max*	*M*	*SD*
Doeltreffende communicatie	1.66	4.49	3.57	0.48
Gedeeld leiderschap	1.82	4.31	3.60	0.41
Gezamenlijke doelgerichtheid	2.62	4.51	3.93	0.34
Innovatief vermogen	2.55	4.58	3.96	0.34
Ondersteunende relaties	2.84	4.43	3.79	0.34
Reflectief vermogen	2.22	4.26	3.57	0.32
Responsief vermogen t.a.v. ouders	2.86	4.41	3.68	0.29
Responsief vermogen t.a.v. samenleving	3.05	4.28	3.91	0.22

FIGUUR 3 SPREIDING VAN DE SCHOOLSCORES BELEIDSVOEREND VERMOGEN



3.5 CONCLUSIE

Om het beleidsvoerend vermogen van scholen in kaart te brengen, werd gebruik gemaakt van de Vragenlijst Beleidsvoerend Vermogen. Deze vragenlijst werd afgenomen bij 1696 leerkrachten uit 77 scholen. Uit de analyse blijkt dat de vragenlijst voldoende valide en betrouwbaar is om de individuele scores van leden van het pedagogische

team te aggregeren naar schoolniveau en dus uitspraken te doen over het beleidsvoerend vermogen van scholen. Tegelijk stellen we vast dat er soms grote verschillen zijn in de percepties van respondenten over het beleidsvoerend vermogen van hun school.

De visie van de directie, beleidsondersteuners of zorgcoördinatoren wijkt voor enkel dragers significant af van de visie van leerkrachten. Directeurs schatten de doeltreffendheid van de communicatie en de mate van gedeeld leiderschap significant hoger in dan de leerkrachten van de school. Het reflectief vermogen van de school schatten directeurs globaal genomen lager in dan de leerkrachten. Beleidsondersteuners en zorgcoördinatoren schatten het gedeeld leiderschap hoger in dan leerkrachten, maar zijn minder positief over het reflectief vermogen, innovatief vermogen en ondersteunende relaties.

De vragenlijst beleidsvoerend vermogen is ontwikkeld als zelfevaluatie-instrument voor scholen. Directeurs en in uitbreiding het hele schoolteam hebben baat bij een accuraat beeld over de sterktes en werkpunten van de school. De vragenlijst beleidsvoerend vermogen kan leden van een schoolteam helpen om hierover in discussie te gaan en prioriteiten vast te stellen.

4 SAMENHANG BELEIDSVOEREND VERMOGEN EN LEERLINGENPRESTATIES

4.1 STEEKPROEF

Voor de uiteindelijke analyses werden enkel scholen weerhouden waarvan ten minste 10 leden of 40 % van het pedagogische team de vragenlijst invulde. Deze criteria verhinderen dat de schoolscores worden berekend op basis van de percepties van slechts enkele leden van het pedagogische team. Uiteindelijk werden 1635 ingevulde vragenlijsten uit 72 scholen weerhouden voor de analyses met betrekking tot de samenhang tussen beleidsvoerend vermogen van scholen en prestaties van leerlingen op TIMSS.

Doordat slechts een deel van de scholen uit de aanvankelijke steekproef deelnam aan het bijkomende onderzoek, zijn de scholen niet meer evenredig verdeeld over de verschillende strata. Tabel 10 biedt een overzicht van het aantal scholen per stratum in de uiteindelijke steekproef. Uit elk stratum namen ten minste twee scholen deel aan zowel TIMSS als het onderzoek naar beleidsvoerend vermogen.

TABEL 10 AANTAL SCHOLEN PER STRATUM

	SES hoog		SES midden		SES laag		*NVT*	Totaal
	Off.	Vrij	Off.	Vrij	Off.	Vrij		
Antwerpen	4	6	2	2	3	3		**20**
Limburg	3		6		5			**14**
Oost-Vlaanderen	5		5		3			**13**
Vlaams-Brabant	3		6		3			**12**
West-Vlaanderen	2		2		4			**8**
Brussel							3	**3**
BuBaO							2	**2**
Totaal	**23**		**23**		**21**		**5**	**72**

Voor de uiteindelijke analyses naar de samenhang tussen beleidsvoerend vermogen van scholen en de prestaties van leerlingen beschikkende we over een dataset met gegevens van 2395 leerlingen uit 72 scholen (zie Tabel 11).

TABEL 11 OVERZICHT VAN DE STEEKPROEF

	TIMSS	Beleidsvoerend Vermogen	Uiteindelijke steekproef
Scholen	147	77	72
Pedagogisch personeel		1696	1635
Leerlingen	4665		2395

4.2 MULTILEVEL ANALYSE

Om de vraag naar de samenhang tussen beleidsvoerend vermogen van scholen en prestaties van leerlingen voor wiskunde en wetenschappen te beantwoorden, zijn multilevelanalyses aangewezen. Multilevelmodellen (ook wel hiërarchische modellen of *mixed effects models* genoemd) zijn gepast omdat leerlingen 'genest' zijn in scholen. Leerlingen uit eenzelfde school gelijken doorgaans meer op elkaar dan leerlingen uit verschillende scholen. Dit heeft onder meer te maken met voorkeuren van ouders en leerlingen bij de keuze van een scholen en

socialisatieprocessen die plaatsvinden op school. Voor dit onderzoek werden bovendien data verzameld op twee verschillende niveaus. De TIMSS-data voor deze onderzoeksvraag werden verzameld op het niveau van de leerlingen, de vragenlijst beleidsvoerend vermogen zorgde voor data op schoolniveau. De manier waarop de steekproef werd getrokken is ook van belang. Leerlingen werden immers geselecteerd op basis van de school waarin ze zitten. De hiërarchie in de data negeren zou kunnen leiden tot een onderschatting van de variantie waardoor we te snel zouden concluderen dat beleidsvoerend vermogen een invloed heeft (type 1 fout). Multilevelanalyses maken het mogelijk om (a) de samenhang tussen de dragers van beleidsvoerend vermogen van scholen en de vaardigheden van leerlingen na te gaan, (b) rekening te houden met de school van de leerling en (c) te controleren voor leerlingkenmerken en schoolkenmerken.

Voor de multilevelanalyses maken we gebruik van he R-pakket EdSurvey. EdSurvey (Bailey et al., 2021) is specifiek gericht op de analyse van internationaal vergelijkend onderzoek zoals TIMSS. EdSurvey maakt het mogelijk om data rechtstreeks uit de TIMSS-datafiles in te lezen en te analyseren, gebruikmakend van gewichten en plausible values. EdSurvey is ontwikkeld door AIR (American Institutes for Research) in opdracht van de National Center for Education Statistics (NCES) dat zich bevindt in de schoot van het departement onderwijs van de Verenigde Staten.

4.3 VARIABELEN

Afhankelijke variabelen

Aangezien we willen weten of het beleidsvoerend vermogen van scholen samenhangt met leerlingeprestaties, nemen we de plausible values op als afhankelijke variabelen in de multilevelmodellen. Hierbij worden de analyses steeds uitgevoerd met alle vijf de plausibe values en wordt het gemiddelde van parameterschattingen voor de vijf plausible values gerapporteerd (voor meer details zie:Fishbein et al., 2021). Omdat we eveneens willen weten of de resultaten anders zouden zijn wanneer we de vaardigheden van leerlingen schatten op basis van enkel de Vlaamse data, voeren we de analyses simultaan uit met de IRT-schattingen. Enkel wanneer de analyses op basis van de IRT-schattingen sterk afwijken van de analyses met de plausible values, wordt over de modellen met de IRT-schattingen gerapporteerd.

Voor de analyses wordt de dataset opgesplitst voor wiskunde en wetenschappen. Drie leerlingen vulden enkel het deel voor wetenschappen in één leerling enkel het deel voor wiskunde. Merk op dat TIMSS plausible values voor wiskunde en voor wetenschappen schat voor alle leerlingen, dus ook voor leerlingen die geen enkele vraag voor wiskunde of voor wetenschappen invulden. Prestaties schatten op basis van achtergrondkenmerken en itemparameters is voor het huidige onderzoek niet opportuun. De betrokken leerlingen werden verwijderd uit de respectievelijke dataframes voor wiskunde en wetenschappen.

