Meerjarenprogramma van het Steunpunt *"Ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde net- en koepeloverschrijdende toetsen in Vlaanderen"*

Inhoudsopgave

Overzicht tabellen

Overzicht figuren

Meerjarenprogramma

1. Inleiding

Onderwijs is bepalend voor de samenleving. Bij alle maatschappelijke actoren leeft de vraag dat scholen hoogstaand, kwaliteitsvol onderwijs aanbieden dat leerlingen zo optimaal mogelijk voorbereidt op het maatschappelijke leven en op de arbeidsmarkt. “Kwaliteit van onderwijs” is echter een ruim begrip en laat zich op verschillende manieren omschrijven. Door het actualiseren van de eindtermen geeft het onderwijssysteem blijk van haar responsiviteit door in te spelen op maatschappelijke vragen en veranderingen. Het adaptief vermogen van een onderwijssysteem is een essentieel onderdeel van onderwijskwaliteit, maar kan niet los gezien worden van de vraag of scholen er ook in slagen om systeembrede geformaliseerde einddoelen bij de leerlingen te bereiken. Een onderwijssysteem heeft bijgevolg behoefte aan meetbare opbrengstindicatoren om gefundeerde uitspraken te kunnen doen over leerprestaties en -vorderingen van leerlingen.

Anders dan in ons omringende landen heeft Vlaanderen geen traditie van centrale toetsing voor het ontwikkelen en verantwoorden van onderwijskwaliteit. In Vlaanderen, maar ook in verschillende andere onderwijssystemen woedt een aanhoudende discussie omtrent de meerwaarde en aard van implementatie van systemen voor het monitoren van leerprestaties van leerlingen. Voorstanders van gestandaardiseerd toetsen zien een meerwaarde in het objectiveren van leerresultaten, in het versterken van interne kwaliteitszorg en een meer gelijke lat voor alle leerlingen. Daarnaast is er de tendens naar het meer gericht aanspreken van onderwijsaanbieders (leraren, scholen en intermediairen) op de wijze waarop ze de autonomie die ze genieten succesvol weten aan te wenden en hoe deze autonomie zich vertaalt naar leerprestaties en leerwinst van leerlingen. Het aantal landen waar het gestandaardiseerd in kaart brengen van leerresultaten en leerwinst van leerlingen gemeengoed is geworden, is de voorbije decennia dan ook gegroeid. Ook in Vlaanderen kan men niet om de vaststelling heen dat steeds meer aandacht wordt geschonken aan de resultaten van de onderwijsprocessen (op micro-, meso- én macroniveau). Het is vanuit dit laatste oogpunt dat scholen sinds het schooljaar 2017-2018 gevalideerde toetsen dienen af te nemen bij alle leerlingen op het einde van de basisschool voor tenminste drie leerdomeinen. Typisch voor het Vlaamse onderwijs zijn daarnaast ook de verscheidene signalen uit wetenschappelijke monitoring die aangeven dat de kwaliteit van het Vlaamse onderwijs tanende is. Zo tonen de resultaten van recente peilingsonderzoeken en internationaal vergelijkende studies (o.a. PIRLS, PISA) aan dat de leerprestaties van de Vlaamse leerlingen van het lager en secundair onderwijs zich in een dalende lijn bevinden, en dit zowel voor Nederlands (lezen) als voor wiskunde (Denis, Janssen & Aesaert, 2019; De Meyer et al., 2020; Dockx, et al., 2019). Daar komt bij dat in alle ons omringende landen (Nederland, Duitsland, Frankrijk, Engeland en Schotland) uiteenlopende vormen van centrale, gestandaardiseerde toetsen worden ingezet om de leerprestaties en -vorderingen van leerlingen in beeld te brengen. Vergelijkingen op basis van PISA-resultaten laten zien dat in schoolsystemen van OESO-landen die op normen gebaseerde externe examens inzetten, leerlingen gemiddeld meer dan 16 punten hoger scoren dan leerlingen in schoolsystemen die dergelijke examens niet inzetten (Schleicher, 2018).

De bovenstaande argumenten hebben zich recent vertaald in concrete intenties in het Vlaams Regeerakkoord en de beleidsnota van minister Weyts om in het Vlaamse onderwijs gestandaardiseerde, gevalideerde, genormeerde en netoverschrijdende proeven in te voeren. Om de

kwaliteit van het Vlaamse onderwijs aan te sturen legt de overheid al geruime tijd eindtermen en ontwikkelingsdoelen vast. Om zicht te krijgen op de mate waarin leerlingen in het Vlaams onderwijs de eindtermen bereiken, worden er sinds 2002 peilingen georganiseerd. Er werd nu beslist te starten met de ontwikkeling van digitale gecentraliseerde proeven Nederlands en wiskunde, terwijl andere vakken/leergebieden in de toekomst overwogen kunnen worden. De proeven zullen worden afgenomen in het vierde en zesde leerjaar van het lager onderwijs en in het secundair onderwijs op het einde van de eerste graad en in het tweede jaar van de derde graad. Daarmee zijn de initiële krijtlijnen van het toekomstige toetsbeleid in Vlaanderen uitgetekend. De reeds uitgesproken intenties dienen verder doorgedacht, geoperationaliseerd en geïmplementeerd te worden. De haalbaarheidsstudies waartoe de Vlaamse overheid opdracht gegeven heeft, hebben de ambitie hierin de eerstvolgende stap te zetten. Dit projectvoorstel neemt de ambities van de overheid en de resultaten van deze haalbaarheidsstudies als een vertrekpunt voor de verdere ontwikkeling en implementatie van een systeem van gecentraliseerd toetsen in Vlaanderen. Hoewel de bakens uitgezet werden zijn er nog veel conceptuele, strategische en praktische onbekenden. Het consortium achter de huidige aanvraag wil academische expertise bundelen en ter beschikking stellen van de overheid om de verdere uitbouw van dit systeem van gecentraliseerd toetsen ten dienste van het Vlaamse onderwijsveld uit te denken en vorm te geven.

Met de verderop toegelichte inhouden en structuren van een nieuw steunpunt wil het indienende consortium de in de oproep opgenomen beleidsdoelstellingen realiseren. Hét centrale beleidsdoel is het genereren van kwaliteitsvolle en wetenschappelijk onderbouwde (beleids-)informatie op leerling-, school- en systeemniveau. Daartoe wordt vooreerst ingezet op de ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde taal- en wiskundetoetsen. Het steunpunt voorziet in een populatiebrede afname in de respectievelijke scholen en leerjaren. De verkregen toetsgegevens worden verwerkt, geanalyseerd en gerapporteerd. Er wordt in kaart gebracht in welke mate leerlingen in het Vlaams onderwijs de eindtermen beheersen. Van zodra mogelijk worden hier trendanalyses aan gekoppeld. Hetzelfde zal later gelden voor leerwinstmetingen. Verder wordt in de diepte beschreven welke verschillen er zijn tussen leerlingen en scholen en hoe deze kunnen verklaard worden. De rapportage blijft niet beperkt tot de klassieke schooloverstijgende onderzoeksrapporten. Het steunpunt zal feedbacksystemen uitdenken die de beschikbare toetsresultaten ontsluiten naar verschillende onderwijsactoren. De overheid zal een ICT-partner aanstellen die een toetsplatform en digitale feedbackmodules zal bouwen. Vanaf de afname van de gecentraliseerde toetsen geeft het steunpunt, gebruik makend van deze systemen, jaarlijks feedback over de toetsresultaten en dit zowel op individueel niveau, klasniveau als op schoolniveau, afhankelijk van de beoogde doelgroep. Ook omtrent de publieke bekendmaking zal het steunpunt een voorstel uitwerken.

Werken met gecentraliseerde toetsen is een complexe evenwichtsoefening, met name op de dimensie autonomie en sturing, en de dimensie ondersteuning en verantwoording. Het behoeft geen betoog dat de implementatie van centrale toetsen in een moderne onderwijscontext die tot hiertoe gekarakteriseerd werd door een niet gecentraliseerd, en sterk school-geïnitieerd ontwikkelingsperspectief, een duidelijk kantelpunt markeert. Het voorliggende projectvoorstel baadt in een geest van het koesteren en bewaken van de autonomie waar leraren en scholen in Vlaanderen over beschikken en wil het systeem van gecentraliseerd toetsen uitwerken vanuit het vertrekpunt dat ondersteuning van het onderwijsveld de primaire doelstelling is. Het draagvlak voor het werken met gestandaardiseerde toetsen is het voorbije decennium gestaag gegroeid in Vlaanderen, bij diverse stakeholders. Dit draagvlak is evenwel uiterst precair en - zoals eerder OBPWO-onderzoek aantoont (Vanhoof, De Maeyer, Van Petegem, Penninckx & Quintelier, 2016) - staat of valt het met de wijze waarop met de informatie uit het systeem van gecentraliseerd toetsen zal worden omgegaan. Het kandiderende consortium ondersteunt de intenties tot verplichte afname door scholen van de gecentraliseerde toetsen, maar is tegelijkertijd zeer bezorgd over de bestaande intenties om resultaten op schoolniveau identificeerbaar openbaar te maken. Wij duiden verderop in dit projectvoorstel de complexiteit aan kennis, vaardigheden en attitudes die doordacht omgaan met

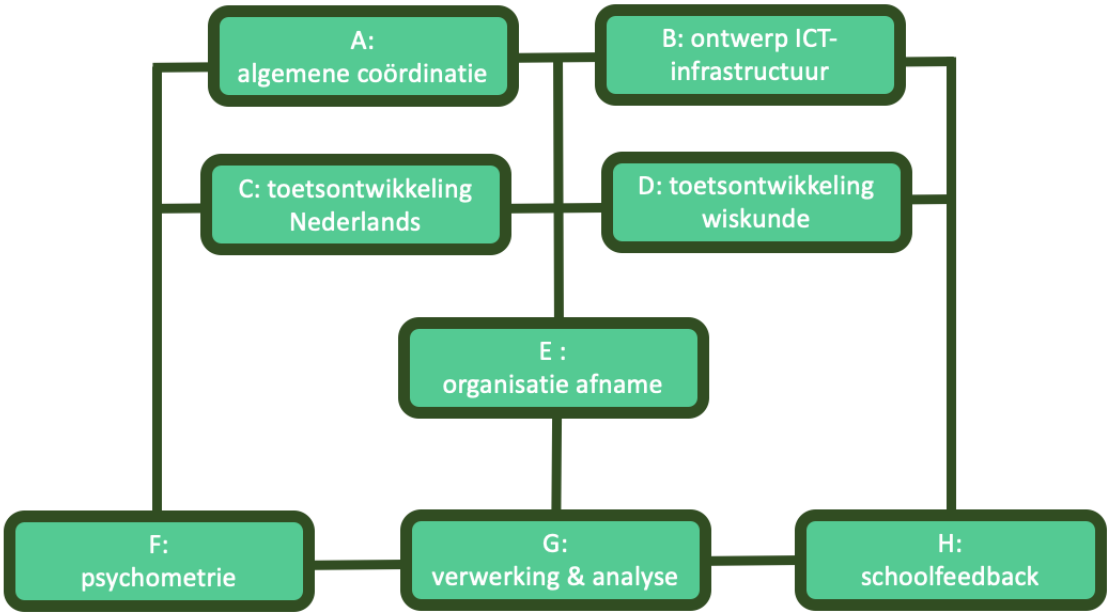
prestatiegegevens van leerlingen, klassen en scholen met zich meebrengen. De kennisname van bevindingen in buitenlandse onderwijssystemen en de vertrouwdheid met het Vlaamse onderwijsveld maken dat partners in dit projectvoorstel pleiten voor het strategische uitgangsprincipe dat het Vlaamse onderwijsveld minstens eerst de kans dient te krijgen om in de veilige context van absolute vertrouwelijkheid te groeien in de maturiteit die gebruik van prestatiegegevens ten gunste van het leren van leerlingen vergt. Finaal zit het kwaliteitsbevorderende mechanisme van gecentraliseerde toetsen immers in het leerwinstondersteunende onderwijsgedrag van leraren en leerondersteunend beleid. De vraag hoe het systeem van gecentraliseerd toetsen net dit kan ondersteunen zit in het DNA en de genese van dit projectvoorstel ingebakken. Het steunpunt is zich bewust van de risico's van gecentraliseerde toetsen wanneer deze 'stakes high' worden voor leerlingen, leerkrachten of scholen. We investeren in feedbacksystemen en ondersteuning in het werkveld om op een transparante en gebruiksvriendelijke wijze resultaten te ontsluiten, te interpreteren en te vertalen naar acties in de praktijk. We vinden het daarbij belangrijk dat er correcte en genuanceerde conclusies getrokken worden en dat de toetsen vanuit een ontwikkelingsperspectief gebruikt worden. De resultaten dienen als input om schoolbeleid te ondersteunen, leerkrachten te professionaliseren en leerlingen te versterken.

Het steunpunt is waakzaam tegenover beperkende en verantwoordingsgerichte perspectieven bij het gebruik van de toetsresultaten. Er zijn immers verschillende risico's verbonden aan gecentraliseerde toetsen, in het bijzonder wanneer de 'stakes high' worden (o.a. Nichols & Berliner, 2008). De vakken die getoetst worden, stijgen bijvoorbeeld in gepercipieerd belang en krijgen vaker en meer aandacht dan vakken die niet gecentraliseerd getoetst worden. Toetsvragen en toetsinhouden kunnen te richtinggevend worden voor de praktijk. En er bestaat evidentie voor selectieve deelname (absenteïsme) van leerlingen aan toetsafnames. Scholen en ouders kunnen selectiever worden in school- en studiekeuze, hetgeen scholen nog meer drijft in de richting van een 'quasi-marktwerking' en ongelijkheid in de hand werkt. Ook leerkrachten worden selectiever in de scholen waarvoor ze gaan werken, waardoor sterke scholen nog vaker sterke leerkrachten aantrekken en scholen in grootstedelijke contexten meer te maken krijgen met een hoog verloop van leerkrachten, waardoor jongeren die extra kansen nodig hebben nog meer kansen ontlopen. De consortiumpartners zijn dan ook vragende partij om met de Vlaamse overheid en het Vlaamse onderwijsveld tot grondige reflectie, debat én besluitname te komen omtrent welke positie best ingenomen wordt op de hoger vermelde dimensies, met het oog op het verbeteren van de onderwijskwaliteit.

De aanbieder van de centrale toetsen (c.q. de opdrachtgever) geeft aan het steunpunt het mandaat om instrumenten te ontwikkelen die geschikt zijn om de vooropgestelde doelstellingen te realiseren. De uiteindelijke validiteit van deze instrumenten ligt niet alleen in handen van de ontwikkelaars (c.q. het steunpunt), maar wordt ook in grote mate bepaald door de gebruikers zelf (c.q. onderwijsprofessionals die ermee aan de slag gaan, maar ook de leerlingen die getoetst worden, ouders en eventuele andere stakeholders die op basis van de resultaten een eigen besluitvormingsproces doorlopen). Waar toetsontwikkelaars relevante gegevens beschikbaar dienen te maken en de nodige randvoorwaarden dienen te creëren om een correcte interpretatie van deze gegevens mogelijk te maken, is het immers de gebruiker zelf die uiteindelijk een interpretatie alsook een eigen vertaalslag maakt (AERA et al., 2014). Aanbieder en ontwikkelaar hebben in dit verhaal de verantwoordelijkheid om zowel gewenste als onbedoelde effecten in kaart te brengen, teneinde erover te waken dat resultaten dezelfde betekenis meedragen en precies die consequenties hebben die initieel vooropgesteld werden (AERA et al., 2014). Een functioneel systeem van gecentraliseerd toetsen vergt dan ook de inzet van zeer diverse perspectieven en expertises. Met het onderscheiden van acht complementaire werkdomeinen, ingevuld door een brede vertegenwoordiging uit universiteiten en hogescholen anticipeert deze projectaanvraag op deze complexiteit.

2. Algemeen overzicht van werkdomeinen en planning

Figuur 3.1 geeft een schematisch overzicht van de acht werkdomeinen. Als overkoepelende domeinen staan bovenaan algemene coördinatie (werkdomein A) en ontwikkeling van het digitale platform voor toetsontwikkeling, toetsafname, dataopslag en schoolfeedback (werkdomein B). Daaronder volgen op gelijke hoogte de werkdomeinen voor toetsontwikkeling taal (werkdomein C) en wiskunde (werkdomein D). Deze toetsontwikkeling gaat de organisatie en ondersteuning van de afname vooraf (werkdomein E). Dit werkdomein neemt een centrale plaats in binnen het steunpunt, gegeven ook de complexe, logistieke uitdagingen die ermee gepaard gaan. Na de dataverzameling volgen in eerste instantie de psychometrische analyses (werkdomein F) voor de ontwikkeling van de meetschalen. Deze meetschalen vormen de afhankelijke variabelen voor een groot deel van de onderzoeksvragen die in werkdomein G beantwoord worden. Als sluitstuk volgt dan de schoolfeedback en het gebruik van de schoolfeedback (werkdomein H). Het raster in de figuur geeft aan dat alle werkdomeinen sterk met elkaar verbonden zijn. Het is immers cruciaal dat de chronologische opeenvolging van de werkdomeinen, zoals hierboven gesuggereerd, continu wordt aangevuld met een voortdurende wisselwerking tussen en afstemming van de verschillende werkdomeinen onderling.



Figuur 3.1 Overzicht van de werkdomeinen

Tabel 3.1 geeft een overzicht van de globale planning van het steunpunt. De geplande kalibraties (K) en afnames (A) van de gecentraliseerde toetsen vormen belangrijke mijlpalen die richtinggevend zijn voor alle werkdomeinen. Deze grote mijlpalen worden vanzelfsprekend verder geconcretiseerd en opgesplitst in deelmijlpalen die voor en na deze afnames binnen elk werkdomein en in onderlinge afstemming tussen de werkdomeinen gerealiseerd moeten worden.

Toetsafnames	2021			2022			2023			2024			2025		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
vierde leerjaar basisonderwijs											K			A	
eerste graad secundair onderwijs					K			A			A			A	
Werkdomein en -pakketten															
A Algemene coördinatie															
A1 Algemeen management HRM															
A2 Financieel en administratief beheer															
A3 Projectmanagement															
B Ontwerp ICT-infrastructuur															
B1 Ontwerp & kwaliteitscontrole digitaal platform															
B2 Datamanagement															
B3 Logistiek beheer (ICT-infrastructuur)															
B4 Ontwikkeling - Beheer website															
C Toetsontwikkeling Nederlands															
C1 Coördinatie toetsen Nederlands															
C2 Toetsontwikkeling Nederlands vierde leerjaar															
C3 Toetsontwikkeling Nederlands zesde leerjaar															
C4 Toetsontwikkeling Nederlands eerste graad															
C5 Toetsontwikkeling Nederlands derde graad															
D Toetsontwikkeling wiskunde															
D1 Coördinatie Wiskundetoetsen															
D2 Toetsontwikkeling Wiskunde vierde leerjaar															
D3 Toetsontwikkeling Wiskunde zesde leerjaar															
D4 Toetsontwikkeling Wiskunde eerste graad															
D5 Toetsontwikkeling Wiskunde derde graad															
E Organisatie en ondersteuning afname															
E1 Ontwikkelen draaiboek afname															
E2 Professionalisering toetsassistenten															
E3 Coördinatie afname, helpdesk															
E4 Coördinatie kalibratie, helpdesk															
F Psychometrie															
F1 IRT-analyses															
F2 Beperkt adaptief toetsen															
F3 Ondersteuning toetsontwikkeling															
F4 Automatisering analyses															
G Verwerking en analyse															
G1 Eindtermen/ Leerwinst															
G2 Verschillen tussen leerlingen en scholen															
G3 Verklaringen															
G4 Aanbod															
H Schoolfeedback															
H1 Inhoudelijk opzet / vormgeving systeem															
H2 Gebruikersonderzoek / effectmonitoring															
H3 Professionalisering feedbackgebruikers															
H4 Schoolfeedback paralleltoetsen BaO															

Tabel 3.1. Algemeen overzicht van de planning

3. Werkdomein A. Algemene coördinatie

De algemene coördinatie van het steunpunt vormt een eerste werkdomein. Terwijl de zeven andere werkdomeinen taakgericht zijn, is het werkdomein van algemene coördinatie in hoofdzaak procesgericht. Het werkdomein Algemene coördinatie omvat drie werkpakketten: A1 Algemeen management, A2 Financieel en Administratief Beheer, en A3 projectmanagement (Tabel 3.6).

	2021			2022			2023			2024			2025		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
A1 Algemeen management HRM															
A2 Financieel en administratief beheer															
A3 Projectmanagement															

Tabel 3.6. Werkpakketten Algemene coördinatie

In Deel 5 van deze aanvraag staat de vooropgestelde organisatie- en overlegstructuur uitvoerig beschreven. In dit luik wordt daarom enkel een beknopte toelichting gegeven bij de besteding van de personeels- en werkingsmiddelen die in functie staan van de algemene beleidsvoering en -organisatie van het nieuwe steunpunt.

Het eerste werkingsjaar staat binnen het werkdomein Algemene Coördinatie volledig in functie van het opzetten van een professioneel arbeids- en organisatiemodel, zowel op strategisch als operationeel niveau. Meer concrete informatie over de taakstelling van de strategisch en operationeel beheerder is beschreven in Deel 5 van deze aanvraag. Een derde aanstelling binnen het werkdomein Algemene Coördinatie is een voltijdse secretariaatsmedewerker voor de financiële en administratieve taken van het steunpunt (A3). Hiervoor wordt een voltijds secretariaatsmedewerker aangesteld op bachelorniveau. Ook hier verwijzen we naar Deel 5 voor een beschrijving van de concrete taken.

Centraal binnen het uittekenen van het organisatiemodel van het nieuwe steunpunt is dat in de startfase een gedetailleerde beschrijving van de werkpakketten en taken binnen elk werkdomein wordt opgesteld, alsook het opstellen van een gedetailleerd overzicht van operationele doelen en verwachte uitkomsten. Het uitwerken van een grondige SWOT-analyse onder leiding van de strategisch en operationeel beheerder zal het steunpunt een realistische inschatting geven van de sterke en zwakke punten van de plaats die het steunpunt in het onderwijslandschap zal innemen, samen met kansen en mogelijke bedreigingen. Bij deze analyse zullen zowel interne medewerkers als externe stakeholders betrokken worden. Vanuit de resultaten van deze analyse en het ontwikkelplan dat hieruit zal volgen, zal het steunpunt zich stevig kunnen verankeren in een steeds veranderende onderwijscontext.

Een van de eerste deeltaken van het consortium is het opstellen van een grondige risicoanalyse. Door de snel veranderende (beleids)context is een dynamisch risicobeheersingsmodel aangewezen. Het eerste tertaal zal een working paper worden voorbereid door het dagelijks bestuur in samenspraak met de domeincoördinatoren en promotoren dat een overzicht geeft van de mogelijke bedreigingen, bijhorende preventie- en/of oplossingsstrategieën en verantwoordelijkheden. Dit proces wordt doorheen de looptijd van het steunpunt geactualiseerd. Het proces zal gecoördineerd worden door de strategisch beheerder. Louter exemplarisch stippen we hieronder een aantal kernelementen aan die deel zullen uitmaken van de risicoanalyse:

- Toetsontwikkeling: afhankelijkheid van onzekere curriculumcontext, cfr. eindtermen lager onderwijs; spanning tussen geautomatiseerd toetsen en toetsen van hogere orde vaardigheden.

