STEUNPUNT
CENTRALE TOETSEN
IN ONDERWIJS

LEERPRESTATIES EN LEERWINST

Ilse Laurijssen
Koen Aesaert

15.09.2022

Conceptnota Steunpunt Centrale Toetsen in Onderwijs
Werkdomein G: Data-analyse en -verwerking
Promotor: Dimokritos Kavadias

Vlaanderen
is onderwijs & vorming

INHOUD

Referentienummer: SCTO/2022.G/2/2

Gelieve als volgt naar deze publicatie te verwijzen:

Laurijssen, Ilse & Aesaert, Koen (2022). *Leerprestaties en leerwinst.* Gent: Steunpunt Centrale Toetsen in Onderwijs. Ref.: SCTO/2022.G/2/2.

Deze publicatie kwam tot stand met steun van de Vlaamse Gemeenschap, Ministerie voor Onderwijs en Vorming. Deze publicatie is ook beschikbaar via www.steunpunttoetsen.be

BELEIDSSAMENVATTING

Door leerwinst centraal te stellen binnen de Vlaamse toetsen, kiest de Vlaamse Regering voor een monitoringstool met het nodige potentieel voor interne kwaliteitszorg door scholen. Leerwinst is echter een concept dat dikwijls inwisselbaar wordt gebruikt ten aanzien van aanverwante begrippen zoals status, groei en toegevoegde waarde. Een eerste doelstelling van deze conceptuele nota is om deze begrippen helder af te bakenen en te verduidelijken op welke wijze deze zich tot elkaar verhouden (deel 1). Tevens bespreekt de nota een aantal statistische en methodologische aandachtspunten waar rekening dient mee gehouden te worden indien we schooleffecten willen schatten (deel 2). Tot slot passen we in deze nota de behandelde concepten en aandachtspunten toe op de centrale toetsen in het Vlaams onderwijs (deel 3). We brengen de mogelijkheden en randvoorwaarden van de centrale toetsen om te rapporteren over status, leerwinst en toegevoegde waarde op de verschillende aggregatieniveaus in kaart voor een aantal voorliggende scenario's. In deel 4 sommen we nog eens de belangrijkste beleidsaanbevelingen op. Deze nota biedt zo de nodige wetenschappelijke ondersteuning voor de verdere concretisering van leerwinstmeting bij de centrale toetsen. In onderstaande samenvatting geven we een minder technische beschrijving van de belangrijkste begrippen en vermelden we enkele van de meest relevante implicaties en met name aanbevelingen voor de analyse en rapportering van de centrale toetsen en stippen eveneens mogelijke beperkingen aan.

Begripsverduidelijking status en groei: de school als trappenhal

Een eerste onderscheid dat gemaakt kan worden is tussen status en groei. Voor een individuele leerling verwijst *status* naar een vaardigheidsniveau dat gemeten wordt door middel van een toets. Het is naar het vaardigheidsniveau dat we verwijzen met de term leerprestatie. Als we het vaardigheidsniveau voorstellen als een trap met verschillende treden, dan verwijst de *leerprestatie* naar de trede op de trap waarop de leerling zich bevindt. *Groei* verwijst naar de vooruitgang die een leerling maakt over de tijd, en vergt een herhaalde meting van dezelfde vaardigheid, aan de hand van toetsen waarvan de resultaten op dezelfde meetschaal staan. Die vooruitgang kunnen we bijvoorbeeld meten aan de hand van het verschil tussen twee opeenvolgende vaardigheidsscores. Het is in deze betekenis van individuele groei dat we spreken van *leerwinst*. Toegepast op het beeld van de trap, gaat het om het aantal treden dat een leerling bijvoorbeeld klimt tussen twee momenten.

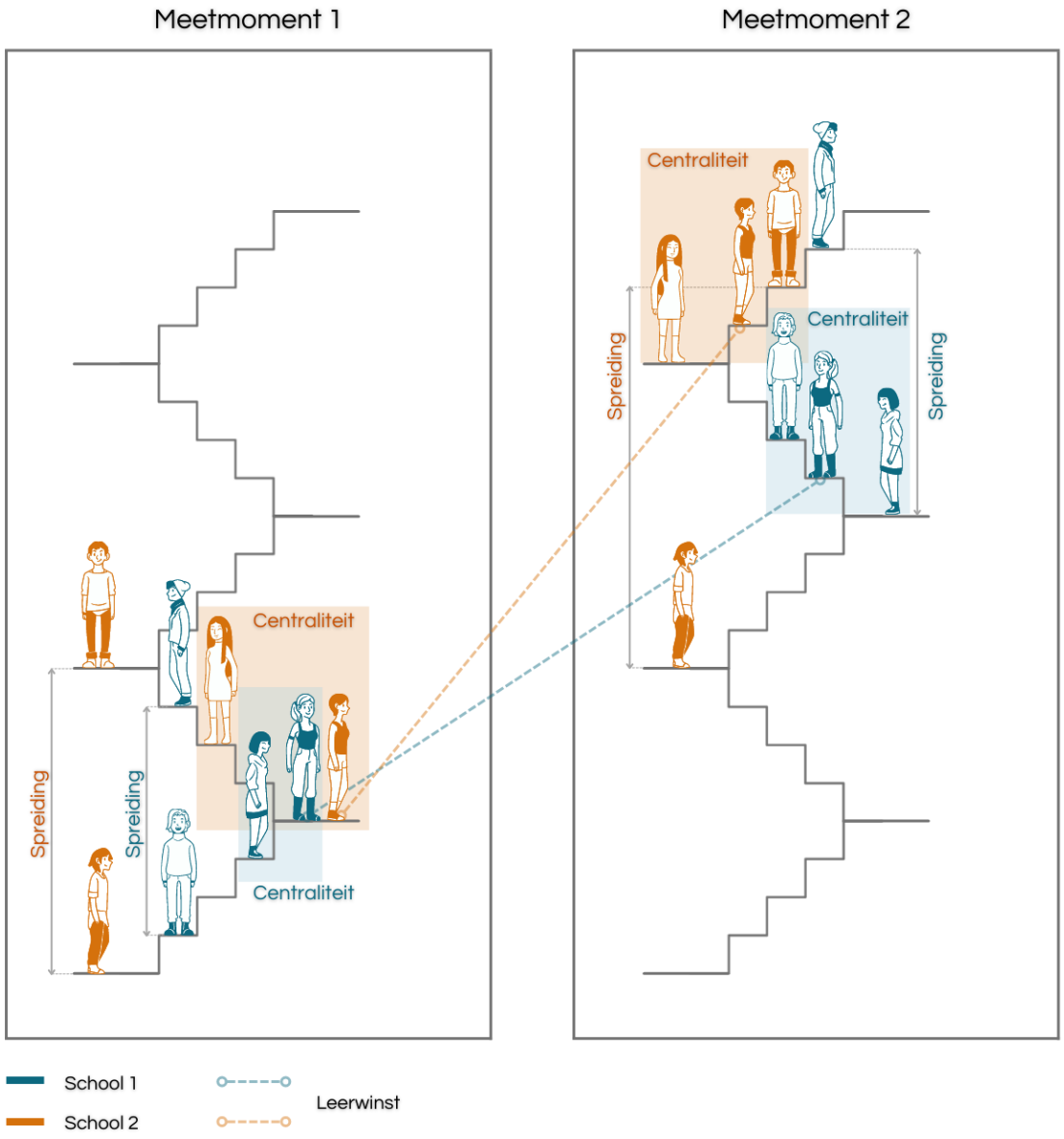
De leerprestaties en leerwinst van individuele leerlingen kunnen ook bekeken worden op *schoolniveau* (of een ander geaggregeerd niveau, zoals klas-, of systeemniveau). Als we het vaardigheidsniveau opnieuw

voorstellen als treden op een trap, dan verwijst status op schoolniveau naar de verdeling van de leerlingen over de trappenhal. De leerprestaties van leerlingen maken zo duidelijk rond welk vaardigheidsniveau de meeste leerlingen zich bevinden (centraliteit), en hoe groot de verschillen zijn tussen de verst en minst gevorderde leerlingen van een school (spreiding). Ze kunnen ook een beeld geven van de mate waarin leerlingen bepaalde (eind)doelen bereiken (eindtermen). Verdergaand op de metafoor van de trappenhal, kan worden vastgesteld welk aandeel van de leerlingen bijvoorbeeld de derde verdieping haalt. Voor leerwinst op schoolniveau neemt de complexiteit toe. Tussen twee momenten verhuist elke individuele leerling op de trap met een aantal treden, sommigen nemen meer treden, anderen minder, de meesten gaan zoals bedoeld omhoog, enkelen struikelen, … Het eindresultaat is opnieuw een verdeling van de leerlingen over de trappenhal, maar wellicht rond een nieuw en hoger gemiddeld niveau, en met een andere (minder of meer) spreiding dan voorheen. In sommige scholen zal de gemiddelde leerwinst groter zijn dan in andere scholen, in sommige scholen zal de leerwinst bijvoorbeeld sterker zijn bij de zwakker presterende leerlingen waardoor de prestatiekloof tussen de verst en minst gevorderde leerlingen afneemt.

Begripsverduidelijking toegevoegde waarde van een school: de school als onderwijsverstrekker

De leerprestaties en leerwinst van leerlingen van een school zeggen weliswaar iets over hoever de leerlingen staan of hoeveel vooruitgang ze boeken tussen twee momenten in hun studieloopbaan, een heel andere vraag is in welke mate de leerlingen effectief *leren door* de school. De *toegevoegde waarde* van een school verwijst naar de bijdrage die de school levert aan het leren van leerlingen, en wordt ook benoemd als het *schooleffect*. De oorsprong van het begrip is economisch, doelend op de toename van de waarde van een product in vergelijking met diens input (grondstoffen) door het productieproces. Het is duidelijk dat het concept van leerwinst hierbij aansluit. Door het centraal zetten van de vooruitgang die leerlingen maken in vergelijking met hun voorafgaande vaardigheidsniveau, vormt een leerwinstmeting een beter vertrekpunt voor de bepaling van de toegevoegde waarde van een school. Een eenmalige statusmeting daarentegen laat niet toe om de toegevoegde waarde van een school vast te stellen. Om het met de trappenhalmetafoor uit te leggen, wanneer leerlingen vanuit verschillende verdiepingen een school kunnen instromen (cf. input), is het niet de eindverdieping die de meeste leerlingen bereiken (eindstatus), maar wel het aantal treden die de leerlingen gemiddeld vooruitgaan (leerwinst) in een school die best aangeeft hoeveel leerlingen leerden tijdens hun verblijf op die school.

Figuur 1: Status en leerwinst op schoolniveau – de trappenhal als metafoor



Niet alle leerwinst van leerlingen is evenwel zomaar toe te wijzen aan de school. Leerlingen leren ook buiten de schoolse context. Bovendien kan er wel een impact zijn van de school als setting, maar heeft die impact niet altijd veel te maken met het onderwijs dat scholen organiseren. Hier houdt dan ook de vergelijking met een industrieel productieproces op: een deel van de verschillen is te wijten aan de context en condities waarin de leerlingen opgroeien. Verschillende elementen komen voor de bepaling van de toegevoegde waarde van een school daarom samen: de causale link met het onderwijs dat scholen aanbieden, het vermogen van scholen om impact te hebben, evenals de faire beoordeling of vergelijking van scholen. Een school wordt (terecht) niet graag afgerekend op resultaten waarvoor ze niet verantwoordelijk is. In dezelfde zin dient bij de vergelijking van scholen rekening gehouden met de factoren die mogelijk wel statistisch samenhangen met gemeten prestaties, maar waarover de scholen zelf (op korte termijn) geen controle hebben.

Scholen opereren immers niet in een vacuüm en via verschillende sporen sluipt de samenleving binnen in wat scholen doen en kunnen doen. Scholen kunnen bijvoorbeeld verschillen in het aantal en de concentratie aan leerlingen uit bijvoorbeeld specifieke kansengroepen. Scholen dienen bijvoorbeeld ook om te gaan met omkaderings- en financieringsmogelijkheden en -beperkingen. Om die redenen is de bepaling van de toegevoegde waarde van een school meer dan het vaststellen van de gemiddelde vooruitgang van haar leerlingen, maar moet deze worden gecontextualiseerd. Er zijn verschillende manieren om die contextualisering te benaderen, welke verschillen in welke leerling-, leerlingsamenstelling- of schoolcontextkenmerken mee in rekening worden gebracht. Vanuit een schoolkeuzeperspectief, kan de vraag zijn wat de leerwinst is die een school zou boeken voor een gemiddelde leerling (type-A schooleffect), of voor een specifiek soort leerling, zoals bijvoorbeeld meisjes tegenover jongens, of een leerling uit een laag SES gezin, … Vanuit een schoolontwikkelingsperspectief kan de vraag zijn welke vooruitgang in leerprestaties de school zou boeken voor een specifieke of gemiddelde leerling gegeven een bepaalde (bv. gemiddelde SES) schoolsamenstelling (type-B schooleffect) of meer algemeen gegeven een bepaalde schoolcontext, bv. schoolgrootte (type-X schooleffect). Door zoveel mogelijk rekening te houden met dergelijke schoolexterne factoren (input- en contextkenmerken), verwijst de toegevoegde waarde van scholen naar de impact die scholen (of een groep van leraren) hebben op de vooruitgang die leerlingen maken als gevolg van het onderwijs dat zij aanbieden aan hun leerlingen (proceskenmerken van de school). Welk type schooleffect wordt gerapporteerd hangt in de eerste plaats af van de doelstelling. Met het oog op schoolverantwoording of -verbetering, zijn de laatste types het meest relevant, met meer uitgebreide controles voor allerlei factoren die voor scholen *niet-manipuleerbaar* zijn. Deze types van schooleffecten benaderen ook het meest de onderwijsprocessen die

zich op school afspelen, hoewel ze nog geen aanwijzingen geven over de concrete onderwijsprocessen die al dan niet leiden tot meer leerwinst.

Een belangrijke kanttekening bij de analyse van toegevoegde waarde tot slot is dat deze wordt uitgevoerd op basis van gemiddelde statistische verbanden, en met name steunt op verschillen tussen scholen. De toegevoegde waarde van een school is immers de hoeveelheid vooruitgang in leerprestaties in vergelijking met de gemiddelde vooruitgang in scholen met dezelfde kenmerken. Dit impliceert in veel gevallen een onderschatting van het (werkelijk) onderwijseffect in de mate dat scholen evenveel of op dezelfde wijze zouden bijdragen tot de vooruitgang van leerlingen (de échte *counterfactual*, namelijk geen onderwijs, behoort immers niet tot de observaties). Dit onderstreept nog eens het belang van een leerwinstmeting en interpretatie van toegevoegde waarde in termen van leerwinst geboekt door leerlingen.

Verschillende indicatoren voor onderwijskwaliteit: status, leerwinst en toegevoegde waarde

Er is een spanningsveld tussen status en leerwinst als het aankomt op het vastleggen van een maatstaf waaraan leerlingen of scholen moeten voldoen. Het Vlaams onderwijssysteem definieert met de eindtermen minimale onderwijsdoelen, wat een duidelijk criterium biedt voor de evaluatie van de (eind)status van leerlingen of scholen. Voor de evaluatie van de leerwinst van leerlingen of scholen, is een dergelijk aan onderwijsdoelen gebonden criterium niet voorhanden. De behaalde leerwinst dient ook steeds beoordeeld te worden in combinatie met de (eind)status. Leerwinst verzoenen met onderwijsdoelen kan door leerwinst te beschouwen in combinatie met de uitgangspositie of het startniveau. Om dezelfde (eind)doelen te bereiken moet de leerwinst (doelstelling) hoger liggen voor lager startende leerlingen. Verder onderzoek moet uitwijzen op welke wijze leerwinst op die manier criteriumgericht geëvalueerd kan worden, op school- en op leerlingniveau. Dat de hoeveelheid gerealiseerde leerwinst afhankelijk is van het startniveau, wordt alvast meegenomen in de (leerwinst)modellering van toegevoegde waarde van scholen (vergelijking scholen met gelijkaardig startniveau). Ook de rapportering van de leerwinst op leerlingniveau wordt ingeval van normering best gerelateerd aan het startniveau (vergelijking met leerlingen met gelijkaardig startniveau).

Methodologische en statistische aandachtspunten

Een fundamenteel aandachtspunt bij het gebruik van statistieken is dat rekening moet worden gehouden met de statistische (on)zekerheid van resultaten, i.e. met de waarschijnlijkheid dat correcte conclusies

kunnen worden getrokken uit de bevindingen. Dit geldt niet alleen voor complexe statistische modellen of analyses, dit geldt evengoed voor eenvoudige indicatoren als gemiddelden. Op basis van een beperkt aantal observaties kunnen geen sterke uitspraken worden gedaan over een geheel. Meer specifiek is de statistische *power* in dat geval te beperkt, wat betekent dat de statistische onzekerheid dan te groot is, de statistische betrouwbaarheid te klein, en de kans op sterke fluctuaties ingeval van een (theoretische) hermeting ervan erg groot is. De kracht van het getal speelt in de context van de centrale toetsen zowel voor de meting van de toetsvaardigheid van een leerling als voor de veralgemening naar grotere gehelen (klas-, school- of systeemniveau) op basis van individuele leerlingen. Zo kan een toets een vaardigheid preciezer meten naarmate deze bestaat uit meerdere toetsopgaven. Zo kan de status, leerwinst of toegevoegde waarde van een klas of school adequater worden bepaald wanneer voor de berekening ervan informatie beschikbaar is voor een groter aantal leerlingen. Door statistische onzekerheid kunnen vastgestelde verschillen niet zomaar ook als echte verschillen worden geïnterpreteerd. Het is daarom aangewezen om statistische betrouwbaarheidsintervallen te rapporteren bij alle resultaten op groepsniveaus evenals bij individuele vaardigheidsscores. Met name zijn deze noodzakelijk voor vergelijkingen ten aanzien van specifieke benchmarkwaarden of om verandering over de tijd te kunnen beoordelen. Een alternatief dat kan worden overwogen bij kleine aantallen, is om de statistische *power* te vergroten door middel van zogenaamde gepoolde data-analyses (op basis van samengevoegde gegevens van opeenvolgende cohortes). De aantallen leerlingen worden dan immers groter. Door het op die manier samennemen van meerdere cohorten verdwijnen evenwel mogelijke veranderingen over de tijd buiten beeld.

Het vooropstellen van een rapportering van resultaten op de centrale toetsen voor een individuele leerling, vergt een meer betrouwbare meting van de te toetsen vaardigheid dan hetgeen gangbaar is in de meeste internationaal vergelijkende onderwijsstudies of de Vlaamse peilingen. Het vooropstellen van leerwinst stelt daarnaast bijkomende uitdagingen. Zo dienen de opeenvolgende toetsen dezelfde vaardigheid op dezelfde meetschaal te meten, en vergt een betrouwbare meting van leerwinst een nog meer betrouwbare meting van de individuele metingen van de toetsvaardigheid op de opeenvolgende momenten. Naast het verhogen van het aantal toetsitems, kan ook een adaptieve afname van de toets hieraan bijdragen. Maar ook bij een sterk betrouwbare meting dienen leerlingresultaten omzichtig geïnterpreteerd te worden, aangezien het slechts om momentopnames gaat. De leerlingresultaten op de centrale toetsen worden door de klassenraad daarom best gebruikt aanvullend op de andere leerlingresultaten die ook meer doorheen een schooljaar worden bekomen.

De status op schoolniveau kan een zinvolle indicator zijn om bijvoorbeeld te beoordelen of de leerlingen van een school voldoende voorbereid zijn voor doorstroom naar een volgend leerjaar of onderwijsniveau.

Status is daarentegen niet geschikt om de effecten van scholen op het leren van leerlingen vast te stellen. Daarom dat we bij toegevoegde waarde uitgaan van een leerwinstmeting. Een schooleffect dat wordt bepaald aan de hand van modellen waarin geen eerdere leerprestaties zijn opgenomen, meet niet de toegevoegde waarde van de school. Bijgevolg is het niet aan te raden beleidsbeslissingen op school- en systeemniveau te nemen op basis van modellen voor één meetmoment. In het geval van een eerste meting (met name in het vierde leerjaar lager onderwijs bij de centrale toetsen), is die eerdere prestatie natuurlijk niet beschikbaar. Voor een analyse van schoolverschillen bij een eerste meting is het zo mogelijk nog belangrijker om zo volledig mogelijk te controleren voor leerling- en schoolkenmerken. Enkel in de mate dat de opgenomen kenmerken het effect van het vaardigheidsniveau van leerlingen bij aanvang (bij instroom) kunnen (weg)verklaren, kan het gecorrigeerde schooleffect worden geïnterpreteerd in termen van toegevoegde waarde. In de praktijk blijkt dat evenwel moeilijk haalbaar. Het aanvangsniveau van leerlingen is immers met voorsprong de belangrijkste determinant van de latere status. De haalbaarheid vergroot wanneer goede indicatoren voor eerdere schoolse prestaties kunnen worden meegenomen. In die zin kan een toegevoegde waarde mogelijk ook worden bepaald op basis van een herhaalde meting van een vaardigheid bij individuele leerlingen die niet helemaal voldoet aan de voorwaarden voor een leerwinstmeting (omdat een gemeenschappelijke meetschaal ontbreekt of omdat de getoetste vaardigheid inhoudelijk niet dezelfde is).

Om de toegevoegde waarde van een school te bepalen is het belangrijk dat de kenmerken die als controlevariabelen in het model worden geïntegreerd zo juist en volledig mogelijk beschikbaar zijn. Zo dient bijvoorbeeld de sociale achtergrond van leerlingen zo correct mogelijk te worden geoperationaliseerd. Een grondige studie van de gewenste kenmerken die in de modellen dienen geïntegreerd te worden is dan ook aangeraden. Indien kenmerken geselecteerd worden waarvan de overheid niet over gegevens beschikt, dan dienen deze bij alle leerlingen in alle scholen bevraagd te worden. Deze zijn immers nodig om een eerlijke vergelijking tussen alle scholen mogelijk te maken.

Anderzijds zijn er ook risico's op overcorrectie wanneer veel factoren mee opgenomen worden in een model. Het is immers niet de bedoeling dat deze de effecten van schoolinterne kenmerken (door scholen beïnvloedbare schoolpraktijken) weghalen uit de geschatte toegevoegde waarde van een school. Daarom is het wenselijk om informatie over relevante schoolinterne kenmerken (bv. organisatiestructuur, schoolleiderschap, schoolklimaat, instructiepraktijken) mee in de modellering te kunnen opnemen. Omdat schoolinterne kenmerken evenwel vaak moeilijker in kaart te brengen zijn, is het aangeraden om sensitiviteitsanalyses uit te voeren voor een steekproef van scholen waarvoor bijkomende indicatoren worden verzameld. Wanneer blijkt dat kennis van bepaalde kenmerken de toegevoegde waarde

schattingen (sterk) verbetert, dient overwogen te worden om die kenmerken op termijn ook meer algemeen in kaart te brengen voor alle scholen.

De schatting van de toegevoegde waarde van scholen steunt op statistische modellering waarvan niet verwacht kan worden dat gebruikers daarmee vertrouwd zijn. Het is dan ook essentieel om de gerapporteerde schooleffecten helder te duiden. Bovendien kan ook de wijze van rapportering van de toegevoegde waarde van scholen worden afgestemd op de behoeften van de gebruikers. Het kan zinvoller zijn om het schooleffect voor te stellen als het voorspelde (gecorrigeerde) resultaat voor een bepaalde referentiegroep van leerlingen en gegeven bepaalde schoolcontextuele kenmerken eerder dan louter verschillen ten opzichte van een globaal gemiddelde te rapporteren. Op die manier komen niet enkel relatieve verschillen tussen scholen in beeld, maar krijgen schoolresultaten mogelijk ook een meer inhoudelijke interpretatie in termen van leerwinst of het behalen van (eind)doelen. Naast dergelijke rapportering op systeemniveau, kan het voor feedback naar scholen toe zinvoller zijn om te kiezen voor een aparte weergave van de deelcomponenten die samen het geschatte schooleffect uitmaken, door zowel het (behaalde) resultaat van de school als het verwachte resultaat van de school (te interpreteren als het resultaat dat scholen met dezelfde kenmerken behalen) weer te geven.

Toepassing voor de Vlaamse centrale toetsen

In het derde deel van de nota passen we de algemene principes toe op de centrale toetsen in het Vlaams onderwijs. Dat verduidelijkt de methodologische en statistische implicaties van bepaalde al gemaakte of nog te maken keuzes.

Zo maakt de toepassing duidelijk dat niet voor alle vooropgestelde toetsen aan de *voorwaarden voor het meten van leerwinst* voldaan zal kunnen worden. Een eerste vraag die zich stelt is voor welke toetsdomeinen de *inhoudelijke basis* van de te meten vaardigheden voldoende stabiel is over de leerjaren of schoolloopbaan van leerlingen heen (voorwaarde 1). Gezien de koppeling van toetsen aan eindtermen, zullen niet alle mogelijke thema's in verschillende leerjaren worden getoetst. Zo zou een thema als machten en logaritmen in het zesde leerjaar secundair onderwijs kunnen worden getoetst maar niet in de eerste graad van het secundair onderwijs noch in het basisonderwijs. Waar de te toetsen domeinen over leerjaren heen wel inhoudelijk voldoende aansluiten, is het voor een leerwinstmeting tevens noodzakelijk dat de toetsen op een *gemeenschappelijke meetschaal* kunnen worden geplaatst (voorwaarde 2). Dat vergt het gebruik van geschikte ankeritems, die beter te bepalen zijn bij een gelijktijdige kalibratie, terwijl de vooropgestelde afnames voorzien in aparte kalibraties over opeenvolgende leerjaren. Zo kunnen de toetsen bij bijvoorbeeld de kalibratie van de toetsen voor het

vierde leerjaar niet geoptimaliseerd worden om te sporen met deze van het zesde leerjaar, die pas twee jaar later voor het eerst worden getest. Niet alleen wanneer toetsen onderwijsniveauspecifiek zijn, ook wanneer ze onderwijsstroom of -finaliteitsspecifiek zijn, kan leerwinst enkel worden gemeten voor alle betrokken leerlingen wanneer een gemeenschappelijke meetschaal mogelijk is.

Het derde deel gaat tevens concreter in op verwachtingen bij de centrale toetsen aangaande het aantal leerlingen, het aantal scholen, en representativiteit van getoetste scholen en leerlingen. Deze elementen bepalen immers de uitspraken die kunnen worden gedaan op verschillende niveaus (leerling, school, systeem). In het ideale scenario wordt een bepaald domein getoetst bij alle leerlingen over alle scholen heen (een uniform of zogenaamd volledig afnamedesign) en, mits leerwinstmeting zinvol is (cf. hierboven), ook over de volledige schoolloopbaan van een leerling. In de meeste gevallen zal bij dergelijk volledig afnamedesign (i.c. de toetsen begrijpend lezen en wiskundige problemen oplossen) aan de *voorwaarden voor rapportering over leerwinst* op de verschillende niveaus worden voldaan, al stellen niet-standaard schooltrajecten de nodige uitdagingen (afwijkende schoolse vordering, schoolmobiliteit, vroegtijdig schoolverlaten).

Om de totale toetsduur voor leerlingen te beperken, zal het toetsdomein wiskunde evenwel in kaart gebracht worden met een gedifferentieerd afnamedesign. Daarbij wordt aanvullend op een toets die bij alle leerlingen wordt afgenomen (i.c. wiskundige problemen oplossen), voor specifieke toetsthema's uitgegaan van een onvolledige afname: een specifiek thema (bv. getalbegrip) wordt dan getoetst bij een deelsteekproef van leerlingen. Bij dergelijke onvolledige afname heeft met name voor leerwinst (en toegevoegde waarde) de keuze van het *longitudinale afnamedesign* belangrijke implicaties. Zo zal bijvoorbeeld enkel wanneer de toetsen over de leerjaren heen gekoppeld worden afgenomen (op school- of individueel niveau) schoolfeedback over leerwinst mogelijk zijn. Voor schoolfeedback zal de leerwinstrapportering van toetsen met onvolledige afname bovendien in eerste instantie beperkt blijven tot de leerlingen met een standaard schoolloopbaan (normaalvorderende, niet-schoolmobiele leerlingen). In de nota worden verschillende, sommige nog te onderzoeken, mogelijkheden voorgesteld om, modelmatig, de leerprestaties van de leerlingen met niet-standaard schoolloopbanen (evenredig) mee te verwerken bij analyses op school- of systeemniveau.

Duidelijk is dat de geschetste voorwaarden, mogelijkheden en beperkingen voor een leerwinstmeting met de centrale toetsen gepaard gaan met uitdagingen voor de analyse voor rapportering en schoolfeedback. Tegenover die complexiteit, biedt leerwinst een duidelijke meerwaarde om onderwijskwaliteit op basis van toetsen te evalueren: a) voor een goede inschatting van de toegevoegde waarde van een school is het noodzakelijk om rekening te houden met de aanvangsprestaties van leerlingen (voordeel van leerwinst tegenover status), b) de toegevoegde waarde die een school realiseert verwerft een heldere

inhoudelijke betekenis in termen van hoeveel vooruitgang de leerlingen maken (voordeel van het kunnen vergelijken van metingen over de schoolloopbaan op dezelfde meetschaal, tegenover een louter scholenvergelijkend perspectief).