Onafhankelijke variabelen

De onafhankelijke variabelen of voorspellers zijn in de eerste plaats de dragers van beleidsvoerend vermogen zelf. Daarnaast worden in de modellen enkele controlevariabelen opgenomen waarvan we weten dat ze mogelijk een rol spelen bij het verklaren van prestaties van leerlingen. Het betreft zowel controlevariabelen op het niveau van de leerling (geslacht, thuistaal en het aantal boeken thuis) als controlevariabelen op het niveau van de school (schoolgrootte en gemiddelde schoolSES). Deze variabelen worden opgenomen om uit te sluiten dat vastgestelde effecten niet zozeer aan een verschil in beleidsvoerend vermogen zijn toe te schrijven maar bijvoorbeeld eerder aan de samenstelling van de leerlingenpopulatie in een school. Hierna volgt een overzicht van de gebruikte onafhankelijke variabelen.

Dragers van beleidsvoerend vermogen. Als onafhankelijke variabelen worden in de eerste plaats de acht dragers van beleidsvoerend vermogen in de modellen opgenomen. Om de interpretatie van de resultaten gemakkelijker te maken, worden deze variabelen gestandaardiseerd.

Geslacht. De TIMSS-data bevatten twee variabelen voor het geslacht van de leerlingen: ITSEX en ASBG01 (leerlingenvragenlijst). Beide variabelen zijn identiek. De variabele ASBG01 komt enkel voor in de student context

data files. De variabele ITSEX wordt ook in de andere bestanden gebruikt. In de TIMSS-bestanden worden de cijfers 1 en 2 gebruikt om te verwijzen naar respectievelijk meisjes en jongens. Om de interpretatie van de resultaten gemakkelijker te maken, wordt de variabele ITSEX gehercodeerd naar 0 (meisjes) en 1 (jongens). Dat betekent dat meisjes de referentiecategorie vormen in deze analyses.

Thuistaal. De variabele thuistaal is afgeleid van de vraag uit de leerlingenvragenlijst “Hoe vaak spreek je thuis Nederlands” (ASBG03). Hierbij kregen de leerlingen vier antwoordmogelijkheden: altijd, bijna altijd, soms en nooit. Deze variabele werd omgezet naar een dummyvariabelen met twee categorieën. Leerlingen die aangaven thuis altijd of bijna altijd Nederlands te spreken vormen de referentiecategorie.

Boeken thuis. Om de sociaaleconomische status van leerlingen te benaderen, maken we gebruik van het aantal boeken een leerling thuis bezit zoals dat werd gerapporteerd door de leerlingen in de leerlingenvragenlijst (ASBG04). Het aantal boeken thuis zoals gerapporteerd door de leerlingen is volgens verschillende onderzoekers een redelijk goede indicator voor SES en hangt samen met zowel wiskundeprestaties als wetenschapsprestaties (Broer et al., 2019). Het aantal boeken een leerling thuis heeft, wordt gezien als een proxy voor de sociaaleconomische achtergrond (ASBG04). De oorspronkelijke variabele bevat vijf categorieën. Deze worden herleid naar een dummyvariabele met twee categorieën. De referentiegroep bestaat uit leerlingen met meer dan 25 boeken thuis.

Schoolgrootte. De schoolgrootte wordt geoperationaliseerd in het aantal leerlingen er op het moment van de testafname ingeschreven waren op school. Informatie over de schoolgrootte werd gehaald uit de scholenvragenlijst die werd ingevuld door de directie van de school. Ontbrekende data over de schoolgrootte werden opgevuld met de informatie uit de sample frame die werd gebruikt voor TIMSS 2019. Deze gegevens weerspiegelen de situatie tijdens het schooljaar 2017-2018 en zijn dus iets minder up-to-date.

SchoolSES. Een indeling van scholen op basis van gemiddelde schoolSES kan op twee manieren gebeuren. In de schoolvragenlijst die werd ingevuld door de directeur van de school werd gevraagd om aan te geven hoeveel procent van de leerlingen uit een economisch welgesteld gezin komen en hoeveel procent uit een economisch achtergesteld gezin. De antwoorden op beide vragen worden samengebracht tot één variabele met drie categorieën die een indicatie geven van de sociaaleconomische achtergrond van de leerlingenpopulatie: eerder welvarend, noch welvarend, noch achtergesteld, eerder achtergesteld. Een tweede mogelijkheid bestaat erin om de schoolSES te baseren op de strata waartoe de scholen behoren. Bij het trekken van de steekproef werden scholen immers niet alleen ingedeeld per provincie, maar ook volgens SES op basis van de Vlaamse indicatoren (thuistaal, opleidingsniveau van de moeder, ontvangen van een studietoelage en de buurt waarin de leerling woont). Scholen worden ingedeeld in drie categorieën: hoog, midden en laag. Uit de vergelijking van beide datasets blijkt echter dat de inschatting van de directeurs in meer dan de helft van de gevallen (n = 32) afwijkt van de indeling op basis van de Vlaamse indicatoren. Van de 61 scholen uit de dataset voor dit onderzoek waarvoor beide gegevens beschikbaar zijn, zijn er 6 scholen waarvan de directeur aangaf dat de school een eerder welgestelde populatie heeft, terwijl de school werd ingedeeld in een stratum met lage SES. 24 andere scholen worden op basis van de schoolvragenlijst één categorie hoger geklasseerd dan blijkt uit strata waartoe de scholen behoren. In twee gevallen schat de directeur de gemiddelde SES van de leerlingen lager in da de indeling op basis van de indicatoren. Zowel de antwoorden van de directeurs als de indeling op basis van de Vlaamse indicatoren zijn slechts een benaderen van de realiteit. Omdat de Vlaamse indicatoren gebaseerd zijn op objectieve criteria en de schattingen van de directeurs eerder op percepties, wordt voor dit onderzoek gebruik gemaakt van de indeling in de strata om scholen in te delen op basis van gemiddelde schoolSES. Voor Brussel en de scholen BuBaO werd geen onderscheid gemaakt op basis van SES. Voor deze scholen wordt de informatie uit de scholenvragenlijst gebruikt. Voor één school BuBaO is geen informatie over de schoolSES beschikbaar. Tabel 12 biedt een overzicht van de gebruikte variabelen en hun herkomst.

TABEL 12 OVERZICHT VAN DE VARIABELEN

	Naam	Herkomst	Soort
Afhankelijke variabelen			
Vaardigheidsscore wiskunde	VaardigWis	IRT-analyse	numeriek
Vaardigheidsscore wetenschappen	VaardigWet	IRT-analyse	numeriek
Plausible Values wiskunde (1-5)	ASMMAT01-05	asabflm7	PV
Plausible Values wetenschappen (1-5)	ASSSCI01-05	asabflm7	PV
Onafhankelijke variabelen			
Leerling			
Geslacht	Jongen	asabflm7	dummy
Meisje	*Ref.*	ITSEX	0
Jongen			1
Thuistaal	Anderstalig	asgbflm7	dummy
(Bijna) altijd Nederlands	*Ref.*	ASBG03	0
(Soms) andere taal			1
Aantal boeken thuis	WeinigBoeken	asgbflm7	dummy
> 25 (> één boeken schap)		ASBG04	0
≤ 25 boeken (≤ één boeken schap)			1
Schoolcontext			
Schoolgrootte	SCHPOP	acgbflm7 sampling frame	numeriek
Gemiddelde schoolSES	SCHSES	acgbflm7	
Hoog	SESHoog	sampling frame	1/0
Midden	SESMidden.		0/0
Laag	SESLaag		0/1
Dragers beleidsvoerend vermogen			
Doeltreffende communicatie	ComZ	BVV	*z-score*
Gedeeld leiderschap	GedZ	BVV	*z-score*
Gezamenlijke doelgerichtheid	GezZ	BVV	*z-score*
Innovatief vermogen	InnZ	BVV	*z-score*
Ondersteunende relaties	RelZ	BVV	*z-score*
Reflectief vermogen	RefZ	BVV	*z-score*
Responsief t.a.v. ouders	ResParZ	BVV	*z-score*
Responsief t.a.v. samenleving	ResSocZ	BVV	*z-score*