- Digitaal platform: afhankelijkheid budget en expertise van de externe ICT-firma; servercapaciteit bij systeembrede toetsing; beperkte mogelijkheden tot laten lopen van testprocedures op finale gebruiksschaal.
- Planning en budget: tijdstip van gunning van opdracht voor IT-partner verhouding investering initiële toetsontwikkeling versus toetsverversing, inzetbaarheid van schoolinterne toetsassistenten.
- Stakeholders: weerstand van onderwijsverstrekkers, scholen, leraren, ouders en leerlingen, strategische meningsverschillen over opzet van systeem van gecentraliseerd toetsen tussen steunpunt en Vlaamse regering.
- Afnamecondities: digitale uitrusting van scholen om gestandaardiseerd toetsen te garanderen;
- Garanties op technische ondersteuning bij calamiteiten; juridische implicaties van gecentraliseerde toetsen op privacy en gegevensbescherming van leerlingen en ouders.
- maatschappelijke impact van gecentraliseerde toetsen, i.c. onbedoelde en ongewenste (neven)effecten: teaching to the test, schaduwonderwijs, mogelijke non-effecten in trendanalyses, …

Het uitwerken van een grondige SWOT-analyse zal het steunpunt een realistische inschatting geven van de sterke en zwakke punten van de plaats die het steunpunt in het onderwijslandschap zal innemen, samen met kansen en mogelijke bedreigingen. Ondanks de best mogelijke voorbereiding loopt elke project het risico niet de vooropgestelde verwachtingen te kunnen realiseren. De algemene coördinatie van het steunpunt heeft als opdracht dit risico te minimaliseren. De coördinatoren en promotoren van het steunpunt zullen systematisch inventariseren welke mogelijk bedreigingen (én kansen) zich aandienen. Oorzaken en gevolgen worden geïdentificeerd. Doel hiervan is om in context van onvolledigheid en onzekerheid van informatie toch de best mogelijke keuzes te maken. Bedreigingen worden geprioriteerd op basis van hun kans op voorkomen en hun verwachte impact. Door te anticiperen op en vervolgens proactief te handelen vermijden we dat ongewenste bedreigingen zich effectief voordoen. Op die manier wil het steunpunt besturen en bijsturen, controle creëren en controle houden. Waar nodig wordt reactief gehandeld. Daarbij worden handelingsopties in een scenariologica uitgedacht, geanalyseerd en aan beslissingscriteria getoetst met het oog op evaluatie. Bij deze analyse zullen zowel interne medewerkers als externe stakeholders betrokken worden, met de opdrachtgever als eerste partner. Vanuit de resultaten van deze analyse en het ontwikkelplan dat hieruit zal volgen, zal het steunpunt zich stevig kunnen verankeren in een steeds veranderende onderwijscontext.

4. Werkdomein B. Ontwerp ICT-infrastructuur

Het Werkdomein Ontwerp ICT-infrastructuur omvat vier grote werkpakketten (Tabel 3.8): B1 Ontwerp & kwaliteitscontrole digitaal platform; B2 Datamanagement; B3 Logistiek beheer (ICT-infrastructuur); en B4 Ontwikkeling en beheer van de website. De vier werkpakketten bestrijken de volledige projectperiode.

	2021			2022			2023			2024			2025		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
B1 Ontwerp & kwaliteitscontrole digitaal platform															
B2 Datamanagement															
B3 Logistiek beheer (ICT-infrastructuur)															
B4 Ontwikkeling - Beheer website															

Tabel 3.8. Werkpakketten Ontwerp ICT-infrastructuur

4.1. B1 Ontwerp en kwaliteitscontrole van het digitaal platform

Een eerste werkpakket binnen dit werkdomein is het ontwerpen van de kenmerken waaraan het digitaal platform moet voldoen. Een digitaal platform voor scholen, leraren en ouders biedt de mogelijkheid om verschillende informatiesystemen te integreren. In het digitaal platform zullen drie verschillende digitale modules geïntegreerd worden:

Module 1. Informatiesysteem voor gebruikers

Module 2. Toetsomgeving voor de leerlingen

Module 3. Feedbacksysteem voor gebruikers

Module 1. Een informatiesysteem voor gebruikers

Naast de mogelijkheid om algemene informatie met betrekking tot de gecentraliseerde toetsen te integreren in het digitaal platform, zullen er ook submodules ontwikkeld worden die gebruikers (scholen, leraren, ouders, …) in staat stellen specifieke gegevens gestandaardiseerd aan te leveren, zoals ouder- of lerarenvragenlijsten waarvan de gegevens mogelijk gekoppeld kunnen worden aan leerlingendata. De eerste module wordt in nauw overleg ontwikkeld met de medewerkers van Werkdomein E. (organisatie van afname) en Werkdomein G. (Verwerking en analyse) en in nauw overleg met de externe ICT-firma.

Module 2. Toetsomgeving voor de leerlingen

Uit de meest recente resultaten uit de MICTIVO-studie in opdracht van het Departement Onderwijs en Vorming (MICTIVO, 2018) blijkt dat slechts de helft van de lagere scholen en twee op drie van de secundaire scholen in Vlaanderen in het bezit zijn van software voor het maken van specifieke oefeningen en toetsen. Digitaal toetsen is in het Vlaamse onderwijs nog niet ingeburgerd. Recent onderzoek, onder meer op basis van data uit grootschalige indicatorenstudies (TIMSS, Fishein et al, 2018; PISA, Zehner, et al. 2020; gestandaardiseerde toetsen in de VS, Bakes & Cowan, 2018), laat zien dat toetsresultaten van leerlingen niet onafhankelijk zijn van het gebruikte medium. Dit impliceert dat een gebrek aan vertrouwdheid met het toetsmedium een mogelijk negatief effect heeft op de betrouwbaarheid van de toets (Backes & Cowan, 2018).

De opdrachtgever kiest voor de digitale afname van gecentraliseerde toetsen voor taal en wiskunde en incorporeert op deze manier de voordelen die digitaal toetsen met zich meebrengen, zoals mogelijkheden tot geautomatiseerde scoring, adaptieve toetsing en geavanceerde mogelijkheden van dataopslag en -koppeling aan andere databronnen. In intensief overleg met de toetsontwikkelaars (werkdomeinen C en D) kan kleinschalig onderzoek opgezet worden bij verschillende leerlingengroepen om na te gaan in welke mate de kenmerken van het digitale toetsplatform van invloed zijn op zowel de gebruikerservaring als de toetsresultaten.

De toetsomgeving moet beantwoorden aan strenge technische kwaliteitscriteria zoals compatibiliteit met gebruikersinfrastructuur, performantie bij tienduizenden gelijktijdige afnames, logging, hosting, informatieveiligheid, encryptie, enzoverder. De toetsomgeving moet tevens beantwoorden aan hoge eisen inzake beheersmogelijkheden (lees- en toegangsrechten), itemtypes en notatiemogelijkheden (i.c. voor wiskunde), navigatiemogelijkheden, communicatie met de servers, representatie van de gegevens, enz.

Naast de intensieve samenwerking met de toetsontwikkelaars, zal het technisch ontwerp van de toetsomgeving voor leerlingen ook in nauw overleg ontwikkeld worden met de coördinatoren van Werkdomein E. (Organisatie van afname) Werkdomein F. (psychometrie) en Werkdomein G. (Verwerking en analyse). De eerste operationele versie van de toetsomgeving is voorzien voor de eerste kalibratie in 2022.

Module 3. Feedbackmodule voor gebruikers

In de beschrijving van Werkdomein H (verder in dit Deel) gaat specifieke aandacht uit naar het belang van accurate, relevante en gebruiksvriendelijk gepresenteerde feedbackgegevens als effectiviteitsvoorwaarde. De ICT-beheerder is actief betrokken bij de methodologie van service-design (Miller, 2015) waarbij de ontwikkeling van het feedbacksysteem rekening houdt met het evenwicht tussen de behoeften van de gebruikers en de noden van de opdrachtgever. Het iteratief ontwerp (Stickdorn et al., 2018) van het feedbacksysteem wordt in steeds toenemende complexiteit opgebouwd door het cyclisch doorlopen van behoefteanalyses, ontwerpactiviteiten en gebruikerstesten. De ICT-beheerder is actief betrokken in dit proces en vertaalt de front-end ontwerpeisen inzake bijvoorbeeld de gebruikersinterface en het content managementsysteem (CMS) van het feedbacksysteem naar de ICT-firma.

Bij uitbreiding geldt voor de drie modules dat de ICT-beheerder verantwoordelijk is voor het inventariseren van de technische vereisten waaraan het digitaal platform moet voldoen. De ICT-beheerder is spilfiguur in de communicatie met de externe ICT-firma die het digitaal platform zal ontwikkelen en maakt de technische vertaalslag van de ontwerpkenmerken waaraan het digitale platform moet voldoen in communicatie met de ICT-firma. Hij / zij dient in overleg met de ICT-firma een efficiënt model van zogenaamd ‘rapid prototyping’ uit te werken, aangezien een streefdoel is om tijdens de eerste kalibratiefase reeds toetsitems, bij aanvang van het derde tertaal van 2022, te integreren in een betaversie van het toetsplatform. Deze versie zal worden opengesteld voor de steekproef van scholen en leerlingen waar de eerste kalibratie zal lopen, met de mogelijkheid tot feedback op het platform zelf. Het toetsplatform zal in het derde tertaal van 2023 operationeel zijn.

4.2. B2 Datamanagement

Een tweede werkpakket van de ICT-beheerder heeft betrekking op het datamanagement. In B1 werd reeds de nadruk gelegd op de strenge technische kwaliteitscriteria waaraan het digitaal platform moet voldoen. Een tweede werkpakket richt zich op datamanagement: de wijze waarop de toets- en feedbackdata worden bijgehouden, beveiligd en bruikbaar gemaakt voor rapportage en gebruik. In overleg met de ICT-firma worden oplossingen gezocht om grote hoeveelheden data op de meest optimale manier te structureren, op te slaan op servers en beschikbaar te stellen aan de verschillende

gebruikers. In overeenstemming met werkpakket B1 is nauwe samenwerking vereist met coördinatoren uit de verschillende werkdomeinen. Het datamanagementsysteem zal eveneens in samenwerking met de domeincoördinatoren worden ontworpen, en in hoofdzaak met de coördinatoren in Werkdomeinen E, F, G en H. Het systeem zal over een kwaliteitsvolle Human-Computer Interface dienen te beschikken om vlotte toegang tot de data voor de onderzoekers te faciliteren.

4.3. B3 Logistiek beheer

De ICT-beheerder treedt tevens op als logistiek beheerder van de ICT-infrastructuur van het Steunpunt.

4.4. B4 Ontwikkeling – beheer website

Het steunpunt maakt de keuze de website in eigen beheer te ontwikkeling. De technische realisatie valt onder de verantwoordelijkheid van de ICT-beheerder. De inhoudelijke opmaak en functionaliteiten van de website worden besproken in samenspraak met alle betrokken partners binnen het steunpunt en in overleg met de opdrachtgever.

4.5. Het belang van een dynamische en gefaseerde aanpak

Aan de implementatie van de digitale modules gaan een grondig ontwerp en ontwikkeling vooraf. De ontwikkeling zelf valt niet onder het takenpakket van het steunpunt maar vereist in de ontwerpfase intense samenwerking tussen de verschillende werkdomeinen en met de externe ICT-firma. Deze cruciale succesfactor geldt voor elk van de modules die deel uitmaken van B1. Gegeven dat het een nieuw project betreft, zonder de mogelijkheden van doorontwikkeling vanuit andere initiatieven, en gegeven de complexiteit van te ontwikkelen modules en submodules, is het aangewezen met een dynamisch model van IT-ontwikkeling te werken en met een systeem van timeboxing, een methode waarbij de ontwikkelen modules worden opgedeeld in onderscheidbare submodules waarvan de technische vereisten duidelijk worden gedefinieerd en duidelijke opleverdata voor de ICT-firma worden vastgelegd.

5. Werkdomeinen C & D. Toetsontwikkeling Nederlands en Wiskunde

5.1. Inleiding

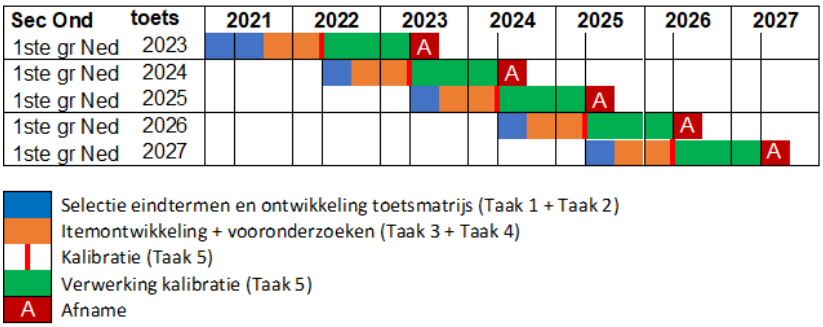
Voor de werkdomeinen Nederlands (C) en wiskunde (D) wordt eenzelfde plan van aanpak gehanteerd. Het werkdomein Nederlands wordt gecoördineerd door de UGent en dat van wiskunde door de VUB, waarbij wiskunde basisonderwijs wordt opgenomen door de KU Leuven. Beide werkdomeinen worden in dit deel samen besproken. De aanpak van beide werkdomeinen wordt uitgewerkt in lijn met wat in het oproepdocument werd uitgetekend. De inhoudelijke afbakening van de toetsen gebeurt tijdens de looptijd van het steunpunt in overleg met de stuurgroep.

De toetsontwikkeling gebeurt via een gefaseerde aanpak van in totaal vijf taken. De eerste drie taken richten zich via een onderzoeksgerichte aanpak op de inhoudelijke ontwikkeling van de toetsen. Meer specifiek omvatten deze taken de selectie van de eindtermen (Taak 1), de opmaak van de toetsmatrijs (Taak 2) en de itemontwikkeling (Taak 3). Na de initiële ontwikkeling van de toetsen volgen twee feedbackrondes aan de hand van vooronderzoeken bij leerlingen en experten/onderwijsdeskundigen (Taak 4) en via een grootschalige kalibratiestudie (Taak 5). Het chronologisch trimestrieel verloop van de verschillende taken binnen het toetsontwikkelingsproces wordt in de figuur hieronder exemplarisch voorgesteld voor de toets Nederlands eerste graad.

In aanloop naar de eerste afname van de gecentraliseerde toetsen wordt een periode ingepland waarin items ontwikkeld worden. Deze periode is minimaal een schooljaar (toetsen eerste graad), maar kan ook langer lopen (toetsen vierde leerjaar waarvan de afname met een jaar is uitgesteld). Het aantal items dat ontwikkeld moet worden is afhankelijk van de eerder opgestelde toetsmatrijs, maar ook van het afnamedesign waarmee de gecentraliseerde toetsen worden afgenomen. De keuze hiervan is ook voorwerp van de haalbaarheidsstudie. In deze aanvraag wordt alvast zoveel als mogelijk geïnvesteerd in itemontwikkeling om met een zo omvangrijk mogelijke itembank te kunnen starten om zo beperkte adaptieve toetsafnames en het werken met verschillende toetsboekjes (indien de gecentraliseerde toetsen volgens de haalbaarheidsstudie best niet op eenzelfde dag zouden worden afgenomen) mogelijk te maken.

Deze periode van itemontwikkeling mondt uit in de kalibratiestudie op basis waarvan de meetschaal en bijhorende itembank wordt opgesteld. Voor de toetsen in de derde graad is de kalibratiestudie pas in het volgende steunpunt voorzien. Merk daarbij op dat al vanaf de eerste afname elke afname tevens dient als de kalibratie voor nieuwe items die pas bij een volgende afname worden ingeschakeld. Daarvoor wordt voor elke toets doorlopend één toetsontwikkelaar voorzien.

Binnen de haalbaarheidsstudie zal aangegeven moeten worden in welke mate de selectie van de eindtermen en de ontwikkeling van de toetsmatrijs constant gehouden kunnen worden over de verschillende afnames heen dan wel dat er telkens ook nieuwe inhouden aan de afnames worden toegevoegd via een variabel deel.



Figuur 3.2. Gantt chart voor de toetsontwikkeling Nederlands eerste graad.

Binnen elk van deze taken wordt een beroep gedaan op relevante stakeholders. Alvorens de verschillende taken hieronder te omschrijven, wordt eerst de samenstelling van deze stakeholders in verschillende adviesorganen besproken. De samenstelling van deze organen kan beschouwd worden als een nultaak en proloog tot de vijf taken van toetsontwikkeling.

5.2. Samenstelling en werking interne en externe expertgroep

5.2.1. Interne expertgroepen

De interne expertgroepen Nederlands en wiskunde zullen de respectievelijke werkdomeinen Toetsontwikkeling mee aansturen en fungeren als een continu aanspreekpunt voor de coördinator en toetsontwikkelaars. Dit zal enerzijds gebeuren aan de hand van informele contacten met individuele leden. Anderzijds wordt voor Taak 1, Taak 2 en Taak 3 minstens één keer een formeel overleg georganiseerd met de expertgroep.

5.2.2. Externe expertgroepen

Zoals in de oproep is voorzien, wordt er per domein, een externe expertgroep opgericht die bestaat uit vertegenwoordigers van het Departement Onderwijs van de Vlaamse Overheid, aangevuld met leden van het onderwijsveld en wetenschappelijke peers. De externe expertgroep zal geconsulteerd worden bij alle fases in de ontwikkeling van de toetsen over de relevantie, haalbaarheid, helderheid en validiteit van de toetstaken en -items met betrekking tot de eindtermen en voor de betreffende leeftijdsgroep. De externe expertgroep zal dus de toetsontwikkeling en de interpretatie van de resultaten nauw mee opvolgen. Het gaat onder andere om de keuze voor de te toetsen aspecten van de eindtermen, de vertaalslag van de geselecteerde eindtermen naar concrete toetsitems, deelname aan de cesuurbepaling en dit vanuit de kennis over wat er precies in de klassen gebeurt, wat de verschillen zijn tussen leerplannen et cetera.

De groep van leden van het onderwijsveld en de wetenschappelijke peers wordt als volgt samengesteld:

- een ruime vertegenwoordiging van de verschillende onderwijsverstrekkers (pedagogische begeleidingsdiensten wiskunde en Nederlands). Zoals voorzien in de oproep, wordt met deze groep ook de mogelijkheid besproken om de centraal ontwikkelde kerntoetsen te complementeren met items (of zelfs aparte toetsen) die zij zelf ontwikkelen op voorwaarde dat hierbij de vereisten op vlak van kwaliteit en haalbaarheid niet in het gedrang komen. Zo blijft het resultaat op de kerntoetsen op deze manier vergelijkbaar, maar het mede-eigenaarschap maakt het draagvlak groter;
- vertegenwoordigers van de onderwijsinspectie (basisonderwijs en secundair onderwijs);
- vertegenwoordigers uit de lerarenopleiding professionele bachelor;

- vertegenwoordigers uit de lerarenopleiding (Educatieve Master) van de Vlaamse universiteiten, geflankeerd door praktijkassistenten en onderwijsbegeleiders.

Specifiek voor Nederlands wordt er daarnaast ook een vertegenwoordiging voorzien uit:

- het Vlaams talenplatform. Het Vlaamse talenplatform verenigt de Vlaamse academische taalopleidingen van de Universiteit Antwerpen, de Vrije Universiteit Brussel, de Universiteit Gent en de KU Leuven. Ook een groeiende groep van docenten en leerkrachten in de hogescholen, talencentra, het volwassenenonderwijs en het basis- en secundair onderwijs sluit zich aan bij het initiatief;
- de resonantiegroep Instaptoets Nederlands voor de Lerarenopleidingen, die werd goedgekeurd door VLHORA en die alle hogescholen (lerarenopleiding professionele bachelor) vertegenwoordigt;
- een vertegenwoordiger van de Nederlandse Taalunie;
- drie tot vier individuele academische experts met taaltoetsing en/of taaldidactische expertise.

Specifiek voor wiskunde wordt de expertgroep aangevuld met:

- Vertegenwoordigers uit Platform Wiskunde Vlaanderen. Dit platform verenigt op een gecoördineerde wijze alle wiskunde-actoren in Vlaanderen: het onderwijs, de onderzoekers, het bedrijfsleven en de overheid. Het heeft als doelstelling onder andere het ondersteunen van degelijk wiskundeonderwijs voor alle jongeren, dat ook op gepaste wijze voorbereidt op het hoger onderwijs, het versterken van de publieke beeldvorming over wiskunde, het aanpakken van het tekort aan wiskundigen in bedrijfsleven en onderwijs, het faciliteren van wiskundige innovatie en valorisatie en het verstevigen van de mogelijkheden voor wiskunde-onderzoek. Omdat het onmogelijk om aan de expertengroep vertegenwoordigers van alle afzonderlijke actoren uit en naast het onderwijs, zoals b.v. vertegenwoordigers van de verschillende onderzoeksgroepen en de diverse (industriële) sectoren en bedrijven toe te voegen, wordt het Platform Wiskunde Vlaanderen uitgenodigd om vertegenwoordigers in de expertgroep af te vaardigen. Deze vertegenwoordigers vormen de brug tussen de expertgroep enerzijds, en de overige belanghebbenden uit het onderwijs, onderzoek en het bedrijfsleven anderzijds. Dat het bedrijfsleven een belangrijke en onderbouwde mening heeft over wiskunde, en de te verwachten kennis van afgestudeerden die de arbeidsmarkt betreden, bleek onlangs uit het artikel *Nodig op school: de wiskunde die stemmen en bankieren veilig houdt* (De Standaard, 27 augustus 2020), waarin verschillende bedrijfsleiders een pleidooi houden voor meer discrete wiskunde in de eindtermen.
- Vertegenwoordigers uit de toelatingsexamens, pretoets en ijkingstoetscommissies wiskunde, die de interne expertgroep met ervaring hierin flankeren. Hoewel de doelgroep van leerlingen die deelnemen aan de toelatingsexamens, pretoets en ijkingstoetsen slechts een deel is van de populatie van leerlingen in het zesde jaar secundair onderwijs, is het toch ook belangrijk om niet alleen de expertise van deze commissies mee te nemen bij de toetsontwikkeling maar om hen ook te betrekken in het verder nadenken van de toekomstige positie van de gecentraliseerde toetsen ten opzichte van de ijkingstoetsen.

5.2.3. Praktijkreflectiegroepen

Naast de expertgroepen, zullen er *ad hoc Praktijk Reflectiegroepen (PR-groep)* worden samengesteld (door de respectievelijke interne en externe expertgroep) waarin leerkrachten en directies die ervaring hebben met de te toetsen leerjaren en niveaus vertegenwoordigd zijn. Een groep van zorgvuldige geselecteerde praktijkmensen zal bijdragen tot de relevantie en gedragenheid van de centrale toetsen.

5.3. Selectie van de eindtermen (Taak 1)

Als eerste taak gebeurt het onderzoek naar de selectie van de eindtermen die zullen getoetst worden waarbij de praktische haalbaarheid en constructvaliditeit geëvalueerd worden. Om enerzijds leerwinst te meten op leerling- en schoolniveau en anderzijds ook te kunnen nagaan of de leerlingen de eindtermen bereiken, inclusief het vaardigheidsniveau, zal de selectie bestaan uit een vaste kern (ongeveer 75% van de toetsitems, hierbinnen zitten alvast alle eindtermen basisgeletterdheid die getoetst kunnen worden) en een variabel deel (ongeveer 25%) dat jaarlijks gewijzigd wordt. Om de haalbaarheid naar het aantal te ontwikkelen items en een redelijke toetstijd voor de leerlingen te garanderen, stelt het oproepdocument dat het voldoende is dat de toetsen uitspraken doen over een set van inhoudelijk samenhangende eindtermen en niet per eindterm afzonderlijk. De toetsitems die peilen naar het variabel deel van de eindtermen worden pas ontwikkeld na de jaarlijkse selectie van deze eindtermen.

De selectie van de eindtermen gebeurt door de wetenschappelijke medewerker(s) in de werkdomeinen onder begeleiding van de domeincoördinatoren en in overleg met de verbindingsmedewerker vanuit het werkpakket psychometrie. Daarnaast gebeurt de selectie in overleg aan de hand van een tweetrapsprocedure met (1) de interne expertgroep, de externe expertgroep en de ad hoc PR-groep en (2) de stuurgroep.

Er zal onderzocht worden in welke mate het nieuwe concept van bouwstenen voor de eindtermen het potentieel hebben om de doelen van leerwinstmonitoring te verzoenen met de vereiste om zich te baseren op de eindtermen. De bouwstenen clusteren eindtermen binnen de sleutelcompetenties, ze vormen de rode draad doorheen het secundair onderwijs en ze gelden voor alle leerlingen, ongeacht finaliteit of onderwijsvorm. In onderstaande figuur worden illustratief de bouwstenen voor wiskunde in het basisonderwijs en in de eerste graad van het secundair onderwijs vermeld. De eindtermen voor de tweede en derde graad bouwen verder op deze bouwstenen.