DOEL VAN DEZE NOTA

Binnen het Vlaamse Regeerakkoord 2019-2024 wordt het bevorderen van kwaliteitsvol onderwijs als de sleutel omschreven om de leerwinst van elke leerling te bevorderen. Daarom worden vanaf het schooljaar 2024 Vlaamse of centrale toetsen in het leerplichtonderwijs geïntroduceerd met als doel de leerwinst op leerling- en schoolniveau in kaart te brengen alsook na te gaan welk aandeel van de Vlaamse leerlingpopulatie de eindtermen voor de leergebieden wiskunde en Nederlands op het einde van een onderwijsniveau beheerst. De verwachting is daarbij dat leerwinst wordt gemeten over de verschillende leerjaren en onderwijsniveaus (basis en secundair) heen (Regeerakkoord, 2019; Oproep Steunpunt, 2020). Aansluitend wordt in de oproep tot universitair steunpunt "Ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde net- en koepeloverschrijdende toetsen in Vlaanderen" de verwachting geuit dat de resultaten van de centrale toetsen kunnen aangewend worden in functie van:

- de beoordeling van de leerling: de resultaten op de toets kunnen worden meegenomen in de globale beoordeling van de leerling, maar zijn niet doorslaggevend in het kader van studievoortgang en - oriëntering. Het schoolteam krijgt een objectief en kwaliteitsvol toetsresultaat dat ze als een van de verschillende informatiebronnen kan meenemen in haar beoordeling;
- zelfreflectie door de leerkracht: net zoals voor de leerlingen, is het niet de bedoeling dat de beoordeling van leerkrachten enkel op basis van de toetsresultaten zou gebeuren. In de eerste plaats is het een instrument voor de betrokken leerkrachten om te reflecteren over de resultaten en zo de kwaliteit te verhogen;
- de opvolging van de onderwijskwaliteit op het niveau van scholen: zwakkere toetsresultaten vormen een knipperlicht. Scholen waarvan de leerlingen significant minder leerwinst genereren op die proeven, moeten in een vrij te kiezen begeleidingstraject stappen om de kwaliteit van hun onderwijs te verhogen
- de opvolging van de onderwijskwaliteit op systeemniveau: (trends in) resultaten op de toetsen bieden aanknopingspunten voor de evaluatie en eventuele bijsturingen van het beleid (Oproep Steunpunt, 2020).

Door leerwinst centraal te stellen binnen de Vlaamse toetsen, kiest de Vlaamse Regering voor een monitoringstool met het nodige potentieel voor interne kwaliteitszorg door scholen. Leerwinst is echter een concept dat dikwijls inwisselbaar wordt gebruikt ten aanzien van aanverwante begrippen zoals

status, groei en toegevoegde waarde. Op schoolniveau verwijst elk van deze begrippen binnen de wetenschappelijke literatuur nochtans naar een specifieke invulling, naar een statistisch model om een schooleffect te schatten en naar de besluiten die uit dat model kunnen getrokken worden.

Hieronder behandelen we eerst leerwinst en de drie aanverwante begrippen 'status', 'groei' en 'toegevoegde waarde'. In Deel 2 bespreken we een aantal statistische en methodologische aandachtspunten waar rekening dient mee gehouden te worden indien we schooleffecten willen schatten. Tot slot passen we de behandelde concepten en aandachtspunten toe op de centrale toetsen binnen de Vlaamse context. Meer specifiek nemen we het afnamedesign van wiskunde voor het secundair onderwijs als vertrekpunt. Zo brengen we de mogelijkheden van de centrale toetsen om te rapporteren over status, leerwinst en toegevoegde waarde op de verschillende aggregatieniveaus in kaart.

1 STATUS – LEERWINST – TOEGEVOEGDE WAARDE

Toetsgegevens worden in onderwijs courant gebruikt om de leerprestaties van leerlingen te meten en te beoordelen. Dit gebeurt niet alleen op leerlingniveau, ook om de resultaten van groepen van leerlingen (klas-, leraar-, school- of systeemniveau) te monitoren of te evalueren. Hoewel de gebruikte terminologie niet altijd eenduidig is, volgen we de aanbeveling van Hollingsworth, Heard en Weldon (2019) om toetsresultaten te onderscheiden van toetsprestaties (respectievelijk *achievement* en *performance*). *Toetsresultaten* worden gemeten op basis van de antwoorden van leerlingen op toetsvragen, terwijl met de term *toetsprestaties* meer specifiek wordt bekeken of de gemeten toetsresultaten voldoen aan bepaalde verwachtingen (norm- of criteriumgericht[1]). Met *leerprestaties* tenslotte doelen we meer algemeen op de (onderliggende) vaardigheid van de leerlingen die we wensen te meten aan de hand van toetsen (bv. begrijpend lezen, of oplossen van wiskundige vraagstukken).

We beginnen dit deel met een bespreking van status (sectie 1.1), d.i. het in kaart brengen van leerprestaties op basis van één toetsafname. Vervolgens gaan we in op leergroei en leerwinst (sectie 1.2), welke over de tijd herhaalde afnames van eenzelfde toets veronderstellen, en toegevoegde waarde (sectie 1.3), waarbij het bepalen van de bijdrage van een school aan het leren van leerlingen centraal staat.

1.1 STATUS

In het domein van onderwijseffectiviteit en -evaluatie verwijst status naar de toetsprestaties van leerlingen in een leer- of vakgebied op één specifiek moment. Deze status kan betrekking hebben op het individuele leerlingniveau of op een groep van leerlingen (klas-, leraar-, school- of systeemniveau). We beginnen met status op het leerlingniveau. De status van een leerling wordt bepaald op basis van de toetsdata die beschikbaar zijn van één enkel meetmoment en representeert het vaardigheidsniveau van de leerling op dat specifiek moment (Castellano & Ho, 2013).

[1] In onderwijstoetsontwikkeling verwijst *criterium* naar een prestatiestandaard die overeenkomt met een bepaald niveau van kennis, van vaardigheden of leerdoelen. Een criteriumgerichte benadering is het meest geschikt om te bepalen of een leerling bepaalde concepten of vaardigheden heeft verworven. Wanneer het toetsresultaat van een leerling wordt vergeleken met dat van een vergelijkingsgroep, bijvoorbeeld alle leerlingen van hetzelfde leerjaar, wordt daarentegen gesproken van een normgerichte benadering.

De toetsprestaties van leerlingen kunnen op verschillende manieren geoperationaliseerd worden, afhankelijk van onder meer het gehanteerde meetmodel of de gehanteerde benchmarks en beheersingsniveaus. Zo kan de status van een twaalfjarige leerling die op het einde van het zesde leerjaar een toets 'begrijpend lezen' van 20 items aflegt, onder het toetsparadigma van de klassieke testtheorie worden weergegeven als een somscore op 20. Gebruik makend van modellen die uitgewerkt werden binnen de zogenaamde itemresponstheorie (IRT), kan diezelfde toetsprestatie worden gerepresenteerd als een gestandaardiseerde vaardigheidsscore (thèta) op een meetschaal met een theoretisch minimum (vb. -3) en maximum (vb. +3) (de Ayala, 2009). Dergelijke vaardigheidsscores zijn gebaseerd op de juistheid van de antwoorden van leerlingen rekening houdend met de itemparameters zoals moeilijkheidsgraad. Lineaire transformaties laten toe dergelijke vaardigheidsscores te transformeren naar makkelijker te hanteren waarden, zoals bij de meetschalen van internationaal vergelijkende onderzoeken zoals PIRLS of TIMSS. Tot slot kan de status van een leerling ook worden uitgedrukt in termen van beheersingsniveaus of ten opzichte van één of meerdere cesuren. Een cesuur kan zowel criterium- als normgericht zijn. Een *criteriumgerichte* cesuur laat bijvoorbeeld toe uitspraken te doen over het al dan niet behalen van welbepaalde leerdoelen. Zo wordt de status van een leerling binnen het peilingsonderzoek na cesuurbepaling niet alleen weergegeven door een geschatte vaardigheidsscore maar ook door de categorisering 'beheerst de eindtermen niet/beheerst de eindtermen'. *Normgerichte* cesuren laten toe om de vaardigheid van een leerling te situeren binnen de verdeling van bijvoorbeeld alle leerlingen van een specifiek leerjaar. Een hoog vaardigheidsniveau kan bijvoorbeeld worden vastgelegd op basis van de 30% hoogst scorende leerlingen. Vaardigheidsscores kunnen ook omgezet worden naar percentielen. Andere gekende relatieve statusscores zijn leerjaar- of leeftijdsequivalenten op basis van gemiddelde (of mediaan) statusscores in leerjaren of leeftijdsgroepen (Cermak, 1989).

De status van een groep van leerlingen (klas-, leraar-, school- of systeemniveau) wordt bepaald op basis van de status van de leerlingen die deel uitmaken van de groep (individueel leerlingniveau). De individuele leerlingscores worden daarvoor *geaggregeerd* op het groepsniveau. Doorgaans wordt de status voor een groep van leerlingen berekend als het gemiddelde van de status van de individuele leerlingen van de betreffende groep (som van de scores gedeeld door het aantal leerlingen). Afhankelijk van kenmerken van de verdeling, kan het ook gaan om de mediaan, of worden de extreme waarden niet meegenomen voor de berekening van het gemiddelde. Op een gegroepeerd niveau kan naast de statuspositie ook de mate van spreiding van de leerlingen worden berekend, bijvoorbeeld aan de hand van de standaardafwijking.

Tabel 1 illustreert op niet-exhaustieve wijze een aantal mogelijke indicatoren van status op leerling- en schoolniveau.

Tabel 1: Mogelijke status-indicatoren

Indicator van Status	Leerling_{xA}	School A	
		Gemiddelde	Standaardafwijking
Somscore	18/20	14,3/20	3,1
Theta (latente vaardigheidsscore)	2,531	0,134	0,20
Beheersing eindtermen	Beheerst ET	67% van de leerlingen beheerst de eindtermen*	

* De spreiding wordt niet apart berekend aangezien leerlingen de eindterm al dan niet beheersen, en bij een tweedeling het percentage meteen ook indicator is voor de hoeveelheid spreiding. Een berekende standaardafwijking voegt in dat geval geen informatie toe, maar hangt louter af van het percentage ($\sqrt{(1-\text{p}) * \text{p}}$).

1.2 LEERWINST

1.2.1 Leerlingniveau

Met leerwinst zetten we de stap van een eenmalige statusmeting naar een herhaalde meting van de leerprestaties van een leerling in een welbepaald leer- of vakgebied op twee of meerdere momenten binnen de studieloopbaan van de leerling. Leerwinst verwijst naar de wijziging in de leerprestaties van individuele leerlingen (McGrath et al., 2015; Rodgers, 2007), naar de vooruitgang of groei in hun toetsvaardigheid (Castellano & Ho, 2013; Penninckx & Quintelier, 2016).

De terminologie wordt niet altijd en door iedereen op dezelfde wijze gebruikt. De *individuele verandering in leerprestaties* wordt in de literatuur verschillend benoemd als *learning progress*, *learning growth* en *learning gain* (Hollingsworth et al. 2019; Zoanetti, 2021). Verschillende auteurs gebruiken deze termen evenwel ook om andere accenten te leggen of voor een meer specifieke of alternatieve interpretatie van leerwinst.[2] Essentieel om van leerwinst te spreken zijn twee elementen: het gaat om een herhaalde meting (1) bij eenzelfde leerling, en (2) van eenzelfde toetsvaardigheid op dezelfde schaal (Castellano & Ho, 2013; Janssens, Rekers-Mombarg, & Lacor, 2014; zie voor verdere toelichting sectie 2.1). Deze beide elementen van leerwinst komen ook samen in de omschrijving of uitdrukking van leerwinst als *individuele groei*.

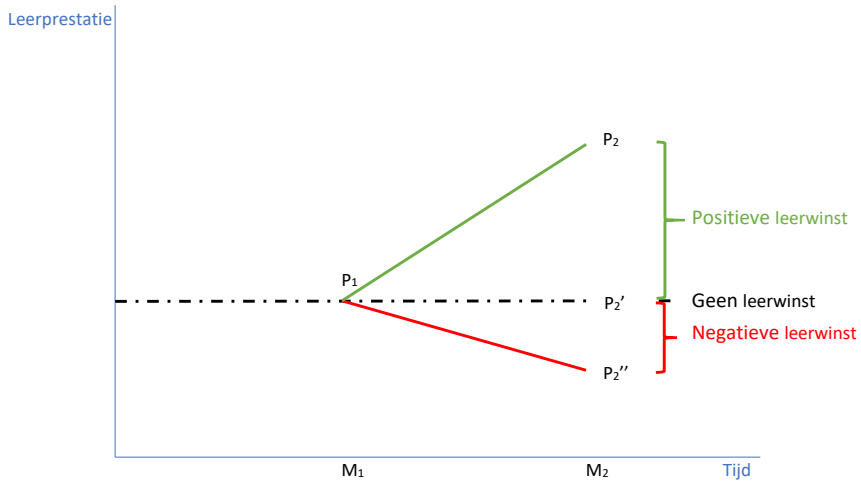
In Figuur 2 wordt de definitie van leerwinst op een schematische manier weergegeven. P_1 verwijst naar de leerprestatie van een leerling op het eerste toetsmoment (M_1), terwijl P_2, P_2' en P_2" verwijzen naar drie mogelijke leerprestaties van diezelfde leerling op een tweede toetsmoment (M_2). In zijn meest eenvoudige vorm wordt de leerwinst die een leerling boekt berekend als het verschil tussen de toetsscores op meetmomenten die verschillen in de tijd (*gain scores;* Gong et al, 2006; Gossman & Powell, 2019). Uitgaande van die eenvoudige operationalisatie van leerwinst als verschilscore, kan de leerwinst van een leerling positief ($LW_{pos}=P_2-P_1$), nul ($LW_0=P_2'-P_1$) of negatief ($LW_{neg}=P_2''-P_1$) zijn. Deze operationalisatie van leerwinst wordt ook omschreven als de *hoeveelheid verandering* of de afgelegde afstand (*distance*

[2] Met groei (*growth*) stellen Castellano & Ho (2013) de individuele verandering in leerprestaties van leerlingen centraal. Met groei (*learning growth*) verwijzen verschillende auteurs ook naar het (volledige) leertraject dat een leerling aflegt om zo het onderscheid met een of twee metingen te maken. Met *learning gain* wordt in de onderzoeksliteratuur doorgaans gefocust op het verschil in toetsscores tussen twee meetmomenten, als de specifieke operationalisatie aan de hand van een verschilscore (*gain scores*, zie ook verder), of om de vergelijkbaarheid van de metingen over de tijd te benadrukken, soms evenwel zonder dat het om een meting bij dezelfde leerlingen gaat. Met *learning progress* wordt de focus dan weer in hoofdzaak gelegd op het herhaald meten bij dezelfde leerling, maar staat daarentegen de vergelijkbaarheid van de metingen niet altijd op de voorgrond.

travelled; McGrath et al., 2015) tussen de eerst gemeten toetsscore en de tweede toetsscore (cf. verticale verschil in Figuur 2). Voor een vergelijking van leerwinst over meerdere meetmomenten, kan het zinvol zijn om de (verschillende) duur tussen (elke) twee opeenvolgende metingen mee in rekening te nemen. Het groeitempo (*growth rate*; Betebenner & Linn, 2009) verwijst naar de *snelheid van verandering*, met name de hoeveelheid leerwinst per tijdseenheid (bijvoorbeeld per leerjaar, cf. de hellingshoek van het verschil in Figuur 2).

De schematische voorstelling van leerwinst maakt duidelijk dat een berekening van individuele leerwinst enkel mogelijk (betekenisvol) is wanneer de metingen over de tijd zinvol met elkaar vergeleken kunnen worden. Dat veronderstelt dat over de tijd dezelfde toetsvaardigheid op dezelfde schaal wordt gemeten (longitudinale meetinvariantie). Groei kan je enkel meten met een meetlint dat niet verandert over de tijd. Omdat toetsen doorgaans aangepast worden aan het ontwikkelingsniveau van de leerlingen, stelt dit bijkomende uitdagingen om de wijziging in leerprestatie over de tijd vast te kunnen stellen. Niet alleen dienen de herhaalde metingen inhoudelijk equivalent te zijn, de toetsresultaten dienen ook gebracht te worden op een gemeenschappelijke meetschaal van de onderliggende vaardigheid (zie ook verder in sectie 2.1).

Figuur 2: Schematische voorstelling leerwinst (aangepast van Gossman & Powell, 2019)



Bovenstaande illustreert eveneens dat niet zomaar kan aangenomen worden dat elke leerling tussen meerdere meetmomenten vooruitgang (positieve leerwinst) maakt, naast verbetering kan ook sprake zijn van stagnatie (geen leerwinst) of verslechtering (negatieve leerwinst). De hoeveelheid leerwinst die een

leerling maakt voor een bepaald leer- of vakgebied kan ook verschillen over de studieloopbaan. Bekeken over de (langere) studieloopbaan van leerlingen tot slot, wordt regelmatig een patroon van afnemende vooruitgang vastgesteld. De meeste leerwinst wordt dan initieel geboekt, waarna het groeitempo (hoeveelheid leerwinst per tijdseenheid) vertraagt. Uit veelvuldig onderzoek blijkt dat dit patroon zich evenwel niet voordoet bij alle leerlingen, noch alle leeftijdsgroepen of op alle toetsdomeinen (Castellano & Ho, 2013).

1.2.2 Schoolniveau

Leerwinst kan ook op het niveau van de school of een onderwijssysteem berekend en beschreven worden (Castellano & Ho, 2013). Dan is de leerwinst van een school gelijk aan de *gemiddelde leerwinst van leerlingen* in de school en de leerwinst van een onderwijssysteem gelijk aan de gemiddelde leerwinst over alle leerlingen en scholen heen. Naar analogie met de karakterisering van status op groepsniveau door middel van de gemiddelde score (of een andere maat voor centraliteit) en de spreiding, kan leerwinst op een groepsniveau worden gezien als een verschuiving in het gemiddelde over de tijd, maar kan daarnaast ook de verandering in spreiding worden bekeken. Zo kan op basis van de dynamiek in twee prestatieverdelingen ook homogenisering dan wel toename van heterogeniteit over de leerlingen binnen een groep worden vastgesteld. De prestaties kunnen over de tijd naar elkaar toegroeien, of daarentegen ongelijker worden tussen de leerlingen.

Tabel 2 illustreert de meting van leerwinst op schoolniveau en hoe deze zich verhoudt tot status en verandering over de tijd en leerjaren. Voor opeenvolgende jaren en leerjaren worden (hypothetische) gemiddelde vaardigheidsscores weergegeven. De meting van leerwinst vergt de opvolging van een cohorte van *dezelfde leerlingen* over de tijd. Het gaat om de wijziging in de leerprestaties voor dezelfde leerlingen, over verschillende meetmomenten (weergegeven op de diagonaal in de tabel, in kleur)[3]. Leerwinst over twee meetmomenten heen kan berekend worden als het verschil tussen de vaardigheidsscores (status) op de twee momenten.[4] Indien alle leerlingen van de school in Tabel 2 een

[3] De berekening van leerwinst op schoolniveau is, omwille van leerlingmobiliteit en afwijkende schoolloopbanen, moeilijker dan hier theoretisch wordt aangegeven in deze figuur ontleend aan Castellano en Ho (2013), zoals we verder aangeven, ondermeer in secties 2.5.2 en 3.3.2.

[4] Noteer dat deze berekening van de leerwinst voor een groep van leerlingen in principe berekend wordt als het gemiddelde van de individuele leerwinst, maar dat hetzelfde cijfer wordt bekomen door het verschil te nemen tussen de gemiddelde status van dezelfde leerlingen op beide momenten (het maakt geen verschil op welk punt de cijfers worden geaggregeerd).

normaalvorderende schoolloopbaan binnen deze school volgen (cf. voetnoot 3), maken de leerlingen die in 2016 in het eerste leerjaar zitten, tegen het vierde leerjaar een leerwinst van gemiddeld 300 punten (620-320=300).

De vergelijking van de vaardigheidsscores over leerjaren in één bepaald jaar (kolom in Tabel 2) is een statusmeting en geen leerwinstmeting, omdat het gaat om metingen op één en hetzelfde moment waarbij leerlingengroepen uit verschillende leerjaren worden vergeleken, maar niet worden opgevolgd over de tijd. Tot slot kunnen ook verschillende leerlingcohorten over de tijd worden vergeleken, in een bepaald leerjaar (rij in de Tabel 2). Opnieuw gaat het dan niet om leerwinst, aangezien het opnieuw een vergelijking van verschillende leerlinggroepen betreft en niet de opvolging van eenzelfde leerlinggroep over de tijd.[5] Wel kan op die manier verbetering over de tijd worden vastgesteld op schoolniveau, ook gekend als een weergave van trends of evolutie (op basis van een vergelijking van meerdere statusmetingen).

Tabel 2: Leerwinst in relatie tot jaar, leerjaar en cohorte (naar Castellano & Ho, 2013)

	Jaar					
Leerjaar	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
1	320	380	350	400	390	420
2	400	450	420	450	480	500
3	510	550	600	650	620	620
4	610	620	630	620	650	660
5	710	780	750	750	800	800
6	810	810	820	820	810	840

[5] Hoewel niet gebruikelijk in onderwijseffectiviteitsliteratuur, is het onderscheid dat soms gemaakt wordt tussen panel- en longitudinale data hier relevant. Bij longitudinale data gaat het dan om meerdere metingen over de tijd, maar niet noodzakelijk voor dezelfde basiseenheden (hier: de verschillende leerlingcohorten). Paneldata zijn een specifiek subtype van longitudinale data waarbij de basiseenheden die herhaaldelijk gemeten worden dezelfde zijn (hier: dezelfde leerlingen). Om leerwinst te kunnen meten is er met andere woorden nood aan paneldata.

Er kan worden opgemerkt dat het begrip trend hier preciezer is dan dat ze soms wordt gebruikt. Zo worden de resultaten van verschillende internationale studies (PISA, TIMSS, PIRLS) en de peilingen ook gepresenteerd als trends. Daarbij dienen we rekening te houden dat het steeds leerlingen zijn van verschillende scholen (want elke meting is gebaseerd op een andere steekproef) met tussenliggende intervallen van meerdere jaren. De trends bieden dan een tentatief beeld van de evolutie op het niveau van het onderwijssysteem, terwijl de hier voorgestelde trendmeting (die aansluit bij deze van de centrale toetsen) betrekking heeft op schoolniveau en een snellere monitoring toelaat.

1.2.3 Meerwaarde van meting van leerwinst

Leerwinst wordt als een meer valide indicator beschouwd dan status om de bijdrage van de school aan het leren van leerlingen vast te stellen. Theoretisch gezien kan een leerling immers bijvoorbeeld heel goede toetsresultaten halen in een bepaald leerjaar zonder dat deze leerling veel vooruitgang maakte ten opzichte van een eerder moment. Omgekeerd kan het zijn dat een school minder goede eindresultaten behaalt met haar leerlingen dan een andere school, maar wel sterkere leerwinst laat optekenen. Uit onderzoek blijkt in elk geval dat leerwinst op schoolniveau inderdaad meer dan status samenhangt met kenmerken van de school die wijzen op de inzet van scholen en minder dan status met achtergrondkenmerken van het leerlingenpubliek (Heck, 2006).

In verschillende landen met een (lange) traditie van evaluatie van leerkrachten of scholen op basis van toetsprestaties van leerlingen, heeft kritiek op een statusgebaseerde evaluatie van de onderwijskwaliteit ertoe bijgedragen dat proefprojecten werden opgestart of stappen werden gezet naar een evaluatie op basis van metingen van de vooruitgang of leerwinst die leerlingen maken in het onderwijs, zoals in de VS (Heck, 2006), Nederland (Janssens et al., 2014), Verenigd Koninkrijk (Leckie & Goldstein, 2017) en Australië (Zoanetti, 2021).

Conceptueel gezien biedt het meten van de vooruitgang die leerlingen maken een duidelijker antwoord op de vraag naar onderwijseffectiviteit dan enkel het meten van het vaardigheidsniveau van leerlingen. Door de vooruitgang van leerlingen als maatstaf te gebruiken, verschuift de focus van de status op een bepaald moment naar het proces van leren en daarmee naar datgene dat een school kan toevoegen aan mogelijk vooraf bestaande vaardigheidsverschillen (zie ook verder sectie 1.3 over toegevoegde waarde van scholen). Dat sluit ook nauw aan bij de onderwijspraktijk waarin leraren aan de hand van toetsen de verwerving en beheersing van leerstof door leerlingen monitoren (Heck, 2006). Al is in dat geval de frequentie van toetsing en opvolging van de leerling van een heel andere orde dan bij de afname van centrale toetsen.

Om verschillende redenen wordt een evaluatie op basis van vooruitgang van leerlingen bovendien in een aantal Angelsaksische landen vaak beschouwd als een meer *faire* manier om de onderwijsresultaten van leraren en scholen te beoordelen. Het erkent immers expliciet dat voor leerlingen die achteroplopen op hun klasgenoten meer onderwijstijd en -middelen moeten worden ingezet om bepaalde (minimum)doelstellingen te halen. Het maakt ook zichtbaar wanneer leerlingen die mogelijk niet een bepaalde (minimum)doelstelling behaalden toch een behoorlijke vooruitgang maakten (Heck, 2006).

Naast het verleggen van de conceptuele focus naar de vooruitgang die leerlingen maken, biedt de benadering op basis van leerwinst van leerlingen ook een aantal meer technische voordelen in vergelijking met trendanalyses op basis van statusmetingen van opeenvolgende cohorten. Concreet gaat het over accuratere schattingen voor kleinere groepen en lagere gevoeligheid voor steekproeffluctuaties (Heck, 2006).

1.2.4 Betekenis en interpretatie van leerwinst (uitdagingen)

Waar bij een statusmeting leerlingen met hogere eerdere vaardigheden gemakkelijker een einddoel of benchmark halen, verschuift de vraag bij een focus op individuele leerwinst naar de hoeveelheid vooruitgang die van leerlingen wordt verwacht. We behandelen hier hoe groot de vooruitgang dient te zijn en de evaluatie van die vooruitgang in relatie tot einddoelen dan wel tot de uitgangspositie.

1.2.4.1 Hoeveel leerwinst is voldoende?

Bovenstaande benadering van leerwinst (verschilscores of groeitempo) laat toe om de grootte van de vooruitgang van een leerling of groep van leerlingen (vb. klas of school) via een eenvoudig getal uit te drukken. Ondanks het voordeel dat men via leerwinst een beeld krijgt van de grootte van de gemaakte individuele of gemiddelde vooruitgang, blijft de vraag hoeveel vooruitgang dan als voldoende kan worden beschouwd. Expertenpanels kunnen zowel norm- als schaalgericht bepalen hoeveel leerwinst als voldoende leerwinst kan worden beschouwd (Castellano & Ho, 2013). In het eerste geval kan bijvoorbeeld worden aangegeven dat de vijf procent laagst scorende scholen onvoldoende leerwinst behalen. In het andere geval kan bijvoorbeeld worden gesteld dat de leerwinst bij elke leerling een minimaal aantal punten op de meetschaal moet bedragen. Ook in het geval van deze laatste schaalgebaseerde doelstelling voor leerwinst, blijft de vooropgestelde waarde toch veeleer arbitrair, aangezien geen conceptueel criterium beschikbaar is op basis waarvan externen de minimale vooruitgang kunnen inschatten.

1.2.4.2 Leerwinst- en einddoelen?

Het is ook onduidelijk hoe een maatstaf voor leerwinst consistent kan worden geformuleerd ten aanzien van te behalen einddoelen (eindtermen) en eventuele vooropgestelde tussenliggende cesuren (bv. basisgeletterdheid B-stroom versus behalen van de eindtermen A-stroom in het tweede jaar van het secundair onderwijs). Die combinatie wordt wel gemaakt in doelgebaseerde interpretaties van leerwinst waarbij (in complexere modellen - die ook niet meteen van toepassing zijn voor de centrale toetsen) wordt nagegaan of leerlingen op een groeitraject zitten waarbij ze ook een bepaald einddoel zullen bereiken, gegeven hun startpositie (Castellano & Ho, 2013; Martineau, 2016). Eindresultaten en leerwinst vormen bijgevolg aparte indicatoren van schoolkwaliteit, die voor een goed beeld van schoolkwaliteit bij voorkeur worden gecombineerd (Heck, 2006). Het gaat om de gemiddelde status die leerlingen uiteindelijk behalen, maar ook hoe groot het groeitraject is geweest voor een groep. Groepen die gemiddeld lager starten maar uiteindelijk eenzelfde gemiddelde eindscore behalen als leerlingen uit een andere school hebben immers een veel langere weg afgelegd.