4.4 GEWICHTEN

TIMSS maakt gebruik van een steekproef van de populatie en elke steekproef impliceert een standaardfout. De standaardfout is een maat die aangeeft hoe accuraat de steekproevenverdeling de populatie van leerlingen beschrijft. De selectie van leerlingen voor het TIMSS-onderzoek gebeurt door een gestratificeerde steekproef (Stratified Two-Stage Cluster Sample Design). Voor Vlaanderen werden scholen ingedeeld in expliciete strata op basis van provincie, SES (hoog, midden, laag) en type (officieel en vrij, enkel Antwerpen). Alle Brusselse scholen en scholen uit het buitengewoon basisonderwijs (BuBaO) worden in één stratum ingedeeld (Martin et al., 2020, p. 9.83). Binnen elk stratum worden scholen willekeurig geselecteerd in verhouding tot het aantal leerlingen. Indien nodig worden binnen elke school willekeurig klassen geselecteerd. Alle leerlingen uit de geselecteerde klassen nemen deel aan TIMSS (voor meer informatie over het trekken van de Vlaamse steekproef zieFaddar et al., 2020).

Het TIMSS-steekproefgewicht van een leerling wordt berekend op basis van de selectiekansen op drie niveaus: de school, de klas en de leerling. Het steekprefgewicht van elk niveau is het omgekeerde van de kans dat een school, klas of leerling geselecteerd werd om deel te nemen aan het onderzoek, gecorrigeerd voor niet deelnemende leerlingen, klassen of scholen. De gewichten worden berekend op basis van het totaal aantal leerlingen in de populatie van het stratum waartoe de school behoort. De som van de gewichten van alle leerlingen uit een stratum is gelijk aan het aantal leerlingen in de populatie.[2] Het totale steekproefgewicht van een leerling (TOTWGT) is het product van de gewichten op de drie niveaus. Dit is het gewicht dat voor de meeste analyses wordt gebruikt.

Voor het onderzoek naar de samenhang tussen beleidsvoerend vermogen en leerlingenprestaties kunnen enkel de data worden gebruikt van scholen en leerlingen die deelnamen aan het onderzoek rond beleidsvoerend vermogen. In de praktijk betekent dit dat we werken met een steekproef van een steekproef. Hierdoor weerspiegelen de TIMSS-gewichten niet meer de kansen dat scholen en hun leerlingen worden geselecteerd en is de som van de gewichten van de deelnemers niet mee gelijk aan de totale populatie. Daarom dienen de gewichten opnieuw berekend te worden.

Bij het herberekenen van de schoolgewichten vertrekken we van het totale gewicht van alle leerlingen van de respectievelijke scholen (STOTWGTU) uit de TIMSS-database. Aangezien de schoolgewichten worden berekend per stratum, worden de scholen opnieuw ingedeeld in hun respectievelijke strata. Dit is belangrijk omdat niet elk stratum evenveel leerlingen in de populatie vertegenwoordigt.

Stap 1: Per stratum (i) wordt voor elke school (ij) de proportie van het totaalgewicht berekend.

$$Prop_BVV_{ij} = \frac{STOTWGTU_{ij}}{\sum_{i=1}^{N} STOTWGTU_i} * 100$$

Stap 2: Vervolgens wordt per stratum (i) een factor berekend waarmee de schoolgewichten van de overblijvende scholen moeten worden vermenigvuldigd na verwijdering van de scholen die niet deelnamen aan het onderzoek BVV.

$$Factor_i = \frac{100}{\sum_{i=1}^{n} Prop_BVV_i}$$

Waarbij $\sum_{i=1}^{n} Prop_BVV$ de som is van de proporties van de scholen die deelnamen aan het onderzoek BVV.

Stap 3: Voor elke school (ij) wordt het totale schoolgewicht vermenigvuldigd met deze factor

$$STOTWGT_BVV_{ij} = STOTWGT_{ij} * Factor_i$$

Stap 4: Dit nieuwe schoolgewicht wordt vervolgens gedeeld door het aantal deelnemende leerlingen van de respectievelijke scholen.

$$SCHWGT_BVV_{ij} = \frac{STOTWGT_BVV_{ij}}{Aantal\ leerlingen_i}$$

Hierbij kennen we alle leerlingen (k) van een school (ij) eenzelfde gewicht toe.

$$TOTWGT_BVV_{ijk} = SCHWGT_BVV_{ij}$$

De som van alle leerlinggewichten uit de subsample voor dit onderzoek is opnieuw (bij benadering) gelijk aan de totale populatie van Vlaamse leerlingen uit het 4de leerjaar.

[2] In de praktijk wijkt de som van de gewichten af van de totale populatie. De informatie over de schoolgrootte, geoperationaliseerd in het aantal leerlingen in het vierde leerjaar, werd verzameld tijdens het schooljaar voor de toetsafnames. Hierdoor zitten er kleine afwijkingen op de som van de leerlinggewichten.

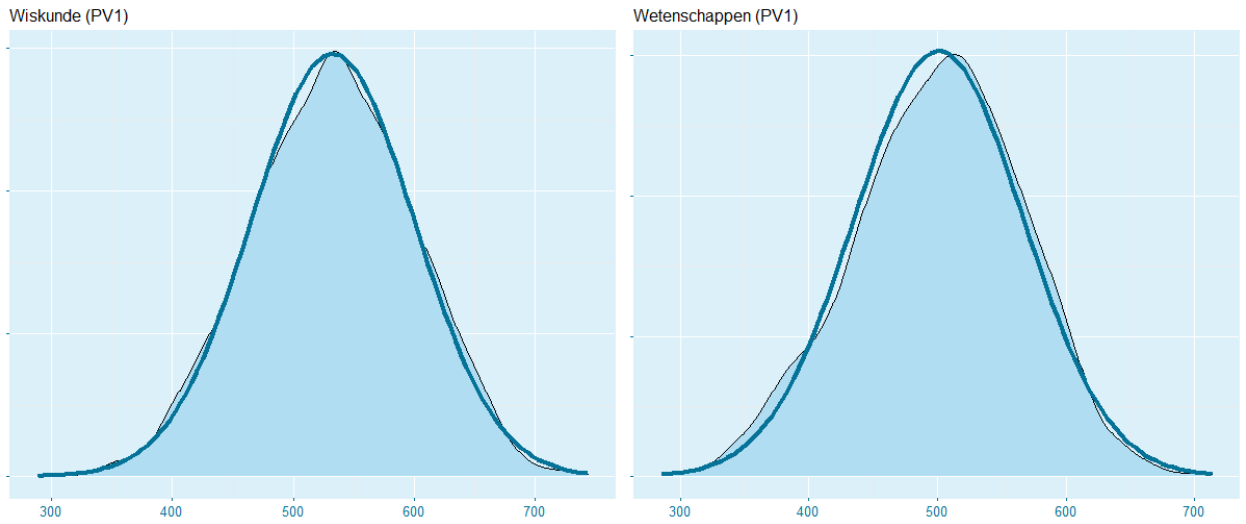
4.5 BESCHRIJVENDE STATISTIEKEN

Vaardigheden van leerlingen

De leerlingen uit de subsample voor het onderzoek naar de samenhang tussen beleidsvoerend vermogen en leerlingenprestaties scoorden gemiddeld 533.18 (*SD* = 66.72*)* voor wiskunde en 500.48 (*SD* = 66.29) voor wetenschappen. Het betreft een gewogen gemiddelde over de vijf plausible values heen waarbij gebruik werd gemaakt van de herberekende gewichten voor deze subsample.

De geschatte vaardigheden voor wiskunde en wetenschappen zijn normaal verdeeld. Figuur 4 toont de verdeling van de prestaties van leerlingen voor wiskunde en wetenschappen (1ste plausible value).