- Inzicht ontwikkelen in en omgaan met getallen en hoeveelheden: getallenleer.
- Inzicht ontwikkelen in en omgaan met ruimte en vorm: meetkunde en metend rekenen.
- Inzicht ontwikkelen in en omgaan met relatie en verandering: zoals algebra, analyse en discrete structuren.
- Inzicht ontwikkelen in en omgaan met data en onzekerheid: zoals kansrekenen en statistiek
- Redeneringen opbouwen en abstraheren rekening houdend met de samenhang en structuur van wiskunde.
- Modelleren en problemen oplossen door analyseren, (de)mathematiseren of aanwenden van heuristieken.

Figuur 3.3. Bouwstenen voor wiskunde in het basisonderwijs en de eerste graad secundair onderwijs

Om het specifieke construct van de toetsen Nederlands en wiskunde voor het lager en secundair onderwijs te ontwikkelen en inhoudsvaliditeit te garanderen, wordt er een inhoudelijke analyse uitgevoerd van de geldende eindtermen voor het lager onderwijs en de eerste en derde graad van het secundair onderwijs. In navolging van het gemaakte onderscheid tussen basisvorming en beroepsvorming in het Regeerakkoord - en voor Nederlands ook in navolging van het Strategisch Plan Geletterdheid Verhogen - hebben we daarbij aandacht voor basisgeletterdheid enerzijds en beroepsgeletterdheid en/of academische geletterdheid anderzijds. Op basis van de inhoudelijke analyse van de eindtermen wordt vervolgens een databank samengesteld waarin de eindtermen voor de volgende criteria worden gecodeerd:

- Constructrelevantie voor de te toetsen domeinen, zoals bijvoorbeeld (1) begrijpend lezen, (2) schrijven en (3) grammatica

- Operationaliseerbaarheid van de competentiegerichte eindtermen in de geplande centrale toetsen
- Cognitief verwerkingsniveau met aandacht voor (delen van) eindtermen op verschillende niveaus
- Gemeenschappelijkheid: eindtermen die voor bso, tso en aso gemeenschappelijk zijn, worden samengenomen. Eindtermen die onderscheidend zijn voor de verschillende onderwijsvormen kunnen afzonderlijk gecodeerd worden.

Met betrekking tot de ontwikkeling van het specifieke construct van de toetsen Nederlands is het belangrijk te vermelden dat we de vier bouwstenen van de sleutelcompetentie Nederlands (Nederlands als communicatiemiddel; Kenmerken en principes van het Nederlands begrijpen om inzetten in communicatie; Inzicht hebben in het Nederlands, als exponent en deel van een cultuur en een maatschappij; Literatuur in het Nederlands beleven) als uitgangspunt nemen. We sluiten daarbij aan bij hedendaagse visies op taal, taalontwikkeling en taalleren die taalcompetentie zien als een complex samenspel van vaardigheden, kennis en attitudes die leerlingen in rijke contexten al doende verwerven, door functionele activiteiten uit te voeren en in interactie te gaan met anderen (Douglas Fir Group, 2016; Ellis, 2009; Long, 2014; Tomasello, 2003).

Binnen dit selectie-onderzoek zal er onder andere een studie gebeuren van de selectie van de "oude" eindtermen in de reeds uitgevoerde peilingen en het internationaal vergelijkend onderzoek (zoals PISA) en het identificeren van de dichtst bijgelegen "nieuwe" eindtermen. De vergelijkende studie die hierbij door het Steunpunt voor Toetsontwikkeling en Peilingen gebeurde kan hierbij als uitgangspunt dienen. Zodoende hebben we ook weet van die eindtermen binnen de nieuwe bouwblokken die bijzondere aandacht verdienen wegens eerdere alarmscores. Voor de A-stroom zal dit onderzoek aangevuld worden met een analoge studie van eindtermen-selecties en probleempunten naar voor gekomen in toelatingsexamens, ijkingstoetsen en pre-toetsen uitgevoerd aan onze universiteiten. Deze resultaten geven waardevolle input voor het bepalen van het vast versus variabel deel en Taak 3.

5.4. Opstellen van een toetsmatrijs (Taak 2)

Op basis van een inhoudelijke analyse van de geselecteerde eindtermen wordt in Taak 2 een toetsmatrijs ontwikkeld. Een toetsmatrijs vormt een blauwdruk voor de toets, en zorgt ervoor dat de samenstelling van de toets representatief is voor de gevraagde kennis en vaardigheden volgens de eindtermen[1]. Door het gebruik van de matrijs garanderen we de inhoudsvaliditeit van de toetsen en zorgen we dat de toetsen jaar na jaar een maximale equivalentie hebben qua samenstelling. In de toetsmatrijs worden niet alleen de inhoudelijke specificaties van de te ontwikkelen toetsen vastgelegd, maar wordt ook schematisch geconsolideerd over welke (groepen van) eindtermen apart gerapporteerd zal worden en waarvoor een specifieke meetschaal zal worden ontwikkeld.

In de toetsmatrijs worden de geselecteerde eindtermen geclassificeerd volgens een aantal dimensies en criteria. Naast een inhoudelijke dimensie en een dimensie die verwijst naar de vereiste cognitieve verwerkingsniveaus, kan bijvoorbeeld ook worden aangegeven hoeveel items van welk taaktype er nodig zijn en naar welke thematische verdeling er gestreefd wordt. Een taak vormt een thematisch of betekenisvol geheel/opdracht waartoe meerdere items behoren. Elk item wordt daarbij gelabeld op basis van een aantal criteria (zie hieronder). Taken zijn bijvoorbeeld een leestekst met verschillende vragen of een plot met statistische data waarover verschillende vragen worden gesteld.

Alle kwaliteitsvolle taken en items worden bewaard in een zogenaamde itembank: een beveiligde database waarin ieder item opgevraagd kan worden aan de hand van een reeks van tags en

[1] Attitudes worden elders gemeten via survey.

psychometrische metadata. Alle bouwblokken die benoemd worden in de toetsmatrijs, zijn getagd in de itembank. Tot die bouwblokken behoren onder andere: selectie van eindtermen, verwerkingsniveau, thema, empirisch bepaalde moeilijkheidsgraad (wordt toegevoegd na kalibratie-studie), meest recente operationeel gebruik, enzovoort.

In deze taak wordt er overleg gepleegd met Werkdomeinen F (Psychometrie), G (Verwerking, analyse en rapportage resultaten) en H (Schoolfeedback). Verder worden ook de interne expertgroep, de externe expertgroep en de ad hoc PR-groep betrokken bij de ontwikkeling van de toetsmatrijs, zoals hierboven al aangegeven.

5.5. Ontwikkelen van de toetsvragen (Taak 3)

Na de beslissing over de definitieve toetsmatrijs worden in een volgende taak de eigenlijke toetsvragen ontwikkeld en is er opnieuw overleg met Werkdomeinen F (Psychometrie), G (Verwerking, analyse en rapportage resultaten) en H (Schoolfeedback). De vragen zijn aangepast aan afname in een digitale omgeving, waar dit laatste gebeurt via overleg met Werkdomein A. De eigenlijke afname gebeurt in samenwerking met Werkdomein E.

Voor de inhoudelijke ontwikkeling van de toetsvragen vormen de leerplannen, de handboeken en bestaande toetsen een inspiratiebron. Een speciaal aandachtspunt daarbij is dat de toetsvragen geen enkel handboek of leerplan bevoordelen. Dit wordt mede bewaakt door een concordantiematrix op te stellen tussen de onderwijsdoelen van de overheid en de leerplandoelen van de onderwijsverstrekkers.

We onderzoeken de meest geschikte toetsvorm om via een digitale toetsafname de kennis en vaardigheden opgenomen in de geselecteerde eindtermen op een betrouwbare en valide manier grootschalig te meten. Op het gebied van item- en antwoordformaat bekijken we innovatieve types die computergestuurde toetsen mogelijk maken in tegenstelling tot die op papier. Zo wordt de toetsontwikkeling geoptimaliseerd en bouwen we expertise hierrond op.

In deze taak onderzoeken we ook op welke manier de beoordeling van de antwoorden op de vragen zal gebeuren. We streven daarbij naar vraag- en antwoordtypes die automatische scoring toestaan én die bovendien voldoende aansluiten bij de eigenheid van de domeinen wiskunde en Nederlands. Er is daarbij zeker een spanningsveld tussen de itemtypes die automatische scoring toelaten en de itemtypes die het meest aanleunen bij de competentiegerichte formulering van de nieuwe onderwijsdoelen. Voor de nieuwe onderwijsdoelen van de eerste graad werd hierover in 2019 zowel door AHOVOKS als door STEP een nota uitgewerkt die beiden aangeven dat competentiegerichte toetsing ook mogelijk is bij grootschalige toetsafnames. Binnen het nieuwe steunpunt wordt deze problematiek verder onderzocht en uitgewerkt voor de andere betrokken onderwijsdoelen.

De ontwikkeling van scorewijzers, waarin wordt aangegeven welke antwoorden juist worden gerekend, maakt integraal deel uit van de itemontwikkeling. We streven naar een uniform scoremodel voor zowel Nederlands als Wiskunde, maar zonder daarbij a priori voor de afzonderlijke gebieden vast te leggen hoe men tot het resultaat komt. De ontwikkeling van het scoremodel gebeurt in nauwkeurig overleg met Werkdomeinen F (Psychometrie), G (Verwerking, analyse en rapportage resultaten) en H (Schoolfeedback). Het scoremodel is immers bepalend voor de mogelijkheden en de manier waarop uitspraken kunnen gedaan worden over het behalen van eindtermen en leerwinst.

Taak 3 wordt opgedeeld in de verschillende deeltaken.

Deeltaak 1: Toetsontwikkeling 1ste graad secundair onderwijs A- en B-stroom (afname vanaf 2023)

- Voor Nederlands gebeurt dit door UGent, voor wiskunde door de VUB.
- Hierbij komen zowel eindtermen basisgeletterdheid als gewone eindtermen aan bod.

- Bij deze toetsontwikkeling houden we rekening met de vraag om leerwinst te kunnen meten tussen het einde van de 1ste graad en het einde van de 3de graad.

Deeltaak 2: Toetsontwikkeling basisgeletterdheid 4de lager (afname vanaf 2024)

- Voor Nederlands gebeurt dit door UGent, voor wiskunde door KULeuven.
- Deze toetsen focussen op basisgeletterdheid. De ontwikkeling en afname van deze toetsen is onder voorwaarde van een tijdige implementatie van de nieuwe eindtermen lager onderwijs. De invoering van de basisgeletterdheid in het basisonderwijs hangt af van de evaluatie van basisgeletterdheid op het einde van de 1ste graad secundair onderwijs. Hiertoe wordt in 2022 een peiling afgenomen voor Nederlands en wiskunde aan het einde van de 1ste graad secundair onderwijs.

Deeltaak 3: Toetsontwikkeling 6de lager (geplande 1ste afname in 2026)

- Voor Nederlands gebeurt dit door UGent, voor wiskunde door KULeuven.
- In deze deeltaak gebeurt de kalibratie met het oog op een eerste afname en leerwinstmeting in schooljaar 2025-2026. De eigenlijke afname van deze toets valt buiten de looptijd van het steunpunt.

Deeltaak 4: Toetsontwikkeling 3de graad secundair onderwijs A- en B-stroom (geplande 1ste afname in 2027)

- Voor Nederlands gebeurt dit door UGent, voor wiskunde door de VUB.

5.6. Vooronderzoeken (Taak 4)

Zoals voorzien in de oproep, beoordelen de vakinhoudelijke experten in deze deeltaak de toetsvragen aan de hand van een aantal criteria. Na herwerking op basis van de feedback van de experten worden de toetsvragen afgenomen bij een beperkte steekproef van leerlingen. Op basis van de analyse van deze resultaten kunnen de toetsvragen verder aangepast of weggelaten worden indien nodig.

Gezien de eerste afnames in 2023 en 2024 wordt een eerste ronde vooronderzoeken voorzien op het einde van 2021/begin 2022 voor eerste graad secundair onderwijs en einde 2022/begin 2023 voor het vierde leerjaar. Dit gebeurt in samenwerking met Werkdomein E.

Vooronderzoek met experten/onderwijsdeskundigen

Met het oog op de kwaliteitsbewaking en het screenen van een eerste pool taken/items zullen de interne expertgroep, de externe expertgroep en de ad hoc PR-groep worden geraadpleegd. Dit met de vraag om al de ontwikkelde taken/items te beoordelen op (a) de relevantie van ten aanzien van de geselecteerde eindtermen en (b) de mate waarin het format van de taak, het antwoordformat, de bewoording en formulering in overeenstemming is met de het leerjaar en niveau van de te toetsen doelgroep. De onderwijsdeskundigen krijgen ook de mogelijkheid algemene opmerkingen te formuleren en aan te geven waar er eventueel nog hiaten zijn in de voorliggende itempool. Deze bevraging loopt bij voorkeur via een digitaal feedbackplatform. Op basis van dit vooronderzoek zal de itempool uit de toetsmatrijs worden bijgesteld waar nodig en zal de formulering van bepaalde items worden bijgestuurd.

Vooronderzoek met leerlingen/piloot

Het pilootonderzoek vindt plaats binnen het ontwikkelde digitale toetsplatform bij een kleine steekproef van leerlingen. In totaal wordt gezorgd dat elk item wordt voorgelegd aan 80 à 100 leerlingen uit het leerjaar en niveau waarvoor de toets wordt ontwikkeld. Bij de samenstelling van de steekproef voor deze pilootafnames wordt gegeven de beperkte steekproefomvang niet zozeer gezocht naar een representatieve steekproef. Wel wordt er bij het secundair onderwijs steeds voor gezorgd dat er een breed scala van studierichtingen bij deze pilootafname betrokken zijn. In het algemeen geldt ook dat bij pilootafnames van toetsen het een goede strategie is om eerder de extreme deelgroepen uit de populatie te bevragen, dan de doorsnee leerlingen.

Er worden een aantal proefafnames opgezet om aan de hand van eenvoudige psychometrische technieken een eerste zicht te krijgen op de moeilijkheidsgraad en afnametijd van de ontwikkelde opgaven. Verder zullen de ontwikkelde taken/items worden uitgetest aan de hand van een cognitief interview. Op deze manier kan worden geverifieerd of de kinderen/jongeren de taken/items begrijpen zoals bedoeld door de toetsontwikkelaars en kan in het geval van vragen met een eenvoudig open antwoord een eerste versie van scoringsregels worden opgesteld. Op basis van dit vooronderzoek zal de itempool uit de toetsmatrijs worden bijgesteld en zal de formulering van bepaalde items alsook de scoringswijzer worden bijgestuurd waar nodig.

5.7. Kalibratieonderzoek (Taak 5)

Tegen het einde van het schooljaar voorafgaand aan de afname van de gecentraliseerde toetsen, worden de ontwikkelde toetsen via het digitale toetsplatform voorgelegd aan een representatieve steekproef van leerlingen. Voor de eerste graad secundair onderwijs zal het eerste kalibratieonderzoek dus plaatsvinden rond mei 2022. De steekproef wordt getrokken als een gestratificeerde meertrapssteekproef uit de betreffende leerlingenpopulatie, zoals dat nu gebeurt bij STEP. De uiteindelijke steekproefgrootte hangt af van (1) het gehanteerde, projectspecifieke afnamedesign, waarbij de lengte van het toetsinstrument doorslaggevend is en eventueel de mate waarin er ankering tussen verschillende schalen dient voorzien te worden; (2) de vereiste stabiliteit van de itemparameters en de daarmee gepaard gaande betrouwbaarheid van de meetschaal. Er wordt ervan uitgegaan om een item voor te leggen aan minimaal 500 à 600 leerlingen. In principe worden alle ontwikkelde items opgenomen in het eerste kalibratieonderzoek voorafgaand aan de eerste afname om zo de grootte van de itembank te maximaliseren. De geheimhouding van de items bij deze studie is mede voorwerp van onderzoek van de haalbaarheidsstudie.

Bij de volgende kalibratieonderzoeken wordt na de afname de nood aan nieuwe items geïnventariseerd. Na ontwikkeling worden deze items eveneens onderworpen aan de hierboven besproken vooronderzoeken. De invoeging van deze items wordt bij de volgende afname zo ingepland dat zoveel mogelijk leerlingen (of alle leerlingen) een beperkt aantal kalibratie-items moeten oplossen, die niet meetellen bij het bepalen van hun eigenlijke toetsresultaat.

Het kalibratieonderzoek beoogt in de eerste plaats de psychometrische kwaliteit van de toetsvragen en de toets in zijn geheel te bepalen. Aan de hand van IRT-modellen worden de verzamelde gegevens van de leerlingen gemodelleerd en wordt de meetschaal ontwikkeld. Deze meetschaal wordt op zijn beurt dan gebruikt als basis voor de cesuurbepaling, het bepalen van de beheersingsniveaus en het eventueel adaptief maken van de toets (Zie werkdomein F).

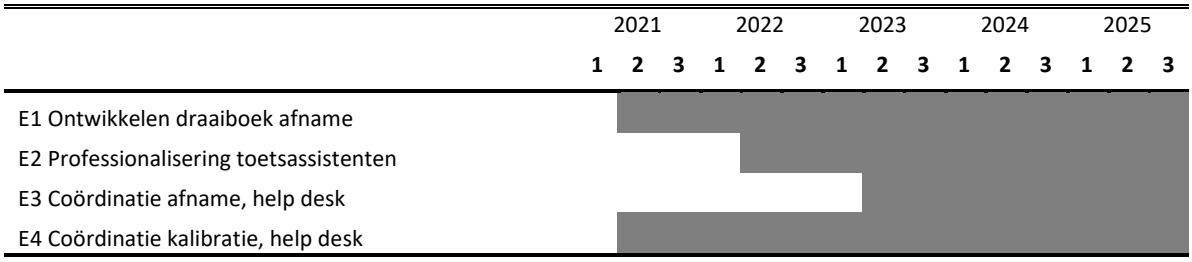
In tweede instantie wordt op basis van de kalibratiestudie ook de werk- en haalbaarheid van het afnameprotocol getest en geëvalueerd (onder meer op eenduidigheid en helderheid van de afnameprocedure, alsook op de haalbaarheid van de praktische organisatie en de voorziene timing). Dit gebeurt via een online bevraging bij alle toetsafnemers en via een steekproef in een aantal

klassen/scholen waar observatoren tijdens de testafname de *fidelity of implementation* zullen beoordelen.

Het kalibratieonderzoek verloopt in samenwerking met Werkdomein E.

6. Werkdomein E. Organisatie en Ondersteuning afname

Het steunpunt organiseert en ondersteunt de afname van de gecentraliseerde toetsen. Concreet neemt het steunpunt volgende taken op: E1 ontwikkelen van de draaiboeken voor afname (en kalibratie); E2 Professionalisering van toetsassistenten; E3 Coördinatie van de afname met help desk ondersteuning; E4 Coördinatie van kalibratie met help desk ondersteuning.



Tabel 3.14. Werkpakketten Organisatie en ondersteuning afname

Data-verzameling

Voor de afnames gaan we uit van een gestandaardiseerde toetsafname die populatiebreed gebeurt op één moment voor het vak Nederlands en één moment voor het vak Wiskunde.

Voor het beantwoorden van onderzoekvraag 1 (leerwinstmeting) wordt toetsdata van leerlingen op populatieniveau verzameld. Om de toegevoegde waarde te kunnen berekenen wordt deze populatiedata gekoppeld aan schoolgegevens (te verkrijgen uit reeds verzamelde gegevens in administratieve databanken) en achtergrondinformatie van de leerlingen (via leerlingbevraging en/of administratieve databanken). Voor het capteren van bijvoorbeeld leerlingenmobiliteit wordt de data gekoppeld aan leerlinggegevens in administratieve databanken. Onderzoeksvraag 2 (beheersing eindtermen) wordt beantwoord op basis van toetsdata op populatieniveau. De onderzoeksvragen 3 (welke leerlingen hebben het moeilijker om de eindtermen te bereiken), 4 (in welke mate verschillen scholen), 5 (in welke mate blijven verschillen tussen groepen leerlingen overeind na controle op achtergrond) en 6 (verklaringen van verschillen groepen leerlingen/scholen) worden beantwoord door toetsdata te koppelen aan achtergrondkenmerken van leerlingen (zie OV1). Onderzoeksvraag 7 (onderwijsaanbod) wordt bevraagd via representatieve steekproeven van leraren. Onderzoeksvraag 8 (tendensen vak/leergebied) wordt beantwoord op basis van toetsdata verzameld op verschillende meetmomenten.

Organisatie van de data-afname

De taken worden verdeeld over meerdere personen om expertise op te bouwen en duurzaam in te zetten. Scholen zullen voor inhoudelijke vragen bij de toetsen en de toetsorganisatie kunnen steunen op het steunpunt. Gedetailleerde draaiboeken zullen uitgewerkt en bezorgd worden aan scholen (E1); toetsassistenten kunnen vorming volgen om hun taak succesvol uit te voeren (E2); en het steunpunt zal actief een rol opnemen in het beantwoorden van vragen van scholen (E3). De begroting is opgesteld voor afnames van de centrale toetsen, met inclusie van de kalibratietoetsen die de afname voorafgaan (E4).

Tijdens de afname zullen er toetsassistenten in elke school aanwezig zijn. Dit zijn bestaande (liefst vastbenoemde) personeelsleden die werkzaam zijn op de school of binnen de scholengemeenschap en opgeleid worden om deze taak adequaat en ethisch verantwoord uit te voeren. Zij engageren zich tot een deontologische code die maakt dat de toetsen op een betrouwbare manier afgenomen worden. Afhankelijk van de grootte van de school (en het aantal leerlingen dat gelijktijdig de toetsen uitvoert) betreft het één of meerdere personen. Zij staan rechtstreeks in contact met het steunpunt,

evt. geflankeerd door de perso(o)n(en) die de technische ICT-ondersteuning op school verzorgen. Er zullen draaiboeken klaarliggen die de procedure van de toetsafname in detail beschrijft. De toetsassistenten worden bijgestaan door een helpdesk in de weken voorafgaand aan de afname en op de dag van afname van de kalibraties en de toetsen. De helpdesk beantwoordt zowel inhoudelijke vragen (Steunpunt), als ICT-gerelateerde vragen (in samenwerking met de IT-partner). De toetsassistenten zullen een eCursus doorlopen met een geautomatiseerde test, die controleert of de inhouden van de eCursus op voldoende wijze begrepen worden. Na het slagen op de test zal de opgeleide toetsassistent een deontologische code ondertekenen alvorens hij of zij deze rol actief op school kan opnemen. Er wordt tevens een coach aangesteld aan wie toetsassistenten vragen kunnen stellen, bv. via video-afspraak. Recurrente vragen worden opgenomen in een FAQ-webpagina en de draaiboeken voor toetsassistenten. Voor de draaiboeken hopen we mede inspiratie te putten uit bestaande materialen (bv. toelatingsproeven arts).

Samen met de IT-partner van het Steunpunt wordt erop toegezien dat de scholen voldoende geëquipeerd zijn om de toetsen digitaal te laten verlopen, zowel qua hardware, software, snelheid en internetbereik. In de voorbereiding van de afnames wordt de situatie in kaart gebracht zodat er – samen met de opdrachtgever van het Steunpunt – bekeken kan worden wat (niet) kan om de context optimaal voor te bereiden zodat een digitale toetsafname in elke school vlot kan lopen. Er wordt tevens een scenario uitgedacht voor scholen die vooraf of op het moment van afname technische problemen ondervinden. Er zullen diverse testmomenten georganiseerd worden met de scholen om de risico's op het moment van afname te beperken. De steekproefgewijze, digitale afnames van de kalibratietoetsen zullen op dat vlak ook belangrijke informatie en ervaringen opleveren.