1.2.4.3 Leerwinst verschillend voor leerlinggroepen?

Waar een (eenzijdige) focus op leerwinst soms bekritiseerd wordt uit vrees voor de verlaging van standaarden voor bepaalde scholen of subgroepen doordat men einddoelen uit het oog verliest, benadrukken anderen dat leerwinst niet dezelfde hoeft te zijn voor verschillende leerlingen, leerlinggroepen of scholen. Zo lijkt er wel consensus dat voor bepaalde vaardigheden of opleidingsonderdelen niet dezelfde leerwinst hoeft te worden verwacht over verschillende leerjaren heen. Dit kan begrepen worden vanuit het onderwijscurriculum en met name de mate waarin in een bepaald leerjaar wordt ingezet op de ontwikkeling van een specifieke vaardigheid. Een gelijkaardige redenering leidt eveneens tot mogelijke verschillende standaarden voor leerwinst voor leerlingen in bijvoorbeeld verschillende studierichtingen of onderwijsvormen.

Leerwinst kan niet alleen verschillen als gevolg van dergelijke verschillen en klemtonen die in het curriculum worden gelegd. Verschillende auteurs wijzen ook op een mogelijke samenhang met de aanvangspositie van leerlingen. De vergelijking van de leerwinst van groepen die verschillen in hun aanvangsprestaties is daardoor minder eenduidig dan het lijkt (Raudenbush, 2013; Harris & Anderson, 2013). Zo is er het regelmatig vastgesteld patroon van (gemiddeld) hogere leerwinst bij leerlingen met lagere aanvangsprestaties dan bij leerlingen met hogere aanvangsprestaties: "students with prior scores higher than the population mean will systematically tend to show lower gains, and vice versa." (Zoanetti 2021: 6). Uit dit patroon kan niet noodzakelijk worden afgeleid dat de ene groep haar achterstand inloopt over de tijd. Convergentie tussen groepen kan ook het gevolg zijn van plafondeffecten voor een bepaalde

(naar boven begrensde) vaardigheid (Van den Broeck, 2014). Eveneens kan het een louter statistisch artefact zijn, als gevolg van de technische kwaliteiten van de meetschaal (Betebinner & Linn, 2009) en het fenomeen van "regressie naar het gemiddelde" als gevolg van meetfouten (Zoanetti, 2021). Om deze redenen wordt het gebruik van eenzelfde standaard voor het vergelijken van de leerwinst van groepen met grote verschillen in hun voorafgaande prestaties vaak ook als *unfair* gepercipieerd. Op basis hiervan bepleiten een aantal auteurs (Zoanetti, 2021; Betebenner & Linn, 2009) normgebaseerde vergelijkingen voor leerwinst waarbij de leerwinst van leerlingen wordt vergeleken met die van leerlingen met vergelijkbare eerdere prestaties. Zo komen we hier tot de conclusie dat leerwinst best wordt geïnterpreteerd in combinatie met het startniveau.

1.2.4.4 Schooleffect?

Tot slot vertelt leerwinst op zich onvoldoende over de bijdrage die een school levert aan het leren van de leerling. De loutere aggregatie van de hoeveelheid gemeten leerwinst op schoolniveau, zegt niet noodzakelijk iets over de rol en effecten van scholen. De leerwinst van leerlingen kan immers ook te wijten zijn aan andere actoren en factoren die buiten de invloedssfeer van de school liggen, zoals het leren buiten de school en de sociaal-economische en demografische kenmerken van de leerlingen. Bijgevolg is leerwinst op zich dus onvoldoende om door scholen behaalde resultaten te evalueren en kan louter via leerwinst geen eerlijke vergelijking gemaakt worden tussen scholen (OECD, 2008). Dit wordt meer uitgebreid besproken in de volgende paragraaf die ingaat op de bepaling van de toegevoegde waarde van scholen.

1.3 TOEGEVOEGDE WAARDE

Toegevoegde waarde wordt binnen effectiviteitsonderzoek gedefinieerd als de mate waarin een school[6] bijdraagt aan het leren of de leerprogressie van haar leerlingen gedurende een bepaalde periode en in

[6] Aansluitend bij het regeerakkoord beperken we ons binnen deze nota tot een omschrijving van toegevoegde waarde op schoolniveau. Voor een beschrijving van de toegevoegde waarde van leraren en de daarmee gepaard gaande potentiële methodologische problemen om deze te modelleren verwijzen we graag naar Amrein-Beardsley en Geiger (2019), Everson (2017), Koedel, Mihaly & Rockoff (2015) en Rothstein (2009).

vergelijking met andere scholen uit dezelfde scholenpopulatie (Sammons et al., 1997). Hoewel deze definitie ook verwijst naar het leren en de progressie van leerlingen, is de toegevoegde waarde van een school toch verschillend van de gemiddelde leerwinst van een school. Bij de bepaling van de toegevoegde waarde van een school gaat het immers primair om de vraag naar de mogelijke oorzaken van de vooruitgang die leerlingen maken tijdens hun schoolloopbaan, en de rol van de school daarin (OECD, 2008). Hoewel het begrip toegevoegde waarde ook gebruikt wordt bij statusmetingen[7], gaan we in de verdere bespreking uit van een toegevoegdewaardebepaling op basis van een leerwinstmeting, waarbij eenzelfde vaardigheid dus werd gemeten op (minstens) twee toetsmomenten (cf. definitie hierboven). In het licht van de hierboven aangehaalde voordelen van een leerwinstmeting, biedt deze immers een beter vertrekpunt om de bijdrage van scholen vast te stellen (zie ook verder bij het type-AA schooleffect). De wijze waarop die leerwinst statistisch wordt gemodelleerd kan verschillen en behandelen we in sectie 2.2 in deze nota. Op de (afwijkende) interpretatie van toegevoegde waarde die niet is gebaseerd op een leerwinstmeting, wordt ingegaan in sectie 2.3. Met name voor de bijzondere case die hiervoor nog niet werd onderscheiden (maar wel gangbaar is in bepaalde landen) waarbij een bepaalde vaardigheid van een leerling wel meerdere keren over de tijd wordt getoetst, maar waarbij de toetsresultaten over de tijd niet op een en dezelfde meetschaal staan en dus niet voldaan is aan de voorwaarde van longitudinale meetinvariantie om van vooruitgang of groei te spreken (zie sectie 1.2.1).

Via een schatting van de toegevoegde waarde wordt in hoofdzaak nagegaan in welke mate een wijziging in leerprestaties van leerlingen over de tijd, toegewezen kan worden aan de school en de inspanningen van die school om vooruitgang te boeken bij haar leerlingen. Het leren van leerlingen wordt immers niet alleen beïnvloed door de school, maar ook door leerling- en contextkenmerken waar de school geen invloed op heeft (Leckie & Goldstein, 2019; Levy et al., 2019). Aangezien scholen geen vat hebben op deze leerling- en contextkenmerken, zijn ze ook niet 'verantwoordelijk' voor de effecten die deze leerling- en contextkenmerken mogelijks hebben op de leerprestaties van leerlingen (Harris, 2011). Dergelijke

Merk ook op dat in tegenstelling tot status en leerwinst, toegevoegde waarde dus geen betrekking heeft op het individueel leerlingniveau. Het betreft immers de bijdrage van een instantie (school of leraar) aan het leren van de leerling.

[7] Om het onderscheid helder te houden, spreken we in dat geval liever niet van toegevoegde waarde maar van een gecontextualiseerde statusmeting (cf. Lenkeit, 2013).

confounding factors[8] vertekenen het beeld dat we krijgen op basis van vastgestelde schoolverschillen. Twee scholen met vergelijkbare leerwinst hebben immers niet noodzakelijk op dezelfde wijze bijgedragen tot die leerwinst. In de ene school realiseerden bijvoorbeeld meer leerlingen een deel leerwinst door buitenschoolse stimulansen of ondersteuning (bv. vanuit het thuismilieu), terwijl een andere school daar minder op kan voortbouwen en dus meer inspanningen moet leveren om dezelfde vooruitgang in leerprestaties te boeken met haar leerlingen.

De term toegevoegde waarde verwijst dan naar dat deel van de vooruitgang in leerprestaties van leerlingen dat gerealiseerd wordt door de effecten van de school zelf. Dat schooleffect, de zogenaamde toegevoegde waarde van een school, is de vooruitgang die leerlingen maken, rekening houdend met leerling- en contextkenmerken die buiten de invloedssfeer van de school liggen. Een manier om die factoren in rekening te brengen is door de leerwinst van scholen te vergelijken met de leerwinst van scholen met dezelfde kenmerken. In de praktijk wordt die vergelijkbaarheid bekomen door te corrigeren voor de impact van de betrokken variabelen door middel van statistische controle. Statistische controle impliceert dat de toegevoegde waarde van een school wordt bepaald op basis van het verschil tussen de gemiddelde vooruitgang in leerprestaties van de leerlingen van de school en de leerwinst die op schoolniveau verwacht kan worden gegeven de leerling- of andere in rekening gebrachte kenmerken. Anders gesteld, de toegevoegde waarde van een school toont of de gemiddelde vooruitgang in leerprestaties of leerwinst van een school beter of minder goed is dan op basis van bijvoorbeeld de achtergrondkenmerken van de leerlingen kan verwacht worden.

Die overblijvende schoolverschillen worden courant uitgedrukt ten opzichte van een globaal gemiddelde. Een school waarin de leerlingen meer vooruitgaan in leerprestaties dan andere scholen met dezelfde kenmerken heeft dan een positieve toegevoegde waarde, terwijl een school met leerlingen die minder vooruitgaan dan gemiddeld een negatieve toegevoegde waarde heeft. De toegevoegde waarde van een school kan ook uitgedrukt worden als de leerwinst die een school behaalt voor een bepaalde benchmark groep van leerlingen met een bepaalde startscore en achtergrondkenmerken, en gegeven bepaalde schoolkenmerken (zie 2.3 over relatieve versus absolute schooleffecten).[9]

---

[8] *Confounding factoren* zijn in dit geval alle factoren die samenhangen met de scholen en een onafhankelijke impact hebben op de leerprestaties van leerlingen (gemeenschappelijke oorzaak voor behoren tot de school en de leerprestaties, maar geen causaal pad van de school via de factoren naar de leerprestaties).

[9] Merk op dat ongeacht de operationalisatie (op basis van verschillen tussen scholen, dan wel de resultaten voor een benchmark groep) het theoretische concept van toegevoegde waarde (wat voegt een school toe aan het leren

Meer inhoudelijk onderscheiden toegevoegdewaardemodellen zich van elkaar op basis van de leerling- en schoolkenmerken waarvoor gecorrigeerd wordt. Een cruciale kwestie is immers de vraag met welke variabelen rekening moet worden gehouden om een eerlijke vergelijking tussen scholen mogelijk te maken. Afhankelijk van de doelstelling wordt beoogd te corrigeren voor specifieke *confounding factors* om het bedoelde schooleffect zo goed mogelijk te isoleren.[10] Richtinggevend daarbij zijn in de praktijk ook het gehanteerde effectiviteitsmodel (cf. het in de literatuur vaak gehanteerde CIPO-model, Scheerens, 1990) en databeschikbaarheid. Afhankelijk van de kenmerken waarmee rekening wordt gehouden om de toegevoegde waarde van scholen te bepalen, definiëren Timmermans, Doolaard en de Wolf (2011) vier verschillende soorten schooleffecten, met name type-AA, type-A, type-B en type-X effecten. Elk type schooleffect dient geïnterpreteerd te worden ten opzichte van de specifieke variabelen die in het model werden opgenomen. De betekenis van het schooleffect - en dus ook de betekenis van de geschatte toegevoegde waarde van een school - verschilt dus naargelang het gebruikte model. Bijgevolg is ook niet elk model geschikt om tot bepaalde besluitvorming te komen.

Tabel 3: Overzicht van de types schooleffecten (naar Timmermans et al., 2011)

Controle voor … (cumulatief)	**Type schooleffect (netto-effect)**
Eerdere prestaties	AA
Achtergrondkenmerken leerlingen (eerdere prestaties, SES, geslacht, thuistaal, …)	A
Compositiekenmerken (gemiddelde SES, verdeling geslacht, thuistaal, …)	B
Niet-beïnvloedbare schoolkarakteristieken (schoolgrootte, financiering, …)	X
Ongecontroleerde kenmerken	
Beïnvloedbare school(proces)kenmerken (vb. instructie)	toegevoegde waarde school

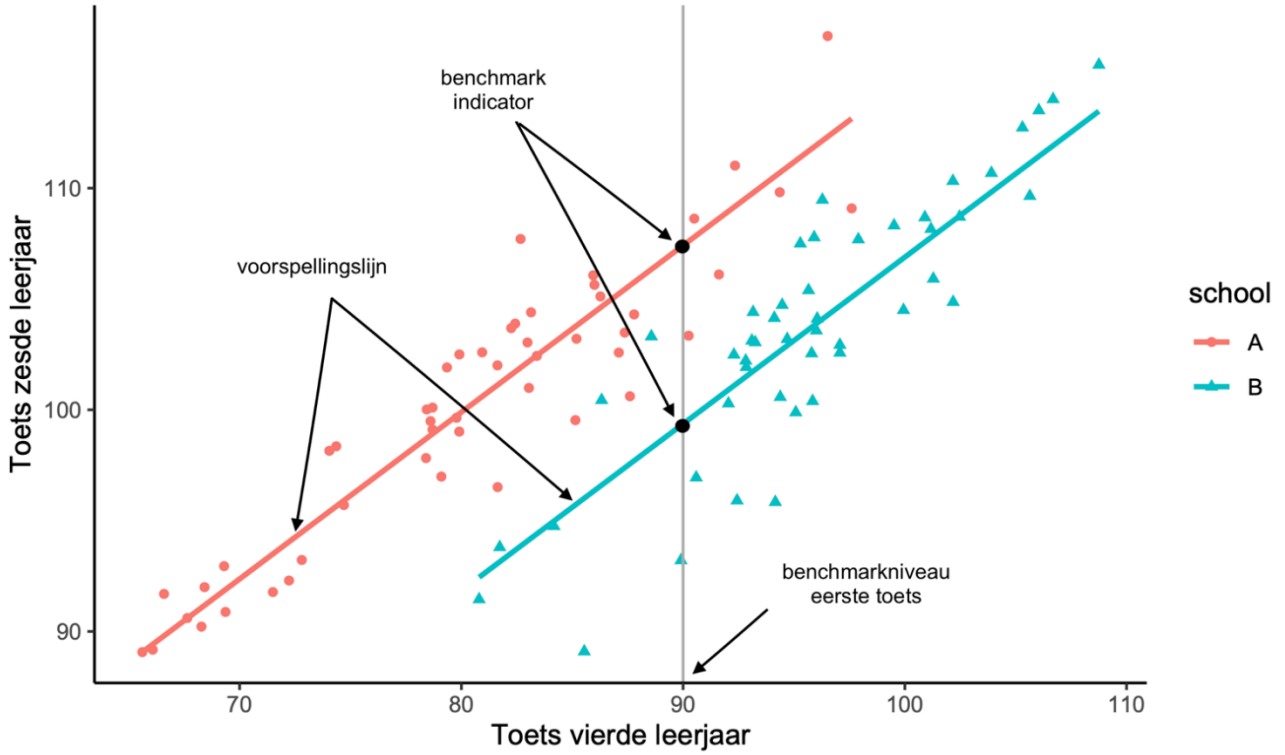
van leerlingen) afwijkt van de operationalisatie van toegevoegde waarde (die vertrekt van een vergelijking met andere scholen). Een vergelijking met hoe leerlingen het zouden doen mochten ze geen onderwijs krijgen, wordt immers niet geobserveerd, cf. *potential outcomes counterfactual framework* als basis voor toegevoegde waarde modellen (Reardon & Raudenbush, 2009).

[10] Merk op dat bij deze aanpak niet de kenmerken van effectieve scholen (kwaliteit en inspanningen) worden opgenomen, enkel de andere factoren die kunnen bijdragen aan de leerwinst (onafhankelijke impact), zie ook Deel 2 punt over bias en predictoren.

1.3.1 Eerdere prestaties en het type-AA schooleffect

Bij een type-AA schooleffect wordt de onderwijsuitkomst van leerlingen op schoolniveau gecontroleerd voor de eerdere prestaties van de leerlingen. Bij dit type schooleffect gaat het om de gemiddelde leerprestatie van leerlingen van een school gecorrigeerd voor hun leerprestaties gemeten op een eerder toetsmoment (nulmeting). Uitgaand van een leerwinstmeting (zoals hierboven), impliceert de modellering die hieraan aan de basis ligt (zie ook Deel 2) dat we het geschatte schooleffect ook kunnen omschrijven als de schatting van de leerwinst die een school boekt voor leerlingen met een specifieke uitgangspositie. Vaak spreekt men van een begintoets en een eindtoets, maar het kan bijvoorbeeld ook gaan om de vooruitgang in leerprestaties van de leerlingen van het zesde leerjaar lager onderwijs ten opzichte van de leerprestaties van diezelfde leerlingen in het vierde leerjaar. We kunnen in dat voorbeeld van een positieve toegevoegde waarde spreken wanneer de leerlingen van een school gemiddeld betere toetsresultaten hebben in het zesde leerjaar dan verwacht (gemiddeld) op basis van hun vaardigheidsscores in het vierde leerjaar.

Figuur 3: Illustratie van leerwinstmeting op basis van vooruitgang ten opzichte van eerdere prestatiemeting op schoolniveau (naar Meyer, 1996)



Het type-AA model biedt dan ook een antwoord op de vraag hoe goed leerlingen in een school het doen ten opzichte van leerlingen met gelijkaardige startprestaties in andere scholen. Deze logica wordt visueel geïllustreerd in Figuur 3. In de figuur worden (hypothetische) toetsscores voor twee metingen weergegeven voor leerlingen van twee scholen. De gemiddelde toetsscore bij de tweede meting is groter voor school B dan die van school A. De leerwinst is echter groter voor school A wat blijkt uit de hogere eindscores in school A dan school B wanneer leerlingen met eenzelfde aanvangscore worden vergeleken (benchmark groep met aanvangsscore van 90). Omdat school A evenwel meer leerlingen heeft met een lagere aanvangsscore, wordt de bijdrage van deze school onderschat op basis van de gemiddelde (ongecorrigeerde) eindscore.

In vergelijking met een statusmeting voor één meetmoment maakt het type-AA schooleffect een meer eerlijke vergelijking tussen scholen mogelijk (Leckie & Prior, 2022). Het feit dat er enkel voor

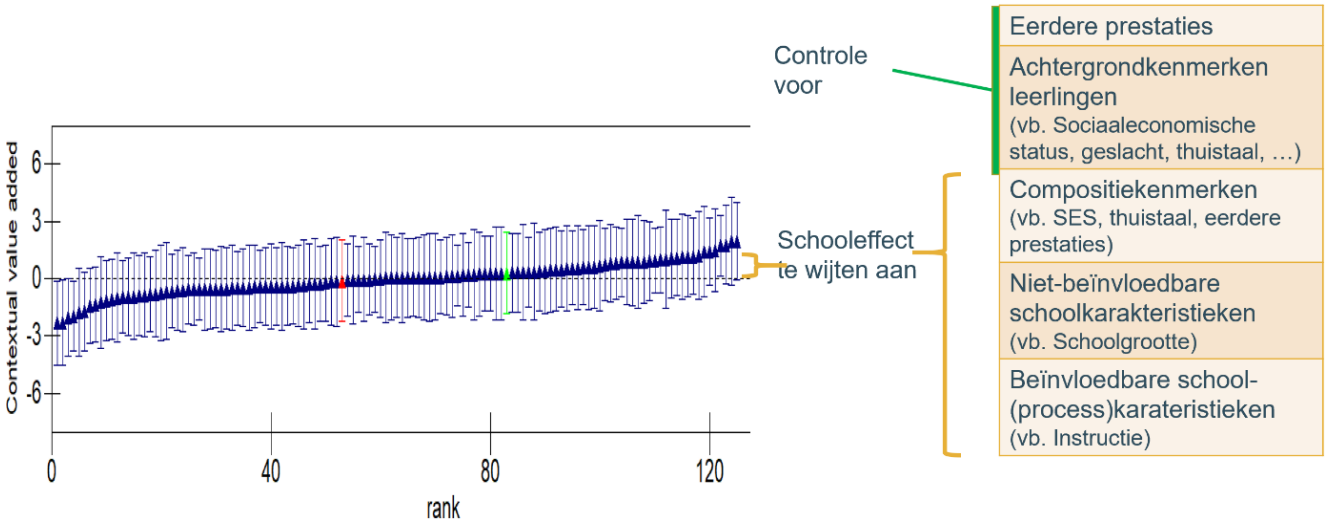
aanvangsprestaties wordt gecontroleerd zorgt er echter ook voor dat de diagnostische waarde van de vastgestelde schooleffecten onder dit model laag is. Deze berekende toegevoegde waarde kan immers niet alleen te wijten zijn aan proceskenmerken of beïnvloedbare karakteristieken op schoolniveau, maar evengoed aan demografische kenmerken van individuele leerlingen, schoolcompositiekenmerken of niet-beïnvloedbare schoolkenmerken. Daarom dat voor een betere schatting van de toegevoegde waarde bijkomend ook wordt gecontroleerd voor andere factoren.

1.3.2 Achtergrondkenmerken en het type-A schooleffect

Bij een type-A schooleffect worden naast de eerdere leerprestaties van de leerlingen ook sociaaldemografische achtergrondkenmerken van de leerlingen toegevoegd zoals geslacht, etnische herkomst of thuistaal, of meer algemeen sociaal milieu (Ballou et al., 2004, Raudenbusch & Willms, 1995). Uitgaande van het eerdere voorbeeld betekent een positieve toegevoegde waarde dat de leerlingen van een school gemiddeld beter scoren in het zesde leerjaar dan gemiddeld gegeven hun prestaties in het vierde leerjaar en hun persoonlijke achtergrondkenmerken. Via het type-A model wordt een antwoord geboden op de vraag hoe goed leerlingen in een school het doen ten opzichte van gelijkaardige leerlingen in andere scholen. Omdat het iets zegt over de gemiddelde leerwinst die een school naar verwachting kan realiseren voor een bepaalde referentieleerling is dit A-type schooleffect, meer dan het AA-type, met name relevant voor leerlingen en ouders in het kader van schoolkeuze (Meyer, 1996, maar zie ook Leckie & Goldstein 2009).

Een veel gehoorde kritiek is dat de correctie voor achtergrondkenmerken het onderpresteren van bepaalde groepen van leerlingen “verbergt” en daardoor legitimeert (Prior et al., 2021). Terwijl de controle voor de sociaaldemografische kenmerken noodzakelijk is om een eerlijke vergelijking tussen leerlingen en scholen mogelijk te maken, zorgt dit er ook voor dat de mate waarin bepaalde groepen (vb. jongens vs. meisjes) onderpresteren wordt weggecorrigeerd (Leckie & Goldstein, 2017). Critici stellen dat deze correctie van totale schoolverschillen lage verwachtingen ten aanzien van bepaalde groepen uitspreekt en daardoor bestendigt (bv. Marks, 2021). Aanwijzingen voor deze effecten werden in onderzoek evenwel niet gevonden (Leckie & Goldstein 2017). Voor een transparante rapportering verdient het daarom wel aanbeveling om de resultaten van een type-A model steeds te combineren met de niet voor achtergrond gecontroleerde schooleffecten (cf. type-AA) (Leckie & Goldstein, 2019; OECD, 2008; Tekwe et al., 2004).

Figuur 4: Illustratie van type-A schooleffecten (Aesaert & Janssen, 2021).



Door rekening te houden met hoe de eerdere toetsscore en persoonlijke achtergrondkenmerken van de leerlingen samenhangen met de prestaties in het zesde leerjaar, wordt de toegevoegde waarde hier als een betere schatting van het effect van de school beschouwd. Maar zelfs bij het in rekening brengen van de eerdere toetsscore en alle mogelijke relevante achtergrondkenmerken, overschat het type-A model mogelijk nog de toegevoegde waarde van een school. De aard van het effect is immers nog steeds meerduidig. Het kan zowel veroorzaakt worden door de schoolcompositie, door niet- of minder beïnvloedbare schoolkenmerken, zoals schoolgrootte, alsook door beïnvloedbare schoolkenmerken, zoals het beleid inzake het onderwijsleerproces voor wiskunde (zie Figuur 4). Omdat schoolcompositie ook schooleffecten kan hebben en meer algemeen omdat een type-A effect niet enkel wordt bepaald door "schoolpraktijken" is het type-A effect niet geschikt in het kader van schoolontwikkeling en verantwoording (Raudenbusch & Willms, 1995).

1.3.3 Schoolcompositie en het type-B schooleffect

Ook de schoolcompositie kan de leerlingprestaties beïnvloeden. Zo is het mogelijk dat niet alleen de sociaaleconomische status (SES) van de individuele leerling, maar ook de schoolcompositie (of klascompositie) op basis van die SES, een invloed heeft op de prestaties of de vooruitgang van de leerlingen (Timmermans, Doolaard, & de Wolf, 2011). In een school waar een groot aandeel van de leerlingen de instructietaal bijvoorbeeld niet beheerst, bestaat de kans dat leerkrachten tot minder goede instructie komen waardoor de effectiviteit van de school de facto lager ligt. Omgekeerd is het ook mogelijk dat het merendeel van de leerlingpopulatie van een school afkomstig is uit gezinnen met een

sterk cognitief stimulerend thuisklimaat. In die gevallen verklaart de compositie waarschijnlijk ook een deel van de type-A schooleffecten (OECD, 2008).

Om een type-B schooleffect te schatten, worden naast de eerdere prestaties van de leerlingen en hun sociaaldemografische kenmerken ook schoolcompositiekenmerken op basis van eerdere prestaties en/of demografische kenmerken als predictoren in het model geïntegreerd. Door rekening te houden met schoolcompositie-effecten, zoals het positieve effect voor een leerling om op een school met hoofdzakelijke sterke medeleerlingen te zitten, wordt de schatting van de toegevoegde waarde van een school binnen dit type model beschouwd als een betere indicator van schoolkwaliteit (Timmermans & Thomas, 2015). Via het type-B model wordt dan ook een antwoord geboden op de vraag hoe goed leerlingen in een school het doen ten opzichte van *gelijkaardige leerlingen* in andere *gelijkaardige scholen*.

In de mate dat een type-B effect iets meer vertelt over de mate waarin de schoolpraktijken bijdragen aan het leren van gelijkaardige leerlingen in gelijkaardige scholen is dit model meer geschikt in functie van schoolontwikkeling en accountability dan het type-A effect (Ehlert et al., 2016). Deze schoolpraktijken kunnen weliswaar nog steeds binnen als buiten de invloedssfeer van de school vallen. Daarom dat dit type-B effect verder gecorrigeerd kan worden voor andere, minstens op korte termijn, schoolexterne factoren.

1.3.4 Externe schoolkenmerken en het type-X schooleffect

In een type-X model wordt tot slot bijkomend rekening gehouden met schoolkenmerken die scholen op korte termijn niet kunnen veranderen, zoals financiering, schoolgrootte of locatie. Deze geschatte toegevoegde waarde of een type-X schooleffect verwijst dan nog louter naar door scholen beïnvloedbare schoolpraktijken, zoals organisatorische structuren, geaggregeerde instructiepraktijken of schoolleiderschap (OECD, 2008). Via het type-X effect beantwoordt men de vraag hoe goed leerlingen in een school het doen ten opzichte van gelijkaardige leerlingen in gelijkaardige scholen met overeenkomstige niet door scholen beïnvloedbare schoolkenmerken.

De vraag welke factoren relevant zijn om rekening mee te houden als we spreken van schooleffecten, is deels een wetenschappelijke, maar deels ook een beleidsmatige vraag. Voor een beleidsanalyse zijn de variabelen met *sterke* verklaringskracht niet steeds de meest interessante, omdat ze zelden aangrijpingspunten bieden om op in te grijpen (bv. SES van de leerlingen van de school). Om verandering te kunnen bewerkstelligen, zijn meer specifiek de *manipuleerbare* sterke variabelen van belang (Ellemers, 1976).

In een aantal onderwijsevaluatiesystemen wordt het type-X, ook het type-B, schooleffect niet geschat, omdat men vreest voor statistische overcompensatie (zie sectie 2.5.1 over bias en predictoren). Net door de kenmerken die deel uitmaken van een vooropgesteld schooleffect te integreren bij de schatting, kunnen schooleffecten evenwel beter geschat worden (cf. Meyer, 1996).

Een fijnmazig begrip van elementen van school en schoolcontext laat ook toe om schoolontwikkeling beter te beheren. Welke factoren gelden als manipuleerbare variabelen is verschillend voor het meso- (klas of school) en het systeemniveau. Binnen de Vlaamse context laat het decretaal kader voor de omkadering van scholen, van het lerarenstatuut en de financiering misschien weinig ruimte voor een individuele school. Door hier rekening mee te houden, krijgen we scherper zicht op schoolkenmerken die als hefboom kunnen dienen.