FIGUUR 4 VERDELING VAN DE VAARDIGHEIDSSCORES WISKUNDE EN WETENSCHAPPEN (PV1)



Dragers van beleidsvoerend vermogen

De beschrijvende statistieken van de dragers van beleidsvoerend vermogen werden toegelicht in deel 2.

Controlevariabelen

Tabel 13 geeft een overzicht van de gewogen (TOTWGT_BVV) gemiddelde prestaties van leerlingen binnen elke categorie van de categorische controlevariabelen. Hieruit blijkt dat de gemiddelde prestaties tussen de categorieën verschilt. Dit geeft een eerste indicatie dat het opnemen van de controlevariabelen in de modellen zinvol kan zijn.

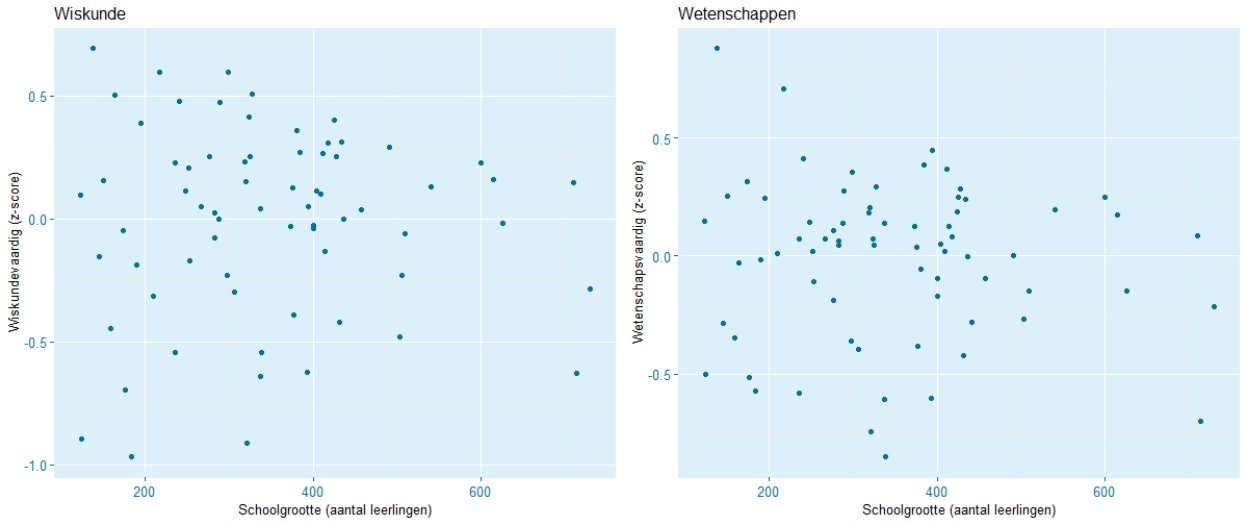
TABEL 13 OVERZICHT VAN HET AANTAL LEERLINGEN PER CATEGORIE EN DE GEMIDDELDE PRESTATIES (GEWOGEN)

Variabele	Categorieën	n_{Wis}	n_{Wet}	NA	M_{Wis}	SD_{Wis}	M_{Wet}	SD_{Wet}
Geslacht	meisje	1234	1236	0	527.50	67.31	497.24	66.39
	jongen	1158	1158		539.37	65.54	504.02	66.02
Thuistaal	(Bijna) altijd Nederlands	1801	1802	9	542.39	64.68	512.00	61.38
	Soms/altijd andere taal	576	577		505.24	64.55	465.29	68.08
Boeken thuis	> 25 boeken thuis	1459	1459	9	548.81	63.16	519.05	59.90
	≤ 25 boeken thuis	908	910		507.58	64.43	470.20	64.92
SchoolSES	SESHoog	801	801	20	545.99	61.05	514.44	56.89
	SESMidden	915	915		535.88	66.84	501.52	65.80
	SESLaag	656	658		519.17	68.57	487.04	71.88

Gewogen gemiddelde en SD *op basis van de 5 plausible values.*

De scholen uit de subsample telden gemiddeld 382.2 leerlingen (*SD* = 137.83). De kleinste school telde 122 leerlingen, de grootste 730. Figuur 5 toont de samenhang tussen de schoolgrootte en de gemiddelde wiskunde- en wetenschapsprestaties van leerlingen. Elk stipje in de grafiek staat voor een andere school. Hieruit blijkt dat er geen samenhang is tussen de schoolgrootte en de wiskundevaardigheden van leerlingen (r = -0.001, p = 0.99). Voor wetenschappen lijkt er een lichte negatieve samenhang, maar deze is niet significant (r = -0.078, p = 0.52).

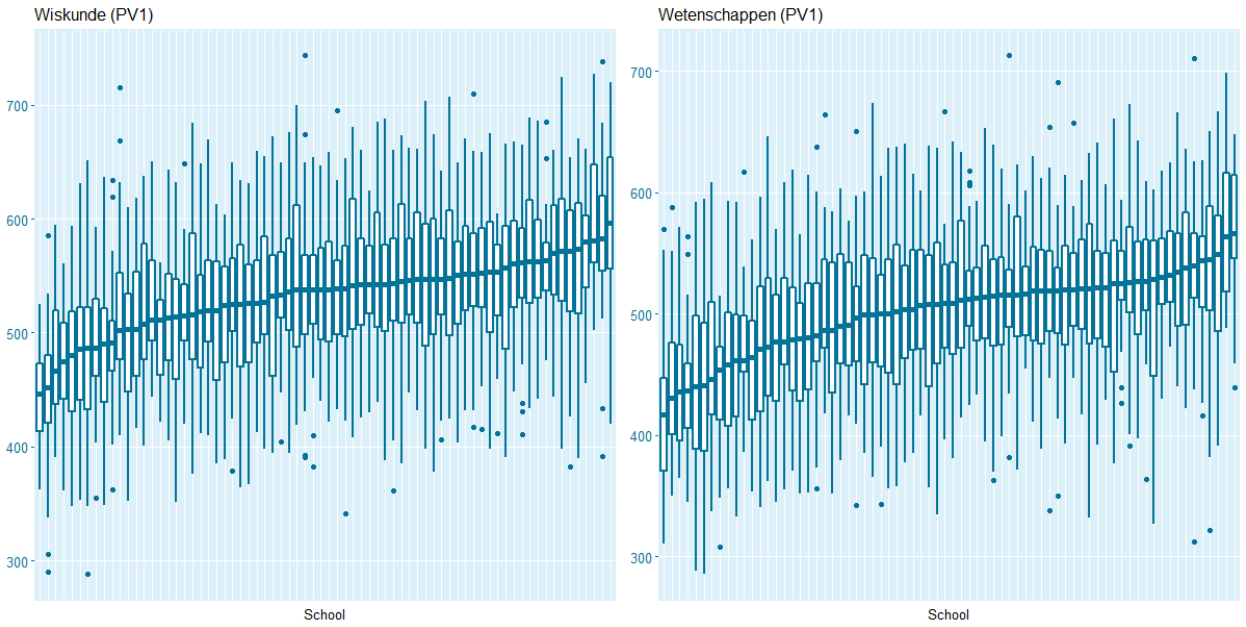
FIGUUR 5 SAMENHANG TUSSEN SCHOOLGROOTTE EN VAARDIGHEDEN VAN LEERLINGEN (IRT)



Verschillen tussen en binnen scholen

Ten slotte bekijken we de spreiding van de vaardigheden van leerlingen binnen elke school. Uit Figuur 6 blijkt dat de vaardigheden van leerlingen zowel tussen als binnen scholen sterk verschillen. Elk staafje toont de spreiding van de vaardigheden van de leerlingen van een school (boxplots). De donkerblauwe stip in het midden van elk staafje is de mediaan. De bolletjes boven en onder de verticale lijntjes zijn de outliers. Het betreft leerlingen die lager scoren dan het eerste kwartiel min anderhalve keer de interkwartielafstand of hoger dan het derde kwartiel plus anderhalve keer de interkwartielafstand.