Om scholen en leerlingen te motiveren om op een authentieke en eerlijke manier deel te nemen aan de toetsen, vinden we het belangrijk om hen voldoende te informeren over het nut van de toetsen en de informatie die het hen oplevert. Dit doen we via reguliere kanalen (bv. website, nieuwsbrief, …) en via meer hippe aanpakken (bv. social media, influencers, promo). Allicht ontdekken we in samenwerking van het team en in dialoog met de opdrachtgever en externe partners verschillende ideeën om motivatie bij deelnemers te stimuleren.

Het steunpunt is bereid de afname te organiseren in overleg met de IT-partner en zorgt er - samen met de medewerkers van het steunpunt - voor dat ook (vak)inhoudelijke en toetsgerelateerde ondersteuning aangesproken kan worden op deze momenten.

Verder is het steunpunt wachtende op de resultaten van de haalbaarheidsstudie om meer info te bekomen over de rol, functie en verwachte expertise van de toetsassistenten. Het steunpunt veronderstelt dat deze personen verbonden zijn aan de scholen als personeel en op die manier verloond worden voor de taak die zij opnemen.

Het steunpunt rekent op haar beurt op de technische expertise en ondersteuning van de ICT-partners (zie aparte call en begroting) voor de logistieke en operationele uitrol van de afnames. Een helpdesk zal operationeel zijn tijdens de periode vooraf aan en tijdens de afname in de scholen.

De kosten inzake bestaffing van tijdelijke medewerkers met IT-technische expertise werden niet begroot binnen het huidige steunpunt. Ook hier levert de haalbaarheidsstudie allicht belangrijke info op in functie van de verdere uitwerking van werkdomein E.

De organisatie en ondersteuning van de afnames wordt gedragen door een beperkt team van medewerkers om continuïteit te garanderen. Reeds bij de opstart van het steunpunt wordt een eerste medewerker aangesteld om diverse afnames te prospecteren (bv. soortgelijke toetsen in binnen- en buitenland, …) en alle noodzakelijke en ondersteunende info te verzamelen om de organisatie en ondersteuning voor te bereiden. De andere werknemers worden aangeworven op het moment dat de eerste kalibraties eraan komen en de grootschalige toetsafnames voorbereid moeten worden, zodat de professionalisering van toetsassistenten ontworpen, uitgetest en op grote schaal kan worden uitgerold. Tevens stelt het team alles in het werk gesteld wordt om draaiboeken, testafnames,

helpdesk, … kortom: de afname van kalibraties en toetsen zo optimaal mogelijk te laten verlopen. Het team van dit werkdomein heeft de nodige expertise en competenties inzake: toetsing, professionalisering van leraren, eLearning, coaching, project- en event-management.

De vaste werkingsmiddelen per personeelslid omvatten een schatting van de kosten die nodig zijn om de job uit te voeren in een kantooromgeving (bv. afschrijving van een laptop, aankoop software, …) en in het onderwijsveld (bv. verplaatsingen in binnenland, evt. naar buitenland) en de academische wereld (bv. deelname aan congressen over evaluatie, …). Projectgebonden middelen zijn voorzien om tijdens de periode van afnames het team tijdelijk te versterken met medewerkers/jobstudenten om extra inhoudelijke ondersteuning te bieden aan de toetsassistenten en de scholen die betrokken zijn. Zoals reeds aangegeven rekenen we hierbij eveneens op een (tijdelijke) versterking vanuit de technische partner waarmee het steunpunt samenwerkt, dewelke de logistieke en operationele uitrol van de afnames realiseren.

7. Werkdomein F. Psychometrie

Tabel 3.16 geeft een overzicht van de vier centrale werkpakketten binnen het werkdomein Psychometrie.

	2021			2022			2023			2024			2025		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
F1 IRT-analyses															
F2 Beperkt adaptief toetsen															
F3 Ondersteuning toetsontwikkeling															
F4 Automatisering analyses															

Tabel 3.16. Werkpakketten Psychometrie

Op de data die tijdens het toetsontwikkelingsproces, de kalibratiestudie en de eigenlijke afname worden verzameld, worden verschillende psychometrische analyses aan de hand van modellen uit de itemresponsetheorie (IRT) uitgevoerd. Deze analyses zijn noodzakelijk voor het ontwikkelen van de meetschaal (F1a), de cesuurbepaling en de koppeling met meerdere beheersingsniveaus (F1b), de berekening van het percentage leerlingen dat de eindtermen haalt (F1c) en het vaststellen van evoluties over de tijd (F1d). Voor de eerste twee taken (F1a en F1b) wordt gebruik gemaakt van de data uit het kalibratieonderzoek. Daarnaast worden de afnames ingeschakeld om continu de toetsitems op de meetschaal te verversen. De analyses tijdens het kalibratieonderzoek vormen ook de basis om adaptiviteit in het toetsdesign te integreren (zie F2). De laatste twee taken (F1c en F1d) gebeuren op basis van data van de eigenlijke afname(s).

7.1. (Door)ontwikkeling meetschaal (F1a)

De kalibratiestudie vindt plaats op het einde van het eerste volledige schooljaar waarin de eerste toetsontwikkeling voor een bepaalde toets en doelgroep plaatsvond. De gegevens uit de kalibratiestudie worden geanalyseerd aan de hand van modellen uit de itemresponsetheorie (IRT). Het doel van deze analyses is om de leerlingen en de opgaven op dezelfde schaal te brengen waarbinnen de vaardigheid van een leerling wordt gemodelleerd als de kans die een leerling met een specifiek vaardigheidsniveau heeft om een specifieke opgave correct op te lossen.

De hoofddoelstelling van de kalibratiestudie is de ontwikkeling van de meetschaal. Tijdens die ontwikkeling wordt tevens de psychometrische kwaliteit van de toetsopgaven en de kwaliteit van de toets geëvalueerd. Dit houdt in dat op beschrijvend niveau gekeken wordt naar het functioneren van de opgaven (voldoende spreiding in moeilijkheidsgraad; discriminatiegraad), nagegaan wordt of er geen sprake is van een plafond- of bodemeffect in de prestaties van de leerlingen (voldoende spreiding in de prestaties van de leerlingen). Daarnaast wordt ook nagegaan in welke mate de opgaven die verondersteld worden dezelfde onderliggende vaardigheid meten (en dus gegroepeerd zijn binnen één toets) dit ook effectief doen. Standaard gebeurt dit aan de hand van een 2PL-model (Two parameter logistic model). Ter verificatie zal op voorhand telkens nagegaan worden of andere IRT-modellen (e.g., Rasch, 3PL, GRM) meer of minder geschikt zijn. Om eventuele multidimensionaliteit van toetsresultaten na te gaan - en dus te controleren of het gepast is de resultaten van de toets op één meetschaal te plaatsen – worden twee soorten factoranalyses uitgevoerd. Ten eerste wordt een factoranalyse uitgevoerd op de tetrachorische correlatiematrix gebaseerd op de resultaten op itemniveau. Hiervoor wordt gebruik gemaakt de irt.fa-functie uit het R-pakket Psych (Revelle, 2017), waarbij voor het nagaan van de dimensionaliteit gebruik gemaakt wordt van ‘parallel analyse’ (Horn, 1965). Ten tweede wordt een nonlineaire factoranalyse op basis van het NOHARM-model (normal ogive harmonic analysis robust method, McDonald, 1997) uitgevoerd via het sirt-pakket in R. Indien

nodig zal ook de geschiktheid van multidimensionele IRT-modellen worden nagegaan (CDM-pakket; Robitzsch, Kiefer, George, Uenlue, 2017). De resulterende meetschaal vormt de basis voor de cesuurbepaling en voor de continue toetsverversing.

In tegenstelling tot het Vlaamse peilingsonderzoek gebeurt de eerste ontwikkeling van de meetschaal dus enkel op basis van de data uit het kalibratieonderzoek en worden de data van de eerste afname van de gecentraliseerde toetsen niet meegenomen. Hoewel de haalbaarheidsstudie hierover nog uitsluitsel moet bieden, zorgt de beperkte periode tussen de toetsafname en de rapportage er immers voor dat de cesuur en de beheersingsniveaus op voorhand gekend dienen te zijn. Het feit dat de data uit de eigenlijke toetsafname niet zullen worden meegenomen, impliceert dat de steekproef van het kalibratieonderzoek groter zal zijn dan de vereiste 550 leerlingen per item. Deze grotere steekproef (minimum 1650 leerlingen/schaal) leidt immers tot meer stabiele itemparameters. Deze stabiele parameters zijn nodig voor de ontwikkeling van een toets met een hoge betrouwbaarheid, wat een vereiste is gezien de meervoudige aard van de rapportage op leerling-, school- en systeemniveau.

Vanaf de tweede afname worden de afnames wel ingezet om de meetschaal verder te ontwikkelen door continu toetsverversing in te bouwen via een gericht afnamedesign bij een deelsteekproef van de leerlingen. Hoewel ook hier de haalbaarheidsstudie nog klaarheid moet scheppen, houden we er in deze aanvraag alvast rekening mee dat er vanaf de tweede afname van de gecentraliseerde toetsen bij een deelsteekproef van leerlingen een beperkt aantal nieuwe items worden toegevoegd aan de eigenlijke toets via een onvolledig kettingdesign. Omdat de afnames populatiebreed zijn, is het haalbaar om de leerlingen die deelnemen aan deze simultane kalibratiestudie elk slechts een beperkt aantal nieuw te kalibreren items aan te bieden, zodat de toetstijd niet sterk toeneemt, maar de meetschaal toch kan worden geüpdatet. Deze te kalibreren items zorgen dus voor een doorontwikkeling van de meetschaal. Daarnaast worden steeds ankeritems voorzien om het linken van de meetschaal bij opeenvolgende afnames te garanderen (cf. infra).

7.2. Cesuurbepaling en beheersingsniveaus (F1b)

Eens de meetschaal is ontwikkeld worden er ijkpunten op vastgelegd die de leerlingen categoriseren naargelang hun vaardigheidsniveau. Enerzijds zullen ijkpunten worden vastgelegd die aanduiden of leerlingen al dan niet de eindtermen halen (cesuurbepaling). Anderzijds zullen ijkpunten worden vastgelegd die toestaan om meerdere beheersingsniveaus van elkaar te onderscheiden.

Cesuurbepaling

De cesuur is dat punt op de meetschaal dat het minimumvaardigheidsniveau aanduidt dat nodig is voor leerlingen om de de eindtermen te halen. Leerlingen die onder de cesuur liggen halen de eindtermen (nog) niet. Leerlingen die op of boven de cesuur liggen halen de eindtermen wel. In de regel wordt per meetschaal één cesuur gelegd. De nieuwe onderwijsdoelen Nederlands en wiskunde voor de eerste graad van het secundair onderwijs worden echter gekenmerkt door een nieuwe structuur. Enerzijds zijn er eindtermen basisgeletterdheid die gemeenschappelijk zijn voor de leerlingen van de A-stroom en B-stroom. Anderzijds zijn er ook aparte eindtermen voor de A-stroom en B-stroom. Naar analogie (en wanneer van toepassing) met de peilingen Nederlands en wiskunde in de eerste graad van het secundair onderwijs in 2022, zullen op de meetschalen drie cesuren gelegd worden: een cesuur voor de eindtermen A-stroom, een cesuur voor de eindtermen B-stroom en een cesuur voor de basisgeletterdheid. Indien blijkt dat deze nieuwe structuur van de onderwijsdoelen ook wordt geïmplementeerd in de overige onderwijsniveaus, zal in het steunpunt worden bekeken in welke mate het ook daar mogelijk is om meerdere cesuren te leggen per meetschaal.

De cesuurbepaling gebeurt met een groep van ongeveer vijfentwintig experten. Deze groep bestaat uit vertegenwoordigers van het departement Onderwijs en vorming, de inspectie, de pedagogische begeleidingsdiensten, de lerarenopleiding, de leerkrachten en andere onderwijsdeskundigen. Zij

hebben allemaal relevante expertise voor het getoetste domein. Er wordt gestreefd dat elke stakeholder door minstens twee personen vertegenwoordigd is. Bij het bepalen van de cesuur kan op vraag van de opdrachtgever gewerkt worden met het differentieel wegen van de groepen los van het aantal aanwezige vertegenwoordigers, zoals dat sinds kort ook gebeurt bij de cesuurbepalingen binnen STEP.

Tijdens de cesuurbepaling wordt gevraagd om het minimumniveau te bepalen voor het halen van de eindtermen voor de ontwikkelde opgavenschalen. Hierbij wordt gewerkt met de zogenaamde 'Bookmark procedure' (Mitzel, Lewis, Patz, & Green, 2001), aangevuld met een visuele weergave van de opgavenschaal. Deze methode heeft uitvoerig bewezen makkelijk implementeerbaar te zijn, laat een combinatie van verschillende itemformats toe, alsook de mogelijkheid om meerdere cesuren op één meetschaal te plaatsen (Davis-Becker et al., 2011; Clauser et al., 2017; Lin, 2006). De toetsopgaven die worden voorgelegd aan de beoordelaars zijn een selectie uit de ontwikkelde meetschaal. Na een individuele beoordelingsronde, volgt een ronde met overleg in kleine deelgroepen en tot slot een ronde met de gehele groep.

Eens de cesuur werd bepaald, wordt deze ook gehanteerd in de daaropvolgende cycli. Op deze manier kunnen evoluties ten opzichte van het behalen van de eindtermen worden vastgesteld (cf. infra - trendanalyses). Indien de inhoud van de meetschaal gewijzigd is door bijvoorbeeld een nieuwe set van eindtermen toe te voegen aan een toets, wordt evenwel een nieuwe cesuur vastgelegd.

Beheersingsniveaus

De eindtermen worden theoretisch geconcipieerd als minimumdoelstellingen en kennen via de cesuurbepaling een empirische vertaling. Naast dit minimumniveau worden echter geen andere beheersingsniveaus gedefinieerd in de eindtermen. Een vertaling naar empirisch bepaalde beheersingsniveaus op basis van de verzamelde data lijkt dan ook een sterk arbitrair karakter te hebben. Daar staat tegenover dat door gebruik te maken van vaardigheidsniveaus op een meer gedetailleerde manier evoluties kunnen worden nagegaan in hoe leerlingen presteren bij opeenvolgende afnames van de gecentraliseerde toetsen. Hoe die beheersingsniveaus kunnen worden toegevoegd, zal in de haalbaarheidsstudie verder onderzocht worden. Hier worden alvast twee mogelijke implementatiepistes geschetst.

De eindtermen worden theoretisch geconcipieerd als minimumdoelstellingen en kennen via de cesuurbepaling een empirische vertaling. Naast dit minimumniveau worden echter geen andere beheersingsniveaus gedefinieerd in de eindtermen. Een vertaling naar empirisch bepaalde beheersingsniveaus op basis van de verzamelde data lijkt dan ook een sterk arbitrair karakter te hebben. Daar staat tegenover dat door gebruik te maken van vaardigheidsniveaus op een meer gedetailleerde manier evoluties kunnen worden nagegaan in hoe leerlingen presteren bij opeenvolgende afnames van de gecentraliseerde toetsen. Hoe die beheersingsniveaus kunnen worden toegevoegd, zal in de haalbaarheidsstudie verder onderzocht worden. Een mogelijke implementatiepiste zou erin kunnen bestaan om op basis van de verdeling van de vaardigheidsniveaus van de leerlingen tijdens de eerste afname (of zelfs tijdens de eerste kalibratie bij een representatieve steekproef van leerlingen) de meetschaal op te delen in opeenvolgende vaardigheidsniveaus, die nog om een inhoudelijke interpretatie vragen. Dit is de aanpak die momenteel in het Vlaamse peilingsonderzoek gehanteerd wordt. Daarbij is gebleken hoe moeilijk het is om te komen tot een consistente inhoudelijke interpretatie van de verschillende percentielpunten.

7.3. Behalen eindtermen (F1c)

Na elke toetsafname wordt op basis van de cesuur bepaald hoeveel procent van de Vlaamse leerlingen de eindtermen halen voor elk van de meetschalen van Nederlands en wiskunde. Daarbij wordt ook gerapporteerd over eventuele verschillen tussen leerlingengroepen.

Om na te gaan in welke mate de leerlingen de eindtermen beheersen worden de resultaten van de eigenlijke toetsafnames op de meetschaal van de kalibratiestudie geplaatst. Dit gebeurt door de data van de kalibratiestudie op te nemen in de analyse waarbij alle items van de kalibratiestudie als ankeritems kunnen worden beschouwd. Deze ankering is een essentiële stap en dient dan ook foutvrij te zijn. Aangezien meeteigenschappen doorheen de jaren veranderen is het noodzakelijk om na te gaan of de items zich “dezelfde manier gedragen” binnen de kalibratiesteekproef als binnen de leerlingenpopulatie van de centrale toetsen. Dit wordt onderzocht aan de hand van een Differential Item Functioning (DIF) analyse naar afnamejaar. Pas na deze controle wordt de cesuur via een lineaire transformatie overgezet op de data van centrale toetsen.

Op een gelijkaardige manier wordt in de resultaten ook een beschrijving gemaakt van hoe de leerlingen zich verhouden ten opzichte van de verschillende beheersingsniveaus.

7.4. Trendanalyses (F1d)

Op basis van een jaarlijkse herhalingsmeting kan per onderwijsniveau onderzocht worden in welke mate er trends zijn met betrekking tot het behalen van de eindtermen en de verdeling van de leerlingen over de verschillende beheersingsniveaus. Zoals hierboven al werd aangegeven wordt via IRT-analyses en het gebruik van ankeritems een koppeling gemaakt tussen de huidige toetsafname en de toetsafname van voorgaande cycli. Ook hier worden de ankeritems onderworpen aan een DIF-analyse naar afnamejaar. Items die DIF vertonen worden vervolgens toegelaten om vrij te variëren tussen de verschillende meetmomenten. De cesuur wordt via een lineaire transformatie overgezet tussen de verschillende meetmomenten.

7.5. Beperkt adaptief toetsen (F2)

Meerdere toetsdesigns lenen zich ertoe gestandaardiseerde instrumenten te ontwikkelen om leerwinst op grote schaal vast te stellen. De meeste van deze designs bevinden zich op een continuüm met als ene uiterste de strikt lineaire toetsen. Het andere uiterste wordt vertegenwoordigd door de (computergestuurde) adaptieve toetsen (Zenisky & Hambleton, 2014). Lineaire toetsen - meestal afgenomen op papier - zijn toetsen die zich niet aanpassen aan het kennis-, vaardigheids- of competentieniveau van de individuele leerling. Dit betekent dat alle leerlingen dezelfde toets of een parallelle vorm van een toets (parallel naar inhoud en moeilijkheidsgraad) oplossen. Elke leerling lost alle items op van de toets, ongeacht of deze heel moeilijk of heel makkelijk zijn (Hambleton & Xing, 2006). Adaptieve toetsen zijn toetsen die zich op itemniveau aanpassen aan het vaardigheidsniveau van de leerlingen. Aangezien de achterliggende algoritmes complexe berekeningen vereisen, worden adaptieve toetsen hoofdzakelijk via de computer afgenomen (CAT of computergestuurd adaptief toetsen).

De resultaten van het eerste perceel van de haalbaarheidsstudie zullen belangrijke input leveren over de mate waarin adaptief of semi-adaptief toetsen binnen de context van de Vlaamse gecentraliseerde toetsen mogelijk is. Het steunpunt zal die input gebruiken en zal daarbij zelf ook de inzet van multistage testing onderzoeken. Multistage toetsen (MST) liggen tussen lineaire en adaptieve toetsen op het bovenvermelde continuüm. Deze toetsen zijn ook adaptief maar in tegenstelling tot CAT gebeurt de aanpassing van de moeilijkheidsgraad van de toets niet na elk afgenomen item maar na de afname van een groep items (= module) (Yan, Lewis, & von Davier, 2014). Een belangrijk voordeel van MST ten

opzichte van CAT is dat de modules kunnen samengesteld worden voorafgaand aan de toetsafname. Dit zorgt ervoor dat de toetsontwikkelaars meer controle hebben over inhoud van de modules, de kwaliteit van de structuur van de toets en de afname. Er zal onderzocht worden in welke mate de structuur van de eindtermen, i.e., basisgeletterdheid, eindtermen en uitbreidingsdoelen in het modulair systeem van MST zou kunnen geïntegreerd worden. Daarnaast kunnen MST ontwikkeld worden op basis van een dataset die via een lineaire afname werd verzameld. Het steunpunt zal onderzoeken in welke mate de dataset van de kalibratiestudie hierbij als input zou kunnen dienen.

7.6. Ondersteuning toetsontwikkeling (F3)

Vanuit het psychometrisch team wordt ook voorzien in de ondersteuning van de verschillende teams van toetsontwikkelaars. Het is immers cruciaal dat de toetsontwikkeling afgestemd wordt op de mogelijkheden en beperkingen van het werken met meetschalen bij de gecentraliseerde toetsen. Deze ondersteuning omvat de volgende taken:

- ondersteuning van coördinatoren toetsontwikkeling m.b.t. toetsontwikkeling (mogelijke item formats, toetsmatrijs, aantal benodigde items, semi-adaptieve afname, …);
- mee bewaken van tijdslijn m.b.t. toetsontwikkeling in nauw overleg met de coördinatoren toetsontwikkeling;
- psychometrische opvolging van de toetsontwikkelingsprojecten bij pilootafnames;
- ondersteuning aanpak cesuurbepaling;
- opmaken van afnamedesigns bij kalibratiestudies en vertaling ervan naar benodigde items;
- terugkoppeling resultaten meetschaal na kalibratie;
- selectie van gekalibreerde items voor elke nieuwe afname (ankeritems);
- inhoudelijk aftoetsen van opzet van semi-adaptieve afnameprocedure met de coördinatoren toetsontwikkeling;
- bespreking van mogelijkheden voor toevoegen van leerplanspecifieke toetsopgaven of toetsen;
- bewaken van de afnametijd voor kalibratiestudies en toetsafnames; en
- vertalen van het proces van toetsverversing naar vereisten m.b.t. nieuw te ontwikkelen items.

7.7. Automatisering analyses (F4)

Gezien de zeer omvangrijke schaalgrootte van de gecentraliseerde toetsen, zeker zodra er afnames zijn voor verschillende leerlinggroepen, is het nodig om binnen het psychometrisch luik te streven naar een zo hoog mogelijke vorm van automatisering van de analyses. Deze taak wordt van bij de start van het steunpunt opgenomen door de verschillende psychometrici die binnen werkdomein F tewerkgesteld zullen worden. De automatisering van de analyses wordt gedurende de hele looptijd verder geoptimaliseerd.

8. Werkdomein G. Verwerking, analyse en rapportage resultaten

In de oproep werden acht onderzoeksvragen weergegeven met betrekking tot de analyse van de data uit de gecentraliseerde toetsen. Deze vragen werden door het aanvragende consortium geclusterd tot vier verschillende werkpakketten waarin deze vragen telkens door een andere onderzoeksgroep worden opgenomen. Zoals aangegeven in de oproep behoren de verwerking, analyse en rapportage van de extra items of aparte toetsen die de onderwijsverstrekkers kunnen toevoegen aan de afnames niet tot de opdracht van het steunpunt.

Vermits een concept als leerwinst verschillende invullingen kent (Janssens, Rekers-Mombarg & Lacor, 2014), dienen we vooraf kort aan te stippen wat we binnen het steunpunt hieronder verstaan. We conceptualiseren *leerwinst* als de groei over de tijd bij dezelfde leerling in het beheersingsniveau van een leerdomein. In het geval van de gestandaardiseerde toetsen gaat de aandacht naar de mate waarin een leerling een eindterm (wiskunde of Nederlands) in hogere mate beheerst. Dergelijke analyses veronderstellen minstens twee meetpunten bij dezelfde leerling. Als we echter uitspraken willen doen over de mate waarin 'vergelijkbare' leerlingen binnen de scholen hogere beheersingsniveaus bereiken op basis van cross-sectionele metingen, dan zullen we het hebben over de *toegevoegde waarde* van scholen (Vanhoof, De Maeyer, Van Petegem, Penninckx & Quintelier, 2016).