2 STATISTISCHE EN METHODOLOGISCHE AANDACHTSPUNTEN

In dit deel gaan we in op de statistische en methodologische aandachtspunten bij de meting van leerwinst en een daarop gebaseerde schatting van toegevoegde waarde van scholen. Zo stelt een meting van leerwinst hoge eisen aan de toetsen (zie sectie 2.1). Gehanteerde analysemodellen kunnen verschillen in de wijze waarop ze de leerwinst of individuele voortgang van leerlingen modelleren (o.m. residuele leerwinst of verschilscore), evenals in de methode van schatting van schooleffecten (o.m. multilevel en shrinkage) (zie 2.2), en in de betekenis daarvan (bv. schaalgerelateerd versus relatieve verschillen) (zie 2.3). Tot slot gaan we in op een aantal factoren die een rol spelen voor de betrouwbaarheid van de schooleffecten (statistische onzekerheid en bias) (in secties 2.4 en 2.5).

2.1 STRENGE EISEN AAN TOETSEN

Hoewel leerwinst of de individuele vooruitgang van leerlingen een aantal conceptuele voordelen heeft ten opzichte van een statusmeting, stelt de herhaalde meting van een bepaalde vaardigheid een aantal bijkomende uitdagingen. Ten eerste is er de voorwaarde dat de gemeten leerprestaties vergelijkbaar zijn over de tijd. Dit betekent dat dezelfde vaardigheid over de verschillende meetmomenten heen op dezelfde schaal kan worden gemeten. Enkel wanneer aan de voorwaarde van longitudinale meetinvariantie is voldaan (in de termen van Item Respons Theorie: de afwezigheid van *differential item functioning* (DIF) over de tijd), kunnen gemeten veranderingen of verschilscores tussen twee momenten eenduidig worden geïnterpreteerd als veranderingen over de tijd (Horn & McArdle, 1992). De eenvoudigste manier om aan deze voorwaarde te voldoen is hetzelfde toetsinstrument gebruiken bij elke afname (hoewel longitudinale meetinvariantie ook dan niet noodzakelijk volgt). Deze aanpak is echter niet wenselijk binnen de centrale toetsen. Voor opeenvolgende statusmetingen van een bepaald (bijvoorbeeld vierde) leerjaar komt bij het gebruik van identieke toetsen de toetsveiligheid in het gedrang. Items kunnen dan immers bekend raken voor de afname en tot (ongewenste) teaching to the test leiden, waardoor de toetsafname niet langer louter de bedoelde vaardigheid van een leerling toetst. Wanneer eenzelfde leerling tijdens de schoolloopbaan identiek dezelfde toets zou afleggen, kunnen door vertrouwdheid met het toetsinstrument gelijkaardige ongewenste leereffecten optreden. Voor statusmetingen van verschillende leerjaren (bijvoorbeeld vierde en zesde leerjaar), kunnen bovendien problemen optreden met betrekking tot de inhoudsvaliditeit. Wanneer de te toetsen leerinhoud verschilt over de twee leerjaren, is eenzelfde toets immers niet mogelijk.

Om leereffecten uit te sluiten en om toetsen beter af te stemmen op de leeftijd of het leerjaar, worden daarom (deels) andere toetsitems voorgelegd aan leerlingen op de verschillende meetmomenten tijdens de schoolloopbaan. Om deze zinvol te kunnen vergelijken, dienen de toetsen en geschatte vaardigheidsscores over de momenten heen (bv. vierde en zesde leerjaar basisonderwijs) op één gemeenschappelijke meetschaal te worden geplaatst. Dat gebeurt aan de hand van modellen ontwikkeld binnen het domein van de item response theorie (IRT). Een identieke toets is dan niet noodzakelijk, wel is er nood aan een aantal items die tijdens beide meetmomenten worden bevraagd, die bovendien ook moeten kunnen fungeren als zogenaamde "ankeritems".[11] Om tot een betrouwbare vergelijking te komen aan de hand van een gemeenschappelijke meetschaal, zijn er voldoende ankeritems nodig. Een algemene richtlijn is dat 20 à 25% van de toetsitems van beide toetsen ankeritems moeten zijn (Hambleton, e.a., 1991; Kolen & Brennan, 2004).

Ten tweede is het van belang om de meetfouten van de toetsresultaten zo klein mogelijk te houden. Zoniet bestaat de kans dat de verschilscore (het verschil tussen twee toetsscores) minder betrouwbaar is dan die van de toetsscores apart (Zimmerman & Williams, 1998). Een grotere meetfout heeft als het nadeel dat de bepaling van de leerwinst op leerlingniveau en daardoor ook op geaggregeerde niveaus minder precies is (i.e. een grotere standaardfout en daardoor een breder betrouwbaarheidsinterval). Een eerste bron van meetfouten doet zich voor omdat er altijd toevallige afwijkingen zijn voor de individuele meting van het vaardigheidsniveau van leerlingen op elk van de meetmomenten (als gevolg van toevallige omstandigheden, en omdat de toets niet 100% accuraat is). Deze toevallige afwijkingen zijn van belang bij de vaststelling van het individuele vaardigheidsniveau (op elk van beide meetmomenten) en bijgevolg ook van de individuele leerwinst. Zolang deze fouten niet systematisch aan elkaar gerelateerd zijn, middelen die effecten zich uit op een geaggregeerd niveau (bv. school of systeem), al neemt de precisie van de schatting van bijvoorbeeld schooleffecten hierdoor wel af. Vaak zijn de scores op IRT-schalen evenwel het meest accuraat gemeten voor het gemiddelde (of rond beoogde vaardigheidsniveaus) en neemt de standaardfout toe voor de scores aan de extremen van de schaal (OECD 2008; Zoanetti, 2021). Een theoretisch gelijke leerwinst gaat hierdoor eveneens minder accuraat gemeten worden bij leerlingen met een erg laag of erg hoog vaardigheidsniveau bij een van de metingen.

Dit alles beperkt de mogelijkheden om bijvoorbeeld de scholen met een beperkte toegevoegde waarde te kunnen onderscheiden van scholen die meer leerwinst realiseren bij hun leerlingen (lagere statistische

[11] Het bepalen van ankeritems is niet vanzelfsprekend. Deze moeten immers voor de afnames worden geselecteerd, maar pas na verschillende statistische analyses blijkt of een item effectief als ankeritem kan fungeren. In de praktijk zullen niet alle vooraf aangeduide ankeritems na deze analyses goede ankeritems blijken te zijn.

power), evenals het adequaat vergelijken van de leerwinst van groepen van leerlingen (bv. studierichtingen of scholen) die sterk van elkaar verschillen in toetsscores bij eerste meting. Meetfouten in toetsscores kunnen ook een vertekend beeld geven (bias) van de schatting van de toegevoegde waarde van scholen (OECD, 2008; Kane, 2017). Om leerwinst en daarop gebaseerde toegevoegde waarde te kunnen vaststellen, is het dan ook noodzakelijk dat de meetfouten op beide toetsmomenten zo klein mogelijk worden gehouden. Dit kan onder meer door adaptief te toetsen of door per leerling meer toetsitems af te nemen. Een adaptieve afname komt tevens tegemoet aan het probleem van de samenhang tussen de grootte van de meetfout en het vaardigheidsniveau van een leerling, omdat het toelaat elke leerling gericht door te toetsen op zijn of haar specifiek vaardigheidsniveau. Een verdere mogelijkheid bestaat er tevens in om via itemadaptieve afname (CAT) de toets af te nemen en een stopcriterium in te voeren op basis van een minimum aantal vragen en een maximale standaardfout. Op die manier kunnen leerlingen ongeacht hun score een gelijkaardige standaardfout krijgen.

2.2 MODELLERING VAN LEERWINST

Aan de basis van een bepaling van de toegevoegde waarde van scholen ligt een modellering van de individuele onderwijsuitkomsten. Omdat we uitgaan van een schatting van de toegevoegde waarde op basis van leerwinstmetingen, vertrekken we daarom van een model voor individuele groei (leerwinst). Een leerwinstmeting heeft een aantal voordelen die het deelt met alle vormen van longitudinale (in casu panel) data. Zo laat dergelijke herhaalde meting over de tijd toe om de invloed van factoren meer eenduidig te ordenen over de tijd (wat is oorzaak, wat is gevolg). Daarnaast biedt het meer mogelijkheden om effecten te isoleren van *confounding factors* in bijzonder ook van de invloed van niet-geobserveerde variabelen (in de mate dat deze niet mee veranderen over de tijd). Deze vormen dan ook twee belangrijke elementen om van causale effecten te kunnen spreken. Hierdoor is een leerwinstmeting ook beter geschikt dan een statusmeting om de toegevoegde waarde van scholen in kaart te brengen.

Om individuele groei (leerlingniveau[12]) te modelleren kunnen verschillende types modellen worden gebruikt. In functie van het bepalen van de toegevoegde waarde van een school gaat het om de toepassing van een statistisch model dat de toetsscore van een leerling op een bepaald tijdsstip relateert aan een (of meerdere) eerdere toetsscore(s), en zo de *individuele wijziging in vaardigheidsniveau* over de tijd (leerwinst) modelleert. Deze statistische modellen kunnen verschillen in de specifieke wijze waarop de toetsscore wordt gerelateerd aan de eerdere toetsscore(s) (lineair, niet-lineair, verschilscore, growth modelling, ...).

De meest eenvoudige modellen die in de onderwijseffectiviteitsliteratuur doorgaans met elkaar gecontrasteerd worden, zijn het *gain score* model en *residualised gain* model. Deze twee modellen zijn een specificatie van respectievelijk een *verschilscore* regressiemodel en anderzijds een *lagged* of *covariate-adjusted* regressiemodel (*residualised change*) die gekend zijn in de meer algemene literatuur om individuele verandering over de tijd te modelleren (bv. Allison, 1990). We lichten deze beide types modellen toe in respectievelijk secties 2.2.2 en 2.2.1. Hoewel vaak de verschillen tussen beide types modellen benadrukt worden, bieden ze onder bepaalde voorwaarden dezelfde inzichten (zie verder in

[12] Dat het vaststellen van een uitkomst op schoolniveau een modellering op leerlingniveau vergt, kan intuïtief begrepen worden. Ten eerste vanuit het vertrekpunt dat het resultaat van de school het geaggregeerd resultaat is van de resultaten van haar leerlingen (net zoals we in deel 1 deden voor status en leerwinst op schoolniveau). De inschatting van de toegevoegde waarde van een school steunt verder op een (modelmatige) voorspelling gebaseerd op de systematische verschillen in resultaten van alle leerlingen over alle scholen heen.

sectie 2.2.3). Tot slot bespreken we complexere modellen die bijvoorbeeld rekening kunnen houden met meetfouten, multilevel schattingen maken, of niet-lineaire en heterogene relaties opnemen (sectie 2.2.4).

2.2.1 Residualised gain model

Het meest gebruikte model voor de schatting van de toegevoegde waarde van scholen, steunt op het gebruik van het *residualised gain* model voor twee meetmomenten (zie ook vergelijking 1a in kaderstuk). Binnen dat model worden de leerprestaties van de leerlingen van een school geschat op basis van hun eerdere prestaties (nulmeting) (en eventueel de leerling- en schoolkenmerken die buiten de invloedssfeer van de school liggen maar die gerelateerd zijn aan de leerprestaties). De wijze waarop dit type model toelaat om leerwinst (verandering) te analyseren gebeurt onrechtstreeks. Het verloopt namelijk via de residuen van het regressiemodel: datgene van de behaalde eindscore dat niet door het model wordt voorspeld of verklaard op basis van de beginscore. Vandaar ook de benaming voor dit type model, door verschillende auteurs ook anders benoemd, ondermeer als *residual gain*, *quasi-gain* of *conditional status* model (bv. Castellano & Ho, 2013; Schochet & Chiang, 2013). Leerlingen verschillen in hoeveel hoger of lager hun effectief behaalde eindscore ($post_{ij}$) is, dan de eindscore die het model voorspelt op basis van hun beginscore ($\beta_0 + \beta_1 * prior_{ij}$), en maakten met andere woorden meer of minder vooruitgang (e_{ij}). In de meest eenvoudige versie wordt een lineaire relatie gemodelleerd tussen de begin- en eindscore (cf. vergelijking 1a in kaderstuk). Leerlingen met een behaalde eindscore gelijk aan de voorspelde eindscore (residu=0) zijn dan leerlingen met gemiddelde leerwinst[13]. Leerlingen met hogere (of lagere) residuen hebben in dat geval een hogere (of lagere) leerwinst dan het gemiddelde.

Onder dit model vormen de individuele verschillen tussen de behaalde score van de leerlingen en hun geschatte (voorspelde) scores (residuen van het regressiemodel) de basis voor het bepalen van de toegevoegde waarde van een school. In haar meest eenvoudige vorm wordt de toegevoegde waarde van een school berekend als het gemiddelde van de residuen van de leerlingen bij dit regressiemodel.[14] De toegevoegde waarde van een school is dan gelijk aan het verschil tussen de daadwerkelijke toetsscore van de leerlingen van een school en de voorspelde toetsscore van de leerlingen, rekening houdend met

[13] Mits aan de voorwaarden voor de toepassing van lineaire regressie is voldaan, en in het bijzonder de assumpties van lineariteit en homoskedasticiteit.

[14] De berekening is complexer ingeval van een multilevel specificatie die de variantie partitioneert tussen het individuele leerlingniveau en het schoolniveau (zie ook verder). Als alternatief kan het schooleffect ook worden geschat door opname van de schoolgroepen als bijkomende variabelen in het regressiemodel (in een zogenaamde fixed effects modellering).

leerlingkenmerken, alsook met de contextkenmerken van scholen die buiten de invloedssfeer van de school liggen (Kolen & Brennan, 2004; McGrath et al., 2015). De toegevoegde waarde van een school is dan positief als de gemiddelde toetsscore van de school hoger is dan verwacht op basis van de eerdere prestaties van de leerlingen en eventueel de leerlingkenmerken en contextkenmerken die in het model werden geïntegreerd.

Overzicht van de basismodellen voor de analyse van leerwinst

residualised gain

$post_{ij} = \beta_0 + \beta_1 * prior_{ij} [+ \beta_x * x_{ij}] + e_{ij}$ (vergelijking 1a)

$post_{ij} - prior_{ij} = \beta_0 + (\beta_1 - 1) * prior_{ij} [+ \beta_x * x_{ij}] + e_{ij}$ (vergelijking 1b)

gain score

$post_{ij} - prior_{ij} = \beta_0 + [\beta_x * x_{ij}] + e_{ij}$ (vergelijking 2a)

$post_{ij} = \beta_0 + 1 * prior_{ij} + [\beta_x * x_{ij}] + e_{ij}$ (vergelijking 2b)

statusmeting

$post_{ij} = \beta_0 + 0 * prior_{ij} + [\beta_x * x_{ij}] + e_{ij}$ (vergelijking 3)

multilevel *residualised gain*

$post_{ij} = \beta_0 + \beta_1 * prior_{ij} + [\beta_x * x_{ij}] + u_j + e_{ij}$ (vergelijking 4)

waarbij
- subscript i verwijst naar de individuele leerling
- subscript j verwijst naar de specifieke school
- prior verwijst naar de vaardigheidsscore bij de eerste meting
- post verwijst naar de vaardigheidsscore bij de tweede meting
- x verwijst naar de leerling- en schoolkenmerken waarvoor (optioneel) wordt gecontroleerd
- β zijn de regressieparameters
- u en e zijn residuen

2.2.2 Gain score model

Het *residualised gain* model van hierboven wordt in de literatuur vaak gecontrasteerd met het *gain score* model om de leerwinst van leerlingen over twee meetmomenten te modelleren. Bij het *gain score* model wordt leerwinst als verschilscore (zie vergelijking 2a) als afhankelijke variabele gespecifieerd in het

toegevoegdewaardemodel. Dit betekent dat de leerwinst van de leerlingen van een school rechtstreeks wordt gemodelleerd. Omdat net zoals bij het residual gain model ook hier verschillende types schooleffecten kunnen geschat worden, wordt de leerwinst eventueel verder gecorrigeerd voor leerling- en schoolkenmerken die buiten de invloedssfeer van de school liggen, maar wel gerelateerd zijn aan de leerprestaties, door deze kenmerken als bijkomende predictoren op te nemen in het regressiemodel.

Bij dit model wordt de toegevoegde waarde van een school berekend op basis van de voorspelde individuele leerwinst (dit is de globaal verwachte verschilscore tussen het tweede meetmoment en de nulmeting, gegeven eventuele controlevariabelen, cf. $\beta_0 + \beta_x * x_{ij}$) en de individuele afwijking (e_{ij} in vergelijking 2a). De toegevoegde waarde van een school is dan positief als de gemiddelde leerwinst van de school hoger is dan verwacht op basis de leerling- en contextkenmerken (van de school) die in het model werden geïntegreerd (Castellano & Ho, 2013).

2.2.3 Vergelijkbaarheid beide modelleringen

Mits de beginscore mee opgenomen wordt als predictor in het *gain score* model, is het model equivalent aan het *residual gain* model (zie vergelijking 1b) en zijn de bekomen resultaten identiek (bv. Dalecki & Willits, 1991). Daarom ook dat we de resultaten van het *residual gain* model kunnen interpreteren in termen van leerwinst. Het *gain score* model lijkt aantrekkelijk omdat het leerwinst rechtstreeks modelleert (Meghir & Rivkin, 2010), toch worden verschilscore modellen voor de analyse van individuele verandering minder gebruikt dan *residualised change* modellen. Het meest gebruikte argument is dat het minder geschikt is om de leerwinst van leerlingen te modelleren ten aanzien van de eerdere prestatiemeting van leerlingen (Castellano & McCaffrey, 2020). Bij een modellering van de verschilscore waarbij de eerste toetsscore mee als predictor in rekening wordt gebracht, vergroot immers het probleem van bias als gevolg van meetfouten, doordat de meetfouten in de eerste toetsscore ook vervat zitten in de meetfouten van de verschilscore. In een standaard *gain score* (of *change score*) model wordt daarom de eerdere toetsscore doorgaans niet mee opgenomen als predictor. Technisch gesproken is een dergelijk *gain score* model evenwel niet langer equivalent aan het *residual gain* model. Het kan eerder beschouwd worden als equivalent aan een *residual gain* model waarin de regressiecoëfficiënt voor de eerste toetsscore wordt vastgezet op 1 (zie hiervoor de vergelijking 2b - Wright, 2018; Castro-Schilo & Grimm, 2018; Meyer, 1996). Schattingen voor deze regressiecoëfficiënt liggen normaliter lager, en zijn doorgaans lager naarmate de metingen verder uit elkaar liggen.

Eenduidige richtlijnen voor de keuze van een *residualised gain*, dan wel van een *gain score* model zonder eerste toetsscore als predictor, zijn er evenwel niet (Köhler, Hartig & Schmid, 2021). Bovendien toont onderzoek dat de geschatte toegevoegde waarde op basis van een *gain score* model weinig tot niet

verschilt van de schooleffecten die via het *residualised gain* model worden geschat (Harris, Ingle & Rutledge, 2014; Zamarro, Engberg, Saavedra & Steele, 2015). Anderzijds blijkt de performantie (in termen van bias) van het *gain score* model in vele gevallen minder goed dan de performantie van het *residualised gain* model (Guarino, Reckase & Wooldridge, 2015; Koedel et al., 2015).

2.2.4 State-of-the-art toegevoegde waarde schattingen

Beide bovenstaande types van statistische modellen blijven onderhevig aan een bias die het gevolg kan zijn van samenhang in de meetfouten van opeenvolgende metingen van vaardigheidsscores. Dit noopt onderzoekers tot de ontwikkeling en het gebruik van meer complexe modellen voor individuele groei. Dergelijke complexe modellen zouden meer mogelijkheden bieden om meetfouten op het individuele (stabiele) niveau te onderscheiden van (werkelijke) individuele verandering. In het bijzonder bieden metingen met meer dan twee momenten op dat vlak heel wat meer mogelijkheden (cf. latent curve en multilevel growth modellen, maar bv. ook latente verschilscore modellen voor twee meetmomenten). Dergelijke modelleringen laten bovendien toe om zowel het status als het leerwinst perspectief te combineren. De keerzijde van deze medaille is dat ze minimaal drie en zelfs meer opeenvolgende metingen vergen van eenzelfde toetsvaardigheid.

Daarnaast kan de performantie van toegevoegdewaardemodellen ook verbeterd worden door in het model zelf rekening te houden met meetfouten. Zo is het gangbaar om slechts één eerdere toetsprestatie als predictor te gebruiken, maar strekt het tot de aanbeveling om meerdere toetsprestaties te integreren in het model (Lockwood & McCaffrey, 2014). Dit reduceert immers het effect van de individuele meetfout op een welbepaald meetmoment (Lockwood & Castellano, 2015). Zo is het binnen de context van de centrale toets wiskunde rond probleemoplossen in het zesde leerjaar bijvoorbeeld mogelijk om naast de toetsscore voor probleemoplossen in het vierde leerjaar ook de toetsscores van meerdere thematoetsen (uiteraard indien beschikbaar) als predictoren in het model op te nemen. Een positieve toegevoegde waarde toont in het voorbeeld aan dat de leerlingen van een school gemiddeld betere prestaties halen in het zesde leerjaar dan men op basis van hun vaardigheidsscores in het vierde leerjaar zou verwachten. Een andere manier om modellen voor individuele groei te verbeteren door meetfouten mee in het model te integreren, is door een simultane schatting van het meetmodel (de vaardigheidsscore) en het structureel model (de verandering over de tijd en de impact van de predictoren) (bv. Kim & Camilli, 2014; Wang & Nydick, 2020).

De meest courant gebruikte techniek om de toegevoegde waarde van een school te bepalen is door middel van een multilevel versie van het *residualised gain* model (cf. vergelijking 4). Omdat leerlingen deel uitmaken van scholen, hangen de leerprestaties van leerlingen van eenzelfde school sterker samen dan

de leerprestaties van leerlingen uit verschillende scholen. De multilevelanalyse voorziet dat de variantie in de toetsscores (op het tweede meetmoment) van leerlingen wordt ingedeeld in variantie op leerlingniveau (e_{ij}) en variantie op schoolniveau (u_j). Binnen deze multilevel-analyse wordt de invloed van een school geschat door middel van een random intercept-model. Het intercept van een regressiemodel staat daarbij voor de waarde van de afhankelijke variabele (c.q. toetsprestatie), gecontroleerd voor de predictoren. Meer concreet is het de geschatte toetsprestatie voor specifieke waarden op de controlevariabelen die, mits de (continue en categorische) predictoren gecentreerd worden, kan worden geïnterpreteerd als de gemiddelde toetsprestatie. Het specifieke aan een multilevelanalyse van leerlingen genest binnen scholen is dat het model in rekening brengt dat scholen verschillen in deze waarde en dat de mate waarin scholen hierin verschillen mee geschat wordt. Op basis daarvan, kunnen vervolgens de random schoolintercepten (elke school haar eigen afwijking op het globale intercept) worden geschat. Technisch gezien worden deze random effecten geschat via empirisch Bayesiaanse predictie, wat resulteert in de zogenaamde 'shrinkage-schattingen' van de random intercepten die de toegevoegde waarde van de scholen representeren (Leckie & Goldstein, 2019). Die *shrinkage* schattingen kunnen worden beschouwd als de schooleffecten (of schoolafwijkingen van het globaal gemiddelde) aangepast aan de precisie waarmee deze schooleffecten geschat kunnen worden. Dit resulteert in conservatieve schattingen van de schooleffecten. In vergelijking met de op schoolniveau berekende gemiddelde verschillen tussen de werkelijke en verwachte scores van leerlingen verschuift door *shrinkage* de voorspelde waarde op schoolniveau naar het globaal gemiddelde. Die verschuiving is minder sterk naarmate de schooleffecten groter zijn (sterkere clustering in de data) en minder sterk voor scholen met grote leerlingaantallen (Leckie, 2018). Door *shrinkage* wordt dan ook vermeden dat kleine scholen overwegend aan de extremen van de verdeling komen te staan louter als gevolg van hun kleine aantal leerlingen. Het beperkt eveneens het risico op overinterpretatie door gebruikers van de vaak extreme resultaten van kleine scholen.

Ook het analytisch model dat gehanteerd wordt, is vaak complexer dan de eenvoudige versie van het *residualised gain* model. Complexere modellen kunnen worden gebruikt wanneer niet voldaan is aan de assumpties van het lineaire regressiemodel (Levy et al., 2019; Meyer, 1996; Reardon & Raudenbush, 2009). Zo kan bijvoorbeeld een niet-lineaire relatie tussen de eerste en de tweede meting mee opgenomen worden voor accuratere schattingen (bv. Leckie & Goldstein, 2019) of zou die relatie kunnen verschillen naargelang de school (heterogeen effect; Meyer, 1996). Zo kan ook heterogeniteit in schooleffecten naargelang achtergrondkenmerken van leerlingen gemodelleerd worden om na te gaan of het schooleffect anders is voor sommige leerlingen dan voor andere leerlingen (bv. Strand, 2016).

2.3 TOEGEVOEGDE WAARDE EN/OF LEERWINST

Toegevoegde waarde en leerwinst worden vaak met elkaar gecontrasteerd. Zoals we het ook in het eerste deel voorstelden, met name de toegevoegde waarde bepaling op basis van leerwinstmetingen, zijn de conceptuele verschillen evenwel niet zo groot. Leerwinst kan als een eerste schatting worden beschouwd van de toegevoegde waarde van een school. Leerwinst is in elk geval een betere indicator van het schooleffect dan een statusbepaling, ondermeer omwille van de neutralisering van de effecten van achtergrondkenmerken die de leerprestatie van leerlingen op de twee tijdsstippen op dezelfde manier beïnvloeden. De integratie van eerdere toetsprestaties zorgt er immers voor dat de sociaaldemografische leerlingkenmerken nog slechts een minimale impact hebben op de geschatte schooleffecten (Newton et al., 2010; Rothstein, 2009). Hierdoor hangt de leerwinst op schoolniveau doorgaans ook sterker samen met de proceskenmerken van scholen dan met de achtergrondkenmerken van de instroompopulatie (Heck, 2006). Maar omdat leerwinst slechts een gedeeltelijke inschatting maakt van de door de school behaalde resultaten voor alle mogelijke leerlingen, of in termen van geleverde inspanningen (cf. deel 1.3), wordt de toegevoegde waarde van een school doorgaans bepaald door verdere correcties door te voeren op de vastgestelde leerwinst.

Toegevoegde waarde en leerwinst worden soms ook voorgesteld als fundamenteel verschillende zaken. Daarbij gaat het vaak om begripsverwarring, want net als bij de term leerwinst, bestaan er ook aangaande toegevoegde waarde heel wat verschillende invullingen en interpretaties.

Soms wordt toegevoegde waarde van scholen begrepen als een eenmalige *statusmeting* gecorrigeerd voor leerlingkenmerken, en zo gecontrasteerd met leerwinstmeting. Omdat een eerdere toetsprestatie doorgaans het belangrijkste kenmerk is dat een toetsprestatie van een leerling verklaart, wordt voor een adequate schatting van toegevoegde waarde echter best een eerdere toetsprestatie meegenomen. Waar Amerikaanse *value-added* schattingen doorgaans leraren evalueren op basis van leerprestaties van hun leerlingen, gebeurt dit in de UK veel meer voor het schoolniveau, maar in beide gevallen vormt de basis nagenoeg steeds de vooruitgang die leerlingen maken (ten opzichte van een eerdere toetsprestatie; Levy et al., 2019). Bij gebrek aan een eerdere toetsprestatie, wordt ook in het Vlaams peilingsonderzoek en bij internationaal vergelijkende studies zoals PIRLS, TIMSS of PISA, de statusmeting van een school gecorrigeerd voor een aantal leerlingkenmerken hier te lande benoemd als toegevoegde waarde. Dit is immers het enig technisch haalbare alternatief bij gebrek aan longitudinale panel-data. Het toevoegen van louter leerlingkenmerken is echter onvoldoende om het effect dat een school heeft op de progressie van leerlingen afdoende te isoleren (OECD, 2008). Ballou en collega's (2004) wijzen er daarbij op dat het belangrijk is om eerdere toetsprestaties als predictoren in het model te integreren. Overeenkomstig tonen

McCaffrey en collega's (2003) aan dat modellen die enkel voor sociaaleconomische en sociaaldemografische kenmerken controleren en niet voor eerdere leerprestaties, er onvoldoende in slagen om het effect van de achtergrondkenmerken van de leerlingen te neutraliseren. Beter zou dan ook zijn om in die gevallen te spreken van een gecorrigeerde statusmeting als (best mogelijke) benadering voor de toegevoegde waarde van scholen (OECD, 2008; Lenkeit, 2013). Op dezelfde wijze kan voor een eerste meting bij de centrale toetsen (in het vierde leerjaar) niet gesproken worden van toegevoegde waarde, omdat men niet beschikt over eerder vastgestelde prestatiemetingen.

Regelmatig wordt toegevoegde waarde en leerwinst van elkaar onderscheiden op basis van de *modellering van leerwinst*, waarbij toegevoegde waarde wordt gelijkgesteld aan het *residualised gain model*, en vergeleken met leerwinst op basis van een *gain score model*. De discussie in de literatuur gaat dan bijvoorbeeld over het verschil tussen verschilscores dan wel de lagged regressie benadering om individuele groei te modelleren. We behandelden beide types modellen hierboven reeds en stipten aan dat het analytisch model niet fundamenteel verschilt tussen een gain score model en een residualised gain model (tenzij het gaat over de rol van de eerdere prestatie). Andere discussies gaan over de wenselijkheid en noodzaak van een gemeenschappelijke meetschaal en over absolute versus relatieve schooleffecten. Maar ook daar blijken verschillen niet steeds zo groot als ze op het eerste zicht lijken.