FIGUUR 6 SPREIDING VAN DE VAARDIGHEDEN VAN LEERLINGEN PER SCHOOL



4.6 MODELLEN

Modelopbouw

Bij het opbouwen van een model gaan we op zoek naar een spaarzaam statistisch model dat de realiteit zo goed mogelijk benadert (Hesser, 2015). Multilevelanalyse start meestal met een nulmodel of *unconditional means model* zonder voorspellers (Model 0) (Roback & Legler, 2021). Dit nulmodel maakt het mogelijk om het aandeel van de variantie te berekenen dat is toe te schrijven aan het niveau van de school. Vervolgens voegen we de controlevariabelen toe op het niveau van de leerling en het niveau van de school (Model 1a). Dit model laat toe om na te gaan welke controlevariabelen significant samenhangen met de prestaties van leerlingen. Met de significante controlevariabelen wordt een basismodel gecreëerd (Model 1b). Dit model dient als referentiepunt voor de modelvergelijking. Om na te gaan welke dragers van beleidsvoerend vermogen samenhangen met de vaardigheden van leerlingen, na controle voor de leerling- en schoolkenmerken, wordt een model geschat met elke drager afzonderlijk (Model 2a-h). Aan deze modellen worden telkens de significante leerling- en schoolkenmerken uit Model 1 toegevoegd. Ten slotte worden de significante dragers samen met de controlevariabelen in één model samengebracht. Tabel 14 geeft een overzicht van de opeenvolgende modellen en de respectievelijke variabelen die worden toegevoegd aan het model.

TABEL 14 OVERZICHT VAN DE MODELLEN

Model	Omschrijving	Fixed effects	Variabelen
Model 0	Nulmodel	*NVT*	*NVT*
Model 1a	Potentiële controlevariabelen	Geslacht Thuistaal Boeken thuis Schoolgrootte SchoolSES hoog SchoolSES laag	ITSEX ASBG03 ASBG04 POPSCH ACBG03
Model 1b	Weerhouden controlevariabelen	Geslacht Thuistaal Boeken thuis SchoolSES laag	ITSEX ASBG03 ASBG04 ACBG03
Model 2a-h	Dragers beleidsvoerend vermogen *Telkens met de significante controlevariabelen uit Model 1 en Model 2*	a. Doeltreffende communicatie b. Gedeeld leiderschap c. Gezamenlijke doelgerichtheid d. Innovatief vermogen e. Ondersteunende relaties f. Reflectief vermogen g. Responsief t.a.v. ouders h. Responsief t.a.v. samenleving	ComZ GedZ GezZ InnZ RelZ RefZ ResParZ ResSocZ
Model 3	Volledig model	Alle significante dragers van beleidsvoerend vermogen Significante controlevariabelen	

Modelvergelijking

Modellen met plausible values kunnen niet met elkaar vergeleken worden op basis van likelihood tests (Bailey et al., 2021). Dat betekent dat er geen gebruik gemaakt kan worden van de -2loglikelihood (-2LL) of deviantie en het daarvan afgeleide Akaike's Information Criterion (AIC) om na te gaan of de complexere modellen beter bij de data passen dat de eenvoudigere. Bovendien bevat het R-pakket EdSurvey geen functies om te controleren of aan de assumpties voor regressieanalyse is voldaan. Om deze probleem te omzeilen maken we voor de modelvergelijking gebruik van de eerste plausible value en de functie lmer() uit het pakket lme4 (Bates et al., 2015) in de fase van de modelopbouw en het pakket performance (Lüdecke et al., 2021) om de kwaliteit van de modellen te beoordelen en de assumpties na te gaan. Hierbij wordt geen gebruik gemaakt van gewichten. Voor de modelopbouw is dat op zich niet zo'n probleem. Bij de eigenlijke parameterschattingen worden wel de vijf plausible values en gewichten gebruikt.

De modellen waarbij telkens één drager van beleidsvoerend vermogen wordt toegevoegd, zijn niet beter dan het basismodel met enkel de significante controlevariabelen (zie Tabel 15). De AIC daalt voor geen enkele drager significant tegenover het basismodel met controlevariabelen. Voor de meeste dragers van beleidsvoerend vermogen stijgt de AIC en de r^2 stijgt niet of amper. Dit betekent dat de modellen met de dragers van beleidsvoerend vermogen niet meer variantie verklaren dan het basismodel met enkel de significante controlevariabelen.

De modellen op basis van de eerste plausible value tonen, na controle voor het geslacht, de thuistaal, het aantal boeken thuis (en voor de modellen wiskunde een lage schoolSES), geen enkel significant verband tussen de dragers van beleidsvoerend vermogen en de wiskundeprestaties van leerlingen. We kunnen dan ook geen model samenstellen met alle significante dragers.

TABEL 15 MODELVERGELIJKING

	AIC	*BIC*	*LogLik*	*ICC*	R^2_{cond}	R^2_{marg}
Wiskunde						
Model 0	26379	26396	-13189	0.149	0.149	*NVT*
Model 1a	26142	26194	-13072	0.092	0.201	0.120
Model 1b	26139	26179	-13073	0.091	0.196	0.116
Model 2						
Doeltr. communicatie	26137	26183	-13073	0.092	0.197	0.115
Gedeeld leiderschap	26136	26182	-13072	0.091	0.196	0.116
Gez. doelgerichtheid	26137	26183	-13073	0.092	0.196	0.115
Innovatief vermogen	26134	26180	-13071	0.087	0.196	0.119
Ondersteunende relaties	26137	26183	-13073	0.092	0.196	0.115
Reflectief vermogen	26135	26181	-13072	0.089	0.197	0.119
Resp. t.a.v. ouders	26137	26183	-13072	0.091	0.196	0.116
Resp. t.a.v. samenleving	26134	26180	-13071	0.088	0.197	0.120
Model 3	-	-	-	-	-	-
Wetenschappen						
Model 0	26242	26259	-13120	0.176	0.176	*NVT*
Model 1a	25900	25952	-12955	0.106	0.250	0.161
Model 1b	25907	25941	-12951	0.114	0.233	0.135
Model 2						
Doeltr. communicatie	25905	25945	-12955	0.115	0.234	0.134
Gedeeld leiderschap	25903	25944	-12954	0.113	0.235	0.137
Gez. doelgerichtheid	25905	25945	-12955	0.115	0.234	0.134
Innovatief vermogen	25902	25942	-12953	0.109	0.235	0.141
Ondersteunende relaties	25904	25945	-12955	0.114	0.234	0.135
Reflectief vermogen	25903	25943	-12954	0.111	0.235	0.140
Resp. t.a.v. ouders	25905	25945	-12955	0.115	0.234	0.135
Resp. t.a.v. samenleving	25901	25942	-12953	0.110	0.236	0.142
Model 3	-	-	-	-	-	-

Assumpties

Vervolgens controleren we of het uiteindelijke model voldoet aan de assumpties (Burnham & Anderson, 2002). Hiermee wordt nagegaan of het model wel een goede benadering is van de realiteit. De modellen voldoen in grote lijnen aan de assumpties. De errortermen zijn normaal verdeeld, er zijn geen outliers en er is geen sprake van multicollineariteit. Voor alle modellen wordt wel een zekere mate van heteroscedasticiteit vastgesteld. Dat wil zeggen dat de variantie van de residuen verschilt al naargelang de onafhankelijke variabelen. Op zich zorgt heteroscedasticiteit niet voor foute schattingen, maar wel voor mogelijke vertekeningen bij de significantietoetsen.