8.1. Leerwinst, Eindtermen en Trendanalyse (G1)

Onderzoeksvraag 1. In welke mate boeken de leerlingen en scholen in het Vlaams onderwijs leerwinst?

Bij de invoering van gestandaardiseerde toetsen binnen het Vlaamse onderwijs staat 'leerwinst' centraal. Het begrip 'leerwinst' en het aanverwante begrip 'toegevoegde waarde' worden echter door zowel beleidsmakers als wetenschappers op verschillende manieren geïnterpreteerd en geoperationaliseerd. Deze verschillende operationaliseringen van leerwinst zijn echter niet zomaar inwisselbaar, ze dienen vaak een ander doel, hebben elk een eigen interpretatie en hebben verschillende praktische vereisten. Als men dus een bepaalde vraag wil beantwoorden over de leerwinst van leerlingen of de mate waarin een school bijdraagt tot de leerwinst van leerlingen, dan is er een passende operationalisering nodig.

Leerwinst wordt beschouwd als de voortgang of toename in kennis, vaardigheden en competenties die een leerling doorheen een bepaalde periode van zijn schoolloopbaan maakt (Harris, 2011). Het is een maat voor de groei van individuele leerlingen en situeert zich in de eerst plaats op het microniveau, het niveau van de individuele leerling (Penninckx & Quintelier, 2016). Leerwinst kan natuurlijk ook op niveau van de school en het niveau van het onderwijssysteem beschreven worden (Castellano & Ho, 2015). Zo 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. Leerwinst van een leerling wordt in de regel bepaald op basis van een vergelijking van minstens twee toetsprestaties van een leerlingen op twee verschillende meetmomenten. Een positief verschil wordt daarbij gezien als leerwinst (Gong, Perie, & Dunn, 2006).

Toegevoegde waarde is daarentegen het verschil tussen de leerwinst van een school en de leerwinst die van deze school verwacht werd. De verwachte leerwinst die een school verwacht is daarbij gebaseerd op de kenmerken van deze leerlingen in de school (Everson, 2017; Koedel, Mihaly, & Rockoff, 2015; Levy, Brunner, Keller, & Fischbach, 2019). Het vergelijken van de leerwinst van scholen

is namelijk onvoldoende om tot een faire vergelijking te komen tussen scholen. Immers, de leerwinst die leerlingen maken wordt immers niet uitsluitend bepaald door de scholen, maar ook door de persoonlijke kenmerken van leerlingen (Leckie & Goldstein, 2019; Levy et al., 2019). Aangezien scholen geen vat hebben op deze leerlingkenmerken, zijn ze ook niet verantwoordelijk voor de effecten die deze leerlingkenmerken mogelijk hebben op de leerwinst. Bijvoorbeeld, leerlingen met lage startprestaties zullen mogelijk minder leerwinst maken dan leerlingen met hoge startprestaties. Ook de thuistaal of SES kunnen mogelijk een invloed hebben (Timmermans, Doolaard, & de Wolf, 2011). Een school waarin leerlingen meer leerwinst maken dan verwacht heeft een positieve toegevoegde waarde. Een school waarin leerlingen daarentegen minder leerwinst maken dan verwacht heeft een negatieve toegevoegde waarde.

Hoewel het conceptuele verschil tussen leerwinst en toegevoegde waarde relatief eenvoudig is, is de operationalisering van beiden minder eenvoudig. Er bestaan dan ook verschillende modellen die leerwinst en toegevoegde waarde beschrijven (Castellano & Ho, 2013; Everson, 2017; Koedel et al., 2015; Levy et al., 2019; Timmermans et al., 2011). Deze modellen zijn het restwinstmodel (residual gain model) voor één meetmoment, het restwinstmodel (residual gain model) voor twee meetmomenten, het leerwinstmodel (gain score model) en het rest-leerwinstmodel (residual gain score model). De verschillen tussen deze modellen bestaan uit: (1) het al dan niet nodig hebben van twee toetsprestaties van leerlingen op twee verschillende meetmomenten, (2) hoe voor de initiële toetsprestaties gecontroleerd wordt, (3) of beide toetsprestaties op dezelfde meetschaal staan (e.g. noodzaak aan een gemeenschappelijke verticale IRT-schaal), (4) of het een model is waarbij de uitkomst vergeleken wordt met een vooraf bepaalde verwachting om de toegevoegde waarde te berekenen.

De eerste longitudinale toetsdata zijn pas beschikbaar in 2026, na de looptijd van dit steunpunt. De vraag in welke mate leerlingen en scholen in het Vlaamse onderwijs leerwinst boeken kan in strikte zin dus niet binnen de eerste termijn van dit steunpunt beantwoord worden. Wel zal gekeken worden naar niet-longitudinale modellen om vooral dan de vraag naar de toegevoegde waarde van scholen te beantwoorden. Het steunpunt zal daarnaast ook onderzoek verrichten om deze leerwinstanalyses zo goed mogelijk voor te bereiden. De resultaten van het eerste perceel van de haalbaarheidsstudie zullen daarbij mee richtinggevend zijn. Pas wanneer duidelijk is hoe leerwinst en toegevoegde waarde worden gedefinieerd kan de juiste analysetechniek worden ontwikkeld. Vragen die daarbij beantwoord moeten worden zijn in welke mate een restwinstmodel dan eerder een leerwinstmodel aangewezen is; welk type effect (AA, A, B, X; zie OECD, 2008; Timmermans et al., 2011) best wordt gerapporteerd om een zo fair mogelijk vergelijking tussen scholen te realiseren; en hoe leerlingenmobiliteit statistisch in rekening kan worden gebracht. Daarnaast is het ook belangrijk om het nastreven van een meting van de leerwinst in strikte zin mee te nemen bij de keuzes van de te toetsen eindtermen bij de inhoudelijke vormgeving van de toetsen op de voorziene meetmomenten. Het spreekt ten slotte voor zich dat de vormgeving en ontsluiting van de schoolfeedback in goede onderlinge afstemming zal gebeuren tussen de werkdomeinen F, G en H.

Onderzoeksvraag 2. In welke mate beheersen leerlingen in het Vlaams onderwijs de eindtermen? Hoe verdelen leerlingen zich over verschillende beheersingsniveaus?

Op basis van de cesuur (procedure cesuurbepaling: zie werkdomein F1b) wordt bepaald hoeveel procent van de Vlaamse leerlingen de eindtermen halen voor elk van de meetschalen voor wiskunde en Nederlands op het einde van het lager onderwijs en de eerste en derde graad van het secundair onderwijs. Indien de nieuwe eindtermen in het lager onderwijs een eigen finaliteit krijgen op het einde van de tweede graad kan dezelfde procedure gevolgd worden voor de meting in het vierde leerjaar. Indien de eindtermen enkel verwijzen naar het minimumniveau op het einde van het lager onderwijs, zullen de resultaten van de leerlingen van het vierde leerjaar via ankeritems op de meetschaal van de leerlingen van het zesde leerjaar worden geplaatst.

Het in kaart brengen van de verdeling van de leerlingen over verschillende beheersings- of vaardigheidsniveaus kan op verschillende manieren, zoals hierboven bij F1b al werd toegelicht. Er zal onderzocht worden (mede op basis van de resultaten van de haalbaarheidsstudie) in welke mate de ontwikkeling van empirisch bepaalde beheersingsniveaus nog mogelijk en wenselijk is binnen een systeem dat reeds wordt aangestuurd vanuit een theoretisch vastgelegde cesuur via het systeem van eindtermen.

Binnen het systeem van cesuurbepaling wordt aan experten gevraagd om het minimale prestatieniveau dat de eindtermen als doel hebben te vertalen naar de opgaven op de meetschaal. Het gaat dus om een locatie van een interpretatie: het cesuurpunt wordt vastgelegd in functie van de betekenis van de eindtermen. In andere metingen zoals TIMSS en PISA worden meerdere beheersingsniveaus op de meetschaal onderscheiden die worden bepaald op basis van percentielen of psychometrische principes. Nadien wordt dan een inhoudelijke interpretatie gegeven aan deze verschillende sectoren op de meetschaal. Hier gaat het dus om een interpretatie van een locatie.

In ieder geval zal er een analyse gemaakt worden van hoe de leerlingen uit verschillende percentielgroepen (ingedeeld op basis van de percentielen 10, 25, 50, 75 en 90) presteren op de meetschaal. Daarbij wordt ook verwezen naar hun verschillen in succeskansen op de voorbeeldopgaven.

Onderzoeksvraag 8. In welke mate is er sprake van een voor- of achteruitgang over de tijd met betrekking tot het beheersen van de eindtermen met betrekking tot een bepaald vak of leergebied?

De opeenvolgende afnames van de gecentraliseerde toetsen in eenzelfde leerjaar bij verschillende cohortes van leerlingen vormen herhalingsmetingen op basis waarvan kan onderzocht worden in welke mate er trends zijn met betrekking tot het behalen van de eindtermen. Dit is een belangrijke kwaliteitsmonitor voor het Vlaamse onderwijs over de jaren heen. Deze herhalingsmetingen hebben ook een belangrijke rol in het aanzetten tot acties ondernomen door verschillende onderwijsactoren omdat mogelijke verschillen in prestatieniveau tussen twee meerdere meetmomenten ook deels het effect van die acties kunnen weerspiegelen.

Voor deze trendanalyses worden over afnamecycli heen ankeritems opgenomen in de gecentraliseerde toetsen. Deze items geven een descriptief beeld over dalingen of stijgingen in de leerlingprestaties, maar zijn vooral cruciaal in het koppelen van de meetschalen over afnamemomenten heen. Daarbij worden bijkomende statistische analyses en controles uitgevoerd met betrekking tot de stabiliteit van de meetschaal over jaren heen en het opsporen van items waarvoor die meetinvariantie niet geldt (het zogenaamde ‘differential item functioning’ of kortweg DIF over afnamemomenten). Door gebruik te maken van multilevel logistische regressiemodellen wordt de multiniveaustructuur van de gegevens mee in rekening genomen.

Naast de algemene niveaustijgingen of –dalingen in het Vlaamse onderwijs kan via een herhalingsmeting informatie worden aangereikt over de mate waarin er veranderingen zijn in de distributie van de leerlingen op de meetschaal. Daarbij zal opnieuw worden gewerkt met dezelfde percentielgroepen als bij de analyse van de eindtermbeheersing (onderzoeksvraag 2) alsook eventueel met de alternatieve vorm van beheersingsniveaus zoals bijvoorbeeld overgenomen uit internationaal vergelijkende studies. Deze analyses bieden inzicht in de mate waarin het aandeel zwakker of sterker presterende leerlingen evolueert doorheen de tijd.

Tot slot zal ook worden nagegaan of een eventuele kloof tussen leerlinggroepen (bijvoorbeeld voor SES, taal, of geslacht; zie onderzoeksvraag 3) al dan niet gewijzigd is doorheen de jaren. Hiervoor wordt gewerkt met een interactie-effect tussen de focusvariabele (bv. geslacht) en het afnamejaar (naast de hoofdeffecten voor beide variabelen uit het interactie-effect). Het al dan niet significant zijn van het interactie-effect geeft informatie over de evolutie van de kloof doorheen de jaren.

8.2. Verschillen tussen leerlingen en scholen (G2)

8.2.1. Algemene aandachtspunten

In dit werkpakket combineren we onderzoeksvragen 3 en 4 over verschillen tussen leerlingen en scholen. Deze onderzoeksvragen kunnen immers *analytisch* wel onderscheiden worden, maar zijn *inhoudelij*k en *empirisch* duidelijk met elkaar verbonden. Alvorens in te gaan op de specifieke onderzoeksvragen bespreken we een paar algemene uitgangs- en aandachtspunten die bij het beantwoorden van elke onderzoeksvraag in dit werkpakket worden meegenomen.

Het huidige steunpunt wil ondersteuning bieden aan de onderwijsactoren in het bewaken en versterken van de onderwijskwaliteit. Het richt zich daarom primair op de 'effecten' van scholen en leraren. Daarbij beogen we enerzijds met een aantal methodologische vernieuwingen deze unieke data optimaal te benutten om zo tot meer beleidsrelevante inzichten te komen, en anderzijds om gevonden effecten beter te kunnen verklaren door deze sterk theoretisch in te bedden.

De gecentraliseerde toetsen (GT) zullen het in Vlaanderen voor het eerst mogelijk maken om op populatieniveau naar verschuivingen in leerprestaties binnen leerlingen te kijken. Meerdere elementen zijn hierbij van belang. Ten eerste, gaat het om toetsen die niet alleen de (continue) kennis- en vaardigheidsniveaus (verder aangeduid als competenties) van leerlingen in kaart brengen, maar deze tevens relateert aan het al dan niet (categorisch onderscheid) behalen van de eindtermen (cf. 7.1). Ten tweede, is het onderzoeksdesign dynamisch – er wordt gekeken naar verschuivingen *binnen* individuen en scholen. Op dat punt sluiten de centrale toetsen aan bij het bestaande longitudinaal onderzoek (LISO, SIBO, enz.). De GT verbeteren, ten derde, ook deze bestaande longitudinale onderzoeken door de volledige populatie van leerlingen en scholen in ogenschouw te nemen. Dat lost onder andere het probleem op waarbij leerlingen die veranderen van school uit het blikveld verdwijnen (een probleem waar de LISO-studie bijvoorbeeld sterk mee kampte). Alleen leerlingen die stoppen met schoolgaan of die naar het buitenland verhuizen verdwijnen uit beeld. Alle scholen blijven per definitie in beeld in de GT-data.

Het potentieel van de GT-data kan pas ten volle gerealiseerd worden als deze gegevens niet alleen gebruikt worden voor monitoring maar ook voor verklarend onderzoek. Dat vereist het in kaart brengen en empirisch toetsen van deze verklaringen.

8.3. Methodologische meerwaarde

Op basis van het voorgaande kunnen drie centrale aandachtspunten gedestilleerd worden die we meenemen in de verdere uitwerken. (1) Zo roept het onderscheid tussen beheersingsniveaus als continue uitkomst en het al dan niet behalen van de eindtermen (categorische uitkomst: J/N) de vraag op of de determinanten van soorten uitkomsten dezelfde zijn (Haile & Nguyen, 2008). (2) Aandacht voor leerwinst impliceert ook dat we op het vlak van de gebruikte voorspellers, aandacht dienen te hebben voor dynamiek. (3) Een grootschalige studie naar leerwinst en schooleffecten dient ten slotte ook niet alleen te kijken naar verschillen in gemiddelden maar ook naar de spreiding binnen scholen/studierichtingen rond het gemiddelde. Op die manier mikken we op methodologische vernieuwingen, gekoppeld aan beleidsrelevantie.

(1) De rol van de positie op de verdeling van competenties

Via het systeem van eindtermen heeft de overheid instrumenten in handen om minimale verwachtingen te formuleren omtrent de te bereiken doelen. Deze eindtermen beogen garanties te bieden op kwalitatief onderwijs die alle onderwijsverstrekkers dienen te borgen en hebben als doel sleutelcompetenties - kennis, inzicht, vaardigheden en attitudes - van de leerlingen in Vlaanderen te versterken. Conceptueel onderscheidt het continuüm van prestatieniveaus (de toetsscores) zich van

de eindtermen, maar via de cesuurbepaling in de toetsscores, worden deze met elkaar in verband gebracht. Vanuit verklarend perspectief stelt zich de vraag of de effecten van voorspellende factoren anders zijn binnen de verschillende categorieën die de cesuurpunten creëren op de vaardigheidsschaal. Daarom dat we in de analyses de afhankelijkheid van de bevindingen van de positie van leerlingen (Onderzoeksvraag 3) en scholen (Onderzoeksvraag 4) op de verdeling van het spectrum van de prestatieniveaus nagaan.

Men kan inderdaad niet a priori aannemen dat de determinanten van de variatie in een uitkomst voor elk deel segment van de verdeling van die uitkomst dezelfde zijn (Eide & Showalter, 1998). Anders gezegd: de kenmerken die verschillen verklaren tussen zwakke en zeer zwakke leerlingen/scholen, zijn niet noodzakelijk dezelfde als degene die het verschil tussen sterke en zeer sterke leerlingen/scholen verklaren. De enorme statistische power van de GT-data in combinatie met het uitgebreid arsenaal aan statistische methoden laat toe onderwijsuitkomsten op een veel fijnmaziger niveau te bestuderen dan tot hiertoe het geval was in Vlaanderen.

(2) Dynamiek van schoolkenmerken

Omdat de meting van leerwinst - de vooruitgang die leerlingen maken over de tijd - centraal staat binnen dit steunpunt, hebben we ook aandacht voor verandering in de scholenkenmerken over de tijd. Niet alleen zijn er aanwijzingen dat sociale segregatie in termen van de achtergrondkenmerken van leerlingen, tussen scholen toeneemt in (sommige regio's in) Vlaanderen (Havermans, Wouters & Groenez, 2018). Een bezorgdheid rond de invoering van GT, luidt precies dat dergelijke toetsen deze segregatie tussen scholen verder zou kunnen versterken. Daarom is het essentieel samen met de onderwijsuitkomsten ook de samenstelling (en de verandering daarin) van het leerlingen- en lerarenpubliek van scholen te monitoren (zie o.m. Xie & Zhang, 2020). Dergelijke monitoring resulteert ook in de constructie van dynamische indicatoren die gerelateerd kunnen worden aan de verandering in toetsprestaties binnen scholen. Dit wordt met name behandeld in het kader van sectie 'Schooleffecten' (7.5).

(3) Schoolverschillen, in gemiddelde én in spreiding

In de schooleffectiviteitsliteratuur leeft steeds meer het idee dat om de effectiviteit van scholen te beoordelen niet alleen het gemiddeld prestatieniveau of gemiddelde leerwinst bestudeerd dient te worden, maar ook de omvang van de spreiding. Twee scholen of studierichtingen kunnen eenzelfde gemiddeld prestatieniveau of leerwinst realiseren maar een heel verschillende spreiding hebben rond dit gemiddelde. Als beleid wenst men dan te weten (1) waar dergelijke verschillen vandaan komen, (2) wat die verschillen impliceren voor de klaspraktijk, en (3) hoe men er op kan ingrijpen. Verschillen tussen scholen in variantie rond gemiddelden kunnen inhoudelijk-theoretisch bovendien rechtstreeks gelinkt worden aan de eerder vermelde segregatiedynamieken (bv negatieve selectie zou kunnen leiden tot meer homogene én betere onderwijsprestaties van leerlingen in een school). Het simultaan bestuderen van gemiddelde en variatie op schoolniveau wordt voorlopig nog maar weinig gedaan onder meer omdat het een andere vorm van statistische modellering vergt en vooral heel veel statistische power vereist (Lester, Cullen-Lester & Walters, 2019). De grootschaligheid van de geplande dataverzameling, biedt op dat vlak unieke opportuniteiten. Dit wordt behandeld in het kader van de paragraaf over schoolverschillen (7.4.1).

Theoretische meerwaarde

Naast de bovenvermelde methodologische vernieuwingen die dit werkpakket vooropstelt, sluit dit werkpakket tevens aan bij een tendens binnen het effectiviteitonderzoek die gericht is op een sterke integratie tussen theorie en empirisch onderzoek. Schooleffectiviteitsonderzoek is nog steeds heel empiristisch en theorie-arm. Dat blijkt ook uit een vroege kritiek: "*Without a theory or set of*

hypotheses about the nature of the processes underlying the statistical models they use, there is no possibility of drawing conclusions which purport to show that the relationship in the models is causal in nature. The lack of an overall theoretical framework led the authors to embark on a fishing expedition in which the size of the holes in the net was determined by the irrelevant criterion of statistical significance" (Cuttace, 1982: 487). Deze bemerking werd tevens drie decennia later herhaald door Scheerens (2013) in een overzichtsartikel van 109 schooleffectiviteitsonderzoeken (zie ook Scheerens 2015). Het probleem dat Cuttace en Scheerens aanhalen, is dat verschuivingen (i.e., leerwinst) en verschillen in studieprestaties maar interessant zijn als ze systematisch zijn, dat wil zeggen dat ze kunnen worden toegeschreven aan en verklaard door gemeten factoren. Alleen afgaan op de statistische significantie wordt inderdaad steeds minder aanvaard als sterk bewijs dat verschillen of verschuivingen ook betekenisvol zijn. Sterk bewijs voor betekenisvolle verschillen/verschuivingen ontstaat wanneer die verschillen/verschuivingen (1) verwacht werden op theoretische gronden en (2) empirisch kunnen worden verklaard door geobserveerde mediatoren (i.e., de kenmerken waarlangs het causaal effect zich voltrekt). Het is immers ook dat mediatieproces waarop een beleid zich noodzakelijkerwijs zal moeten enten.

Tegen de achtergrond van die algemene aandachtspunten gaan we hieronder dieper in op de specifieke deelonderzoeksvragen.

8.4. Verschillen tussen leerlingen of het belang van achtergrond

Onderzoeksvraag 3. Welke leerlingen(groepen) hebben het moeilijker om de eindtermen te behalen?

(1) Achtergrondkenmerken

Zowel internationaal onderzoek naar onderwijsongelijkheid, als nationaal onderzoek tonen dat naast cognitieve verschillen sociale, demografische en economische verschillen een impact hebben op het verwerven van competenties en op de uiteindelijke leerprestaties van jongeren. Daarbij blijkt wel dat de relevantie van kenmerken varieert in functie van de specificiteit van de bestudeerde uitkomst. Zo blijken meisjes, bijvoorbeeld, gemiddeld beter te scoren op taalgerelateerde vaardigheden (bv. leesvaardigheid, Logan & Johnston 2010), maar is het genderverschil op het vlak van wiskundeprestaties minder groot en niet in alle landen hetzelfde (Salvi del Pero & Bytchkova 2013). Andere studies vonden onder meer systematische verschillen in onderwijsprestaties op basis van etnisch-culturele achtergrond (Obgu 1987; Marx, Ko & Friedman 2009), de sociaaleconomische status van leerlingen (Danhier & Jacobs, 2017; OECD, 2016a; Sirin, 2005), de migratiestatus (Danhier & Jacobs, 2017, Jacobs, Rea & Hanquinet, 2007), en de thuistaal (Bellens et al., 2013; Vandenbroeck et al., 2016). De sociaal-economische en etnisch-culturele achtergrond van leerlingen beïnvloedt bovendien niet alleen leerprestaties op individueel niveau, maar ook op klas- en schoolniveau (bv. Danhier & Martin 2014; Van Ewijk & Sleegers 2010a,b). Het monitoren en bestuderen van de opgesomde verschillen is belangrijk, niet alleen omdat ze relevant zijn vanuit een sociaal rechtvaardigheidsperspectief, maar ook omdat ze verband houden met gemakkelijk toegankelijke indicatoren voor scholen en overheden en dus beleidsmatig interessant zijn (zie Kavadias, 2008).

Het bestuderen van deze verschillen behelst twee zaken. Ten eerste, het blijvend monitoren van verschillen op leerling en schoolniveau naar de indicatoren die gebruikt worden voor financieringsdoeleinden (i.e., de onderwijs kansarmoede-indicatoren). Ten tweede, onderzoek doen naar de relevantie van die indicatoren door hun verklaringskracht te vergelijken met andere indicatoren. De GT bieden inderdaad ook de mogelijkheid om bijkomende informatie te verzamelen over jongeren én scholen die vandaag nog niet aanwezig is in de administratieve databanken. Langs die weg kan onderzoek uitgevoerd worden die de performantie van de huidige onderwijs kansarmoede-indicatoren monitort en aanscherpt.