Zo wordt leerwinst soms onderscheiden van toegevoegde waarde door het gebruik van een *gemeenschappelijke meetschaal* voor de metingen over de tijd. Onze bespreking van toegevoegde waarde vertrekt steeds van een leerwinstmeting omdat individuele groei het meest helder het leren van leerlingen uitdrukt en dan ook een goed vertrekpunt vormt om de bijdrage van scholen aan het leren van hun leerlingen te bepalen. Een leerwinstmodellering door middel van het *gain score* model is enkel mogelijk (zinvol) wanneer de metingen over de tijd zinvol met elkaar vergeleken kunnen worden. Bij een *residualised gain* model daarentegen hoeven de vaardigheidsscores van de nulmeting en de afhankelijke variabele niet op dezelfde meetschaal te staan (al kan het dan niet langer een *gain* model worden genoemd, cf. Castellano & Ho, 2013 die spreken van een *conditional status* model). Dit type model is zowel inzetbaar wanneer de metingen inderdaad op een gemeenschappelijke meetschaal kunnen geplaatst worden, maar ook wanneer dat niet het geval zou zijn omdat de toetsscores over de tijd niet vergelijkbaar zijn. In strikte zin spreken we in dat laatste geval evenwel niet meer van leerwinst. Wanneer de toetsen over de tijd niet hetzelfde meten of de resultaten niet naar eenzelfde vaardigheidsschaal kunnen worden omgezet, kunnen we niet langer spreken van individuele groei. Het gaat dan nog louter

om de verwachte prestatie gegeven een eerdere prestatie[15]. De resultaten van een dergelijke niet-leerwinstgebaseerde toegevoegde waarde kunnen dan nog wel inzicht bieden in relatieve verschillen tussen scholen, maar hebben dan niet langer een meer absolute interpretatie in termen van leerwinst. Die grotere flexibiliteit van het lagged regressiemodeltype verklaart wellicht mee het gebruik van dat modeltype in het nadeel van modellen op basis van verschilscores (en laat bijvoorbeeld ook eenvoudig toe om meerdere eerdere toetsprestaties mee te nemen).

Tot slot wordt ook gewezen op het verschil in referentie waarop leerwinst dan wel toegevoegde waarde betrekking heeft. Het geschatte schooleffect van een *gain score* model (in essentie een gemiddelde verschilscore) wordt voorgesteld als een eenvoudig te interpreteren indicator van hoeveel vooruitgang een leerling maakte (gegeven eventuele andere kenmerken). Het gaat om een verschuiving van een specifiek aantal punten op de (gezamenlijke) meetschaal. Leerwinst kan dan ook beschouwd worden als een absolute en descriptieve maat die de grootte aangeeft van de progressie die een leerling, klas of school maakt tussen meerdere meetmomenten. Dit wordt vaak als een belangrijk voordeel beschouwd ten opzichte van toegevoegde waarde modellen waarbij scholen worden vergeleken met het (relatieve) gemiddelde of de verwachte prestaties.

Het geschatte schooleffect van een *residualised gain* model wordt dan bekritiseerd omdat het niet die absolute interpretatie heeft, maar enkel aangeeft of een school meer dan wel minder leerwinst boekte bij haar leerlingen dan gemiddeld of dan andere scholen. Van een school met een negatieve toegevoegde waarde zegt het alleen dat de vooruitgang die de leerlingen maakten minder is dan gemiddeld (relatief), maar dat maakt niet duidelijk of en in welke mate de school (absolute) leerwinst wist te boeken bij de leerlingen. Dat verschil is evenwel niet gebonden aan het toegepaste model, wel aan de onderliggende data. Mits leerwinstmetingen gebruikt worden, kunnen resultaten van beide modeltypes geïnterpreteerd worden in termen van leerwinst (cf sectie 2.2). Het is wel gebonden aan het feit dat op basis van een *residualised gain* model vaak enkel de schoolafwijkingen van het globaal verwacht gemiddelde gerapporteerd wordt. Dat probleem kan evenwel worden verholpen door de gerapporteerde indicator voor toegevoegde waarde van een school aan te passen. Op basis van het *residualised gain* model kan ook een *voorspelde gemiddelde prestatie* voor een school worden berekend. Modelmatig gaat het dan

[15] In de literatuur wordt in een aantal van die gevallen toch gesproken over individuele vooruitgang (bv. Progress 8 in de UK). Hoewel de de scores over de tijd niet op eenzelfde meetschaal staan, en dus strikt genomen niet vergelijkbaar zijn, zijn ze dan vaak wel op een alternatieve manier vergelijkbaar gemaakt, maar dan in relatieve zin (bv. transformatie van beide testen in percentielen, standaardiseren naar zelfde gemiddelde en standaardafwijking, of op basis van cesuren of benchmarkpunten; Wright, 2018).

om de som van het globaal verwacht gemiddelde en de schoolafwijking (bij een random-interceptmodellering). Dat schooleffect is dan niet langer louter relatief, maar eveneens te interpreteren op dezelfde schaal als de toetsprestatie (bij de tweede meting). Die operationalisatie van toegevoegde waarde van een school kan dan gezien worden als de (geschatte) gemiddelde leerprestatie van een school voor een gegeven benchmark van leerlingen (cf. aanbeveling Meyer, 1996). Dergelijke benchmark verwijst naar een (hypothetische) groep leerlingen met een bepaalde aanvangsscore en specifieke achtergrondkenmerken, waarbij dezelfde benchmark wordt gebruikt voor alle scholen. Typisch worden voorspelde waarden geschat op basis van de referentiewaarden of -categorieën van de controlevariabelen in de analyse. Maar het is evengoed mogelijk de geschatte leerprestatie van een school te rapporteren voor een specifieke of ook verschillende benchmark groepen zoals jongens versus meisjes, leerlingen met thuistaal Nederlands dan wel anderstalige leerlingen, …. Op die manier is de toegevoegde waarde van een school niet alleen te interpreteren in termen van hoeveel vooruitgang wordt geboekt, maar bovendien te interpreteren in termen van te bereiken doelen.

2.4 STEEKPROEFGROOTTE EN STATISTISCHE ONZEKERHEID

Om de bovenstaande toegevoegdewaardemodellen te schatten wordt doorgaans gebruikt gemaakt van multilevel lineaire regressie waarbij de schooleffecten als random effecten worden gemodelleerd. Bij een multilevel lineaire regressie worden eisen gesteld aan de steekproef(grootte) op zowel leerling- als op schoolniveau. Er dienen met name zowel voldoende scholen als voldoende leerlingen per school te participeren. Dit is niet alleen noodzakelijk in functie van een stabiele en accurate schatting van de variantieparameters (schooleffecten) en regressieparameters (verklarende variabelen) maar ook om voldoende betrouwbare schoolfeedback te kunnen geven. Anders gesteld: zowel het aantal deelnemende scholen als het aantal leerlingen dat deelneemt per school dient voldoende groot te zijn om een betrouwbaar beeld te krijgen van de verschillen tussen scholen en van de mate waarin de leerling- en schoolkenmerken samenhangen met de gemeten toetsprestatie. Indien er uitspraken worden gedaan op klasniveau, dient ook hier per toets het aantal deelnemende leerlingen per klas voldoende groot te zijn. Uitspraken over het effect van bepaalde (groeps)variabelen, kunnen ook een invloed hebben op de vereiste steekproefgrootte,bv. wanneer men verschillen tussen basisopties in de eerste graad wil nagaan. Tot slot zijn uitspraken over school- en klaseffecten en uitspraken over het effect van bepaalde variabelen pas betekenisvol wanneer deze voor elke toets gesitueerd kunnen worden binnen een representatieve steekproef.

Het aantal scholen dat deelneemt aan het onderzoek is vooral van belang om uitspraken te kunnen doen die betrekking hebben op de variantie of verschillen tussen scholen, met name om de impact van schoolkenmerken adequaat te kunnen inschatten. Hoewel 30 tot 50 scholen gangbaar zijn binnen onderwijskundig onderzoek, is dit onvoldoende om adequate standaardfouten van varianties op klas- en schoolniveau te schatten (Maas & Hox, 2005). Meer specifiek is een minimum van 100 scholen vereist om te vermijden dat de standaardfouten van de varianties onderschat worden (Van der Leeden, Busing & Meijer, 1997). Deze richtlijn is met name relevant voor het op voorhand kunnen inschatten van de analysemogelijkheden bij een onvolledig afnamedesign (zie de toepassing bij de centrale toetsen, in sectie 3.4.1 en verder).

Naast het aantal scholen is het ook belangrijk dat er per school voldoende leerlingen deelnemen aan de toetsen. Dit is noodzakelijk voor een precieze en accurate schatting van de specifieke schooleffecten. Een schooleffect dat geschat wordt op basis van een beperkt aantal leerlingen is weinig betrouwbaar. De score van elke leerling heeft dan immers een groot effect op de algemene schatting van het schooleffect (Jakubowski, 2008). Bijgevolg kan in dergelijke scholen de geschatte toegevoegde waarde sterk afwijken van het daadwerkelijke schooleffect (Goldstein, 1997). Studies van onder meer McCaffrey e.a (2009) tonen

inderdaad aan dat de standaardfout van de geschatte toegevoegde waarde van een school sterk toeneemt met een dalend aantal participerende leerlingen.

In de regel wordt een minimum van 20 tot 30 leerlingen per groep (d.w.z. per school, of per klas binnen elke school) per uitkomst (toets) aangeraden om betrouwbare resultaten te bekomen. Indien men via de centrale toetsen bijvoorbeeld klaseffecten wil rapporteren over een bepaalde toets, dan strekt het tot aanbeveling dat er minstens twintig leerlingen van de klas voor die specifieke toets deelnemen. De praktijk leert ons dat dit echter niet altijd mogelijk is. Voor een goed begrip van de resultaten is het daarom wenselijk om schattingen van school- of klasgemiddelden of van toegevoegde waarde steeds te rapporteren met informatie over de statistische betrouwbaarheid van de schattingen, bijvoorbeeld door middel van betrouwbaarheidsintervallen. Deze richtlijn wordt tevens als good practice meegegeven als alternatief voor het louter rapporteren van statistische significantie-toetsen (met p-waarden kleiner dan ,05 of 0,01) (zie o.m. Altman et al. 2000, Kline 2013). Een visuele weergave door middel van foutenbalken lijkt bovendien voor leraren het duidelijkst (Hopster-den Otter et al., 2019). Bijkomend bieden de *shrinkage* schattingen (Leckie, 2018; cf. sectie 2.2.4) bescherming tegen de overinterpretatie van bepaalde extreme schooleffecten door kleine leerlingaantallen.

2.4.1 Stabiliteit van een schooleffect

Een belangrijke voorwaarde opdat schoolfeedback gebruikt wordt, is niet alleen dat de geschatte schooleffecten betrouwbaar zijn maar ook als betrouwbaar beschouwd worden door de scholen. Vooral voor scholen die minder bijdragen aan het leren van hun leerlingen dan kan worden verwacht, is het belangrijk dat er voldoende vertrouwen is in de resultaten. Deze gepercipieerde betrouwbaarheid hangt sterk samen met de mate waarin de geschatte schooleffecten stabiel zijn over verschillende cohorten van leerlingen heen (Dumay et al., 2014).

Het is echter volstrekt normaal (en soms wenselijk) dat schooleffecten doorheen de jaren kunnen veranderen. De kwaliteit van scholen kan immers toenemen door bepaalde maatregelen die de school heeft genomen. Het is evenwel belangrijk om werkelijke verandering over de tijd te kunnen onderscheiden van andere vormen van verandering die te wijten zijn aan instabiliteit en random fluctuaties van schooleffecten (dit is de kern van trendanalyse). In de vorige sectie werd al duidelijk dat schooleffecten minder betrouwbaar zijn (in de zin van grotere statistische onzekerheid) bij scholen met minder leerlingen. Bijgevolg zijn schooleffecten die worden geschat op basis van een beperkt aantal deelnemende leerlingen ook minder stabiel over verschillende cohortes heen (OECD, 2008). Ook andere factoren kunnen bijdragen tot instabiliteit, zoals het soort model, verandering in de toetskarakteristieken doorheen de jaren, het aantal predictoren dat wordt opgenomen in het model en het aantal jaren tussen

het eerste en het tweede toetsmoment.[16] Ondermeer door de impact van het aantal leerlingen op de gevoeligheid voor het *shrinkage*-effect, bieden aparte schattingen van de toegevoegde waarde van scholen voor opeenvolgende cohorten, bovendien geen correct beeld (bias) van de stabiliteit van schooleffecten over de tijd. Leckie (2018) toont aan dat de stabiliteit van schooleffecten beter wordt geschat op basis van simultane modellering van toegevoegde waarde over cohorten.

Om de instabiliteit en random fluctuaties van schooleffecten doorheen de tijd "glad te strijken" wordt soms aanbevolen het gemiddelde schooleffect van drie opeenvolgende cohortes van leerlingen te hanteren. Het aggregeren van schooleffecten over een periode van drie jaar reduceert dan ook de kans op een type-I fout, d.w.z. de kans dat een school ten onrechte als significant beter of minder goed dan gemiddeld wordt beschouwd (Schochet & Chiang, 2010). Hoewel deze aanpak tot meer stabiele schooleffecten leidt, impliceert dit dat daadwerkelijke veranderingen in schoolkwaliteit moeilijker kunnen vastgesteld worden. Het is ook niet voor alle scholen noodzakelijk om de geschatte schooleffecten over meerdere jaren uit te middelen. Zo kan voor scholen met grote leerlingaantallen het schooleffect afdoende betrouwbaar worden geschat, waardoor ook de werkelijke verandering sneller onderscheiden kan worden van instabiliteit die te wijten is aan statistische onzekerheid. Ook zal een loutere aggregatie van de jaarlijks (aan *shrinkage* onderhevige) geschatte schooleffecten minder goede resultaten geven dan een gepoolde analyse.

[16] Een bespreking van elk van deze factoren gaat voorbij aan het doel van deze nota, we verwijzen dan ook graag naar Koedel e.a. (2015) en OECD (2008).

2.5 BIAS EN SCHOOLEFFECTEN

Bias verwijst naar het fenomeen waarbij schooleffecten systematisch over- of onderschat worden. In tegenstelling tot een inaccurate schatting van een schooleffect, kan bias niet opgelost worden door de leerlingensteekproef binnen een school te vergroten. Systematisch onder- of overschatte schooleffecten kunnen door meerdere potentiële bronnen worden veroorzaakt. De mogelijke rol van meetfouten in de toetsscores en keuzes op het vlak van modellering van leerwinst kwamen hierboven reeds aan bod. We gaan hier nog specifiek in op bias als gevolg van de predictoren die in de modellen worden opgenomen (in sectie 2.5.1), en als gevolg van de leerlingen voor wie de leerwinstmeting beschikbaar is of voor wie de leerwinstmodellering wordt uitgevoerd (in 2.5.2).

2.5.1 Bias en predictoren

In de allereerste plaats kan bias ontstaan doordat er te weinig of verkeerde predictoren in het model worden opgenomen. De leerlingkenmerken die worden opgenomen in toegevoegdewaardemodellen representeren meestal niet alle achtergrondkenmerken die gerelateerd zijn aan de prestaties van de leerlingen. Dit probleem stelt zich in het bijzonder wanneer enkel administratieve indicatoren van de kenmerken van leerlingen beschikbaar zijn (Meyer, 1996). Zo is het bijvoorbeeld onvoldoende om SES uitsluitend via de proxy opleidingsniveau van de ouders te operationaliseren. Dergelijk model zou leiden tot een te beperkte correctie voor leerlingkenmerken en bijgevolg voor een overschatting van de schooleffecten. Schooleffecten van scholen met een hoge SES-populatie worden dan gekenmerkt door een opwaartse bias (het geschatte schooleffect is hoger dan het echte schooleffect) terwijl schooleffecten van lage SES-scholen dan net een neerwaarts bias (het geschatte schooleffect is kleiner dan het echte schooleffect) kennen. Bijgevolg dienen de achtergrondkenmerken die gerelateerd zijn aan de prestaties van de leerlingen zo volledig mogelijk, d.w.z. door voldoende en de juiste proxies, in het model geïntegreerd te worden (bv. Ehlert et al., 2016).

Anderzijds kan door het in rekening brengen van compositiekenmerken (type-B schooleffect) de toegevoegde waarde van een school worden onderschat. Dit is het geval wanneer sprake zou zijn van omgekeerde causaliteit, met name wanneer de schoolpraktijken zelf een effect hebben op de schoolcompositie. Zo is het mogelijk dat een heel goede school de betere leerlingen aantrekt, net omdat de school zeer goed onderwijs levert. Door te controleren voor compositie wordt in dit geval ook gecontroleerd voor een deel van de schoolpraktijk. Als gevolg lijken de schooleffecten (geschatte netto-effecten) kleiner en wordt de daadwerkelijke bijdrage van de school aan het leren van de leerling onderschat (Castellano, Rabe-Hesketh & Skrondal, 2014).

Net zoals bij het type-B schooleffect bestaat bij het type-X schooleffect het risico om te "overcontroleren" en bijgevolg het netto-effect van de school te onderschatten (Timmermans et al., 2011). Niet alleen een rechtstreekse causale samenhang van schoolexterne met schoolinterne factoren, maar ook louter statistische samenhang die bijvoorbeeld te maken heeft met niet-gekende factoren, vergroot het risico op overcontrole en daardoor onderschatting van de (overblijvende) schooleffecten. Terwijl het niet gangbaar is bij toegevoegde waardemodellen, zou het expliciet mee opnemen van indicatoren voor de schoolinterne factoren in de modellen, die neerwaartse bias voor type-X schooleffecten kunnen verminderen (Meyer, 1996).

2.5.2 Bias en schoolloopbanen

Niet alle leerlingen zijn normaalvorderende leerlingen met een standaard traject binnen een school. Indien bij het schatten van de schooleffecten geen rekening wordt gehouden met afwijkende schoolloopbanen van leerlingen, kan er ook bias ontstaan. Afwijkende schoolloopbanen verwijzen niet alleen naar zittenblijvers, maar ook naar leerlingen die vroegtijdig de school verlaten en naar leerlingenmobiliteit. Dit laatste verwijst naar de één- of meermalige mobiliteit van leerlingen tussen scholen op andere momenten dan de reguliere leeftijd wanneer ze een opleiding starten of beëindigen (Dockx, De Fraine & Stevens, 2016; Strand, 2002). Afwijkende schoolloopbanen vinden echter niet at random plaats. Zo hangt leerlingenmobiliteit niet alleen samen met eerdere prestaties van leerlingen, maar ook met hun achtergrondkenmerken, zoals gezinsinkomen of etnische herkomst. Bovendien zijn er ook verschillen op schoolniveau: leerlingenmobiliteit manifesteert zich meer binnen scholen met een leerlingenpubliek met voornamelijk lage SES en scholen met een hogere verstedelijkingsgraad (Rumberger, 2003). Het feit dat de determinanten van afwijkende schoolloopbanen en van leerprestaties deels overlappen, impliceert dat het verwijderen van deze leerlingen uit een school of geen rekening houden met hun mobiliteit in het model kan resulteren in bias. Anders gezegd, indien er geen rekening wordt gehouden met de onvolmaakt hiërarchische datastructuur ten gevolge van afwijkende schoolloopbanen, zullen zowel de variantie op schoolniveau als de schooleffecten biased zijn (Chung & Beretvas, 2012; Leckie, 2009; Timmermans et al., 2012).

Om die bias te vermijden kunnen verschillende aanpassingen gedaan worden aan de multilevel regressie modellen. Zo kan leerlingenmobiliteit bijvoorbeeld in rekening worden gebracht aan de hand van een multiple membership model (Timmermans, Snijders & Bosker, 2013) terwijl een vertekende schatting van schooleffecten onder invloed van vroegtijdige schoolverlaters kan worden gereduceerd door middel van multiple imputation op basis van achtergrondkenmerken van deze leerlingen (Leckie, 2009; Schafer & Graham, 2002).

3 CENTRALE TOETSEN

In dit deel koppelen we de doelstellingen van de centrale toetsen aan de concepten en aandachtspunten die hierboven werden besproken. Meer specifiek gaan we na in welke mate er op leerling-, klas-, school- en systeemniveau betrouwbare uitspraken gedaan kunnen worden inzake status, leerwinst en toegevoegde waarde. Het ontwikkelde toetsdesign van wiskunde secundair onderwijs wordt daarbij als uitgangspunt genomen. Dit toetsdesign creëert immers een aantal contouren die bepalen in welke mate status, leerwinst en toegevoegde waarde op de verschillende aggregatieniveaus kunnen vastgesteld worden. Meer specifiek bespreken we de implicaties van de breedtetoets en van scenario's voor de (longitudinale) afname van focustoetsen wiskunde in het secundair onderwijs, welke veralgemeend kunnen worden naar toetsen in andere onderwijsniveaus en te toetsen domeinen ingeval van respectievelijk een volledig dan wel onvolledig afnamedesign. Analoge redeneringen zijn geldig voor toetsen in het lager onderwijs, maar met de vereenvoudiging dat in het lager onderwijs er nog geen sprake is van verschillende onderwijsstromen en -finaliteiten waarmee bij de toetsen in het secundair onderwijs wel rekening moet worden gehouden omdat deze samengaan met verschillende eindtermen.

In wat volgt wordt eerst kort ingegaan op de veralgemeenbaarheid van de inzichten op basis van het toetsdesign van wiskunde secundair onderwijs voor de toetsen Nederlands. Vervolgens geven we een overzicht van de basisprincipes van het toetsdesign voor de centrale toetsen wiskunde secundair onderwijs. Voor een uitgebreide toelichting van dat toetsdesign verwijzen we naar de Nota *Centrale Toetsen Wiskunde* van het Steunpunt (2021).

3.1 TOETSDESIGN NEDERLANDS

Het toetsdesign van de centrale toetsen Nederlands wordt in deze nota niet behandeld. Gezien het grote aantal toetsen en het gedifferentieerd afnamedesign is het toetsdesign wiskunde vrij complex waardoor de inzichten transfereerbaar zijn naar het toetsdesign van Nederlands. Met name zijn de implicaties van de breedtetoets wiskunde het best veralgemeenbaar naar de toetsen Nederlands lezen, gezien beide uitgaan van een volledige afname. Voor wat betreft de toetsen Nederlands schrijven, kunnen we momenteel nog geen uitspraken doen, aangezien er nog onvoldoende zicht is op de aard van de informatie die deze gaan opleveren.

3.2 TOETSDESIGN WISKUNDE

Elk schooljaar legt elke leerling in het vierde en zesde leerjaar lager onderwijs en in het tweede en zesde leerjaar secundair onderwijs een breedtetoets 'wiskundige problemen oplossen' af. Leerlingen die in 2024 een breedtetoets afleggen in het tweede leerjaar secundair onderwijs, zullen – in het geval van een normaalvorderende leerling – in 2028 in het zesde leerjaar opnieuw een breedtetoets afleggen. Het daaropvolgende schooljaar zal een nieuwe cohorte van leerlingen de breedtetoets afleggen in het tweede jaar secundair onderwijs. Ook van deze tweede cohorte van leerlingen zal vier jaar later in het zesde leerjaar een breedtetoets worden afgenomen. Bovendien zullen op dat moment ook de leerlingen van de eerste cohorte met een jaar schoolvertraging in het zesde leerjaar de breedtetoets maken. De breedtetoets 'wiskundige problemen oplossen' is specifiek per onderwijsniveau, onderwijsstroom en onderwijsfinaliteit. De breedtetoets is adaptief.

Naast de breedtetoets zijn er binnen de A-stroom 10 thematoetsen die jaarlijks allemaal op systeemniveau worden aangeboden; binnen de B-stroom zijn er 6 thematoetsen. Per school worden jaarlijks drie verdiepende thematoetsen (uit drie verschillende categorieën) aangeboden, die ook alle drie in elke klas van die school aan bod komen. In het daaropvolgende jaar ontvangt een school één van de drie thematoetsen opnieuw. Daarnaast worden alle dieptethema's binnen een termijn van vijf jaar aan een school toegekend. De thematoetsen zijn specifiek per onderwijsfinaliteit en onderwijsniveau. Elke leerling van een klas legt twee van drie thematoetsen af. Dit betekent bijvoorbeeld dat in een klas van 21 leerlingen, elke thematoets door 14 leerlingen wordt opgelost.

3.3 BREEDTETOETS

3.3.1 Status

Systeemniveau

Het feit dat de breedtetoets 'wiskundige problemen oplossen' jaarlijks bij alle leerlingen wordt afgenomen, impliceert dat de gerealiseerde steekproef op leerling- en schoolniveau zo goed als representatief is voor Vlaanderen. Mogelijks zijn er kleine afwijkingen ten opzichte van de populatie omwille van afwezigheden of (niet-toevallige) vrijstellingen. Op systeemniveau is het dan ook mogelijk de toetsprestaties van de leerlingen afzonderlijk voor het tweede leerjaar secundair onderwijs en het zesde leerjaar secundair onderwijs weer te geven. Om voor elk leerjaar trendanalyses op basis van de opeenvolgende cohortes van leerlingen te kunnen uitvoeren is het daarbij noodzakelijk om (1) per leerjaar

de breedtetoets van de opeenvolgende cohortes van leerlingen van ankeritems te voorzien, en (2) om de toetsprestaties weer te geven aan de hand van een geschatte IRT-score. In functie van interpreteerbaarheid ondergaat de IRT-score het best een lineaire transformatie. Op de meetschaal kunnen één of meerdere cesuren worden gelegd overeenkomstig het aantal beheersingsniveaus waarbinnen men de resultaten van de Vlaamse leerlingen wil situeren.

Daarnaast wordt in het toetsdesign vermeld dat de breedtetoets 'wiskundige problemen oplossen' specifiek zal zijn per onderwijsniveau, onderwijsstroom en onderwijsfinaliteit. Aangezien we ons hier beperken tot uitspraken over status, wordt niet dieper ingegaan op de specificiteit per onderwijsniveau. Binnen het basisonderwijs worden geen specifieke toetsen voorzien. In het secundair onderwijs betekent specificiteit naar onderwijsstroom dat er een aparte breedtetoets zal ontwikkeld worden voor de leerlingen van de A-stroom en een aparte breedtetoets voor de leerlingen van de B-stroom. Dit is ook wenselijk gezien de specifieke vereisten van het curriculum en de onderwijscontext van de A-stroom en de B-stroom, hoewel basisgeletterdheid ook als gemeenschappelijk kan worden beschouwd voor zowel de A- als de B-stroom. De mate waarin de mogelijkheid wordt gecreëerd om de breedtetoets van de A-stroom en de breedtetoets van de B-stroom via ankeritems op dezelfde meetschaal te plaatsen heeft implicaties voor de uitspraken inzake leerprestaties van de leerlingen van het tweede leerjaar secundair onderwijs.

Indien wordt beslist om de toetsen volledig gescheiden te houden, dan zal ook de rapportering over de leerprestaties van de leerlingen van de A-stroom en leerlingen van de B-stroom volledig gescheiden blijven. Er is immers geen gemeenschappelijke meetschaal die een vergelijking mogelijk maakt. Indien er ankeritems in de breedtetoetsen van de A-stroom en de B-stroom worden geïntegreerd én men slaagt erin een gemeenschappelijke meetschaal te creëren, dan kunnen de leerprestaties van de leerlingen van de A-stroom ook gesitueerd worden ten opzichte van de leerlingen van de B-stroom en omgekeerd.

Wat betreft de specificiteit naar onderwijsfinaliteit in de derde graad van het secundair onderwijs geldt dezelfde redenering. Indien de onderwijsvorm-specifieke breedtetoetsen volledig gescheiden van elkaar worden ontwikkeld of de toetsen ondanks ankeritems niet op dezelfde meetschaal kunnen worden geplaatst – dit laatste is mogelijk gezien de zeer diverse context van het curriculum en de wiskundeoefeningen – kan enkel per onderwijsfinaliteit over de leerprestaties van de leerlingen gerapporteerd worden. In dat geval is het ook niet mogelijk een vergelijking te maken tussen de verschillende onderwijsfinaliteiten.

School- en klasniveau

Naast de representativiteit van de steekproef dient ook het aantal leerlingen binnen een school of klas voldoende groot te zijn. In sommige scholen is het aantal leerlingen dat is ingeschreven in de A-stroom of B-stroom echter vrij beperkt waardoor voor die scholen de uitspraken over de prestaties van de leerlingen op schoolniveau minder betrouwbaar zullen zijn. De leerlingenaantallen van het schooljaar 2020-2021 geven bijvoorbeeld aan dat de A-stroom in 40 van de 725 scholen minder dan 20 leerlingen telt. In de B-stroom gaat het om 1 op de 3 scholen (nl. 141 van de 415 scholen). Daarnaast is ook mogelijk dat in sommige klassen het aantal deelnemende leerlingen te klein zal zijn om betrouwbare uitspraken te doen over de toetsprestaties. Dit is bijvoorbeeld mogelijk in de derde graad BSO waar sommige klassen uit heel kleine leerlinggroepen bestaan.