4.7 RESULTATEN

De hier gepresenteerde resultaten zijn gebaseerd op schattingen van de vijf plausible values, rekening houdend met de herberekende gewichten. De ICC van het nulmodel voor wiskunde bedraagt 0.15. Dat betekent dat 15 % van de variantie in scores verklaard wordt door verschillen tussen scholen. Tabel 16a toont de parameterschattingen voor de modellen met betrekking tot wiskunde voor elke drager van beleidsvoerend vermogen. Geen enkele drager van beleidsvoerend vermogen hangt significant samen met de wiskundeprestaties van leerlingen. Leerlingkenmerken en

een lage schoolSES doen er wel toe. Voor een leerling uit de referentiecategorie (een meisje dat thuis (bijna) altijd Nederlands spreekt, thuis meer dan 25 boeken heeft en naar een school gaat met gemiddelde of hoge schoolSES) verwachten we dat ze ca. 549 scoort voor wiskunde (intercept). Over alle modellen heen scoren jongens voor wiskunde ongeveer 13 punten meer dan meisjes. Leerlingen die in de vragenlijst aangaven thuis vaak of altijd een andere taal te spreken, scoren gemiddeld bijna 20 punten minder dan hun klasgenoten die thuis (bijna) altijd Nederlands spreken. Het aantal boeken de leerlingen thuis hebben hangt het sterkst samen met wiskundeprestaties. Leerlingen die aangeven dat ze thuis hooguit 25 boeken hebben, scoren ongeveer 32.5 lager op wiskunde. Ook leerlingen uit scholen met een gemiddeld lage schoolSES scoren beduidend lager (ca. -13) dan hun jaargenoten uit scholen met een gemiddelde of hoge schoolSES.

TABEL 16A PARAMETERS VOOR DE MODELLEN VOOR ELKE DRAGER VAN BELEIDSVOEREND VERMOGEN (WISKUNDE)

	Com		Ged		Gez		Inn	
Random effects	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
School	332.23	18.23	324.92	18.03	333.93	18.27	312.90	17.74
Resterend	3519.79	59.33	3519.91	59.33	3519.84	59.33	3520.97	59.34
Fixed effects	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Intercept	549.02	3.39	548.80	3.42	549.13	3.38	549.31	3.33
Drager (kolomnaam)								
Jongen	13.31	3.62	13.31	3.62	13.33	3.62	13.28	3.62
Anderstalig	- 19.68	4.11	-19.65	4.12	-19.64	4.13	-19.79	4.15
Weinig boeken thuis	-32.46	3.95	-32.52	3.95	-32.45	3.95	-32.51	3.95
Lage schoolSES	-13.32	6.44	-12.56	6.38	-13.59	6.48	549.31	3.33
	Rel		Ref		ResPar		ResSoc	
Random effects	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
School	332.38	18.23	320.59	17.94	331.91	18.22	317.79	17.83
Resterend	3519.94	59.33	3520.31	59.33	3519.98	59.33	3519.77	59.33
Fixed effects	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Intercept	549.02	3.39	549.14	3.37	549.20	3.37	548.99	3.35
Drager (kolomnaam)								
Jongen	13.31	3.62	13.29	3.62	13.34	3.62	13.28	3.62
Anderstalig	-19.70	4.10	-19.71	4.14	-19.67	4.14	-19.70	4.16
Weinig boeken thuis	-32.44	3.95	-32.43	3.95	-32.44	3.95	-32.41	3.95
Lage schoolSES	-13.02	6.34	-13.67	6.46	-13.67	6.38	-13.32	6.42

Enkel significante effecten worden opgenomen ($t > |1.96|$)

Wanneer we de analyses uitvoeren met de geschatte wiskundevaardigheden op basis van de IRT-analyse met de Vlaamse data, krijgen we gelijkaardige resultaten. De ICC van het nulmodel is een beetje lager (0.12). Bij deze modellen is een lage schoolSES niet significant. Opmerkelijk is dat bij deze schattingen het innovatieve vermogen van scholen lichtjes negatief samenhangt met leerlingenprestaties. (β_{Inn} = -0.06 (z-score), SE = 0.03, t = -2.25). Dit effect is echter verwaarloosbaar klein en is weinig informatief.

Voor wetenschappen ziet het plaatje er gelijkaardig uit. De ICC van het nulmodel voor wetenschappen bedraagt 0.18. Dat betekent dat 18 % van de variantie in scores verklaard wordt door verschillen tussen scholen. Tabel 16b toont de parameterschattingen voor de modellen met betrekking tot wiskunde voor elke drager van beleidsvoerend vermogen. Ook voor wetenschappen hangt geen enkele drager van beleidsvoerend vermogen significant samen met de prestaties van leerlingen. Van een leerling uit de referentiecategorie verwachten we een wetenschapsscore van ca. 517.7 (intercept). Voor wetenschappen scoren jongens opnieuw hoger dan meisjes (β_{Jongen} = ca. 8 punten). Ook voor wetenschappen is het effect van een andere thuistaal of maximaal 25 boeken thuis significant. Kinderen die aangeven dat ze thuis maximaal 25 boeken hebben, scoren ongeveer 38 punten lager dan de referentiegroep en kinderen die aangeven thuis regelmatig een andere taal te spreken scoren 27 punten minder.

TABEL 16B PARAMETERS VOOR DE MODELLEN VOOR ELKE DRAGER VAN BELEIDSVOEREND VERMOGEN (WETENSCHAPPEN)

	Com		Ged		Gez		Inn	
Random effects	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
School	440.09	20.98	428.63	20.70	442.33	21.03	425.30	20.62
Resterend	3185.12	56.44	3185.10	56.44	3185.12	56.44	185.67	56.44
Fixed effects	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Intercept	517.69	3.49	517.73	3.49	517.72	3.49	517.77	3.44
Drager (kolomnaam)								
Jongen	8.32	3.70	8.31	3.70	8.34	3.71	8.30	3.70
Anderstalig	-27.23	3.70	-27.19	3.70	-27.20	3.69	-27.35	3.70
Weinig boeken thuis	-38.37	3.63	-38.42	3.63	-38.36	3.63	-38.43	3.63
	Rel		Ref		ResPar		ResSoc	
Random effects	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
School	440.97	20.98	430.69	20.77	442.02	21.02	423.01	20.57
Resterend	3185.23	56.44	3185.35	56.44	3185.15	56.44	3185.03	56.44
Fixed effects	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Intercept	517.76	3.50	517.70	3.47	517.73	3.50	517.64	3.40
Drager (kolomnaam)								
Jongen	8.33	3.70	8.30	3.70	8.34	3.70	8.30	3.70
Anderstalig	-27.22	3.70	-27.28	3.70	-27.20	3.70	-27.29	3.72
Weinig boeken thuis	-38.36	3.63	-38.37	3.63	-38.37	3.63	-38.36	3.63

Enkel significante effecten worden opgenomen

Wanneer we de analyses uitvoeren met de geschatte wetenschapsvaardigheden op basis van de IRT-analyse met de Vlaamse data, krijgen we gelijkaardige resultaten. De ICC van het nulmodel voor wetenschappen is lager (0.11). Net als voor wiskunde is er een beperkte negatieve samenhang tussen het innovatief vermogen van scholen en de wetenschapsprestaties van leerlingen (β_{Inn} = -0.05 (z-score), *SE* = 0.02, *t* = -2.12). Hetzelfde geldt voor reflectief vermogen van scholen (β_{Ref} = -0.05 (z-score), *SE* = 0.02, *t* = -2.22) en de responsiviteit van scholen tegenover de samenleving (β_{ResSoc} = -0.06 (z-score), *SE* = 0.02, *t* = -2.12). Dit effect is echter verwaarloosbaar klein en is weinig informatief.

5 CONCLUSIES

Met dit onderzoek konden we geen significant effect vaststellen tussen het beleidsvoerend vermogen van scholen en de prestaties wiskunde en wetenschappen van leerlingen zoals deze werden gemeten tijdens het TIMSS-onderzoek. Leerlingenkenmerken zoals het geslacht, de thuistaal en sociaaleconomische status (gemeten door het aantal boeken thuis) hebben wel een significant effect. De nulhypothese (beleidsvoerend vermogen heeft *geen* effect) kan op basis van de onderzoeksresultaten niet worden verworpen. De bevinding dat het gepercipieerde beleidsvoerend vermogen van scholen niet significant samenhangen met leerlingenprestaties betekent nog niet dat het beleidsvoerend vermogen van scholen er niet toe doet. Hieronder formuleren we enkele mogelijke verklaringen en enkele aanbevelingen die voortvloeien uit het onderzoek.