(2) Relevantie van distributie

De koppeling van de eindtermen aan het beheersingsniveau van wiskunde en Nederlands roept interessante vragen op, in bijzonder of een analyse van welke leerlingen en leerlingengroepen de eindtermen behalen (categorische onderscheid) dezelfde of andere factoren aan het licht brengt dan een analyse van het beheersingsniveau van de getoetste kennis en vaardigheden (continu kenmerk). Of nog: is er variatie in de verklaringen voor verschillen in onderwijsuitkomsten naargelang het specifieke deel van de distributie dat men bestudeert? Het is inderdaad plausibel dat de verklaringen voor het verschil tussen zwakke en zeer zwakke leerlingen anders zijn dan deze voor het verschil tussen sterke en zeer sterke leerlingen (OECD, 2016b). Een beleid dat vaststelt dat het aandeel toppresteerders achteruitgaat, heeft nood aan verdiepend onderzoek dat zich specifiek richt op de groep van goede presteerders. Dezelfde aanpak kan tevens gebruikt worden om andere determinanten van ongelijke prestaties in kaart te brengen (Constanzo & Desmoni, 2017). Een school met een specifiek instroomprofiel heeft nood aan onderzoek dat zo goed mogelijk aansluit bij hun leerlingenpubliek. Er zijn met andere woorden genoeg beleidsmatige redenen om op een veel fijnmazigere wijze met behulp van multilevel (Snijders & Bosker, 2011) en quantile regressietechnieken (Konstantopoulos, et al., 2019) naar verklaringen in prestatieverschillen te kijken dan wat mogelijk was met data op basis van steekproefgegevens. Het is dat toegevoegd potentieel van de GT dat deze onderzoekslijn beoogt te realiseren.

8.4.1. Schoolverschillen

Onderzoeksvraag 4. In welke mate bestaan er verschillen tussen scholen in het Vlaams onderwijs in de mate dat ze leerwinst realiseren en in de mate van het beheersen van de eindtermen door hun leerlingen?

Voor het bestuderen van schoolverschillen volgen we een drieledig pad: (1) we focussen op relatieve in plaats van absolute effecten, houden daarbij rekening met mogelijke verschillen in verklaringen naargelang de positie in het ‘competentiespectrum’, (2) brengen naast gemiddelde verschillen tussen scholen tevens verschillen in spreiding in kaart, en doen dat alles vanuit (3) de vraag naar verklaring voor verschillen. We nemen daarbij aan dat verklaring overtuigender wordt indien we tevens een sterke theoretische basis kunnen bieden voor de gevonden effecten.

(1) Relatieve effecten – vergelijken en focus op verschil

Bij het bestuderen van onderwijseffecten (zowel in termen van toegevoegde waarde als leerwinst) dient een onderscheid gemaakt te worden tussen de ‘absolute’ en ‘relatieve’ effecten van onderwijs. Bijna alle hedendaags onderwijseffectiviteitsonderzoek naar studieprestaties bestudeert de relatieve veeleer dan de absolute impact van onderwijs (Goldstein, 1997). Dat is een rechtstreeks gevolg van het feit dat dit type onderzoek zich primair richt op onderwijs verschaft in scholen. De absolute impact van onderwijs is daarbij alleen meetbaar indien we jongeren die schoollopen vergelijken met gelijkaardige jongeren die dat niet doen. Dat is uiteraard niet mogelijk. Wat we hier bestuderen zijn de relatieve effecten van onderwijs. Het gaat over de vraag welke scholen of klassen effectiever zijn dan anderen (Scheerens, 2015). Het is uiteraard mogelijk dat de effectiviteitverschillen klein of onbestaande zijn, terwijl onderwijs globaal genomen wel een zeer groot effect heeft. We documenteren niet alleen de relatieve verschillen. We onderzoeken vooral ook de mate waarin verschillen tussen scholen systematisch blijken samen te gaan met kenmerken van de onderwijscontext zoals de leerlingenpopulatie, het onderwijsaanbod, of andere contextuele kenmerken (zoals bv. de stedelijke context, Eizaguirre, 2019).

(2) Relevantie van de distributie

Net zoals voor verschillen tussen leerlingen, is het ook relevant na te gaan of verschillen tussen scholen dezelfde zijn naargelang we ze beoordelen op basis van de continue kennis- en vaardigheidsscores dan wel wanneer we deze indelen in categorieën, en meer algemeen of schoolkenmerken dezelfde verklaringskracht hebben over het gehele continuüm van de afhankelijke variabelen.

Bijkomend kunnen we scholen groeperen naargelang hun gemiddelde score op het continuüm en nagaan of het effect van leerlingenkenmerken hetzelfde is in de scholen met hogere en met lagere scores. Indien dit blijkt, biedt deze analyse eveneens een aanzet tot de meer causale uitdieping van schooleffecten.

(3) Simultane modellering van gemiddelde en spreiding

Zoals eerder verduidelijkt focussen schooleffectiviteitsonderzoekers steeds meer op zowel de gemiddelde prestatieniveaus als op de variatie binnen scholen op dergelijke gemiddelden. Op die manier kan onderzocht worden of scholen niet alleen gemiddeld goede prestaties leveren, maar ook of ze dat voor al hun leerlingen doen. Dat betekent ook dat een bepaald klas/school/studierichting kenmerk zowel een effect kan hebben op het gemiddelde als op de variantie rond dat gemiddelde (bv. Raudenbush & Bryk 1987; Kim & Choi 2008; Culpepper 2010; Leckie et al. 2014). Zogenaamde heteroscedasticiteitsmodellen laten toe om verschillen in residuele variantie te modelleren. In volle ontwikkeling zijn heteroscedasticiteitsmodellen die een dubbele voorspellingsfunctie hanteren, of 'joint modelling of mean and dispersion', met statistische modellen die toelaten om heterogene varianties te relateren aan verklarende variabelen op schoolniveau en zelfs aan over scholen variërende (random) effecten van verklarende variabelen op leerlingniveau (Double Hierarchical Generalized Linear Models – Lee & Nelder 2006; Mixed Effects Location Scale modellen – Lester, Cullen-Lester & Walters, 2019). Dergelijke modellen zijn in staat een eventuele veranderende leerlingensamenstelling tussen twee meetmomenten in een school in rekening te brengen en kunnen in die zin rechtstreeks gelinkt worden aan de eerder besproken segregatiedynamieken. Het is door dit alles simultaan te bestuderen dat én de potentiële meerwaarde van GT benut kan worden én potentiële perverse effecten gedetecteerd, gemonitord en geremedieerd kunnen worden.

Aandacht voor voorspellers in spreiding op schoolniveau in studieprestaties en leerwinst is tot op heden in Vlaanderen nagenoeg afwezig. Dat komt ten dele omdat dergelijke modellen veel statistische power vragen, dewelke panelstudies op basis van steekproeven van scholen (bv. LISO) niet altijd kunnen bieden. De data van de GT, precies omdat het om populatiedata gaat, bieden die wel. Kortom, aandacht voor de zogenaamde 'joint modelling of mean and dispersion' brengt een methodologische vernieuwing in het Vlaams onderwijsonderzoek met directe, en substantiële beleidsmatige relevantie.

8.5. Schooleffecten of de zoektocht naar verklaringen (G3)

Onderzoeksvraag 5. In welke mate blijven de verschillen tussen (groepen van) leerlingen en scholen overeind nadat er is rekening gehouden met de achtergrondkenmerken van leerlingen en de context van de scholen?

en

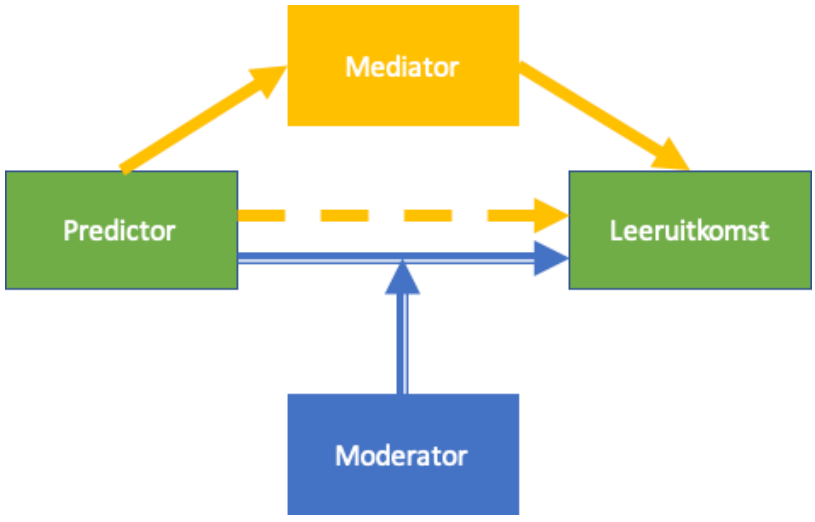
Onderzoeksvraag 6. Hoe kunnen we de verschillen tussen (groepen) van leerlingen en scholen verklaren?

Niet alle statistisch gevonden verschillen tussen scholen zijn noodzakelijk ook 'schooleffecten'. Om dit te kunnen uitmaken is het belangrijk op zoek te gaan naar theoretische verklaringen. Een belangrijke

leidraad in het hedendaagse effectiviteitsonderzoek wordt gevormd door het CIPO-model. Effectiviteit wordt gezien als de mate waarin een school de gewenste output produceert. Deze output komt altijd tot stand binnen een zekere Context (op macroniveau), is de resultante van een welbepaalde Input die vervolgens op school- (meso) en klasniveau (micro) wordt verwerkt in een Proces (Creemers & Kyriakides, 2015). Standaardpraktijk in effectiviteitsonderzoek beschouwt 'schooleffecten' als de schoolverschillen die overblijven nadat rekening is gehouden met achtergrondkenmerken van leerlingen en de context van scholen in een multilevel setting. Zo'n aanpak biedt niet altijd aanknopingspunten voor beleid; beleidsmakers dienen immers te weten *waarop* men *om welke reden* dient in te grijpen. De grote uitdaging met betrekking tot het identificeren van 'schooleffecten' is dan ook dat wat er binnen scholen gebeurt in termen van leeractiviteiten nooit honderd procent kan afgezonderd worden van wat er buiten de school gebeurt. Dit impliceert de facto een samenhang tussen de onderzoeksvragen 3 en 4, alsook tussen de onderzoeksvragen 5 en 6. Een vergelijkbaar probleem uit zich in het huidig onderzoeksveld heel expliciet wanneer het gaat om de rol van schoolcompositie-effecten. Immers, de etnische en sociale compositie van scholen blijkt vaak ook effecten te hebben onafhankelijk van de (statistische effecten van) leerlingkenmerken (Van Ewijk & Sleegers 2010a,b). De vraag is dan in welke mate de sociale compositie beschouwd kan worden als een contextkenmerk dan wel eerder wijst op procesverschillen waar het beleid van scholen vat op kan hebben.

8.5.1. De focus op mediatoren

Meer dan vroeger is het effectiviteitonderzoek geneigd alleen die verschillen te beschouwen als effectiviteitverschillen die ook kunnen worden toegewezen aan die proceskenmerken (bv. Liu et al. 2015; Timmermans & Thomas 2015; Agirdag 2018; Wenger, Gärtner en Brunner 2020; zie ook Berkowitz et al 2017). Een effectverklaring wordt inderdaad overtuigender indien men niet alleen kan tonen dat er een verschuiving is in de leerresultaten (leerwinst/verlies) maar ook dat die verschuiving door veranderingen in de mediator kan worden verklaard. Naast mediatoren bestaan ook moderatoren; dat zijn kenmerken die ervoor zorgen dat een bepaald (leer)proces anders verloopt voor bepaalde types leerlingen (Wu & Zumbo, 2007). De moderatoren kunnen verwijzen naar socio-demografische kenmerken (SES, gender, migratieachtergrond) maar evenzeer naar schoolkenmerken (samenstelling, wijze van onderwijsaanpak) of leerproceskenmerken (onderwijsmethode, handboek).



Figuur 3.4. Schematische weergave mediator versus moderator

Het identificeren van dergelijke mediatoren en moderatoren verhoogt de zekerheid dat men inderdaad vat krijgt op het achterliggende causale proces voor de verschuiving in leeruitkomst. De selectie van deze potentiële mediatoren/moderatoren gebeurt op basis van een (leer)theorie (zie voor

een overzicht Hattie, 2009; Kyriakides, Creemers & Panayiotou, 2020). Doorheen dit werkpakket van het steunpunt besteden we aandacht aan dit aspect door de gebruikte indicatoren voor het toetsen van verklaringen voor geobserveerde verschillen in de toegevoegde waarde van de school, te linken aan een specifieke theorie. Op die manier vangen we ook het probleem op dat elke indicator altijd hoogstens een proxy is van het achterliggende fundamentele proces.

Om te komen tot nieuwe mediatoren voor de analyse zullen we ook gebruik maken van inzichten die worden uitgewerkt in werkpakket G2 (7.2) en gegevens die in het kader van werkpakket G4 (7.6) zullen worden verzameld omtrent aanbod.

8.5.2. De ontwikkeling van dynamische indicatoren voor leerlingen en leerkrachtenstromen

Als we kijken naar het huidige schooleffectiviteitsonderzoek in Vlaanderen, valt op dat heel vaak gebruikt wordt gemaakt van statische indicatoren (Opdenakker, 2020). Zo wordt bijvoorbeeld onderzocht in welke mate de sociaaleconomische leerlingencompositie in een school een effect heeft op de leerprestaties van leerlingen. Niettegenstaande dergelijk onderzoek zeker zijn waarde heeft, kampt het met de beperking dat men er impliciet en ten onrechte van uitgaat dat de samenstelling van het leerlingenpubliek van een school een statisch gegeven is. Recent onderzoek van Havermans, Wouters en Groenez (2018) naar de evolutie van de schoolse segregatie in het Nederlandstalig onderwijs in België tussen de schooljaren 2001-2002 en 2015-2016 toont dat dit niet zo eenduidig is. Zo toont een analyse van de evolutie van de Hutchens-index – deze index meet homogeniteit als de mate waarin leerlingen volgens een geselecteerd achtergrondkenmerk op een gelijke manier over scholen verspreid zijn – een algemene toename van segregatie naar sociale samenstelling in het secundair onderwijs. Deze bevinding roept de vraag op in welke mate de evolutie van de compositie van de leerlingenbevolking van een school, in sociaaleconomische, culturele, academische… termen, een zelfstandig effect kan hebben op de leerresultaten en leerwinst van leerlingen.

Het belang van bovenstaande vraag schuilt niet enkel in de vaststelling dat het huidige schooleffectiviteitsonderzoek in Vlaanderen er weinig over kan vertellen. Het is ook nauw verbonden met mogelijke ongewenste effecten die gepaard kunnen gaan met het invoeren van GT. Zoals eerder verduidelijkt luidt een van de in de literatuur vaak aangehaalde kritieken op centrale toetsing dat *high stakes testen* – testen waar gevolgen aan verbonden zijn voor een school of een individuele leerling – scholen aanzet tot praktijken die gericht zijn op het aantrekken van goed presterende leerlingen en het afstoten van zwakker presterende leerlingen (Knoester & Au 2017). In het gedifferentieerd Vlaamse onderwijslandschap zal dit zorgen voor een mogelijke versterking van de 'selectie bias', waarbij de reeds bestaande academische en sociale verschillen in leerlingensamenstelling van scholen dreigen te vergroten. Dit ondergraaft de facto de primaire doelstelling van de toetsen, namelijk kwaliteitsbewaking en ondersteuning.

Om bovenstaande mogelijke negatieve effecten in kaart te brengen zijn twee opdrachten cruciaal binnen het huidige werkpakket: (1) een monitoring van leerlingenstromen en leerkrachtenstromen tussen scholen en (2) het toetsen of mogelijke negatieve selectiedynamieken zich ook effectief stellen. Voor de eerste opdracht maken we gebruik van de rijke administratieve data. Deze bieden de mogelijkheid om op zoek te gaan naar dynamische indicatoren die ons kunnen helpen om verschuivingen in verschillen tussen de scholen te begrijpen (Xie & Zhang, 2020). In navolging van Havermans, Wouter en Groenez (2018) ontwikkelen we op basis van een literatuurstudie indicatoren die per school aangeven in welke mate de leerlingensamenstelling (bv. Hutchens index) en de leerkrachtsamenstelling (bijv. op basis van anciënniteit, ziekteverzuim – zie Van Droogenbroeck et al., 2020) (1) over de tijd en (2) doorheen verschillende leerjaren veranderden. Het doel is indicatoren te ontwikkelen die een zicht bieden op de *dynamiek* die zich in scholen voordoet in termen van de leerling- en leerkrachtsamenstelling. De ontwikkeling van dergelijke monitoring-instrumenten heeft op zich een waarde (bijvoorbeeld om eventuele ongewenste effecten van GT in kaart te brengen),

maar de primaire doelstelling is uiteraard de relevantie van deze dynamieken voor de verschuiving in leerprestaties te bestuderen.

Ook al starten de werkelijke leerwinstmetingen pas in 2026, kunnen we de externe validiteit en verklaringskracht van de dynamische indicatoren alvast toetsen op de toetsscores en de toegevoegde waarde van scholen, die we met de eerste metingen bekomen. In een latere fase kunnen modellen ook toegepast worden op leerwinst. Om de externe validiteit en de fine-tuning van de ontwikkelde indicatoren te toetsen maken we gebruik van een koppeling met de LISO-gegevens. Op die manier beschikken we, lang voor we de effectieve populatiedata ontvangen, over de mogelijkheid om zeer nauwkeurig de werking van de indicatoren voor een subset van Vlaamse scholen te analyseren.

8.6. Het Onderwijsaanbod (G4)

Onderzoeksvraag 7. Wat is het onderwijsaanbod in de klas met betrekking tot een bepaald vak of leergebied?

Als het gaat om verklaringen voor onderwijsuitkomsten, dan zijn niet enkel school- en leerlingenkenmerken zoals cognitieve bekwaamheid, de motivatie of de samenstelling van de leerlingenpopulatie en de schoolorganisatie van belang. Leerlingen leren immers op basis van de informatie die hen wordt aangereikt in het curriculum (inhoud), de kwaliteit van het aanbod (vorm) en de wijze waarop het onderwijs wordt aangeboden (didactiek), kortom, het onderwijsaanbod in de klas en school. Dit zijn ook de kenmerken die de eerder vermelde effecten van de achtergrondkenmerken van leerlingen en scholen mediëren (Figuur 3.4). Die laatste staan ook centraal in het door Creemers en Kyriakides voorgestelde *multilevel integrated model of educational effectiveness* (Kyriakides, Creemers & Panayiotou, 2020) en verwijzen naar het Proces-element in het veelgebruikte CIPO-model voor onderwijsuitkomsten. Overzichtsanalyses (bv. Muijs, 2014) en meta-analyses (e.g. Hattie, 2009) leveren overtuigende evidentie aan dat naast individuele verschillen de meeste variantie in leerprestaties verklaard wordt door factoren die zich op klas- en leraarniveau bevinden.

Onderwijsaanbod behelst verschillende dimensies. Kenmerken gerelateerd aan de vorm (o.a. gebruikte handboeken) en didactiek (o.a. wijze van instructie) worden in kaart gebracht via de vragenlijsten bij leerkrachten die afgenomen worden op het moment van de afname van de GT's (analoog aan de huidige peilingsproeven). In dit werkpakket toetsen we hun empirische voorspellingskracht samen met wat men inhoud of het 'geïnstitutionaliseerd onderwijsaanbod' noemt. In sterk en vroeg gedifferentieerde onderwijssystemen zoals het Vlaamse wordt het onderwijsaanbod voor een leerling inderdaad in de eerste plaats bepaald door de specifieke studierichting waarvoor hij/zij koos. Dat element is vooral van toepassing op het secundair onderwijs, al wordt in sommige basisscholen ook met informele of formele niveaugroepen gewerkt. We weten dat toetsprestaties inderdaad variëren naargelang de studierichting die de leerlingen volgen of de vorm van binnengroepsdifferentiaties die scholen hanteren (Laurijssen & Glorieux, 2020). Hoewel het bestaande onderzoek op dit vlak relevante inzichten biedt, zijn er twee lacunes die met de data van de GT's weggewerkt kunnen worden. Ten eerste, wordt het geïnstitutionaliseerd aanbod vaak verengd tot de onderwijsvorm waarin een leerling les volgt. Dit kan echter maar pas vanaf het derde jaar van het secundair. Bovendien hebben we nood aan een fijnmazigere blik die op beschrijvend niveau afdaalt tot de specifieke studierichtingen en op analytisch-verklarend vlak werkt met meer groepen dan een verdeling naar onderwijsvormen. Dergelijke, meer fijnmazige, benadering biedt een drieledige meerwaarde: (1) het maakt het mogelijk studierichtingen te identificeren die systematisch beter of slechter doen in vergelijking met studierichtingen met een vergelijkbaar onderwijsaanbod; (2) het laat ook toe beter te identificeren welke aspecten van het aanbod uiteindelijk cruciaal zijn voor de geobserveerde verschillen in uitkomsten; (3) een beter zicht op de relevantie van studierichting zorgt er ook voor dat leerlingen en ouders meer zicht krijgen op de impact van een studiekeuze en verhoogt langs die weg de kans op het maken van goed geïnformeerde studiekeuzes. Aandacht voor de

studierichting zal, ten tweede, en vooral in de latere fase waarbij echt gewerkt kan worden met gegevens over leerwinst, ook meer inzicht bieden in de impact van veranderingen in studierichting voor leerlingen doorheen hun onderwijstraject. Dat laatste blijft in Vlaanderen een blinde vlek en dat ondanks het zeer grote aantal heroriënteringen als gevolg van B- en C-attesten.

Zoals aangeven zullen we dus zowel beschrijvende gegevens rond studieprestaties en later leerwinst naar studierichting geven, als verklarende analyses uitvoeren op het niveau van studierichtingen. We gaan ervan uit dat het onderwijsaanbod van studierichtingen vooral bepaald wordt door het officiële curriculum (o.a. het lessenpakket) en in het bijzonder dat de combinatie van specifieke lesuren relevant is veeleer dan bijvoorbeeld uren wiskunde of Nederlands als afzonderlijke dimensies te gebruiken. Onderzoek naar tracking leerde op dat vlak bijvoorbeeld, dat onderwijsvormen niet alleen een effect hebben omdat ze variëren in termen van het aanbod in specifieke vakken, maar omdat de combinatie van vakken leidt tot geïnstitutionaliseerde verwachtingen bij leerlingen en scholen rond het beheersingsniveau in een bepaalde onderwijsvorm (Domina et al., 2017). Alleen door het gehele aanbod van een studierichting in rekening te brengen, kan men met andere woorden de volle relevantie van dergelijke richtingen vatten.

Naast het onderzoek naar structurele schooleffecten, zullen in het gekoppelde vragenlijstonderzoek ook kenmerken worden bevraagd die constitutief zijn voor de kwaliteit van de leraar (teacher quality). Onderwijsonderzoek levert overtuigende evidentie aan dat leerprestaties van leerlingen in sterke mate afhankelijk zijn van factoren die aan de leraar en de klas kunnen worden toegeschreven (Goe, 2007; Muijs, et al, 2024; Nye et al., 2007). Nye en collega's (2004) bijvoorbeeld tonen aan dat 7-21% van de verschillen in leerprestaties toe te schrijven zijn aan lerareneffecten. De kwaliteit van de leraar wordt in dit steunpuntonderzoek gekwantificeerd aan de hand een samengestelde maat waaronder 1) inputkenmerken (opleiding, onderwijservaring, verwachtingen, onderwijsopvattingen), 2) het onderwijsaanbod (het aanwezige aanbod van leermaterialen in de klas waaronder de leermethoden), en 3) de kwaliteit van het verstrekte onderwijs aan de leerlingen (teaching quality) (zelfgerapporteerd pedagogisch-didactisch handelen, pedagogische relatie, klasmanagement). De door dit steunpunt onderzochte kenmerken van leraren kunnen bijdragen tot het verklaren van verschillen tussen (vergelijkbare) scholen, wat tegemoetkomt aan één van de centrale uitgangspunten van de opdracht, namelijk het ondersteunen van een schoolintern kwaliteitszorgsysteem in een datarijke omgeving.

Vanuit praktisch oogpunt zal het vragenlijstonderzoek digitaal verlopen via het te ontwerpen digitale platform. Dit heeft verschillende voordelen, waaronder het gebruiksgemak voor de leraar, de mogelijkheid tot rechtstreekse koppeling van de data aan leerlingendata, het minimaliseren van ontbrekende waarden. Het vragenlijstonderzoek bij leraren zal niet systeembreed worden uitgevoerd aangezien het werken met representatieve steekproeven generaliseerbare resultaten oplevert en op die manier de gemiddelde belasting van leraren aanzienlijk wordt verminderd. Dit minimaliseert tevens de kans dat het onderzoek door de leraren en scholen zelf als "high stakes" wordt geïnterpreteerd.