Het Steunpunt en de opdrachtgever dienen bijgevolg een aantal richtlijnen te ontwikkelen over de vereisten waaraan de rapportage voor kleine leerlingengroepen dient te voldoen. Dergelijke richtlijnen kunnen bijvoorbeeld het minimum aantal vereiste leerlingen omschrijven voor rapportage op school- en klasniveau. Een alternatief bestaat erin om alle leerlinggroepen – inclusief kleine klassen - op te nemen in de rapportage, mits daarbij voldoende aandacht wordt besteed aan de communicatie inzake betrouwbaarheid. Dit kan bijvoorbeeld door het betrouwbaarheidsinterval mee te geven en te bespreken (zie aanbevelingen in sectie 2.4). De rapportage van een betrouwbaarheidsinterval maakt de feedback eenvoudiger en meer eenvormig. Er is immers geen aangepaste rapportering meer nodig voor verschillende groepen. Daarnaast toont de rapportering met een betrouwbaarheidsinterval ook of een school of klas wel of niet duidelijk verschilt ten aanzien van een of meerdere cesuren.

Leerlingniveau

Onder ideale omstandigheden kan de toetsscore en het daarbijhorende beheersingsniveau van iedere leerling op de breedtetoets zonder problemen gerapporteerd worden. Die rapportage en de interpretatie van de resultaten vraagt echter in de praktijk om de nodige voorzichtigheid. Elke toetsscore wordt naast de te meten onderliggende vaardigheid immers ook beïnvloed door toevallige fouten en meetfouten. Toevallige fouten hebben te maken met de toetscondities en specifieke leerlingkenmerken die zich manifesteren tijdens de toetsafname, zoals een leerling die minder gemotiveerd is of zich wat ziek voelt. Meetfouten zijn omgekeerd evenredig met de betrouwbaarheid van de toets en bijgevolg sterk gerelateerd aan de precisie waarmee de te meten vaardigheid kan vastgesteld worden. Binnen de centrale toetsen is het noodzakelijk dat de geschatte toetsscore een heel hoge betrouwbaarheid heeft. De meetfout van de geschatte toetsscore van elke individuele leerling dient met andere woorden zo klein mogelijk te zijn (zie ook sectie 2.1). We beschouwen dit als noodzakelijk om uitspraken te doen over

toetsprestaties op leerlingniveau. De resultaten kunnen immers worden meegenomen in de beoordeling van de leerling door de klassenraad.

Indien de breedtetoets 'wiskundige problemen oplossen' adaptief wordt aangeboden, wordt alvast een noodzakelijke stap genomen om de meetfout zo klein mogelijk te houden. Naast de adaptiviteit van de toetsen strekt het tot de aanbeveling om op basis van de resultaten van het kalibratieonderzoek na te gaan of de breedtetoets(en) voldoet aan de vereiste betrouwbaarheidscoëfficient van minimum .85 tot .90 (Johnson, Penny, & Gordon, 2009; NRC, 2014). Het gebruik van een algemene betrouwbaarheids-coëfficient is daarbij onvoldoende om individuele leerlingresultaten weer te geven. Dit dient te gebeuren aan de hand van een standaardmeetfout die de (on)zekerheid van de gerapporteerde toetsscores (en ook leerwinst) van een leerling in kaart brengt (Brennan, 2001). Bijgevolg is het ook op leerlingniveau aan te raden de betrouwbaarheidsintervallen rond de geschatte toetsprestatie te rapporteren.

3.3.2 Leerwinst

Leerlingniveau

Leerwinst op leerlingniveau kan beschreven worden als het verschil in de toetsscore van een leerling op de breedtetoets 'wiskundige problemen oplossen' in het zesde leerjaar secundair onderwijs en de toetsscore van diezelfde leerling op de breedtetoets in het tweede leerjaar secundair onderwijs. Nog meer dan bij de individuele toetsprestaties op één meetmoment, dient ook hier nagegaan te worden in welke mate de berekende leerwinst van een leerling voldoende betrouwbaar is. De leerwinst van de leerlingen inzake wiskundige problemen oplossen zal immers het verschil zijn van twee geschatte toetsscores waaraan ook twee meetfouten verbonden zijn (Harvill, 1991). Enkel als de betrouwbaarheid van de verschilscore ≥.85, zijn de uitspraken over de leerwinst vanuit psychometrisch standpunt geschikt om mee te nemen in de beoordeling van die leerling. Het strekt tot de aanbeveling dit op basis van de data van het kalibratieonderzoek na te gaan.

De tweede belangrijke voorwaarde waaraan voldaan dient te worden in functie van de rapportage van individuele leerwinst is dat de breedtetoetsen van verschillende leerjaren doorheen de schoolloopbaan op een gemeenschappelijk meetschaal kunnen geplaatst worden. Om tot een gemeenschappelijke meetschaal te kunnen komen over verschillende leerjaren, is het in de eerste plaats belangrijk dat er een gemeenschappelijke inhoudelijke basis is voor de toetsen in bijvoorbeeld het tweede en het zesde leerjaar secundair onderwijs. In die zin kan een gemeenschappelijke meetschaal als ontwikkelingsgericht worden beschouwd (Lissitz & Huynch, 2002). Concreet betekent dit dat het domein 'wiskundige problemen oplossen' voldoende moet kunnen losgekoppeld worden van de vakspecifieke inhouden die in de

verschillende leerjaren (en bij uitbreiding onderwijsniveaus en onderwijsfinaliteiten) aan bod komen. Gezien het transversale karakter van wiskundig probleemoplossen lijkt dit inhoudelijk mogelijk.

Daarnaast vormt de afstand die overbrugd dient te worden tussen met name het tweede en zesde leerjaar van het secundair onderwijs mogelijks wel een probleem om leerwinst aan de hand van de breedtetoetsen vast te stellen. Meer algemeen kan immers worden gesteld dat hoe verder de leerjaren uit elkaar liggen hoe minder betrouwbaar, d.w.z. hoe groter de equating error, de gemeenschappelijke schaal is (Wu, 2010). Aangezien ongeveer 83% van de equating error wordt bepaald door de ankeritems die worden gebruikt om de toetsen aan elkaar te linken (Michaelides & Haertel, 2004), is het aan te raden te onderzoeken welke items uit de breedtetoets 'wiskundige problemen oplossen' het meest geschikt zijn als ankeritems. Verkeerd gekozen ankeritems kunnen er immers toe leiden dat een aanzienlijk deel van de vastgestelde leerwinst te wijten is aan equating error. De schaalontwikkeling evenals de selectie van geschikte ankeritems om leerwinstmeting toe te laten, vergt ideaal gezien een simultane kalibratie over leerjaren heen, terwijl dit niet voor alle leerjaren voorzien is in de eerste kalibraties[17]. Daarnaast zou via een kleinschalige studie kunnen onderzocht worden in welke mate de schalen van het tweede en zesde leerjaar via de tussenliggende leerjaren aan elkaar zouden kunnen gekoppeld worden. Dit zou mogelijks kunnen via een common item design op basis van de kalibratiedata en een bijkomende afname in het derde, vierde en vijfde leerjaar (Kolen & Brennan, 2004).

Ten derde dienen de breedtetoetsen over de onderwijsfinaliteiten en onderwijsstromen heen op dezelfde meetschaal geplaatst te kunnen worden. Zoals aangegeven worden de breedtetoetsen specifiek ontwikkeld voor elke onderwijsfinaliteit en -stroom. Indien de breedtetoets van het zesde leerjaar doorstroom- en dubbele finaliteit bijvoorbeeld niet op dezelfde meetschaal kunnen geplaatst worden, krijgt de berekende leerwinst van een leerling uit een doorstroomrichting en een leerling uit een richting met dubbele finaliteit een andere betekenis. In het geval geen gemeenschappelijke meetschaal kan ontwikkeld worden, zorgt de combinatie van onderwijsstromen en onderwijsfinaliteiten ervoor dat er tot zes (2x3) basistrajecten zijn waarvoor leerwinst van leerlingen kan berekend worden. Voor elk van deze basistrajecten kan dan niet gegarandeerd worden dat de geschatte leerwinst met elkaar kan worden vergeleken. Bovendien kunnen de individuele schoolloopbanen van de leerlingen langsheen deze basistrajecten ervoor zorgen dat de leerwinstresultaten van de leerlingen van eenzelfde klas of school niet met elkaar te vergelijken zijn. Zo is het mogelijk dat niet alle leerlingen van een klas (dus zelfde

[17] Werkdomein F bereidt momenteel een nota voor waarin de voorwaarden voor verticale (en horizontale) equivalering verder worden uitgewerkt voor de centrale toetsen.

onderwijsfinaliteit) ook leerwinstresultaten hebben (afhankelijk van onderwijsstroom). Bovendien zullen de leerwinstresultaten van leerlingen van een school of klas niet noodzakelijk vergelijkbaar zijn (afhankelijk van de combinatie onderwijsstroom/-finaliteit. Zo zal de interpretatie van de individuele leerwinst van een leerling die gestart is in de A-stroom en uiteindelijk in klas X terecht komt in de arbeidsmarktfinaliteit verschillen van de interpretatie van een leerling die gestart is in de B-stroom en uiteindelijk in klas X is terechtgekomen. Beide leerlingen hebben immers een ander basistraject gevolgd, elk met zijn eigen gemeenschappelijke meetschaal.

Indien zou blijken dat het zowel inhoudelijk als statistisch mogelijk is om de leerwinst inzake wiskundige problemen oplossen voor elke leerling in kaart te brengen, kan eventueel ook een benchmark voor voldoende leerwinst ontwikkeld worden. Zo kan onder meer worden vastgelegd hoeveel punten op de gemeenschappelijke meetschaal of standaarddeviaties een leerling erop vooruit moet gaan om van voldoende leerwinst te spreken. Die mate waarin de leerwinst van een leerling zich situeert ten opzichte van de ontwikkelde benchmark dient daarbij steeds in combinatie te worden bekeken ten aanzien van de einddoelen (eindtermen) (zie ook 1.2 Leerwinst). Zo is het bijvoorbeeld mogelijk dat een leerling meer dan voldoende leerwinst maakt tussen het tweede en zesde leerjaar maar toch de eindtermen niet bereikt.

Daarnaast dient onderzocht te worden in welke mate een leerwinstbenchmark universeel dient te zijn voor alle leerlingen. Onderzoek toont bijvoorbeeld aan dat leerlingen met lagere aanvangsprestaties (gemiddeld genomen) meer leerwinst maken dan leerlingen met hogere aanvangsprestaties. Bijgevolg zouden leerlingen met hoge aanvangsprestaties benadeeld worden ten opzichte van leerlingen met lage aanvangsprestaties indien een universele benchmark voor leerwinst wordt gehanteerd. Ook hier komt de noodzaak aan een combinatie van een leerwinstbenchmark en de einddoelen duidelijk naar voor. De leerlingen met de hoge aanvangsprestaties zullen naar alle waarschijnlijkheid wel de eindtermen beheersen op het einde van het zesde leerjaar ondanks het gegeven dat ze minder leerwinst maakten. Analoog kan men zich de vraag stellen in welke mate een universele leerwinstbenchmark voor de verschillende onderwijsfinaliteiten zinvol is. Om een benchmark te ontwikkelen voor de periode tweede leerjaar – zesde leerjaar kan eventueel gecontroleerd worden of de leerwinst van de leerlingen inzake wiskundig probleemoplossen zich lineair ontwikkelt over de verschillende leerjaren van het secundair onderwijs. De eerder vermelde bijkomende data-afname in het derde, vierde en vijfde leerjaar alsook de kalibratiedata zouden hiervoor gebruikt kunnen worden.

Schoolniveau

Wanneer – onder de hierboven beschreven voorwaarden - leerwinst op individueel niveau kan berekend worden voor 'wiskundige problemen oplossen', kan deze in principe ook geaggregeerd worden naar het schoolniveau. De vraag stelt zich echter onder welke omstandigheden deze aggregatie zinvol is. Naast de vereiste dat het aantal leerlingen per school voldoende groot is (zie ook 3.2.1 Status) is het niet altijd mogelijk om leerlingen eenduidig aan een school toe te wijzen voor de leerwinstberekening of ontbreken er leerwinstgegevens van sommige leerlingen. Meer specifiek zijn het fenomenen zoals leerlingenmobiliteit, vroegtijdig schoolverlaten en zittenblijven die de aggregatie van leerwinst op leerlingniveau naar leerwinst op schoolniveau bemoeilijken. Zittenblijven is in de context van de centrale toetsen in principe (en op termijn) niet problematisch, mits de leerwinst wordt berekend op basis van de toetsscores van leerlingen op individueel niveau (niet op cohorteniveau). Wat betreft leerlingenmobiliteit, wisselen heel wat leerlingen van school tussen het tweede en het zesde leerjaar secundair onderwijs en bijgevolg tussen de twee momenten waarop ze de breedtetoets afleggen. Wanneer de breedtetoets populatiebreed is afgenomen, brengen leerlingen die van school veranderen wel een toetsscore mee. Maar wanneer een leerling bijvoorbeeld op twee of drie scholen heeft gezeten tussen beide meetmomenten, stelt zich de vraag of en aan welke school/scholen deze leerling dan dient toegewezen te worden om de leerwinst op schoolniveau in kaart te brengen. Met name stelt deze vraag zich voor de interschoolse mobiliteit op niet-reguliere momenten in de schoolloopbaan (cf. ook sectie 2.5.2).[18] Daarnaast stelt zich ook de vraag hoe rekening kan gehouden worden met leerlingen die vroegtijdig, i.e., voor de afname van de breedtetoets in het zesde leerjaar, de school verlaten. Voor zover we weten is er tot op heden geen standaardmethode voorhanden om leerwinst op schoolniveau te aggregeren, rekening houdend met de bovenstaande problemen (maar zie ook verder bij toegevoegde waarde in sectie 3.3.3). Binnen het Centrum voor Onderwijseffectiviteit en -evaluatie (KU Leuven) loopt op dit ogenblik een doctoraatsonderzoek waarbinnen een dergelijke indicator van leerwinst op schoolniveau die rekening houdt met leerlingenmobiliteit en vroegtijdig schoolverlaten wordt ontwikkeld.

[18] De schoolveranderingen van leerlingen op de typische momenten waarop leerlingen van school veranderen (de overgang van lager naar secundair onderwijs, de overgang na de eerste graad ingeval van middenscholen) vormen geen onmiddellijk probleem voor de toewijzing van de gemeten leerwinst aan een school, door de combinatie van twee elementen. Ten eerste sluiten de afnamemomenten van de centrale toetsen nauw aan op die typische overgangsmomenten. Ten tweede kan de gemeten leerprestatie in een bepaald leerjaar benaderend ook geïnterpreteerd worden als het aanvangsniveau van het volgende leerjaar. Zo kan de meting in het zesde leerjaar lager onderwijs worden beschouwd als een beginmeting bij het begin van het secundair onderwijs.

Het Steunpunt en de opdrachtgever kunnen anderzijds beslissen om de geaggregeerde leerwinst enkel te berekenen op basis van de normaalvorderende leerlingen die binnen dezelfde school blijven. In dat geval zal de berekende leerwinst in veel gevallen een overschatting zijn van de daadwerkelijk leerwinst van de school.

Systeemniveau

Hoewel leerlingenmobiliteit hier minder aan de orde is, zijn er ook op systeemniveau problemen die vergelijkbaar zijn met die van het schoolniveau. Zo is het niet duidelijk in welke mate er rekening dient gehouden te worden met vroegtijdige schoolverlaters.

Samenvattend kan gesteld worden dat – mits technisch haalbaar – de resultaten van de breedtetoetsen 'wiskundige problemen oplossen' in het secundair onderwijs kunnen gebruikt worden om uitspraken te doen over leerwinst van individuele leerlingen. Uitspraken over leerwinst op school- en systeemniveau zijn mogelijk maar dienen steeds gecontextualiseerd te worden binnen een bepaalde leerlingenpopulatie (vb. zonder vroegtijdig schoolverlaters). De uitspraken op school- en systeemniveau zullen daarom steeds een (licht) vertekend beeld geven. Modelmatige schattingen kunnen (deels) tegemoet komen aan de opgesomde problemen, wat ons brengt bij toegevoegde waarde.

3.3.3 Toegevoegde waarde

Schoolniveau

De breedtetoetsen 'wiskundige problemen oplossen' worden over de onderwijsniveaus en onderwijsfinaliteiten heen populatiebreed afgenomen. Daarmee is voldaan aan de voorwaarde van een representatieve steekproef op leerling- en schoolniveau.[19] Daarnaast zullen het aantal scholen en het aantal leerlingen per school naar alle waarschijnlijkheid ook voldoende groot zijn voor de toegevoegdewaarde-analyses (maar gelijkaardige beperkingen gelden als hierboven vermeld voor statusmeting op schoolniveau). In het geval van kleinere analyse-eenheden zoals klassen, studierichtingen of onderwijsfinaliteiten, stelt zich echter de vraag in welke mate er voldoende leerlingen per analyse-eenheid zijn in functie van een betrouwbare schatting van de toegevoegde waarde. Ook in het geval van kleinere analyse-eenheden betreft het hier nog steeds de TW van een school. De doorlooptijd van vier

[19] De gerealiseerde steekproef kan evenwel nog steeds afwijken van de populatie, wanneer sprake is van selectieve non-respons. Concreet zullen er steeds niet-deelnemende leerlingen (of scholen) zijn, als gevolg van vrijstellingen, ziekte, technische problemen, … (cf. sectie 3.3.1). In het bijzonder voor de vergelijking van scholen en voor de bepaling van trends dient de mate en wijze van non-respons bewaakt te worden.

jaar tussen de beide toetsmomenten in het secundair onderwijs en het feit dat leerlingen met meerdere leerkrachten worden geconfronteerd zorgen ervoor dat de bijdrage aan het leren niet kan worden toegewezen aan een specifieke klas of leerkracht. In dat geval betreft het dus de toegevoegde waarde van een school voor een specifieke klas.

De centrale toetsen werden in het leven geroepen vanuit de perspectieven van verantwoording en onderwijsverbetering. Om de resultaten van de breedtetoetsen maximaal en zo valide mogelijk te kunnen inzetten voor beide perspectieven strekt het tot de aanbeveling om de toegevoegde waarde van de scholen te schatten aan de hand van een type-B model of een type-X effect. Beide modellen controleren immers voor sociaaldemografische en -economische leerlingkenmerken alsook voor schoolcompositie-effecten waardoor de toegevoegde waarde van een school binnen deze beide types modellen beschouwd kan worden als een indicator van schoolpraktijken. Beide modellen sluiten ook het meest aan bij de principes van accountability en schoolverbetering. Zoals eerder vermeld is het belangrijk dat de covariaten en factoren die in het model worden geïntegreerd zo juist en volledig mogelijk zijn om de sociaaldemografische en -economische kenmerken zo correct mogelijk te operationaliseren. Een grondige studie van de gewenste kenmerken die in de modellen dienen geïntegreerd te worden is dan ook aangeraden. Indien kenmerken geselecteerd worden waarvan de overheid niet over gegevens beschikt, dan dienen deze bij alle leerlingen in alle scholen bevraagd te worden. Deze zijn immers nodig om een eerlijke vergelijking tussen alle scholen mogelijk te maken.

Zoals ook hierboven voor leerwinst werd aangehaald, is de leerlingpopulatie van een school die deelneemt in het tweede leerjaar niet dezelfde als de leerlingpopulatie in het zesde leerjaar. Zo zijn er bijvoorbeeld leerlingen die veranderen van school (interschoolse mobiliteit), blijven zitten of ongekwalificeerd uitstromen. Bij het schatten van de toegevoegde waarde van de school dienen deze vormen van irreguliere schoolloopbanen in rekening te worden gebracht tijdens de modellering. Dit kan onder meer door gewichten toe te kennen aan het aantal jaar dat een leerling op een school vertoeft of het gebruik van cross-classificatie en multiple-membership modellen (zie ook sectie 2.5.2). Indien dit niet gebeurt, zal de geschatte toegevoegde waarde van een school enkel betrekking hebben op de normaalvorderende leerlingen die binnen dezelfde school blijven gedurende het secundair onderwijs. In de meeste gevallen zal dit een vertekening van het echte schooleffect met zich meebrengen (zie boven in sectie 3.3.2).

Om de toegevoegde waarde van de scholen te schatten is het niet vereist dat de toetsen van het tweede en het zesde leerjaar op een gemeenschappelijke schaal worden geplaatst. De mate waarin binnen een leerjaar de verschillende toetsen op een schaal kunnen worden geplaatst zal echter wel een invloed uitoefenen op de mogelijkheden om schooleffecten te rapporteren. Gezien de – weliswaar wenselijke -

specificiteit waarmee de breedtetoetsen worden ontwikkeld, is er geen garantie dat de breedtetoetsen voor de A-stroom en B-stroom op dezelfde schaal kunnen geplaatst worden. Hetzelfde geldt voor de specifieke breedtetoetsen van de verschillende onderwijsfinaliteiten in het zesde leerjaar. De mate waarin dit al dan niet lukt zal bepalend zijn voor het groeperingsniveau waarbinnen uitspraken kunnen worden gedaan over schooleffecten, zoals bijvoorbeeld op het niveau van de totale school of, enkel op niveau van het ASO.

In een eerste scenario kunnen de breedtetoetsen van het tweede leerjaar op een gemeenschappelijke meetschaal worden geplaatst en die van het zesde leerjaar ook. In dat geval is het mogelijk om zowel uitspraken te doen over de toegevoegde waarde van de school op het niveau van de totale school als voor de specifieke onderwijsfinaliteiten. De toetsscores van het tweede leerjaar A-stroom en B-stroom – die als predictor in het model worden ingebracht – maken dan immers onderdeel uit van dezelfde meetschaal. Bijgevolg is de betekenis van eenzelfde toetsscore van een leerling uit de A-stroom en een leerling uit de B-stroom dan hetzelfde. Hetzelfde geldt voor de toetsscores van de leerlingen uit het zesde leerjaar ASO, TSO en BSO. Enkel dan hebben dezelfde toetsscores over de verschillende onderwijsvormen heen ook dezelfde betekenis. Bij het schatten van de toegevoegde waarde van de school op het niveau van de totale school en per onderwijsfinaliteit is het belangrijk om rekening te houden met de interschoolse leerlingenmobiliteit.

Naast het schatten van de toegevoegde waarde van een school voor een onderwijsfinaliteit, kan ook de toegevoegde waarde van een onderwijsfinaliteit van een school worden geschat. In dat laatste geval is het belangrijk om naast de interschoolse leerlingenmobiliteit ook rekening te houden met de intraschoolse leerlingenmobiliteit. Zo zijn er leerlingen die starten in de A-stroom en dan uiteindelijk via ASO en TSO terechtkomen in het BSO. Voor dergelijke leerlingen dient in rekening gebracht te worden wat dan precies het aandeel van het BSO-onderwijs is in de progressie die de leerling heeft gemaakt.

In een tweede scenario is het niet mogelijk om de breedtetoetsen van het tweede leerjaar op een gemeenschappelijke meetschaal te plaatsen en die van de verschillende onderwijsvormen in het zesde leerjaar wel. Het feit dat de toetsscores van de leerlingen in de A-stroom en de leerlingen in de B-stroom dan niet op dezelfde schaal staan, impliceert dat ze als afzonderlijke predictoren in het model dienen ingebracht te worden. In dat geval kan de toegevoegde waarde van een school in zijn totaliteit niet bekeken worden. Wel kunnen per stroom, en per combinatie van onderwijsfinaliteit en onderwijsstroom schooleffecten geschat worden. Ook hier strekt het tot de aanbeveling de leerlingenmobiliteit in rekening te brengen tijdens de modellering. Schooleffecten die geschat worden voor een specifieke combinatie van onderwijsstroom en onderwijsvorm kunnen in sommige gevallen weinig betrouwbaar zijn omwille

van het beperkt aantal leerlingen in die combinatie. Denk bijvoorbeeld aan leerlingen uit de B-stroom die uiteindelijk in het KSO eindigen.

In een derde scenario kunnen de breedtetoetsen van de verschillende onderwijsvormen in het zesde leerjaar niet op een gemeenschappelijke meetschaal geplaatst worden. Dit impliceert dat de toegevoegde waarde van een school enkel per onderwijsfinaliteit (mits gemeenschappelijke meetschaal in tweede leerjaar) of per combinatie van onderwijsstroom en -finaliteit kan berekend worden.

3.4 THEMATOETSEN

3.4.1 Status

Systeemniveau

Naast de breedtetoets die jaarlijks Vlaanderenbreed wordt afgenomen, worden eveneens jaarlijks 10 dieptethema's getoetst in de A-stroom en 6 in de B-stroom. Per school worden daarbij telkens drie dieptethema's aangeboden in een zogenaamde focustoets, met een dieptethema uit elk van drie inhoudelijke categorieën (voor meer details verwijzen we naar de nota *Centrale Toetsen Wiskunde* van het Steunpunt, 2021). Elk van de thematoetsen dient in minimum 100 scholen te worden afgenomen om variantieparameters zonder bias te bekomen (cf. sectie 2.4). Uitgaande van een eenvoudig (niet geblokt[20]) afnamedesign betekent dit dat er 100 scholen nodig zijn per cluster van drie thematoetsen. Dit resulteert in een minimum van 300 scholen (in het geval één van de drie clusters uit vier thematoetsen zou bestaan) voor de A-stroom en 200 scholen voor de B-stroom. In Vlaanderen zijn 725 secundaire scholen waarbinnen de A-stroom wordt aangeboden en 415 scholen bieden de B-stroom aan. Bijgevolg is aan de minimumvereiste van 100 scholen per toets ruimschoots voldaan.

Een tweede belangrijke voorwaarde is dat de steekproef voor elke thematoets ook representatief is voor de Vlaamse leerlingen- en scholenpopulatie. Daarbij dient op voorhand onderzocht te worden voor welke criteria men die representativiteit wenst in te bouwen en in welke mate die criteria per toets voldoende omvangrijke subpopulaties van scholen en leerlingen creëren om nog betekenisvolle uitspraken te doen. Onderwijsnet is bijvoorbeeld een criterium dat in het peilingsonderzoek en de internationaal vergelijkende onderzoeken in Vlaanderen traditioneel wordt gebruikt om de steekproef zo representatief mogelijk te houden. Van de 415 scholen die de B-stroom aanbieden worden er echter slechts 10% vertegenwoordigd door het officieel gesubsidieerd onderwijs (OGO). Dit betekent dat de zes thematoetsen dienen verdeeld te worden over dit beperkt aantal scholen. Op zich is het beperkte aantal OGO-scholen geen probleem zolang de resultaten van de verschillende onderwijsnetten niet met elkaar worden vergeleken. Indien dit wel het geval is of indien het OGO een representatief beeld wenst van de thematoetsen in de B-stroom, dan dient het aantal deelnemende OGO-scholen per thematoets voldoende groot te zijn.

[20] Een geblokt afnamedesign verwijst naar een design waarbij er overlap is tussen de toetsen die worden afgenomen. Concreet wordt dan een reeks van toetsvragen voorgelegd aan een bepaalde groep en krijgt een andere groep een andere reeks toetsvragen, en krijgen beide groepen daarbovenop eenzelfde set van ankervragen.

Onderwijsnet is naar alle waarschijnlijkheid niet het enige stratificatiecriterium dat zal gebruikt worden om de representativiteit van de steekproef van elke toets voor deelgroepen te garanderen, andere criteria zijn bijvoorbeeld provincie of percentage indicatorleerlingen op een school. De combinatie van deze criteria kan voor heel kleine subpopulaties van scholen zorgen. Een doordachte keuze van stratificatievariabelen en de mogelijkheden van een geblokt afnamedesign dienen dan ook bekeken te worden om de representativiteit van de steekproeven per toets te garanderen. Dit is noodzakelijk om op systeemniveau de toetsprestaties van de leerlingen voor het tweede leerjaar secundair onderwijs weer te geven en deze resultaten jaar na jaar met voldoende zekerheid te kunnen vergelijken. Daarnaast dient diezelfde oefening ook gemaakt te worden voor het zesde leerjaar secundair onderwijs.

Naast de jaarlijkse leerprestaties van de leerlingen kunnen voor elk leerjaar ook trendanalyses op basis van de opeenvolgende cohortes van leerlingen (d.w.z. eenzelfde leerjaar over de schooljaren heen) uitgevoerd worden. Net zoals bij de breedtetoets is het noodzakelijk om (1) per leerjaar elke thematoets van de opeenvolgende cohortes van leerlingen van ankeritems te voorzien, en (2) om de toetsprestaties weer te geven aan de hand van een geschatte IRT-score. Bij de toewijzing van de thematoetsen over de opeenvolgende cohortes dient daarbij ook rekening te worden gehouden met het roterend afnamedesign van de toetsen. Tabel 4 toont (omwille van didactische redenen) een sterk vereenvoudigd design voor negen thematoetsen binnen het tweede leerjaar van de A-stroom. Hierbij ontvangt elke school die tot een specifieke groep (A, B of C) behoort alle thematoetsen over de vijf afnamejaren heen: alsook ontvangt elke school het daaropvolgende jaar één van de drie thematoetsen opnieuw. Daarnaast zijn alle toetsen vertegenwoordigd in de opeenvolgende afnamejaren (over de drie groepen heen) wat trendanalyses op systeemniveau mogelijk maakt. De voorwaarde voor dit scenario is dat de groepen A, B en C representatief zijn tegenover de populatie en elkaar en dat deze groepen gedurende een cyclus van vijf jaar worden aangehouden. We benadrukken dat Tabel 4 louter illustreert dat het binnen het huidige toetsdesign van wiskunde mogelijk is om voor de verschillende thematoetsen trendanalyses uit te voeren die rekening houden met rotatie. De tabel verwijst dus niet naar een definitief afnamedesign noch naar een ontwerpdesign dat door het steunpunt werd ontwikkeld.