5.1 DUIDING BIJ DE RESULTATEN

VERKLARING 1. LEERLINGENKENMERKEN ZIJN EEN BELANGRIJKERE VOORSPELLER VOOR PRESTATIES DAN SCHOOLKENMERKEN

Leerlingenkenmerken zoals het geslacht, de thuistaal en het aantal boeken thuis spelen een belangrijkere rol dan schoolkenmerken zoals de schoolomvang, gemiddelde schoolSES of het beleidsvoerend vermogen van scholen. Enkel voor scholen met een gemiddeld lage schoolSES werd een significant negatief effect vastgesteld voor wiskunde. Dat leerlingenkenmerken een grotere impact hebben dan schoolkenmerken kon worden verwacht aangezien leerlingenkenmerken een directie invloed uitoefenen op de prestaties en de invloed van schoolkenmerken indirect is.

VERKLARING 2. DE RELATIE TUSSEN BELEIDSVOEREND VERMOGEN VAN SCHOLEN EN LEERLINGENPRESTATIES IS INDIRECT

Het effect van beleidsvoerend vermogen van scholen op leerlingenprestaties is indirect. De exacte samenhang tussen beleidsvoerend vermogen van scholen en prestaties van leerlingen blijft tot op heden onduidelijk. De Maeyer et al. (2007) gaan ervan uit dat het beleidsvoerend vermogen een invloed heeft het schoolklimaat en dat dit schoolklimaat op zijn beurt de prestaties van leerlingen kan beïnvloeden. Leithwood et al. (2020) wijzen op 11 mogelijke tussenliggende variabelen, waaronder organisatorische, relationele en emotionele. Door het veronderstelde indirecte verband tussen beleidsvoerend vermogen van scholen en de prestaties van leerlingen is het ook moeilijker te meten. Onderzoek naar de samenhang tussen intermediaire variabelen die samenhangen met enerzijds beleidsvoerend vermogen en anderzijds leerlingenprestaties kan hier meer licht op werpen.

VERKLARING 3. HET GEVOERDE BELEID IS BELANGRIJKER DAN HET BELEIDSVOEREND VERMOGEN

Er bestaat mogelijk een discrepantie tussen het beleidsvoerend vermogen van scholen en het effectief gevoerde beleid. Het is niet omdat in scholen een sterk beleidsvoerend vermogen aanwezig is, dat dit ook ten volle wordt benut. Enkel wanneer het beleidsvoerend vermogen ook effectief wordt vertaald in een sterk onderwijskundig beleid, kunnen we verwachten dat het beleidsvoerend vermogen indirect een effect heeft op de prestaties van leerlingen. We kunnen beleidsvoerend vermogen van scholen weliswaar beschouwen als een conditio sine qua non voor een doeltreffend schoolbeleid, maar er niet vanuit gaan dat de aanwezigheid van een sterk beleidsvoerend vermogen ook automatisch leidt tot een gepast beleid en betere prestaties van leerlingen. Daarvoor zijn ook andere elementen noodzakelijk zoals bijvoorbeeld een breed gedragen academisch optimisme (de opvatting dat de school een verschil kan maken op basis van doelmatigheid, vertrouwen en een academische focus) en voldoende kennis en inzicht in leerprocessen van kinderen en effectieve leeractiviteiten en -omgeving om de juiste prioriteiten te stellen.

VERKLARING 4. DE AUTONOMIE VAN VLAAMSE LEERKRACHTEN

Uit het onderzoek van Boderé et al. (2018) blijkt de grote hang naar autonomie van Vlaamse leerkrachten. De leerkrachten die deelnamen aan dat onderzoek gaven aan dat wat er in de klas gebeurt hun verantwoordelijkheid is. Dit sluit niet uit dat er een zekere mate van sturing nodig en wenselijk is. Maar zeker op het vlak van didactische expertise wordt een te strenge sturing door leerkrachten afgewezen. Dit maakt meteen het spanningsveld duidelijk tussen beleid (sturing) en autonomie. Beleid kan pas effectief zijn wanneer het vertaald wordt naar de praktijk.

VERKLARING 5. HET METEN VAN BELEIDSVOEREND VERMOGEN IS MOEILIJK

Om het beleidsvoerend vermogen van scholen te meten, werd gebruik gemaakt van een vragenlijst die werd ingevuld door de leden van het pedagogische team van de deelnemende scholen. Vragenlijsten peilen naar percepties van respondenten. Deze percepties worden niet alleen gevormd door het 'werkelijke' beleidsvoerend vermogen van scholen, maar ook door individuele ervaringen en overtuigingen. Bovendien wordt van respondenten verwacht at ze de vragen lezen, interpreteren, informatie oproepen, een oordeel vellen en het antwoord aanduiden. In elke stap van dit cognitieve proces kan ruis optreden. Respondenten kunnen vragen anders interpreteren dan bedoeld, hebben mogelijk geen zicht op het totale plaatje, dichten recente gebeurtenissen meer belang toe, neigen naar sociaal wenselijke antwoorden, vermijden de extremen of zoeken deze juist op etc. Uit het valideringsonderzoek bleek dat de schalen (dragers) voor beleidsvoerend vermogen voldoende consistent zijn en dat de antwoorden van respondenten binnen eenzelfde school voldoende overeenstemming vertonen om uitspraken te kunnen over het beleidsvoerend vermogen van scholen. Maar percepties geven slechts een indicatie van het beleidsvoerend vermogen van scholen. Verder onderzoek is nodig om de precieze relatie te achterhalen tussen gepercipieerd beleidsvoerend vermogen en het 'ware' beleidsvoerend vermogen van scholen.

5.2 AANBEVELINGEN

Als sluitstuk van dit onderzoeksrapport formuleren we aanbevelingen die voortvloeien uit de vaststellingen die we deden tijdens het onderzoek.

AANBEVELING 1. STIMULEER DE DIALOOG OVER BELEIDSVOEREND VERMOGEN BINNEN SCHOLEN

Tijdens de validering van de vragenlijst stelden we vast dat er in sommige scholen grote verschillen bestaan in de percepties van het beleidsvoerend vermogen tussen personeelsleden met een verschillende functie binnen de school. Directeurs schatten de effectiviteit van de communicatie en de mate van gedeeld leiderschap significant hoger in dan leerkrachten. Deze twee dragers behoren tot de kern van beleidsvoerend vermogen (Van Petegem et al., 2004) en kunnen sterk worden beïnvloed door het leiderschap van de directeur. Devos et al. (2013) toonden aan dat wanneer directeurs hun leiderschapskwaliteiten hoger inschatten dan hun personeelsleden, dit een negatieve invloed heeft op gedeelde besluitvorming, toewijding van leerkrachten en de mate waarin hervormingen worden doorgevoerd. Het is dan ook belangrijk dat schooldirecteurs zich bewust zijn van de verschillen in percepties over het beleidsvoerend vermogen van hun school. De vragenlijst beleidsvoerend vermogen werd ontwikkeld als zelfevaluatie-instrument. Door het beleidsvoerend vermogen van scholen in kaart te brengen en directies te stimuleren hierover in dialoog te gaan met het schoolteam, kunnen sterktes en werkpunten worden vastgesteld en aangepakt. Gezien de significante verschillen in de percepties van directeurs en leerkrachten over het beleidsvoerend vermogen van scholen lijkt het aangewezen om het gebruik van de vragenlijst als intern evaluatie-instrument verder aan te moedigen.

AANBEVELING 2. VERSTERK DE INTERNE (FORMELE EN INFORMELE) COMMUNICATIE

Van Petegem et al. (2004) beschouwen effectieve communicatie, samen met effectief en gedeeld leiderschap en ondersteunende relaties, als kernfactoren en noodzakelijke voorwaarden voor een beleidseffectieve school. De huidige analyses tonen aan dat doeltreffende communicatie inderdaad sterk samenhangt met de andere dragers van beleidsvoerend vermogen. Een goede communicatie kan worden beschouwd als een fundament voor een effectief, gedeeld en geïntegreerd schoolbeleid. Toch zien we dat de communicatie in een aantal scholen schort. 16 % van de respondenten geeft een gemiddelde score lager dan 3 (oneens of helemaal oneens) op de positief geformuleerde stellingen die te maken hebben met de doeltreffendheid van de communicatie op hun school. Voor een kleine 10 % van de scholen is de gemiddelde schoolscore over alle respondenten en stellingen over doeltreffende communicatie heen lager dan 3. Het versterken van de interne (formele en informele) communicatie kan de beleidseffectiviteit van scholen ten goede komen.