Het verzamelen van informatie over teacher quality via kwantitatieve methoden is niet zonder beperkingen. Om een rijk beeld te krijgen van het pedagogisch-didactisch handelen van leraren geven observatiegegevens een meer valide representatie van de werkelijkheid. Het is in het kader van het steunpunt echter niet mogelijk deze effectiviteitsvraag via interpretatieve methoden te beantwoorden aangezien dit onderzoekscapaciteit vereist die niet kan begroot worden zonder impact op toetsontwikkeling, ondersteuning bij afname en schoolgerichte feedback. Het steunpunt engageert zich echter de mogelijkheden van flankerend onderzoek te faciliteren door samenwerkingsverbanden aan te gaan met de bredere onderzoeksgemeenschap.

9. Werkdomein H. Schoolfeedback(gebruik)

9.1. Situering werkdomein

Feedbackgebruik is in wezen het alfa en omega van elk systeem van werken met gecentraliseerde toetsen. Toetsontwikkeling en toetsafname staan immers in het teken van de eruit voortvloeiende mogelijkheid tot gebruik van de verzamelde informatie. Doelgericht en systematisch gebruik van leerlingoutputdata is in Vlaanderen evenwel vooralsnog beperkt (Van Gasse, Vanhoof, Mahieu, & Van Petegem, 2015; Vanlommel, Vanhoof, & Van Petegem, 2017), niettegenstaande geweten is dat dataverzameling, -analyse en -gebruik een positieve impact kunnen hebben op de kwaliteit van onderwijsbeslissingen (Rossi, Lipsey, & Freeman, 2004; Schildkamp, Lai, & Earl, 2012). Net deze onderwijsbeslissingen, en hoe het gebruik van feedback uit de gecentraliseerde toetsen deze kunnen informeren, staan vanuit verschillende perspectieven centraal in dit werkdomein van het steunpunt.

Centrale toetsen geven leerlingen, ouders, leerkrachten, scholen en overheidsinstanties een instrument om te beschrijven, reflecteren en bij te sturen. Wij verwijzen verderop in dit werkdomein 'Schoolfeedback(gebruik)' naar deze groepen als 'gebruikers', in de betekenis van zij die actief aan de slag kunnen gaan met de informatie die het instrument oplevert. Gebruikers worden dus niet als passieve ontvangers beschouwd. Vanuit een argument-based approach op toetsontwikkeling en validiteitsonderzoek hebben we in dit steunpunt - naast aandacht voor de evaluatie van de psychometrische kenmerken van de toetsen zoals beschreven in de andere werkdomeinen van dit voorstel (validity argumenten) - doelbewust ook aandacht voor de evaluatie van de interpretaties en gebruik van de toetsresultaten (interpretative/use argumenten) voor diverse doelgroepen (Kane, 1992, 2006, 2013). Het gebruik en de gevolgen van een toets (i.c. de eruit resulterende feedback) vormen immers een onontbeerlijk onderdeel van de validering. Ook bij testen die in hoofdzaak bedoeld zijn als beleidsinstrument zijn er belangrijke redenen om een evaluatie van het gebruik en de gevolgen mee in beschouwing te nemen. Indien een primair doel van de toetsen is om te dienen in cycli van interne kwaliteitszorg bij vakgroepen, scholen, scholengemeenschappen etc. is het belangrijk te bewaken dat de juiste interpretaties gegeven worden aan de toetsresultaten (evalueren om te leren) in functie van optimalisatie en verbetering van de praktijk.

Opdat de al toetsend verkregen informatie op een correcte en gecontextualiseerde manier kan ingebracht worden in beslissingen van leerlingen, leraren, schoolleiders en eventueel ook het ruimere publiek, dient eerst een informatief en performant feedbackplatform geconceptualiseerd en ontwikkeld te worden. Dat platform heeft de ambitie de beschikbare informatie gebruiksvriendelijk te ontsluiten. Een eerste werkpakket binnen het Werkdomein 'Schoolfeedback(gebruik)' is dan ook gericht op de opzet en ontwikkeling van een feedbacksysteem (cf. Werkpakket H1). In dit werkpakket wordt - in de logica van service design (Miller, 2015) - via intensieve consultatie van en aftoetsing bij opdrachtgevers en toekomstige gebruikers tot een conceptuele blauwdruk van het te ontwikkelen feedbackdashboard gekomen. Werkpakket H2 is gericht op het zogenaamde 'Gebruikersonderzoek en effectmonitoring'. Waar Werkpakket H1 in het teken staat van de ontwikkeling en optimalisering van feedbacksystemen, is dit tweede werkpakket erop gericht in kaart te brengen hoe gebruikers de informatie gedeeld via het feedbacksysteem feitelijk hanteren en welke randvoorwaarden belangrijk zijn om optimaal gebruik mogelijk te maken. In essentie gaat het om een continue monitoring van de processen en (on)bedoelde effecten van de implementatie van de centrale toetsen. Werkpakket H3 richt zich op professionalisering en heeft de ambitie om enerzijds scholen voor te bereiden op de organisatie en betekenis van de toetsafnames voor de school en anderzijds om het Vlaamse onderwijsveld uit te rusten met de vereiste kennis om het feedbackdashboard correct te hanteren en om aan expertiseontwikkeling te doen m.b.t. good practices van hoe scholen in partnerschappen met bv. lerarenopleidingen en pedagogische begeleidingsdiensten de verkregen informatie kunnen integreren in een breder verhaal van interne kwaliteitszorg. Dit werkpakket voorziet in verschillende

soorten van ondersteuning en professionalisering voor diverse groepen van gebruikers op verschillende tijdstippen van het implementatieproces (denk aan creatie van een draagvlak vooraf aan afnames, verwerving van datageletterdheid vooraf en na de afnames en ondersteuning i.f.v. feedbackgebruik) van de gecentraliseerde toetsen.

Werkpakketten H1, H2 en H3 krijgen gedurende de volledige looptijd van het steunpunt vorm en lopen over de vijf projectjaren. Dat is anders voor Werkpakket H4 'Schoolfeedback paralleltoetsen'. Dit is een tijdelijk werkpakket waarin een medewerker van Werkdomein H de continuïteit van schoolfeedback voor de dan nog bestaande maar uitdovende paralleltoetsen basisonderwijs zal verzorgen. Tot einde 2022 blijft de coördinatie hiervan in handen van het huidige Steunpunt Toetsontwikkeling en Peilingen (STEP). In het najaar van 2022 voorzien we de nodige afstemming met STEP om deze feedback in het nieuwe steunpunt na de afname van de paralleltoetsen aan het einde van schooljaar 2022-2023 aan de betrokken scholen te kunnen blijven aanbieden.

	2021			2022			2023			2024			2025		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
H1 Inhoud. opzet / vormgeving feedbacksyst.															
H2 Gebruikersonderzoek / effectmonitoring															
H3 Professionalisering feedbackgebruikers															
H4 Schoolfeedback paralleltoetsen BO 22/23															

Tabel 3.20. Werkpakketten Schoolfeedbackgebruik

De realisatie van de verschillende werkpakketten binnen Werkdomein H vergt voortdurende afstemming en integratie. Doorheen de looptijd wordt daartoe op regelmatige basis intern overleg georganiseerd, in aanvulling bij de steunpuntbrede overlegstructuren. De postdoc-medewerker in dit werkdomein krijgt hierin een coördinerende functie toebedeeld. Samen met de promotoren staat deze persoon ook in voor de strategische sturing en kwaliteitsbewaking van de vier werkpakketten.

1.1. Geïntegreerde kennisbasis voor feedbackgebruikgerelateerde werkpakketten

Met de geïntroduceerde set aan werkpakketten (H1, H2 en H3) zetten we in op een systemische benadering van schoolfeedback(gebruik). Binnen het werkdomein hanteren we een geïntegreerde versie van modellen die informatiegebruik bij besluitvorming in het onderwijs in kaart brengen (in het bijzonder: Mandinach, 2012; Schildkamp & Lai, 2013; Visscher & Coe, 2002).

Empirische evidentie leert dat informatiegebruik beïnvloed wordt door een breed scala aan factoren die stimulerend of juist beperkend kunnen werken (Schildkamp & Poortman, 2015). Het gaat om factoren die inherent zijn aan de informatie of het informatiesysteem zelf, alsook factoren op het niveau van de gebruiker, zijn of haar school, en de onderwijs- en maatschappelijke context waarin hij of zij zich beweegt, met inbegrip van ondersteuning en professionalisering (Schildkamp et al., 2017; Schildkamp, Poortman, Luyten, & Ebbeler, 2016; Schildkamp & Kuiper, 2010; Van Gasse et al., 2015; Wayman, Jimerson, & Cho, 2012). Deze elementen worden in de verschillende werkpakketten conceptueel verder uitgediept en in concrete projectdoelen en activiteiten vertaald.

In de vooropgestelde doelstellingen van de gecentraliseerde toetsen herkennen we opportuniteiten voor verschillende soorten van informatiegebruik zoals die in de literatuur beschreven worden (Hellrung & Hartig, 2013; Rossi et al., 2004; Verhaeghe et al., 2010; Visscher & Coe, 2003). Input geven voor de leerlingenevaluatie is een gebruiksdoel op instrumenteel niveau. Leraren laten reflecteren over hun eigen effectiviteit, is een voorbeeld van voornamelijk conceptueel informatiegebruik. Kwaliteitsindicatoren toekennen aan scholen bevindt zich deels op de symbolische dimensie van

informatiegebruik. Schoolresultaten ook breder ontsluiten, opent tot slot de deur naar strategische overwegingen die door sommige stakeholders als ongewenst worden gezien. De verschillende dimensies van informatiegebruik kunnen naast elkaar bestaan, maar een goede balans vinden zal een uitdagende evenwichtsoefening zijn. Bovendien is het niet vanzelfsprekend om één informatiebron voor verschillende doeleinden in te zetten. Om effecten van informatiegebruik te ontrafelen, gaan we onder meer na welke specifieke rol standaard-gebaseerde, groepsgerichte en zelfgerelateerde indicatoren (kunnen) spelen in de verschillende besluitvormingsprocessen van diverse gebruikers n.a.v. de feedback.

Instrumenten voor geïnformeerde besluitvorming dienen door de gebruiker als betrouwbaar, valide, en relevant gezien te worden (Pierce & Chick, 2011; Schildkamp & Teddlie, 2008; Van Gasse et al., 2015; Visscher, 2002; Visscher & Coe, 2003). Gebruikers dienen ook controle te ervaren over de acties die zij (kunnen) ondernemen op basis van de data, en van mening te zijn dat hun acties vervolgens weer impact hebben op de data (i.e. toetsresultaten) die de instrumenten opleveren (Prenger & Schildkamp, 2018; Schildkamp & Kuiper, 2010). Toch geven het aanbieden van sterke, valide instrumenten en het creëren van eigenaarschap, op zichzelf onvoldoende garantie op succes. Schoolleiders en leraren hebben vaak het gevoel over veel informatie te beschikken waarvan zij niet meteen zien hoe deze kan bijdragen aan de school- en klaspraktijk (Van Gasse et al., 2015). Hun zelfvertrouwen om informatie te gebruiken voor veranderingsdoeleinden moet dus worden ontwikkeld (Mandinach & Gummer, 2016; Van Gasse et al., 2017; Verhaeghe et al., 2010). Efficiënt gebruikmaken van informatiebronnen zoals outputdata vergt van gebruikers immers kennis, vaardigheden en attitudes die niet als vanzelfsprekend verondersteld mogen worden (Datnow et al., 2007; Kerr et al., 2006).

Om de vooropgestelde doelstellingen van de centrale toetsen te realiseren, wordt van gebruikers minstens een doorgedreven datageletterdheid gevraagd. Datageletterdheid omvat een palet aan verschillende competenties dat verder reikt dan statistisch inzicht (Chick & Pierce, 2013) en visuele geletterdheid of de vaardigheid om grafische (data)voorstellingen correct te interpreteren (Lee et al., 2017). Om correcte interpretaties mogelijk te maken is ook een vorm van ‘assessment literacy’, of een inzicht in de mogelijkheden en beperkingen van verschillende toetsvormen (AERA et al., 2014) noodzakelijk. Om een gedegen vertaalslag te maken, moeten onderwijsprofessionals verder voldoende pedagogische datageletterdheid aan de dag leggen: de competenties om data te analyseren en op grond daarvan veranderingen te implementeren in de school- en klaspraktijk (Mandinach, 2012; Mandinach & Gummer, 2016; Mandinach & Schildkamp, 2020). De aangeleverde data correct lezen is slechts een eerste stap. Gebruikers dienen daarenboven in staat te zijn om kritisch naar de instrumenten en de informatie te kijken, en relaties te leggen met het door hen verzorgde onderwijs (Faber & Visscher, 2014). Ook in Vlaams onderzoek werd reeds aangetoond dat gebruikers sterk verschillen in hun datageletterdheid, i.e. hun capaciteiten om data op een gepaste manier te interpreteren, diagnoses te stellen en er vervolgacties aan te koppelen (Vanlommel et al., 2017). Dat maakt dat ze verschillend omgaan met informatie uit feedbackrapporten (Vanthournout et al., 2009). Feedbacksystemen moeten bijgevolg niet enkel data presenteren, maar gebruikers ook ondersteunen bij het interpreteren en gebruik ervan. Die ondersteuning dient derhalve voldoende gedifferentieerd te zijn voor diverse gebruikersprofielen. De nood aan de aanwezigheid en verwevenheid van de verschillende werkpakketten in dit werkdomein wordt vanuit deze inzichten treffend aangetoond.

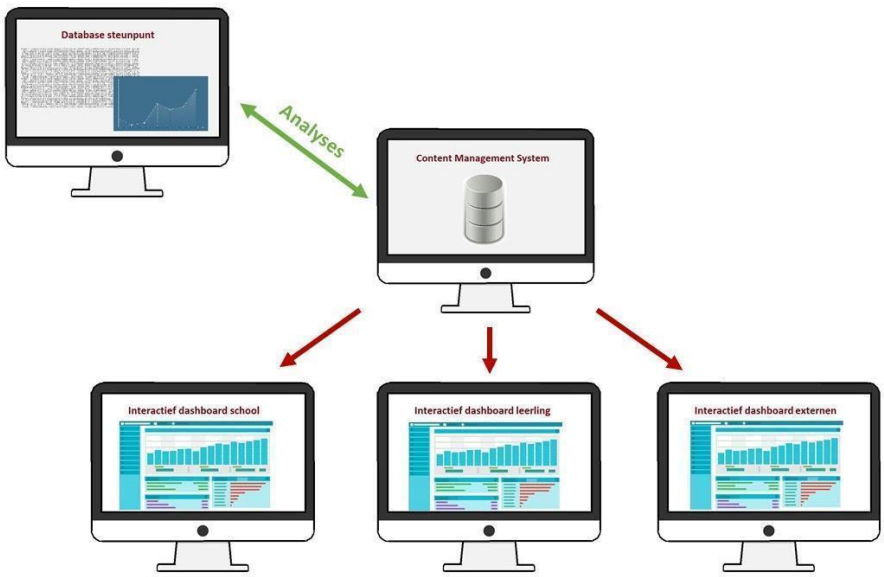
In de literatuur rond geïnformeerde besluitvorming in het onderwijs, worden verschillende aspecten van informatiegebruik en datageletterdheid ook bestudeerd door een ‘sensemaking’ lens (Ketelaar et al., 2012; Schildkamp, 2019). Wanneer gebruikers aan de slag gaan met data, wenden zij niet alleen bepaalde competenties aan, maar komen ook eigen overtuigingen en mentale modellen (Bertrand & Marsh, 2015; Jimerson, 2014) en beslissingsstijlen (Vanlommel et al., 2017; Vanlommel & Schildkamp,

2019) in het spel. Die bepalen voor een stuk mee welke interpretaties en gevolgtrekkingen uit de data gehaald worden, welke waarde men eraan hecht, en hoe men uiteindelijk ageert. 'Sensemaking' heeft een sterke cognitieve dimensie, maar is ten dele ook een collectief proces (Coburn, 2001; März & Kelchtermans, 2013). In een schoolteam krijgt collectieve 'sensemaking' bijvoorbeeld vorm binnen samenwerkingsverbanden en wordt het beïnvloed door de rollen die individuele actoren daarin spelen. Inzichten als deze werken sterk door in de voorziene aanpak en keuzes binnen Werkpakket H3 'Professionalisering feedbackgebruikers'. Als het evident is dat ook de onderwijscontext een invloed heeft op (zowel individuele als collectieve) 'sensemaking', vertrekt ondersteuning best vanuit de vraag hoe die ruimere context kan vormgegeven en aangewend worden met het oog op professionalisering.

1.2. Opzet en ontwikkeling feedbacksysteem (H1)

Een cruciale effectiviteitsvoorwaarde voor een feedbacksysteem is dat het de verschillende gebruikersgroepen juiste en relevante informatie op een gebruiksvriendelijke manier aanbiedt en hen ondersteunt bij het interpreteren van deze informatie. Een kernelement in het (goed) gebruiken van informatie in scholen is de perceptie van gebruikers met betrekking tot de relevantie van informatie (Van Gasse, Vanhoof, Mahieu, & Van Petegem, 2015). Daarnaast wordt informatiegebruik sterk beïnvloed door de mate waarin informatiesystemen gebruiksvriendelijk zijn opgebouwd (Van Gasse et al., 2015). De informatie die in informatiesystemen beschikbaar wordt gemaakt moet niet enkel betrouwbaar en valide zijn, maar ook actueel en op een behapbare manier gepresenteerd worden.

Het doel van dit werkpakket is bijgevolg om een feedbacksysteem vorm te geven dat zich kenmerkt door (1) het weergeven van informatie die relevant is voor verschillende groepen van gebruikers (bijvoorbeeld, schoolleiders, leraren, leerlingen, ouders), (2) een gebruiksvriendelijk design, (3) kwaliteitsvolle informatie die de interpretatie van informatie ondersteunt, en (4) die aangepast is aan de verschillende noden van diverse soorten gebruikers (bijvoorbeeld, leerlingen ten opzichte van schoolleiders). Centraal in het feedbacksysteem staat daarom een content managementsysteem, dat de rechten en rollen van verschillende soorten gebruikers bewaakt en dat informatie uit de centrale steunpunt-databank ophaalt en vertaalt naar de dashboards voor diverse gebruikers (zie Figuur 3.5).



Figuur 3.5. Feedbacksysteem in functie van diverse gebruikers

Om deze doelen te verwezenlijken wordt gebruik gemaakt van de methodologie van servicedesign (Miller, 2015). Dit is een aanpak om diensten te ontwikkelen waarbij er wordt gestreefd naar een evenwicht tussen de behoeften van de gebruikers en de noden van de opdrachtgevers. Servicedesign kenmerkt zich onder meer door een intensieve samenwerking tussen toekomstige gebruikers en ontwerpers en door een iteratieve aanpak (Stickdorn et al., 2018). Daarbij wordt een product of dienst in steeds toenemende complexiteit opgebouwd door het cyclisch doorlopen van behoefteanalyses, ontwerpactiviteiten en gebruikerstesten. Toegepast op het feedbacksysteem, wil dit zeggen dat verschillende groepen van gebruikers (bijvoorbeeld, schoolleiders, leerkrachten, leerlingen) betrokken worden in het selecteren van gepaste informatie voor de feedback en het uittesten van de interactieve dashboards. Gebruikers worden betrokken in verschillende stappen van het designproces: van concept naar prototype tot het finale design (Bergvall-Kåreborn et al., 2009). Telkens treden de stakeholders op als co-designers en als testers van de prototypes (Pierson & Lievens, 2005).

9.1.1. Doelstellingen en activiteiten

Om een *gebruiksvriendelijk* feedbacksysteem met *relevante* informatie voor de verschillende gebruikers te creëren, staan de weergave van de data, ondersteunende informatie voor het interpreteren van de informatie, de inhoudelijke feedback, de lay-out van het feedbacksysteem en de gebruikersinterface centraal in dit werkpakket. Via interactieve dashboards zal *aangepaste feedbac*k teruggekoppeld worden voor verschillende groepen van gebruikers (bijvoorbeeld schoolbesturen, schoolleiders/leraren en leerlingen, maar ook het brede publiek en in het bijzonder toekomstige schoolkeuzemakers), met inbegrip van hun resultaten ten opzichte van referentiegroepen en hun evolutie doorheen de tijd.

Een aandachtspunt in de ontwikkeling van het feedbacksysteem is vertrouwelijkheid van informatie. Afhankelijk van verschillende rechten en rollen in het content managementsysteem, zal informatie wel of niet beschikbaar gemaakt worden voor verschillende gebruikers (*conform* de *GDPR-wetgeving*). Daarnaast zal er ook aandacht zijn voor de publieke toegankelijkheid van gegevens, onder meer door de ontwikkeling van een publiek toegankelijk dashboard. In dit publiek toegankelijk dashboard wordt beschikbare informatie zodanig gepresenteerd dat deze in lijn is met juridische vereisten inzake openbaarheid van bestuur (zoals deze door de haalbaarheidsstudie en de opdrachtgever zullen geëxpliciteerd worden) en zich ondertussen zo beperkt als mogelijk leent tot het creëren van schoolrangschikkingen door derden.

Om deze doelen te bereiken, legt dit onderzoeksluik zich toe op verschillende kerntaken. Essentieel zijn vooreerst behoefteanalyses bij verschillende gebruikersgroepen om zicht te krijgen op hun noden en wensen t.o.v. een interactief dashboard bij centrale toetsing. Dit zal vorm krijgen door middel van een doordachte combinatie van focusgroepen met gebruikers, het in kaart brengen van user journeys, interviews met experts, literatuurverkenning, consulteren van good practices en living labs. Deze aanpak geldt zowel voor de besloten als de publiek toegankelijke delen van het interactief dashboard, met als verschil dat voor dit laatste ook de juridische consequenties van de openbaarheid van bestuur (zoals de opdrachtgever deze wenst in te vullen) mee bepalend zijn voor de inhoudelijke vormgeving.

Daarnaast zal het steunpunt een aantal ontwerpactiviteiten uitvoeren. Het ontwerpen van interactieve dashboards met onder meer data-visualisaties, ondersteunende teksten en weergave van analyses doorheen de tijd zal gebeuren in samenwerking met gebruikers per onderwijsniveau. Dit geldt ook voor de inhoud van publieke dashboards. Daarnaast zal het content managementsysteem ontworpen worden, wat de administratieve module is die de rechten en rollen voor de interactieve dashboards zal waarborgen. In nauwe samenwerking met de softwareontwikkelaars zullen functionele analyses opgesteld worden om de wensen van gebruikers technisch te vertalen. Om deze ontwerpactiviteiten vorm te geven, zullen verschillende persona's gecreëerd worden die het denken sturen. Daarnaast worden designrichtlijnen uitgeschreven, en prototypes en mock-ups gecreëerd en geëvalueerd.

Naast behoefteanalyses en eigenlijke ontwerpactiviteiten legt dit werkpakket zich toe op uitgebreide gebruikerstesten. Deze testen houden in dat de resultaten van de ontwerpactiviteiten uitgeprobeerd en verder verfijnd worden in samenwerking met gebruikers. Focusgroepgesprekken en het analyseren van user statistics zullen daarbij generieke inzichten geven. Voor meer gedetailleerd inzicht in de gebruiksvriendelijkheid van de interactieve dashboards kan de combinatie van think-aloud met eye-tracking protocollen uitweg bieden. De gebruikte technieken dienen vooreerst doelgeschikt te zijn en volgens de geldende kwaliteitsstandaarden uitgevoerd te worden. Daarnaast wordt telkens de vraag gesteld of de te investeren inspanningen in verhouding staan tot de verzamelde beleidsinformatie en onderzoeksdata. Think-aloud protocollen zijn eerder makkelijk implementeerbaar terwijl ze de kans geven tot de kern van de bevraagde processen door te dringen. Onderzoeksgroep Edubron (UAntwerpen) beschikt over een lab met eye-tracking hard- en software, waardoor deze methodologie ook binnen dit steunpunt eerder eenvoudig kan geïntegreerd worden.