Tabel 4: Illustratie trendmeting bij rotatie van thematoetsen

Afnamejaar	Steekproef schoolniveau SO2								
	A (n=242)			B (n=242)			C (n=241)		
2024	1	2	3	4	5	6	7	8	9
2025	1	5	6	4	8	9	7	2	3
2026	5	4	9	8	7	3	2	1	6
2027	4	8	3	7	2	6	1	5	9
2028	8	7	6	2	1	9	5	4	3

Net zoals bij de breedtetoets, wordt in het toetsdesign vermeld dat elke thematoets specifiek zal zijn per onderwijsniveau, onderwijsstroom en onderwijsvorm. De principes die van toepassing zijn bij specificiteit in de breedtetoets gelden ook voor de thematoetsen.

School- en klasniveau

Het probleem inzake het vereiste aantal leerlingen per school en klas is voor de thematoetsen pertinenter dan bij de breedtetoets. Op basis van het huidige toetsdesign zal immers twee derde van de leerlingen van een school een thematoets afleggen. Dit betekent dat de standaardfout van het gemiddelde met 23% toeneemt ten opzichte van een school waar alle leerlingen zouden deelnemen waardoor minder betrouwbare uitspraken mogelijk zijn voor het schoolniveau (en tevens sterkere shrinkage van de schoolresiduen optreedt ingeval van multilevel modelschattingen, cf. sectie 2.2.4). Uitgaande van de (weliswaar strenge) vereiste van 20 leerlingen per groep, betekent dit dat voor ongeveer 88 van de 725 scholen die de A-stroom aanbieden mogelijk onvoldoende betrouwbare uitspraken zullen kunnen gedaan worden op schoolniveau inzake de gemiddelde prestaties van de leerlingen op de thematoetsen. In de B-stroom klimt dit aantal naar 254 (of 61%) van 415 scholen. Op klasniveau zal dit probleem zich nog meer manifesteren. Uitgaande van het feit dat 2/3 van de leerlingen van een klas een thematoets aflegt, tonen de steekproefdata van de peiling wiskunde en Nederlands SO1AB 2022 aan dat 99% van de klassen van de A-stroom en B-stroom onvoldoende groot zijn om aan de vereiste van 20 leerlingen per groep te voldoen. In de A-stroom zullen gemiddeld 12 leerlingen per klas een thematoets maken. In de B-stroom betreft dit gemiddeld 8 leerlingen per klas. In de derde graad zal dit probleem zich nog meer manifesteren aangezien de klasgroepen daar doorgaans uit een beperkter aantal leerlingen bestaan.

Leerlingniveau

Hier gelden dezelfde principes zoals vermeld bij de breedtetoets onder 3.3.1.

3.4.2 Leerwinst

Leerlingniveau

De voorwaarden inzake meetnauwkeurigheid van de geschatte leerwinst en gemeenschappelijke meetschalen die gelden voor de breedtetoetsen zijn ook van toepassing op de thematoetsen (zie 3.3.2). Het probleem met betrekking tot de gemeenschappelijke meetschalen is hier mogelijks nog pertinenter aanwezig dan bij de breedtetoetsen. Omdat de curricula in de eerste en derde graad sterk verschillen, is de kans klein dat voor elk van de verschillende thematoetsen een gemeenschappelijke meetschaal kan geconstrueerd worden en individuele leerwinst kan worden berekend. In veel gevallen verschillen de te toetsen eindtermen hiervoor te sterk over de graden heen. Enkel voor specifieke thematoetsen voor bepaalde modeltrajecten is er mogelijk sprake van voldoende inhoudelijke aansluiting. Voor focustoetsen wiskunde in het secundair onderwijs gaat het dan met name in het bijzonder en enkel over de eindtermen voor de B-stroom en 6BSO. Binnen het lager onderwijs lijkt inhoudelijke aansluiting evidenter, terwijl voor de overgang lager naar secundair onderwijs eveneens moeilijkheden op dat vlak verwacht kunnen worden. Maar ook wanneer inhoudelijke aansluiting mogelijk is, moet nog steeds bekeken worden of een gemeenschappelijke meetschaal mogelijk is.

Systeemniveau

Naast de constructie van de gemeenschappelijke meetschaal dient ook een beslissing genomen te worden of voor bepaalde thematoetsen een gekoppelde afname wordt voorzien voor (groepen van) leerlingen tussen het tweede en zesde leerjaar. Door een gekoppelde afname vergroot het aantal leerlingen voor wie een leerwinstberekening mogelijk wordt. Dit moet evenwel zorgvuldig worden afgewogen tegen mogelijke nadelen van een gekoppeld afnamedesign. De implicaties van een gekoppelde afname verschillen naargelang de wijze van koppeling (het longitudinale afnamedesign), waarbij we een gekoppelde afname voor leerlingencohorten in een school kunnen onderscheiden van een gekoppelde afname voor individuele leerlingen.

- Bij gekoppelde afname op *cohorteniveau* binnen een school worden de thematoetsen die in het tweede leerjaar in een bepaalde school worden afgenomen als uitgangspunt genomen voor de thematoetsen die in die school nadien worden afgenomen. Meer specifiek betekent dit dat scholen die in de eerste graad een thematoets A ontvingen, vier jaar later in het zesde leerjaar de thematoets uit hetzelfde domein dienen te ontvangen. Voor leerlingen met standaard

schoolloopbanen wordt in het tweede en zesde leerjaar daardoor dezelfde thematoets(en) afgenomen. Als gevolg van het rotatiedesign, legden de niet-normaalvorderende leerlingen en schoolmobiele leerlingen evenwel niet (noodzakelijk) dezelfde thematoets(en) af in het tweede leerjaar dan dewelke ze in het zesde leerjaar voorgelegd krijgen.[21]

- Bij een koppeling op *leerlingniveau* worden de thematoetsen die elke individuele leerling in het tweede leerjaar aflegde als uitgangspunt genomen. Concreet betekent dit dat elke leerling die in de eerste graad de thematoets A ontving, minstens vier jaar later in het zesde leerjaar een thematoetsen uit hetzelfde domein zal ontvangen. Als gevolg van het rotatiedesign, krijgen de niet-normaalvorderende leerlingen en schoolmobiele leerlingen dan niet (noodzakelijk) dezelfde thematoets(en) in het zesde leerjaar als hun klas/schoolgenoten met standaard schoolloopbanen.

Indien leerwinstmeting op individueel niveau mogelijk is (cf. inhoudelijke aansluiting en gemeenschappelijke meetschaal), zorgt een afname van een thematoets gekoppeld op leerlingniveau ervoor dat leerwinst kan berekend worden voor een zo maximaal mogelijk aantal leerlingen, maar met belangrijke nadelen voor schoolfeedback (zie verder). Een gekoppelde afname op cohorteniveau maximaliseert daarentegen de schoolfeedback voor leerwinst (zie verder), maar heeft als nadeel dat op systeemniveau de steekproef van leerlingen met een leerwinstmeting niet langer representatief is en met name normaalvorderende niet-mobiele leerlingen sterk oververtegenwoordigd zijn. Voor de bepaling van leerwinst op een thematoets op systeemniveau lijkt een individueel gekoppelde afname daarom te verkiezen boven een cohortegekoppelde afname. Op deze manier worden ook de niet-normaalvorderende leerlingen evenredig opgenomen in de berekening van leerwinst op systeemniveau.

Tot slot maakt een gekoppeld afnamedesign voor thematoetsen, en het meest bij een gekoppelde afname op cohorteniveau, 'teaching to the test' mogelijk in de derde graad. De berekening van leerwinst vereist immers dat de toetsen van beide afnamemomenten zich binnen hetzelfde domein situeren. Anders gezegd: de toetsen die in de eerste graad worden afgenomen informeren de scholen over de toetsen die ze vier jaar later in het zesde leerjaar zullen ontvangen. Voor het bepalen van leerwinst op een thematoets op systeemniveau, is daarom een niet-gekoppeld afnamedesign ook een mogelijk alternatief.

[21] Wanneer binnen een school een toets bij afname slechts aan een deel van de leerlingen wordt voorgelegd (cf. scenario van 2 thematoetsen per leerling en 3 per school), impliceert deze longitudinale toewijzing van toetsen op schoolniveau dat leerwinstmeting ook ontbreekt voor ruim de helft van de normaalvorderende leerlingen. Het lijkt dan aan te bevelen om bij de toewijzing van de thematoets rekening te houden met het thema waarop de leerling voordien werd getoetst. In zekere zin wordt dan een combinatie gemaakt van de cohortegebonden gekoppelde afname en de op leerlingniveau gekoppelde afname.

Gezien de uitspraken over leerwinst zich hier situeren op systeemniveau gelden de beperkingen inzake het vereiste aantal leerlingen immers hier minder. Daar tegenover staat dat er wel dient bekeken te worden of de steekproef nog voldoende representatief is voor de Vlaamse leerlingenpopulatie.

School- en klasniveau

De leerwinst op school- en klasniveau is niks anders dan een geaggregeerde vorm van de leerwinst op leerlingniveau. Individuele leerwinstmeting is dan ook een voorwaarde om de leerwinst op school- en klasniveau mogelijk te maken.

In het geval leerwinstmeting op individueel niveau mogelijk is, gelden dezelfde beperkingen inzake het vereiste aantal leerlingen per klas en school zoals die werden besproken voor de status van de thematoetsen op school- en klasniveau (sectie 3.4.1). Het probleem van het aantal leerlingen stelt zich evenwel nog pertinenter. De wijze waarop hangt opnieuw af van of en hoe de thematoetsen gekoppeld worden afgenomen. Wanneer de afname van de thematoetsen niet gekoppeld wordt, is de kans groot dat een specifieke school voor een bepaalde thematoets onvoldoende leerlingen heeft met een leerwinstmeting om uitspraken te doen op schoolniveau.

Een gekoppelde afname op cohorteniveau heeft als voordeel dat het aantal leerlingen met een leerwinstmeting op school- en klasniveau toeneemt. De berekening van de leerwinst op school- en klasniveau blijft evenwel (in hoofdzaak) beperkt tot de normaalvorderende, niet-schoolmobiele leerlingen van de school. Als gevolg van het rotatiedesign, legden de niet-normaalvorderende leerlingen en schoolmobiele leerlingen immers een andere focustoets en dus niet (noodzakelijk) dezelfde thematoets(en) af in het tweede leerjaar dan dewelke ze in het zesde leerjaar voorgelegd krijgen.

Bij een gekoppelde afname op leerlingniveau kan weliswaar voor het grootste aantal leerlingen leerwinst berekend worden, toch blijft ook dan de leerwinstberekening op school- of klasniveau (in hoofdzaak) beperkt tot de normaalvorderende, niet-mobiele leerlingen. Voor de niet-normaalvorderende leerlingen is de leerwinstmeting (doorgaans) niet voor dezelfde thematoetsen als voor de normaalvorderende leerlingen van de school. Leerlingen die bijvoorbeeld een jaar dubbelen kregen in het tweede leerjaar een andere focustoets en bijgevolg in veel gevallen een andere thematoets (zie bv. Tabel 4). Voor leerlingen die tussen beide afnamemomenten van school wisselen heeft de leerwinst betrekking op de thematoets(en) die werd afgenomen in de school waar de leerling zat in het tweede leerjaar. Dit alles maakt dat ook bij een individueel gekoppelde afname van een thematoets, de leerwinstanalyse op school- of klasniveau voor de meeste scholen beperkt blijft tot de normaalvorderende leerlingen van een school. Een individueel gekoppelde afname van een thematoets heeft verder als belangrijk nadeel dat ook de statusmeting op schoolniveau voor een thematoets in het zesde leerjaar (cf. 3.3.1) enkel nog

(hoofdzakelijk) betrekking heeft op de normaalvorderende niet-mobiele leerlingen. Sommige leerlingen met een irreguliere schoolloopbaan zullen in het zesde leerjaar immers een andere thematoets toegewezen krijgen dan hun klas- en schoolgenoten.

Bovenstaande toont aan dat er aan verschillende voorwaarden voldaan zal moeten worden opdat het mogelijk zou zijn om de individuele leerwinst van de Vlaamse leerlingen op de thematoetsen tussen het tweede en zesde middelbaar op een vlotte en betrouwbare manier in kaart te brengen. Onder meer de keuze over de manier waarop de focustoetsen van het tweede en zesde leerjaar aan elkaar worden gekoppeld (individueel – cohorte) zal daarbij doorslaggevend zijn. Meer specifiek zal de aard van de koppeling bepalen in welke mate er leerwinst op individueel dan wel op school- of klasniveau kan berekend worden. Daarnaast zal de berekening van leerwinst op school- en klasniveau steeds beperkt blijven tot de normaalvorderende leerlingen van een school, ongeacht de koppeling tussen de focustoets van het tweede en zesde leerjaar gebeurt op leerling- of cohorteniveau.

3.4.3 Toegevoegde waarde

Schoolniveau

De principes inzake toegevoegde waarde van scholen op basis van de breedtetoetsen gelden ook voor de thematoetsen (zie 3.2.3). Daarnaast zijn er een aantal aandachtspunten die de mogelijkheden voor het schatten van de toegevoegde waarde van scholen voor de thematoetsen kunnen beïnvloeden. Meer specifiek – en net zoals bij leerwinst - betreft het de invloed van de wijze waarop de thematoetsen gekoppeld worden, i.e. koppeling op leerlingniveau, koppeling op cohorteniveau en geen koppeling.

Terwijl de berekening van leerwinst een gemeenschappelijke meetschaal vereist, is dit voor het schatten van de toegevoegde waarde van een school niet nodig. Hoewel niet noodzakelijk, gaat men er meestal wel van uit dat de toetsscores van beide meetmomenten betrekking hebben op hetzelfde onderdeel van het curriculum. Concreet betekent dit dat een thematoets van het tweede en zesde leerjaar over hetzelfde wiskundige domein gaat. We herhalen hier dat een herhaalde toetsing van eenzelfde inhoudelijk domein over leerjaren heen niet voor alle onderwijsniveaus, -domeinen of onderwijsfinaliteiten evident is. Voor wiskunde bieden de eindtermen over het tweede en zesde leerjaar heen bijvoorbeeld enkel voldoende aanknopingspunten voor bepaalde dieptethema's en louter binnen de arbeidsmarktfinaliteit (B-stroom en 6BSO).

Zowel de koppeling op leerlingniveau als de koppeling op cohorteniveau houden mogelijkheden en beperkingen in voor het bepalen van de toegevoegde waarde van een school voor een bepaald thema. Het is dan ook aan het steunpunt en de opdrachtgever om hier een keuze in te maken.

In het geval de thematoetsen worden gekoppeld op cohorteniveau, zal een specifieke thematoets beschikbaar zijn voor alle leerlingen van het zesde leerjaar terwijl dezelfde thematoets van het tweede leerjaar enkel volledig beschikbaar zal zijn voor de normaalvorderende leerlingen (zie 3.3.2). Concreet betekent dit dat de status of toetsscore van elke leerling in het zesde leerjaar (de afhankelijke variabele in het model) beschikbaar is. De toetsscore van het tweede leerjaar (die als predictor in het model dient te worden geïntegreerd) is daarentegen enkel beschikbaar voor de normaalvorderende leerlingen en een beperkt deel van de niet-normaalvorderende leerlingen. Een mogelijkheid zou zijn om de ontbrekende toetsscores van het tweede leerjaar (voorafgaande meting) via imputatie te integreren in het model. Imputatie van ontbrekende toetsscores wordt heel ruim toegepast in internationale prestatie-onderzoeken zoals PISA. Een meer precieze imputatie kan in het geval van de centrale toetsen gebeuren op basis van de andere, gekende toetsscores (o.m. voor de breedtetoets van het tweede leerjaar) en achtergrondkenmerken van de niet-normaalvorderende leerlingen. Indien voor de piste van imputatie wordt gekozen, dient een bijkomend onderzoek na te gaan wat de meest geschikte vorm van imputatie is en in welke mate deze invloed heeft op de geschatte toegevoegde waarde.

Indien de koppeling gebeurt op leerlingniveau, zal de thematoets van het zesde leerjaar die dat jaar centraal staat in de school enkel beschikbaar zijn voor normaalvorderende leerlingen (zie 3.3.2). Specifiek zullen sommige leerlingen met een irreguliere schoolloopbaan in het zesde leerjaar een andere thematoets toegewezen krijgen dan hun klas- en schoolgenoten. Bijgevolg kunnen de data van deze irreguliere leerlingen niet meegenomen worden om de gemiddelde status (de afhankelijke variabele in het model) van de leerlingen in de klas en school te schatten. De traditionele aanpakken om tijdens de schatting van de toegevoegde waarde van scholen rekening te houden met alternatieve schoolloopbanen, i.e., multiple membership modeling en multiple imputation bieden hiervoor echter geen oplossing. Het fundamentele verschil tussen de reguliere leerlingen en leerlingen met een alternatieve schoolloopbaan zit hem immers in de meest fundamentele variabelen van het toegevoegde waardemodel, nl. de toetsscores in het zesde leerjaar. Wel kan onderzocht in welke mate een multivariate schatting - waarbij meerdere outcomes van het zesde leerjaar worden gecombineerd (i.c. thematoetsen en breedtetoets) - tot een goede benadering van de (multidimensionele) toegevoegde waarde van de school leidt.

Net zoals bij leerwinst, impliceert de koppeling van de focustoetsen nog steeds dat de kans op ‘teaching to the test’ in het zesde leerjaar wordt vergroot. De scholen krijgen in het tweede leerjaar immers slechts drie thematoetsen en in functie van de toegevoegde waarde-berekening dienen deze in het zesde leerjaar tot hetzelfde domein te behoren.

Bovenstaande toont aan dat de schatting van de toegevoegde waarde van een school niet probleemloos zal verlopen, en dit zowel onder het scenario van een koppeling op cohorteniveau als onder het scenario

van een koppeling op leerlingniveau. De vraag stelt zich dan ook naar de meerwaarde en noodzaak van gekoppelde toetsen om de toegevoegde waarde van een school te berekenen. Een mogelijk alternatief is om de toegevoegde waarde van een school te schatten op basis van niet-gekoppelde toetsen. Concreet betekent dit dat de toetsscores van de focustoets in het zesde leerjaar niet worden gecontroleerd voor de toetsscores van een focustoets van hetzelfde domein in het tweede leerjaar, maar wel voor de toetsscores op een andere toets. In die zin zou kunnen nagegaan worden in welke mate kan gecontroleerd worden voor de toetsscores van de breedtetoets in het tweede leerjaar. De veronderstelling daarbij is weliswaar dat er voor de breedtetoets een gemeenschappelijke schaal voor de A-stroom en B-stroom kan ontwikkeld worden. Een belangrijke voorwaarde daarbij is dat de toegevoegde waarde van een school niet fundamenteel verandert naargelang de gebruikte predictor, nl. een schooleffect op basis van een gekoppelde focustoets mag in principe niet verschillen van het schooleffect op basis van de breedtetoets. Dit vraagstuk zou kunnen worden onderzocht aan de hand van een bijkomende studie op basis van de TIMSS-repeat data.

Geschatte schooleffecten (toegevoegde waarde) kunnen over de tijd verschillen. Wijzigingen over de tijd kunnen wijzen op werkelijke veranderingen op schoolniveau, maar kunnen eveneens te maken hebben met instabiliteit van schattingen die te wijten is aan statistische onzekerheid als gevolg van kleine aantallen en bijvoorbeeld toevallige fluctuaties in leerlingsamenstelling (Koedel, Mihaly, & Rockoff, 2015). Om weerstand te bieden aan het probleem van instabiele schooleffecten wordt daarom soms de gemiddelde toegevoegde waarde van een school genomen van bijvoorbeeld drie aaneensluitende schooljaren (beter is om stabiliteit te analyseren via gepoolde data-analyse, zie ook sectie 2.4.1). Het roterende afnamedesign zorgt er echter voor dat een thematoets geen twee of drie aaneensluitende schooljaren binnen dezelfde school wordt afgenomen. Dit beperkt de mogelijkheden voor het (accurater) schatten van toegevoegde waarde van een (kleine) school op basis van meerdere opeenvolgende metingen. De cohortes die dezelfde thematoets binnen een school zullen afleggen, liggen immers verschillende schooljaren uit elkaar. Dat beperkt ook de mogelijkheden voor monitoring van trends op schoolniveau, waarbij wordt nagegaan in welke mate verschillen over de tijd te wijten zijn aan echte verandering dan wel aan schattingsfouten.

3.5 BESLUIT

In de eerste twee delen van deze nota werden de voorwaarden, richtlijnen en aanbevelingen aangaande de meting en analyse van leerprestaties en leerwinst besproken op basis van een uitgebreide screening van de internationale literatuur. Dit laatste deel verduidelijkt de methodologische en statistische

implicaties van bepaalde al gemaakte of nog te maken keuzes door de meer algemene principes concreter toe te passen op de centrale toetsen in het Vlaams onderwijs. Met name gaat het om de aanwezigheid van de voorwaarden voor leerwinstmeting enerzijds en de implicaties van mogelijke afnamedesigns op steekproefgroottes en representativiteit voor rapportering of analyse op verschillende niveaus: leerling, school en systeem.

Voorwaarden voor leerwinstmeting

Waar in de eerste delen aanbevelingen omtrent de meetnauwkeurigheid van de toetsen en leerwinst centraal staan, maakt de toepassing in het derde deel duidelijk dat aan de voorwaarden voor het meten van leerwinst niet altijd voldaan zal kunnen worden. Zo dient om te beginnen de *inhoudelijke basis* van de te meten vaardigheden voldoende stabiel te zijn over de leerjaren of schoolloopbaan van leerlingen heen. Bij de centrale toetsen worden de te meten vaardigheden gekoppeld aan te meten eindtermen. Naarmate eindtermen sterker verschillen over onderwijsniveaus en tussen -stromen en onderwijsvormen, wordt het moeilijker om eenzelfde inhoudelijk domein meermaals te toetsen bij een leerling. Voor meer algemene vaardigheden met een duidelijk ontwikkelingsperspectief (zoals wiskundige problemen oplossen of leesvaardigheid) lijkt dat evidenter terwijl dat voor meer specifieke thema's (eindtermen) niet zondermeer zinvol is. De inhoudelijke breuk is met name groot voor specifieke thema's tussen het lager en het secundair onderwijs, evenals tussen de eerste en derde graad secundair onderwijs.

Waar de te toetsen domeinen over leerjaren heen wel inhoudelijk voldoende aansluiten, is het voor een leerwinstmeting tevens noodzakelijk dat de toetsen op een *gemeenschappelijke meetschaal* kunnen worden geplaatst. Daarvoor is het noodzakelijk dat gebruik wordt gemaakt van een geschikte set van ankeritems. Kalibratiestudies zijn in deze essentieel. Kalibratiestudies zijn noodzakelijk om vast te stellen of een gemeenschappelijke meetschaal mogelijk en valide is over verschillende varianten van de toetsen, om de beste set van items en ankeritems te kunnen selecteren voor elk van de varianten. Net als bij bijvoorbeeld de peilingen kan het gaan om toetsvarianten voor afname bij verschillende leerlinggroepen binnen één afnamemoment (bv. A-stroom en B-stroom, of meer onderwijsfinaliteiten meer algemeen) of om afnamemomenten over verschillende schooljaren heen (cf. trends). Naast deze zogenaamde horizontale equivalering vraagt leerwinstmeting evenwel ook om een verticale equivalering: het koppelen van de meetschaal over leerjaren heen. Om de toetsen van twee verschillende leerjaren op eenzelfde meetschaal te kunnen zetten, zullen de kalibratiestudies tevens daarvoor de juiste set van items en ankeritems dienen te identificeren. In het vooropgesteld tijdspad van de centrale toetsen is evenwel geen gelijktijdige kalibratie van de meetschaal over de leerjaren heen voorzien voorafgaand aan de eerste

afnames. Dit heeft als risico dat enkele jaren later zou kunnen blijken dat leerwinstmeting toch niet mogelijk is door een gebrek aan geschikte ankeritems. Bovendien kan worden verwacht dat door de langere duur tussen de toetsen van de eerste en derde graad secundair onderwijs de vaardigheidsniveaus verder uit elkaar liggen, waardoor het moeilijker wordt om de toetsen op eenzelfde meetschaal te plaatsen. Bijkomend onderzoek moet uitwijzen of, en hoe dit kan gerealiseerd worden. Mogelijk is een afname van uitgebreidere toetsen bij kalibratie of van bijkomende toetsen bij een doelgroep uit een tussenliggend leerjaar noodzakelijk om de resultaten van dergelijke toetsen op een gemeenschappelijke meetschaal te kunnen zetten.

Representativiteit en grootte steekproeven

Waar het in de eerste twee delen gaat over de wijze van verwerking en analyse van de leerwinstgegevens, wordt in dit derde deel concreter ingegaan op verwachtingen bij de centrale toetsen aangaande het aantal leerlingen, het aantal scholen, en representativiteit van getoetste scholen en leerlingen. Deze elementen bepalen immers de uitspraken die kunnen worden gedaan op verschillende niveaus (leerling, school, systeem). Daarbij heeft met name voor leerwinst (en toegevoegde waarde) de keuze van het longitudinale afnamedesign belangrijke implicaties.

In het ideale scenario wordt een bepaald domein getoetst bij alle leerlingen over alle scholen heen en, mits leerwinstmeting zinvol is (cf. hierboven), ook over de volledige schoolloopbaan van een leerling. Om de totale toetsduur voor leerlingen te beperken, wordt voor bepaalde specifieke toetsen uitgegaan van een onvolledig afnamedesign (aanvullend op toetsen met een volledige afname) om een toetsdomein volledig in kaart te brengen. Daarbij zullen scholen en leerlingen verschillen in een aantal specifieke toetsen die bij een bepaalde afname worden voorgelegd. Voor eenvoudig te interpreteren resultaten op systeemniveau (bv. trends in jaarlijkse leerprestaties) is het daarbij van belang dat, ingeval van een onvolledig afnamedesign, de verdeling van de toetsen over de scholen en leerlingen zodanig gebeurt dat elke specifieke toets wordt voorgelegd aan een representatieve groep leerlingen en scholen.

Een onvolledig afnamedesign voor een specifieke toets reduceert het aantal leerlingen en scholen voor die toets. Uitgaand van een redelijk aantal te verdelen toetsen *over* scholen, blijven deze aantallen evenwel groot genoeg voor betrouwbare analyse van school- en leerlingverschillen. Een onvolledige afname *binnen* scholen zal evenwel voor het merendeel van de klassen resulteren in een te beperkt aantal leerlingen voor een betrouwbare analyse op klasniveau. Dit dient zorgvuldig te worden afgewogen tegen het voordeel van de bredere feedback voor scholen en de snellere rotatie die zo wordt bekomen voor de

specifieke toetsen op schoolniveau. Analoog is het tevens voor een kleine school aangewezen om alle leerlingen op dezelfde domeinen te toetsen.

Tabel 5: Schematische weergave van implicaties van verschillende longitudinale afnamedesigns

	volledige afname	onvolledige afname: 3 scenario's voor longitudinaal gekoppelde afname		
		geen longitudinaal gekoppelde afname	gekoppelde afname	
			op niveau van cohorte binnen de school	op niveau van individuele leerling
Leerwinst op systeemniveau				
Leerwinst op schoolniveau (feedback)				
Status op schoolniveau (feedback)				
Status op systeemniveau				

Legende:

	voldoende aantallen en representatieve groep
	voldoende aantallen, maar selectieve groep
	onvoldoende aantallen

Wanneer het over leerwinstmeting gaat tot slot, bieden onvolledige afnamedesigns bijzondere uitdagingen. Zo is de eerste vraag voor hoeveel leerlingen met een begintoets voor een specifiek toetsdomein ook een eindtoets voor hetzelfde toetsdomein wordt afgenomen. Bij een volledig afnamedesign op beide meetmomenten, zijn beide toetsen beschikbaar voor de volledige leerlingpopulatie, uitgezonderd non-respons, leerlingen die het Vlaams onderwijsssysteem in- of

uitstromen, en sommige leerlingen met een versnelde studieloopbaan. Wanneer de afname van een specifieke toets in een later leerjaar, net zoals bij het eerdere leerjaar, op toevallige wijze wordt verdeeld over scholen (= *scenario ongekoppelde afname*), neemt de kans dat een leerling twee keer wordt getoetst op hetzelfde domein sterk af. Leerwinst in een specifiek domein kan dan nog wel op systeemniveau worden geanalyseerd, maar voor een zinvolle analyse of rapportering op schoolniveau zullen aantallen snel te klein worden.