AANBEVELING 3. LAAT HET HELE SCHOOLTEAM PARTICIPEREN AAN HET SCHOOLBELEID

Beleid is een gedeelde verantwoordelijkheid van het hele schoolteam. Net als communicatie wordt gedeeld leiderschap beschouwd als een kernfactor van een beleidseffectieve school (Van Petegem et al., 2004). De aanwezigheid van beleidsvoerend vermogen betekent nog niet dat dat beleidsvoerend vermogen ook ten volle

wordt benut. Om beleidsvoerend vermogen om te zetten naar effectief beleid moeten de leden van het schoolteam bereid en bekwaam zijn om aan het beleid te participeren en dat beleid uit te voeren. Leerkrachten meer laten participeren aan het schoolbeleid kan hun bereidheid vergroten om dit beleid in de praktijk om te zetten. In die optiek is het mogelijk zinvol om te investeren in professionaliseringsinitiatieven die de participatie van leerkrachten te versterken.

AANBEVELING 4. STIMULEER ONDERZOEK NAAR INTERMEDIAIRE FACTOREN TUSSEN BELEIDSVOEREND VERMOGEN EN LEERLINGENPRESTATIES.

Met dit onderzoek konden we geen significant effect vaststellen tussen beleidsvoerend vermogen van scholen en leerlingenprestaties. Dit betekent niet noodzakelijk dat er geen invloed is. De relatie tussen beleidsvoerend vermogen van scholen en leerlingenprestaties is indirect. Onderzoek naar de samenhang tussen intermediaire variabelen en leerlingenprestaties kan scholen helpen om bij hun beleidsbeslissingen de aandacht te richten op factoren die er het meest toe doen. Hiervoor kan mogelijk gebruik gemaakt worden van de beschikbare data van de onderwijsinspectie, pedagogische begeleidingsdiensten en CLB's.

AANBEVELING 5. MOEDIG SCHOLEN AAN OM DEEL TE NEMEN AAN GROOTSCHALIG ONDERZOEK

De bereidheid van Vlaamse scholen om deel te nemen aan grootschalig onderzoek is in vergelijking tot andere landen beperkt. Slechts 66 % van de scholen uit de oorspronkelijke steekproef was bereid deel te nemen aan TIMSS (Martin et al., 2020). Hiermee laat Vlaanderen enkel Nederland achter zich. De meeste scholen die niet wensten deel te nemen konden vervangen worden door de eerste of tweede vervangschool. Daarmee haalde Vlaanderen net de opgelegde participatiegraad voor TIMSS. Van de scholen die deelnamen aan TIMSS nam de helft ook deel aan het bijkomende onderzoek naar beleidsvoerend vermogen van scholen. Deelnemen aan wetenschappelijk onderzoek wordt door sommige scholen als belastend ervaren. Voor onderzoekers en beleidsmakers is de bereidheid van scholen om deel te nemen aan onderzoek cruciaal om meer inzicht te verwerven in de onderwijsrealiteit in Vlaanderen. Daarom is het wenselijk dat scholen worden aangemoedigd om deel te nemen aan onderzoek.

REFERENTIES

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., ... Christensen, A. A. (2021). *EdSurvey: Analysis of NCES education survey and assessment data. R package version 2.6.9*. https://CRAN.R-project.org/package=EdSurvey

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Biemann, T., Cole, M. S., & Voelpel, S. (2012). Within-group agreement: On the use (and misuse) of rWG and rWG(J) in leadership research and some best practice guidelines. *The Leadership Quarterly, 23*(1), 66-80. https://doi.org/10.1016/j.leaqua.2011.11.006

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions.* (pp. 349-381). San Francisco, US: Jossey-Bass.

Boderé, A., Van Gasse, R., Sassenus, S., & Van Petegem, P. (2018). *De leraar in de 21ste eeuw: Een kwalitatief onderzoek naar de bereidheid van leraren om de verwachtingen van de 21ste eeuw te realiseren*

Broer, M., Bai, Y., & Fonseca, F. (2019). Methodology: Constructing a Socioeconomic Index for TIMSS Trend Analyses. In *Socioeconomic Inequality and Educational Outcomes* (Vol. 5). Verenigde Staten, Cham: Springer. https://doi.org/10.1007/978-3-030-11991-1_3.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference. A practical information-theoretic approach*. Springer. https://doi.org/10.1007/b97636

Chang, I. H. (2011). A study of the relationships between distributed leadership, teacher academic optimism and student achievement in Taiwanese elementary schools. *School Leadership & Management, 31*(5), 491-515. https://doi.org/10.1080/13632434.2011.614945

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2 ed.). New York, United States: Lawrence Erlbaum Associates.

De Maeyer, S., Rymenans, R., Van Petegem, P., van den Bergh, H., & Rijlaarsdam, G. (2007). Educational leadership and pupil achievement: The choice of a valid conceptual model to test effectsin school effectiveness research. *School Effectiveness and School Improvement, 18*(2), 125-145. https://doi.org/10.1080/09243450600853415

Devos, G., Hulpia, H., Tuytens, M., & Sinnaeve, I. (2013). Self-other agreement as an alternative perspective of school leadership analysis: an exploratory study. *School Effectiveness and School Improvement, 24*(3), 296-315. https://doi.org/10.1080/09243453.2012.693103

Faddar, J., Appels, L., Merckx, B., Boeve-de Pauw, J., Delrue, K., De Maeyer, S., & Van Petegem, P. (2020). *Vlaanderen in TIMSS 2019. Wiskunde- en wetenschapsprestaties van het vierde leerjaar in internationaal perspectief en doorheen de tijd*. Antwerpen: Universiteit Antwerpen.

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 User Guide for the International Database*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-database/

Hesser, H. (2015). Modeling individual differences in randomized experiments using growth models: Recommendations for design, statistical analysis and reporting of results of internet interventions. *Internet Interventions, 2*(2), 110-120. https://doi.org/10.1016/j.invent.2015.02.003

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: what did they really say? *Organizational Research Methods, 9*(2), 202-220. https://doi.org/10.1177/1094428105284919

Leithwood, K., Harris, A., & Hopkins, D. (2020). Seven strong claims about successful school leadership revisited. *School Leadership & Management, 40*(1), 5-22. https://doi.org/10.1080/13632434.2019.1596077

Lüdecke, D., Ben-Shachar, M., S., Patil, I., Waggoner, P., & Makowski, D. (2021). Assessment, testing and comparison of statistical models using R. *Journal of Open Source Software, 6*(59). https://doi.org/10.31234/osf.io/vtq8f

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Boston, US: TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods

Roback, P., & Legler, J. (2021). *Beyond multiple linear regression. Applied generalized linear models and multilevel models in R*. CRC Press. https://bookdown.org/roback/bookdown-BeyondMLR/

Valcke, M., & Standaert, R. (2020). *Onderwijsbeleid in Vlaanderen*. Leuven, Belgium: Acco.

Van Petegem, P., Verhoeven, J. C., Buvens, I., & Vanhoof, J. (2004). *Zelfevaluatie en beleidseffectiviteit in Vlaamse scholen. Het gelijke onderwijskansenbeleid als casus*. Gent, Belgium: Academia Press.

Vanhoof, J., Deneire, A., & Van Petegem, P. (2011). *Waar zit beleidsvoerend vermogen in (ver)scholen? Aanknopingspunten voor zelfevaluatie en ontwikkeling [Where is policy-making capacity in schools (hidden)? Starting points for self-evaluation and development*. Mechelen, Belgium: Plantyn.

Vanhoof, J., & Van Petegem, P. (2017). *Doeltreffend schoolbeleid. Praktijkboek beleidsvoerend vermogen in scholen*. Leuven, Belgium: Acco.