Een longitudinaal gekoppeld afnamedesign kan die aantallen op het schoolniveau vergroten. Wanneer de afname van een specifieke toets in het volgende leerjaar op schoolniveau wordt afgestemd op de specifieke toets die werd aangeboden in het eerdere leerjaar binnen de betrokken school (= *scenario cohortegekoppelde afname*), wordt rapportering van leerwinst voor de specifieke toetsen op schoolniveau wel mogelijk. Dit tweede scenario heeft evenwel als nadeel dat leerwinst niet langer gemeten wordt voor een representatieve steekproef. Leerlingen met een niet-standaard schoolloopbaan (i.c. niet-normaal vorderende en schoolmobiele leerlingen) zullen sterk ondervertegenwoordigd zijn, waardoor leerwinst wordt overschat. Wellicht zijn ook de meest kwetsbare groepen leerlingen oververtegenwoordigd bij de groep leerlingen met een niet-standaard schoolloopbaan die in dit scenario grotendeels buiten de radar zou vallen. Voor heel wat leerlingen zal in dit scenario geen leerwinst vastgesteld kunnen worden. Dit scenario van cohortegekoppelde afname heeft dan ook weinig meerwaarde wanneer leerlingen sterk uitwaaieren over scholen tussen de twee betrokken meetmomenten, in het bijzonder voor de overgang tussen het lager en het secundair onderwijs.

Een scenario waarbij een specifieke toets in het volgende leerjaar wordt toegewezen op basis van de specifieke toets die elke individuele leerling aflegde in het eerdere leerjaar (= *scenario leerlinggekoppelde afname*) combineert de voordelen van voldoende aantallen op schoolniveau en van het behoud van een representatieve steekproef op systeemniveau van de twee andere scenario's. Het voordeel van de representatieve groep die wordt bekomen voor leerwinst met dit longitudinaal afnamedesign, gaat evenwel ten koste van de feedbackmogelijkheden op schoolniveau voor de eindstatus (beperkt tot de specifieke toetsen afgenomen in de school in het eerdere leerjaar voor in hoofdzaak leerlingen met een standaardloopbaan).

Ongeacht het type longitudinale koppeling bij onvolledige afname, zal de feedback over leerwinst op schoolniveau grotendeels beperkt blijven tot de leerlingen met een standaard schoolloopbaan. Immers, leerlingen die van school veranderden tussen beide meetmomenten, evenals leerlingen die sneller of minder snel vorderden dan de leerlingen van hun cohorte, deden ofwel niet de begintoets die in de betrokken school werd aangeboden in het eerdere leerjaar (scenario cohortekoppeling), ofwel niet de eindtoets die aansluit op de begintoets van de school (scenario individuele koppeling). Daarom kunnen

zelfs bij longitudinaal gekoppelde afname van toetsen in een onvolledig afnamedesign de aantallen leerlingen nog steeds onvoldoende groot zijn voor schoolfeedback. Dit is afhankelijk van het aantal over leerlingen te verdelen verschillende specifieke toetsen bij een gedifferentieerd afnamedesign en hoeveel leerlingen uitwaaieren over scholen tussen het eerste en het tweede gemeten moment (concreter: de kans voor een leerling om op de twee meetmomenten de specifieke toets af te leggen die centraal staat in de school bij tweede meting). Een tweede gevolg op het vlak van feedback op schoolniveau van leerwinst doordat deze beperkt is tot de leerlingen met een standaard schoolloopbaan bij toetsen met onvolledige afname, is dat de gemeten leerwinst (geaggregeerd) niet zonder kwalificatie als indicator voor onderwijskwaliteit kan worden geïnterpreteerd. Het lijkt aannemelijk dat net leerlingen met niet-standaard schoolloopbanen extra onderwijsinspanningen vergen; omgekeerd kan hun voorkomen ook duiden op de beperkingen van het onderwijs.

Telkens wanneer sprake is van een selectieve groep waarbij leerwinst wordt gemeten (op systeemniveau) of waarover wordt gerapporteerd (op schoolniveau), kan verwacht worden dat de gemiddelde leerwinst wordt overschat. Dit bemoeilijkt de vergelijking van scholen die verschillen in de mate waarin ze leerlingen hebben met niet-standaardloopbanen, evenals de inschatting van trends op systeemniveau wanneer over de tijd veranderingen zouden optreden in bijvoorbeeld het voorkomen van die afwijkende schoolloopbanen. Dit zou ook ongewenste effecten in de hand kunnen werken. *Om al die redenen lijkt het aangewezen om leerlingen met atypische loopbanen maximaal mee te nemen bij verwerking, analyse en rapportering*. Voor de systeemanalyses impliceert dat bijkomende correctie ingeval van het cohortegekoppelde onvolledige afnamedesign. Bij de schatting van de toegevoegde waarde van scholen kan door gebruik te maken van complexere modellen (multi-group multilevel modellen) de leerwinst van mobiele leerlingen worden verrekend (scenario volledige afname). Voor leerlingen zonder volledige leerwinstmeting (scenario onvolledige afname), kunnen de (analysetechnische) mogelijkheden tot correctie voor die potentiële bias worden onderzocht. Dit is met name van belang ingeval van schoolfeedback bij de longitudinaal gekoppelde afnames, omdat de groep leerlingen waarvoor de leerwinstmeting beschikbaar is dan selectief is. We suggeren voor beide types van longitudinaal gekoppelde onvolledige afname een mogelijke, maar nog verder te onderzoeken, piste. De toewijzing of correctie voor dergelijke leerlingen met een bijzonder traject maakt de leerwinst- of toegevoegdewaardebepaling in ieder geval analytisch complexer, en maakt mogelijk ook snelle feedback moeilijk. Het is daarom ook niet zeker is of dergelijke gecorrigeerde feedback haalbaar is om van bij de start van de centrale toetsen toe te passen.

4 BELEIDSAANBEVELINGEN

Doorheen de nota worden veel aanbevelingen geformuleerd. Een aantal zijn expliciet geformuleerd in de tekst. Andere vloeien meer impliciet uit de aangehaalde argumenten. Sommige volgen vrijwel onmiddellijk uit de internationale literatuur, terwijl andere voortkomen uit een confrontatie van die literatuur met de doelstellingen en context van de centrale toetsen in Vlaanderen. Hieronder zetten we de belangrijkste aanbevelingen nogmaals expliciet op een rij, voor wat betreft communicatie, de uitgangspunten, het toetsontwerp, het afnamedesign en het omgaan met niet-standaard schoolloopbanen.

Transparante communicatie over leerlingprestaties en schoolverschillen

Schoolverschillen en geschatte schooleffecten dienen helder toegelicht en begrijpelijk gemaakt te worden. Gekoppeld aan die doelstelling formuleert het steunpunt volgende aanbevelingen:

Rapportering van betrouwbaarheidsintervallen

Omdat resultaten niet allemaal even accuraat of veelzeggend zijn, is het aanbevolen om de statistische onzekerheid expliciet op te nemen bij de rapportering van resultaten, in het bijzonder bij schoolfeedback aan de hand van betrouwbaarheidsintervallen. Dit is van belang voor de resultaten op leerlingniveau, met name om mee te geven dat niet altijd duidelijk is of een vaardigheidsscore boven of onder een bepaalde cesuur ligt. Dit is eveneens relevant voor de interpretatie van resultaten op hogere niveaus als lesgroep of schoolniveau en dan met name bij het maken van vergelijkingen tussen groepen of over de tijd.

Duiding individuele leerlingresultaten

De toetsresultaten op leerlingniveau mogen niet als op zichzelf staande diagnostische instrumenten worden gehanteerd. Ze worden immers met onzekerheid gemeten. Dit moet ook worden teruggekoppeld naar de scholen of onderwijsverstrekkers die de toetsen mee als evaluaties willen meenemen. Ze dienen zich bewust te zijn dat de resultaten steeds tezamen met de andere prestaties van leerlingen dienen te worden bekeken in hun samenhang door de klassenraad in elke evaluatie van een individuele leerling.

Eenduidig begrippenkader rond geschatte schooleffecten

Het begrip van toegevoegde waarde is essentieel wanneer het gaat om schoolverschillen. Omdat de reikwijdte van een schatting van een schooleffect helder dient te zijn, bevelen we aan om de term toegevoegde waarde enkel te hanteren wanneer de gemeten schoolresultaten gecorrigeerd worden op

basis van (onder meer) een beginmeting. Wanneer een beginmeting niet beschikbaar is, gaat het om gecontextualiseerde schoolresultaten maar niet om toegevoegde waarde.

Doelgerichte rapportering van schooleffecten

Geschatte schooleffecten kunnen op verschillende wijze worden gerapporteerd. Het lijkt aanbevolen dit aan te passen aan het doelgebruik of -publiek. Voor schoolfeedback biedt het te verwachten resultaat voor die school gegeven haar leerling- en schoolkenmerken het meest direct inzicht om de resultaten van de school mee te vergelijken (te interpreteren als het resultaat van vergelijkbare scholen).

Op systeemniveau daarentegen, of wanneer men scholen wil vergelijken, is eerder het resultaat dat de school behaalt voor een bepaalde referentiegroep (qua leerling- en schoolkenmerken) relevanter.

Bij toegevoegde waarde op basis van een leerwinstmeting, kunnen deze schattingen van het schooleffect tevens uitgedrukt worden in termen van de hoeveelheid geboekte leerwinst.

Modelmatige uitgangspunten

De keuze van het model is van belang. Uit de literatuur volgen een aantal aanbevelingen over de kenmerken die mee in rekening worden gebracht bij de schatting van toegevoegde waarde.

Leerwinst als basis voor toegevoegde waarde

Voor een goede inschatting van de toegevoegde waarde van een school dient vertrokken te worden van een leerwinstmeting. Hierdoor kan het aanvangsniveau van leerlingen mee in rekening worden gebracht. Op basis van een inschatting van het schooleffect die gebaseerd is op een enkele of eerste statusmeting, kunnen daarentegen geen sterke conclusies worden geformuleerd over de bijdrage van de school. Er kan dan immers onvoldoende uitgemaakt worden welke factoren bijdragen tot de toetsresultaten van individuele leerlingen binnen scholen.

Kenmerken die in rekening worden gebracht bij de schatting van de toegevoegde waarde

Vanuit een schoolontwikkelingsperspectief is het meest geschikte model er een dat rekening houdt met leerlingkenmerken, evenals met compositie op schoolniveau en andere schoolkenmerken.

Modellering van schooleffecten

Aan de basis van de bepaling van een schooleffect, ligt een modellering van het verwachte schooleffect. Hoe goed die modellering is, bepaalt mee hoe zinvol de geschatte schooleffecten zijn. In het bijzonder dienen veronderstellingen met betrekking tot het gebruikte schattingsmodel (zoals bv. het al dan niet

lineair verband tussen twee opeenvolgende prestatiemetingen) daarom grondig geëvalueerd te worden en indien nodig dient de modellering bijgesteld of verfijnd te worden.

Uitdagingen voor leerwinstmetingen: toetsontwerp

Een leerwinstmeting zal niet mogelijk zijn voor alle toetsdomeinen over de hele schoolloopbaan.

Welke toetsdomeinen?

Een eerste voorwaarde is dat een te toetsen domein over meerdere schooljaren heen, inhoudelijk voldoende stabiel moet zijn. Dit zal alleen het geval zijn voor de meer algemene en ontwikkelingsgerichte toetsdomeinen. Het is van belang dat deze toetsdomeinen waarvoor een leerwinstmeting zinvol kan zijn, worden geïdentificeerd.

Gemeenschappelijke meetschaal?

Vervolgens dient ook de tweede voorwaarde vervuld te worden dat de metingen in verschillende leerjaren op een gemeenschappelijke schaal kunnen geplaatst worden. Aanbevelingen op dat vlak zijn de opname van geschikte ankeritems in de toetsen, wat in principe een gelijktijdige kalibratie vooraf van de meetschalen over de leerjaren heen vergt. Een tweede aanbeveling is een uitgebreid kalibratieonderzoek voor de overbrugging van de eerste naar de derde graad secundair onderwijs. Dit verhoogt immers de kans op het kunnen uitvoeren van adequate leerwinstmetingen op termijn.

Betrouwbare meetschaal

Met het oog op individuele leerwinstmeting, zijn betrouwbare metingen onontbeerlijk. Niet alleen dient de meetschaal op de twee meetmomenten voldoende betrouwbaar te zijn, het is ook sterk aan te bevelen dat de meetnauwkeurigheid niet afhangt van de positie op de meetschaal. Daarvoor lijkt adaptief toetsen een belangrijke voorwaarde.

Uitdagingen voor leerwinstmetingen: welk longitudinaal afnamedesign?

Het longitudinaal afnamedesign heeft belangrijke implicaties voor welke leerlingen een leerwinstmeting beschikbaar zal zijn. Bij een volledig afnamedesign stellen zich geen problemen op het vlak van steekproefgrootte, representativiteit of schoolfeedback. Bij een onvolledig afnamedesign ligt dat anders.

Steekproefgrootte en representativiteit

Bij een onvolledig afnamedesign biedt geen enkel scenario de mogelijkheid om op alle niveaus (leerling, school, systeem) én voldoende aantallen én een representatieve steekproef voor leerwinst én eindstatus te bekomen. De voor- en nadelen voor elk niveau, evenals het risico op curriculumvernauwing bij longitudinaal gekoppelde toetsafname, dienen bij de keuze van een longitudinaal afnamedesign zorgvuldig overwogen te worden.

Schoolfeedback over leerwinst

Om bij een onvolledig afnamedesign, schoolfeedback van leerwinst mogelijk te maken, dient de afname longitudinaal gekoppeld te worden. Het scenario om op schoolniveau een cohorte leerlingen meermaals op hetzelfde domein te toetsen in verschillende leerjaren lijkt dan het meest aangewezen (het alternatief van individuele koppeling lijkt praktisch minder haalbaar en daarbij komt schoolfeedback van de eindtoets in het gedrang).

Omgang met niet-standaard schoolloopbanen

Bij een onvolledig afnamedesign is duidelijk dat schoolfeedback over de leerwinst van leerlingen op schoolniveau of zelfs de meting van leerwinst op systeemniveau beperkt dreigt te worden tot enkel de groep leerlingen met een standaard schoolloopbaan (te weten: normaalvorderend en geen schoolse mobiliteit). Deze situatie komt in hogere mate voor in (groot-)stedelijke contexten en bij leerlingen die lager presteren. Dit heeft gevolgen voor de geschatte leerwinst van scholen (overschatting indien geen rekening wordt gehouden met de selectiviteit van de betrokken leerwinstmeting). Bovendien kan die selectiviteit ongewenste effecten inhouden voor oriëntering van leerlingen. Om die redenen bepleiten we om verder te onderzoeken hoe in voorkomend geval kan worden gecorrigeerd voor (de ondervertegenwoordiging van) niet-standaard schoolloopbanen.

Meer algemeen, ook buiten de context van een onvolledig afnamedesign, verdient het aanbeveling om aan de hand van complexere modellen voor toegevoegdewaardeschatting de resultaten van leerlingen met niet-standaard schoolloopbanen zo goed als mogelijk mee te nemen (cf. in- en uitstroom uit Vlaams onderwijs, schoolmobiliteit via multigroep multilevel modellen). Dit is met name noodzakelijk indien men een goed representatief beeld wil bekomen, wanneer men scholen wil vergelijken en om trends over de tijd in te schatten.

Schema: keuzes en implicaties / stappenplan, met het oog op leerwinstmeting

1. Inhoudelijk zinvol?
- ja => technisch leerwinstmeting mogelijk maken: ankeritems, simultane kalibratiestudie

2. Afname: volledig?
- ja => okee, leerwinstmeting inclusief niet-standaardloopbanen (zowiezo wel correcties nodig voor in/uitstroom uit Vlaams onderwijssysteem)
- neen

3. Longitudinale koppeling?
- ja => schoolfeedback voor leerwinst mogelijk + systeemanalyse (bij cohortekoppeling bijkomend correcties nodig voor ondervertegenwoordiging van niet-standaardloopbanen)
- neen => enkel systeemanalyse voor leerwinst

5 BIBLIOGRAFIE

Aesaert, K., & Janssen, R. (2021). *Externe kwaliteitszorg in het Vlaamse onderwijs*. Gepresenteerd op het Leuvense Onderwijs Onderzoek Symposia (LOOS), Leuven, 15 september 2021.

Allison, P. D. (1990). Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology*, *20*, 93–114.

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with Confidence*. British Medical Journal Books.

Amrein-Beardsley, A., & Geiger, T. J. (2019). Potential sources of invalidity when using teacher value-added and principal observational estimates: Artificial inflation, deflation, and conflation. *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-019-09311-w

Ayala, R. J. de (2009). *The Theory and Practice of Item Response Theory*. Guilford Press.

Ballou, D., W. Sanders and P. Wright. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics,* 29.

Betebenner, D. W., & Linn, R. L. (2009). *Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability* [Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009]. Center for K–12 Assessment & Performance Management.

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Castellano, K. E., & Ho, A. D. (2013). *A Practitioner's Guide to Growth Models*. Council of Chief State School Officers.

Castellano, K. E., & McCaffrey, D. F. (2020). Comparing the Accuracy of Student Growth Measures. *Journal of Educational Measurement*, *57*(1), 71–91. https://doi.org/10.1111/jedm.12242

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons, *Journal of Educational and Behavioral Statistics*, *39*(5), 333-367. https://doi.org/10.3102/1076998614547576

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*, *35*(1), 32–58. https://doi.org/10.1177/0265407517718387

Cermak, S. (1989). Norms and scores. In L. J. Miller (Ed.), *Developing Norm-Referenced Standardized Tests* (1st ed., pp. 91–123). Routledge. https://doi.org/10.4324/9781315859811

Chung, H., & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology, 65*(2), 185-200.

Dalecki, M., & Willits, F. K. (1991). Examining change using regression analysis: Three approaches compared. *Sociological Spectrum*, *11*(2), 127–145. https://doi.org/10.1080/02732173.1991.9981960

Dockx, J., De Fraine, B., & Stevens, E. (2016). De rol van de eerdere schoolloopbaan bij de overgang naar het secundair onderwijs. *Pedagogische Studiën, 93*(4), 223-240.

Dumay, X., Coe, R., & Anumendem, D. N. (2014). Stability over time of different methods of estimating school performance. *School Effectiveness and School Improvement, 25*(1), 64-82.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter? *Educational Policy*, *30*(3), 465–500. https://doi.org/10.1177/0895904814557593

Ellemers, J. E. (1976). Veel kunnen verklaren of iets kunnen veranderen: Krachtige versus manipuleerbare variabelen. *Beleid & Maatschappij*, *11*, 281–290.

Everson, K. C. (2017). Value-Added Modelling and Educational Accountability: Are We Answering the Real Questions? *Review of Educational Research, 87*(1), 35-70.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*, 369-395.

Gong, B., Perie, M., & Dunn, J. (2006). *Using student longitudinal growth measures for school accountability under No Child Left Behind: An update to inform design decisions*. http://www.nciea.org/publications/GrowthModelUpdate_BGMAPJD07.pdf

Gossman, P., & Powell, S. (2019). Learning Gain: Can It Be Measured? In A. Diver (Ed.), *Employability via Higher Education: Sustainability as Scholarship* (pp. 37–51). Springer International Publishing. https://doi.org/10.1007/978-3-030-26342-3_3

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, *10*(1), 117–156. https://doi.org/10.1162/EDFP_a_00153

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newberry Park, CA: Sage.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Harvard Education Press.

Harris, D. N., & Anderson, A. (2013). *Does Value-Added Work Better in Elementary Than in Secondary Grades?* Knowledge Brief. Carnegie Knowledge Network.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*(1), 73–112.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*, 33–41.

Heck, R. H. (2006). Assessing School Achievement Progress: Comparing Alternative Approaches. *Educational Administration Quarterly*, *42*(5), 667–699. https://doi.org/10.1177/0013161X06293718

Hollingsworth, H., Heard, J., & Weldon, P. R. (2019). *Communicating Student Learning Progress: A Review of Student Reporting in Australia.* Australian Council for Educational Research (ACER).

Hopster-den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2019). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. Assessment in Education: Principles, Policy & Practice, 26(2), 123–142. https://doi.org/10.1080/0969594X.2018.1447908

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical to measurement invariance in aging research. *Experimental Aging Research*, 18, 117-144.

Janssens, F., Lyset Rekers-Mombarg, & Lacor, E. (2014). *Leerwinst en Toegevoegde Waarde in het Primair Onderwijs*. CED-Groep. https://doi.org/10.13140/2.1.2575.4563

Jakubowski, M. (2008). Implementing value-added models of school assessment. *EUI Working Papers RSCAS 2008/06*, European University Institute.

Johnson, R., Penny, J., & Gordon, B. (2009). *Assessing performance: Designing, Scoring, and Validating Performance Tasks.* The Quildford Press.

Kane, M. T. (2017). *Measurement error and bias in value-added models* (Research Report No. RR-17-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12153

Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-Scale Assessments in Education*, *2*(1), 1. https://doi.org/10.1186/2196-0739-2-1

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences, 2nd Edition.* American Psychological Association.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Köhler, C., Hartig, J., & Schmid, C. (2021). Deciding between the Covariance Analytical Approach and the Change-Score Approach in Two Wave Panel Data. *Multivariate Behavioral Research*, *56*(3), 447–458. https://doi.org/10.1080/00273171.2020.1726723

Kolen, M. J. & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking. Methods and Practices*, second edition. New York: Springer.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(3), 537–554.

Leckie, G. (2018). Avoiding Bias When Estimating the Consistency and Stability of Value-Added School Effects. *Journal of Educational and Behavioral Statistics*, *43*(4), 440–468.

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal*, *43*(2), 193–212. https://doi.org/10.1002/berj.3264

Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, *45*(3), 518–537. https://doi.org/10.1002/berj.3511

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(4), 835–851.

Leckie, G., & Prior, L. (2022). A comparison of value-added models for school accountability. *School Effectiveness and School Improvement.* https://doi.org/10.1080/09243453.2022.2032763

Lenkeit, J. (2013). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*, *24*(1), 1–25. https://doi.org/10.1080/09243453.2012.680892

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2019). Methodological issues in value-added modeling: An international review from 26 countries. *Educational Assessment, Evaluation and Accountability*, *31*(3), 257–287. https://doi.org/10.1007/s11092-019-09303-w

Lissitz, R. W., & Huynh, H. (2002). Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability. *Practical Assessment, Research, and Evaluation, 8*, Article 10. https://doi.org/10.7275/npzw-wd59

Lockwood, J. R., & Castellano, K. E. (2015). Alternative Statistical Frameworks for Student Growth Percentile Estimation. *Statistics and Public Policy, 2*(1), 1-9. https://doi.org/10.1080/2330443X.2014.962718

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22–52. https://doi.org/10.3102/1076998613509405

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Marks, G. N. (2021). Should value-added school effects models include student- and school-level covariates? Evidence from Australian population assessment data. *British Educational Research Journal*, *47*(1), 181–204. https://doi.org/10.1002/berj.3684

Martineau, G. J. (2016). *A Guide to Understanding and Selecting Measures of Growth for Smarter Balanced Members*. Smarter Balanced Assessment Consortium.

McCaffrey, D. F., J. R. Lockwood, D. M. Koretz and L. S. Hamilton. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: The RAND Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606.

McGrath, C. H., Guerin, B., Harte, E., Frearson, M., & Manville, C. (2015). Learning gain in higher education. Cambridge, UK: RAND Corporation.

Meghir, C., & Rivkin, S. G. (2010). Econometric Methods for Research in Education. NBER Working Paper No. 16003. In *National Bureau of Economic Research*. National Bureau of Economic Research.Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's Schools: The Role of Incentives* (pp. 197–224). National Academies Press. https://doi.org/10.17226/5143

Michaelides,M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating*. Los Angeles: Center for the Study of Evaluation (CSE), CRESST, University of California, LA.

Newton, X., Darling-Hammond, L., Haertel, E., &Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 1–27. https://doi.org/10.14507/epaa.v18n23.2010

National Research Council (2014). *Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12.* The National Academies Press.

OECD (2008). *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools*. Paris: OECD Publishing. https://doi.org/10.1787/9789264050259-en

Penninckx, M., Vanhoof, J., Quintelier, A., De Mayer, S., & Van Petegem, P. (2017). *Zicht op leerwinst: Scenario's voor gestandaardiseerd toetsen*. Leuven: Acco.

Prior, L., Jerrim, J., Thomson, D., & Leckie, G. (2021). A review and evaluation of secondary school accountability in England: Statistical strengths, weaknesses and challenges for 'Progress 8' raised by COVID-19. *Review of Education*, *9*(3). https://doi.org/10.1002/rev3.3299

Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335.

Raudenbush, S., W. (2013). *What do we know about using value-added to compare teachers who work in different schools?* [Knowledge Brief]. Carnegie Knowledge Network.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of Value-Added Models for Estimating School Effects. *Education Finance and Policy*, *4*(4), 492–519. https://doi.org/10.1162/edfp.2009.4.4.492

Rodgers, T. (2007). Measuring Value added in Higher Education: A Proposed Method for Developing a Performance Indicator Based on the Economic Value added to Graduates. *Education Economics, 15*(1), 55–74.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: selection on observables and unobservables. *Education Finance and Policy, 3,* 537–571. https://doi.org/10.1162/edfp.2009.4. 4.537

Rumberger, R. W. (2003). The causes and consequences of student mobility. *The Journal of Negro Education, 27*(1), 6-21.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Scheerens, J. (1990). School Effectiveness Research and the Development of Process Indicators of School Functioning. *School Effectiveness and School Improvement*, *1*(1), 61–80. https://doi.org/10.1080/0924345900010106

Schochet, P. Z., & H.S. Chiang (2010). Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. Washington: National Center for Education Evaluation and Regional Assistance.

Schochet, P. Z., & Chiang, H. S. (2013). What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142–171. https://doi.org/10.3102/1076998611432174

Steunpunt Centrale Toetsen in Onderwijs (2021). *Centrale toetsen wiskunde. Eindtermenselectie en toetsdesign voor de centrale toetsen wiskunde*. 23.12.2021.

Strand, S. (2002). Pupil mobility, attainment and progress during key stage 1: A study in cautious Interpretation. *British Education Research Journal, 28*(1), 63-78.

Strand, S. (2016). Do some schools narrow the gap? Differential school effectiveness revisited. *Review of Education*, *4*(2), 107–144. https://doi.org/10.1002/rev3.3054

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–36. https://doi.org/10.3102/10769986029001011

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, *22*(4), 393-413. https://doi.org/10.1080/09243453.2011.590704

Timmermans, A. C. , Snijders, T. A. B., & Bosker, R. J. (2012). In search of value added in the case of complex school effects. *Educational and Psychological Measurement, 73*(2), 210–228.

Timmermans, A. C., Snijders, T. A. B., & Bosker, R. J. (2013). In Search of Value Added in the Case of Complex School Effects. *Educational and Psychological Measurement*, *73*(2), 210–228. https://doi.org/10.1177/0013164412460392

Timmermans, A. C., & Thomas, S. M. (2015). The Impact of Student Composition on Schools' Value-Added Performance: A Comparison of Seven Empirical Studies. *School Effectiveness and School Improvement*, *26*(3), 487–498. https://doi.org/10.1080/09243453.2014.957328

Van den Broeck, W. (2014). Vlaams onderwijs, let op uw zaak! In *Visie(s) op onderwijs* (pp. 145–198). Pelckmans.Van der Leeden, R., Busing, F., & Meijer, E. (1997). *Applications of bootstrap methods for two-level models*. Paper presented at the Multilevel Conference, Amsterdam.

Vlaamse Regering (2019). *2019-2024 Regeerakkoord*. Brussel: Vlaamse Overheid.

Vlaamse Regering (2020). *Oproep Steunpunt "Ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde net- en koepeloverschrijdende toetsen in Vlaanderen"*. Brussel: Vlaamse Overheid.

Wang, C., & Nydick, S. W. (2020). On Longitudinal Item Response Theory Models: A Didactic. *Journal of Educational and Behavioral Statistics*, *45*(3), 339–368. https://doi.org/10.3102/1076998619882026

Wright, D. B. (2018). Estimating school effectiveness with student growth percentile and gain score models. *Journal of Applied Statistics*, *45*(14), 2536–2547. https://doi.org/10.1080/02664763.2018.1426742

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice 29*(4), 15–27.

Zamarro, G., Engberg, J., Saavedra, J. E., & Steele, J. (2015). Disentangling disadvantage: Can we distinguish good teaching from classroom composition*? Journal of Research on Educational Effectiveness*, *8*(1), 84–111. https://doi.org/10.1080/19345747.2014.972601

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology*, *51*, 343-351.

Zoanetti, N. (2021). Interpreting learning progress using assessment scores: What is there to gain? *Research Conference 2021: Excellent Progress for Every Student: Proceedings and Program*, 77–88. https://doi.org/10.37517/978-1-74286-638-3_17