9.2. Gebruikersonderzoek en effectmonitoring (H2)

Centraal in het aanmoedigen van het gebruik van objectieve gestandaardiseerde toetsen staat de assumptie dat gebruikers in staat zijn om op basis van de toetsresultaten te komen tot begrip, diagnose, en de formulering van beslissingen en acties in de praktijk. In dit werkpakket is de vraag hoe gebruikers hierbij feitelijk te werk gaan het uitgangspunt. Het gebruikersonderzoek wordt opgezet als instrumenteel voor de vormgeving van de andere werkpakketten in dit werkdomein. Het staat ten dienste van de ontwikkeling van het feedbacksysteem en van de vormgeving van professionalisering. Hoewel het gebruikersonderzoek hier los van de andere werkpakketten beschreven wordt zal de inhoudelijke en methodologische vormgeving van het gebruikersonderzoek dus in sterke mate mee bepaald worden door de vraagstukken uit de andere werkpakketten. Toch zijn gebruikersonderzoek en effectmeting ook an sich een expliciet doel met het oog op het informeren van strategische keuzes binnen het steunpunt en door de opdrachtgever. Het werkpakket beschrijft hoe onderwijsprofessionals in Vlaamse scholen de gecentraliseerde toetsen percipiëren, hoe schoolinterne en -externe stakeholders concreet met de resultaten aan de slag gaan, en hoe het gebruik van de instrumenten in het onderwijsveld strookt met de intenties die eraan ten grondslag liggen bij de inrichters van het systeem.

De opdrachtgever formuleert ambitieuze doelstellingen voor het gebruik van de gecentraliseerde toetsen. In die doelstellingen benoemt hij in de eerste plaats de onderwijsprofessional op de klasvloer als een primaire stakeholder. Enerzijds bieden de toetsen een spiegel aan leraren en schoolleiders die zij kunnen aanwenden om hun eigen beleid en praktijk tegen het licht te houden. Anderzijds krijgen deze onderwijsprofessionals ook het vertrouwen en de verantwoordelijkheid om de informatie uit de toetsen te gebruiken als een instrument voor de evaluatie van leerlingen. Omdat outputdata uit de gecentraliseerde toetsen voor leerlingenevaluatie gebruikt kunnen worden, wordt de individuele Vlaamse leerling voor het eerst ook een expliciete stakeholder of ‘gebruiker’. En aangezien het, tot slot, in principe ook mogelijk is om schoolresultaten openbaar te maken met het oog op het informeren van schoolkeuze, beschouwen we ook ouders als belanghebbenden.

Om ervoor te zorgen dat de vooropgestelde doelstellingen van de centrale toetsen geen dode letter blijven, dringt een doorgedreven gebruikersonderzoek zich op. Een continue monitoring van de effecten van de implementatie van de centrale toetsen stelt de overheid in staat om de vinger aan de pols te houden, de randvoorwaarden voor effectief gebruik van de centrale toetsen te monitoren en te optimaliseren naargelang de werkelijke acties en percepties op de praktijkvloer. Het stelt ons in staat kinderziektes en struikelblokken te identificeren, bij te sturen en vervolgens verder te omkaderen en professionaliseren waar nodig. Vanuit hetzelfde perspectief informeert het gebruikersonderzoek ook de andere werkpakketten binnen dit werkdomein.

Door effectmonitoring vanaf de start van de opdracht in te bouwen, erkennen we dat de implementatie van centrale toetsen in het Vlaamse onderwijsveld niet in een vacuüm gebeurt. Hoewel over informatiegebruik naar aanleiding van schoolfeedback al een behoorlijke academische kennisbasis bestaat, mag niet uit het oog verloren worden dat diezelfde kennisbasis ook net leert dat feedbackgebruik in situ en dus pas gecontextualiseerd vorm krijgt en kan begrepen worden. Alle gebruikers die aan de slag zullen gaan met de toetsen, nemen immers voorkennis, bestaande vaardigheden en percepties mee, en verwachtingen die onder meer gekleurd zijn door eerdere ervaringen met en narratieven rond gestandaardiseerde toetsen (bijvoorbeeld koepelgebonden toetsen en programma's uit het buitenland). Door bewust vanuit deze beginsituatie te vertrekken om (de effecten van) het gebruik van centrale toetsen in kaart te brengen, kan Vlaanderen een uniek momentum aangrijpen om het systeem geïnformeerd uit te bouwen. Dat maakt tevens dat Vlaanderen de kans krijgt om een wezenlijke en opgemerkte bijdrage te leveren aan de wetenschappelijke kennisbasis rond geïnformeerde besluitvorming in scholen. Daarmee realiseert het steunpunt ook de beleidsdoelstelling rond het verder uitbouwen van wetenschappelijke expertise die de opdrachtgever vooropstelt.

9.2.1. Doelstellingen en activiteiten

Het gebruikersonderzoek wil in de eerste plaats op basis van een brede én diepgaande analyse van bestaande percepties, gebruikersintenties, en verwachte en gerealiseerde effecten, het beleid (i.e. de overheid als opdrachtgever, en de strategische sturing van het steunpunt) informeren. Daarnaast stelt deze analyse ons in staat om 'good practices' te identificeren om het veld inspiratie en houvast te geven. Verder levert de studiemateriaal op om verdere onderzoeksexpertise uit te bouwen. Ook de integratie in de andere werkpakketten van dit werkdomein rechtvaardigt de aanwezigheid van het gebruikersonderzoek en de effectmonitoring. Zo zullen de design-based studies die ingezet worden om het feedbacksysteem vorm te geven (Werkpakket H1) baat hebben bij een blik door de ogen van gebruikers op de data zoals die gedissemineerd worden, en schetst het gebruikersonderzoek een kader waarbinnen de professionalisering van leraren en schoolleiders verder vorm kan krijgen (Werkpakket H3).

Om (effecten van) het gebruik van feedback bij de centrale toetsen te beschrijven en te verklaren, is een multimethodeaanpak (Tashakkori & Teddlie, 2002) aangewezen. Een combinatie van kwantitatieve en kwalitatieve bouwstenen laat toe om het gebruikersonderzoek zowel in de breedte als in de diepte te voeren. In de eerste plaats stellen we voor om survey-onderzoek te laten plaatsvinden. Surveys laten toe om heersende percepties over de centrale toetsen in het Vlaamse onderwijsveld in kaart te brengen, om een globale status op te maken van de randvoorwaarden voor een efficiënt gebruik van deze toetsen, en om generaliseerbare uitspraken te doen over gebruikersintenties (in een eerste fase) en gerealiseerde effecten (nadat de toetsen daadwerkelijk geïmplementeerd zijn). Door (minstens) de kern van de survey meermaals te bevragen, is het mogelijk om evoluties te monitoren in zowel opvattingen als gedrag. We realiseren voor de kwantitatieve monitoring een bevraging bij een representatief staal van gebruikers, onderwijsprofessionals in de eerste plaats. Uitgangspunt is het trekken van een steekproef van scholen gewoon basisonderwijs en scholen gewoon secundair onderwijs, met - afhankelijk van hoe breed de afname door de overheid zal ingevuld worden - ook een (over)vertegenwoordiging van buitengewoon onderwijs. We streven daarbij een representatieve verdeling per onderwijsnet na. In basisscholen bevragen we de schoolleiding en leraren van het vierde en het zesde leerjaar. In secundaire scholen vragen we responses van de schoolleider(s), leraren Nederlands en wiskunde in het tweede jaar, en vanaf schooljaar 2024 voegen we daar leraren Nederlands en wiskunde in het zesde jaar aan toe. De grote lijnen die met het survey-onderzoek geschetst worden, zullen verder ingekleurd te worden met behulp van diepgaande bevragingen bij doordacht geselecteerde gebruikersgroepen. Een jaarlijkse cyclus van interviews met leraren en schoolleiders, waar zinvol in de vorm van focusgroepen, stelt ons in staat

werkpunten en opportuniteiten concreet te maken rond kwesties zoals datageletterdheid en de creatie van een draagvlak.

Het survey-onderzoek en de kwalitatieve verdieping gaan van start bij het begin van de steunpuntopdracht in 2021, en zijn nadrukkelijk gericht op het bevragen van onderwijsprofessionals. Vanaf de implementatiefase (2023 voor secundair onderwijs, 2024 voor het basisonderwijs) worden ook ouders en leerlingen bevraagd. Deze bevragingen kunnen ad hoc vormgegeven worden, of we kunnen ervoor opteren om surveys en interviewstudies uit te voeren die parallel lopen met het kerntraject, maar afgestemd zijn op de eigenheid van deze gebruikersgroepen.

Los van dit kerntraject zijn er ook specifieke extensies mogelijk. Om het kwalitatieve luik van het gebruikersonderzoek te versterken kan overwogen worden om bij de start, maar evenzeer ook tijdens de verdere looptijd van het steunpunt, bepaalde Delphi-technieken in te zetten. Deze onderzoeksmethode is exploratief van aard en bestaat erin om een panel van doordacht gekozen stakeholders samen te brengen die zich op een gestructureerde manier, en elk vanuit eigen ervaring, opinie en expertise, over een complex vraagstuk buigen (Day & Bobeva, 2005). In de eerste plaats denken we daarbij aan een panel van “typische gebruikers”: leraren, schoolleiders en eventuele andere onderwijsprofessionals die geacht worden om concreet met de resultaten aan de slag te gaan zoals pedagogisch begeleiders. Een andere mogelijkheid bestaat erin om microprocesstudies (cf. suggestie in Schildkamp, 2019) uit te voeren waarbij we door middel van observaties op de vloer nagaan hoe specifieke processen zoals individuele en collectieve ‘sensemaking’ vorm krijgen. We voorzien dergelijke inhoudelijke keuzes te maken in lijn met de strategische keuzes van het steunpunt en in overleg met de opdrachtgever.

9.3. Professionalisering feedbackgebruikers (H3)

In de inleiding bij het werkdomein werd reeds aangehaald dat het doelgericht en systematisch gebruik maken van leerlingoutputdata een positieve impact kan hebben op de kwaliteit van onderwijs, maar dat deze ‘good practice’ in Vlaanderen nog erg beperkt is. Uit onderzoek bleek eerder dat het gebruik van schoolfeedback en leerlingoutputdata kan versterkt worden door het aanbieden van ondersteuning van de gebruikers (Wayman, Jimerson, & Cho, 2012; Schildkamp, Poortman, Luyten, & Ebbeler, 2016). Op verschillende niveaus is er groeiende aandacht voor datagebruik in onderwijs wat zich de voorbije jaren heeft vertaald in cursussen en nascholingstrajecten rond thema’s zoals de informatierijke schoolomgeving, evidence-informed/evidence-based onderwijs, data-driven decision making, onderzoekende leraar en school, designteams … Als parapluterm gebruiken we hier ‘datageletterdheid’ om te verwijzen naar de “competentie om verschillende vormen van gegevens te verzamelen, analyseren en interpreteren en om de verkregen informatie te vertalen naar concreet inzetbare kennis en praktijken voor de vormgeving van onderwijsleeromgevingen” (Gummer & Mandinach, 2015, p. 2).

Vanzelfsprekend neemt dit werkpakket de generieke kennis over effectieve professionalisering van onderwijsprofessionals als vertrekpunt. Deze zal doorontwikkeld worden naar de context van datagebruik én naar het Vlaamse onderwijsveld. Over het ondersteunen en professionaliseren van onderwijsprofessionals op vlak van datageletterdheid doen Henderson en Corry (2020) op basis van een doorlichting van de bestaande internationale onderzoeksliteratuur vier aanbevelingen die ook geïntegreerd worden in de ondersteuningsactiviteiten van het steunpunt centrale toetsen. Ten eerste wordt aanbevolen om niet alleen kennis bij te spijkeren, maar ook om concrete vaardigheden te oefenen en attitudes te verwerven. Een hands-on aanpak, die de diverse gebruikers in staat stelt om resultaten te consulteren, te ontsluiten en te gebruiken in functie van de praktijk van het leren en lesgeven is daarbij onontbeerlijk. Het doel is om gebruikers van de toetsresultaten te ondersteunen bij de analyse en de vertaalslag ervan naar de les-, vak- en schoolpraktijk. Ten tweede wordt

aangemoedigd om naast individueel, ook in samenwerkingsverbanden met data aan de slag te gaan. Samenwerking binnen lerarenteams, vakgroepen en scholengroepen kan versterkt worden door partnerschappen met hoger onderwijsinstellingen als een belangrijke strategie om de datageletterdheid bij onderwijsprofessionals te verhogen. Ook het inbrengen van voldoende verschillende perspectieven wordt aangehaald als belangrijke strategie om te garanderen dat data op een bedachtzame manier gebruikt worden en de kans op onbewuste bias tegenover kansengroepen en minderheden te verkleinen. Datageletterdheid neemt sterker toe in scholen waar schoolleiders kansen creëren en ondersteuning bieden voor collaboratief onderzoek. Ten derde wordt op basis van eerder onderzoek aangeraden om onderwijsprofessionals niet alleen vertrouwd te maken met het gebruik van kwantitatieve data, maar om ook het verzamelen en analyseren van kwalitatieve data onder de aandacht te brengen. Kwantitatieve data hebben een grote meerwaarde, al vergt een genuanceerde en doorgedreven analyse een verdere aanvulling met kwalitatieve gegevens om beslissingsprocessen en onderwijspraktijken optimaal te informeren. Ten vierde wijzen de auteurs op het belang van het trainen van technologische vaardigheden. Gebruikers die met data aan de slag gaan, moeten dashboards en andere databases en informatiesystemen kunnen hanteren. Om die reden staat een hands-on aanpak centraal waarbij de data actief geconsulteerd worden in alle initiatieven die we organiseren.

9.3.1. Doelstellingen en beoogde resultaten

Het steunpunt beoogt om individuen en teams op te leiden en te ondersteunen om de informatie over de prestaties van leerlingen op de centrale toetsen (schoolfeedback) te gebruiken om de kwaliteit van hun onderwijspraktijken te verbeteren. Verschillende soorten van ondersteuning en professionalisering voor diverse groepen van gebruikers op verschillende tijdstippen worden voorzien.

Vóór de afnames voorzien we in informerende kennisclips die duiding geven bij de doelen en de opzet van de toetsen en afnames, bedoeld voor een breed publiek van stakeholders. Deze worden aangevuld met aanbodgestuurde e-courses voor leerkrachten en schoolleiders die meer gedetailleerd duiding geven bij de doelen en opzet van de centrale toetsen. Vanaf de publicatie van de eerste resultaten voorzien we aanbodgestuurde e-tutorials om individuele gebruikers te introduceren in de werking van het feedbacksysteem en om gebruikers te ondersteunen bij een correcte interpretatie van figuren en tabellen. We blijven tevens werken met aanbodgestuurde e-courses voor leerkrachten en schoolleiders die meer gedetailleerd duiding willen bij de feedback van de centrale toetsen, de mogelijkheden en beperkingen van de resultaten en de vertaalslag naar de onderwijspraktijk. Binnen deze e-courses worden eveneens de ‘advanced functies’ uitgelegd en worden gebruikers geïnformeerd over GDPR-wetgeving. Er worden dus door het steunpunt geen systematische fysieke één op één of nascholingen in groep georganiseerd. We zetten ook in op vraaggestuurde blended pilootprojecten voor schoolteams die in samenwerking met minstens één externe partner zoals een lerarenopleiding, pedagogische begeleidingsdienst, CLB, ... via collaboratief onderzoek aan de slag willen met de leerlingoutputdata om de onderwijskwaliteit te verhogen. Het steunpunt voorziet een blueprint van hoe een traject eruit kan zien en voorziet een flankerend programma voor de deelnemers en coördinatoren van de projecten waarin ze gecoacht en begeleid worden. Die blueprint in geënt op academische inzichten inzake professionalisering van onderwijsprofessionals in het algemeen en inzake informatiegebruik in het bijzonder. Op die manier zetten we in op het principe van train-the-trainer en op het responsabiliseren van tussenstructuren voor de professionalisering op het vlak van datageletterdheid. Het steunpunt ziet organisaties als de pedagogische begeleidingsdiensten en private aanbieders van begeleiding als cruciale partners in het creëren van randvoorwaarden voor succesvol gebruik van de toetsinformatie. Hoewel er geen systematische training voorzien wordt van deze partners, zal het steunpunt de relaties met deze partners binnen dit werkdomein intensief behartigen. Dat vertaalt zich in een uitgesproken engagement om deze partners te informeren over

keuzes en activiteiten van het steunpunt en om met hen te dialogeren over hoe zij flankerend de doelstellingen van het steunpunt mee kunnen helpen realiseren.

Met de e-courses en e-tutorials kan een onbeperkt aantal gebruikers bereikt worden. Met de vraaggestuurde pilootprojecten waar een teamgerichte aanpak centraal staat wordt een beperkt aantal scholen bereikt. Bedoeling van de pilootprojecten is vooral om aan expertiseontwikkeling te doen en inspirerende 'good practices' te verzamelen van hoe een school de toetsresultaten duurzaam kan inbedden in de eigen PDCA-cyclus. Scholen kunnen zich kandidaat stellen voor een blended traject onder begeleiding van het steunpunt. Een zo diverse set van scholen wordt uit de kandidaturen geselecteerd om ook zo rijke en gevarieerde voorbeelden van good practices te produceren waar later een brede waaier van scholen zich in kan herkennen. De expertise die door het steunpunt geproduceerd wordt in de pilootprojecten kan door lerarenopleidingen en pedagogische begeleidingsdiensten ingezet worden. Ons inziens en mede op basis van de suggesties van eerder wetenschappelijk onderzoek zijn het de lerarenopleidingen en pedagogische begeleidingsdiensten die scholen via partnerschappen en collaboratief onderzoek kunnen ondersteunen bij het analyseren en betekenis geven van CT-resultaten om onderwijspraktijken te verbeteren.

De e-courses en de pilootprojecten worden volgens de principes van ontwerponderzoek op systematische basis geëvalueerd zodat de e-courses en de blueprint voor het collaboratief onderzoekstraject kunnen verbeterd worden. Ook wordt hier de link met het gebruikersonderzoek gelegd.

10. Referenties

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Agirdag, O. (2018): The impact of school SES composition on science achievement and achievement growth: mediating role of teachers' teachability culture, *Educational Research and Evaluation*, *24* (3-5): 264-276.

Bakes, B. & Cowan, J. (2018). Is the Pen Mightier Than the Keyboard? The Effect of Online Testing on Measured Student Achievement. Working Paper, National Center for Analysis of Longitudinal data in Education research.

Bellens, K., Arkens, T., Van Damme, J. & Gielen, S. (2013). *Sociale ongelijkheid en ongelijkheid op basis van thuistaal inzake wetenschapsprestaties in het Vlaamse onderwijs. Veranderingen tussen 2003 en 2011 op basis van TIMSS, vierde leerjaar*. Leuven: Centrum voor Onderwijseffectiviteit.

Berkowitz, R., H. Moore, R.A. Astor, R. Benbenishty (2017): A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement. *Review of Educational Research*, *87* (2): 425-469.

Bertrand, M., & Marsh, J. A. (2015). Teachers' Sensemaking of Data and Implications for Equity. *American Educational Research Journal*, *52*(5), 861–893. https://doi.org/10.3102/0002831215599251

Bosker, R. J., & Scheerens, J. (1997). *The foundations of educational effectiveness*. Elsevier.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics,* 40(1), 35–68. https://doi.org/10.3102/1076998614548485

Chick, H., & Pierce, R. (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development*, *15*(2).

Coburn, C. E. (2001). Collective Sensemaking about Reading: How Teachers Mediate Reading Policy in Their Professional Communities. *Educational Evaluation and Policy Analysis*, *23*(2), 145–170. https://doi.org/10.3102/01623737023002145

Creemers, B., & Kyriakides, L. (2015). Developing, testing, and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement*, *26*(1), 102–119. https://doi.org/10.1080/09243453.2013.869233

Culpepper, S.A. (2010): Studying individual differences in predictability with gamma regression and nonlinear multilevel models. *Multivariate Behavioral Research*, *45* (1): 153-185.

Cuttace, P. (1982). Essay Review. *Urban Education*, *16*(4), 483–492. https://doi.org/10.1177/004208598201600406

Danhier J. & É. Martin (2014): Comparing Compositional Effects in Two Education Systems: The Case of the Belgian Communities, *British Journal of Educational Studies*, *62* (2): 171-189.

Danhier, J. & Jacobs, D. (2017): *Segregatie in het onderwijs overstijgen. Analyse van de resultaten van het PISA 2015 onderzoek in Vlaanderen en in de Federatie Wallonië-Brussel*. Brussel: Koning Boudewijnstichting.

Datnow, A., Park, V., & Wohlstetter, P. (2007). Achieving with data: How high-performing School systems use data to improve instruction for elementary students. San Francisco: Center on Educational Governance University of California.

Day, J., & Bobeva, M. (2005). A generic toolkit for the successful management of delphi studies. *Electronic Journal of Business Research Methods*, *3*(2), 103–116.

Domina, T., A. Penner & E. Penner (2017): Categorical Inequality: Schools As Sorting Machines. *Annual Review of Sociology*, *43* (1): 311-330.

Eide, E., & Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, *58*(3), 345–350. https://doi.org/10.1016/S0165-1765(97)00286-3

Eizaguirre, S. (2019). Urban Education. In *The Wiley Blackwell Encyclopedia of Urban and Regional Studies* (pp. 1–5). American Cancer Society. https://doi.org/10.1002/9781118568446.eurs0433

Everson, J. (2017). The implications and impact of 3 approaches to health information exchange: community, enterprise, and vendor-mediated health information exchange. *Learning Health Systems, 1*(2), e10021. https://doi.org/10.1002/lrh2.10021

Faber, M., & Visscher, A. (2014). Leidt het gebruik van digitale leerlingvolgsystemen tot betere leerprestaties? 4W: weten wat werkt en waarom, 3, 14-21.

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: examining mode effects for computer-based assessment and im- plications for measuring trends. *Large-scale Assessments in Education, 6*(1), 11.

Goe, L. (2007). *The Link between Teacher Quality and Student Outcomes: A Research Synthesis*. Washington: ETS.

Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, *8*(4), 369–395. https://doi.org/10.1080/0924345970080401

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. Center for Assessment--NCLB growth update. Center for Assessment - NCLB growth update.

Gummer, E., & Mandinach, E. (2015). Building a conceptual framework for data literacy. *Teachers College Record, 117*(4), 1-22.

Harris, D. N. (2011). Value-added measures and the future of educational accountability. *Science, 333*(6044), 826–827. https://doi.org/10.1126/science.1193793

Hattie, J. (2009). Visible learning. A synthesis of over 800 meta-analyses relating to achievement. London: Rouledge.

Havermans, N., Wouters, T., & Groenez, S. (2018). *Schoolse segregatie in Vlaanderen: Evolutie van 2001- 2002 tot 2015-2016.* (SONO/2017.OL3.2/2; Steunpunt Onderwijsonderzoek, p. 72). SONO. http://steunpuntsono.be/wp-content/uploads/2018/10/SONO_2017.OL3_.2_2_finaal.pdf

Hellrung, K., & Hartig, J. (2013). Understanding and using feedback - A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190. https://doi.org/10.1016/j.edurev.2012.09.001

Henderson, J., & Corry, M. (2020). Data literacy training and use for educational professionals. *Journal of Research in Innovative Teaching & Learning,* 2397-7604.

Heymans, P. J., Godaert, E., Elen, J., van Braak, J., & Goeman, K. (2018). MICTIVO2018. Monitor voor ICT-integratie in het Vlaamse onderwijs. Eindrapport van O&O-opdracht: Meting ICT-integratie in het Vlaamse onderwijs (MICTIVO). KU Leuven / Universiteit Gent.

Jacobs, D., A. Rea & L. Hanquinet, L (2007): Prestaties van de leerlingen van buitenlandse herkomst in België volgens de PISA-studie: vergelijking tussen de Franse Gemeenschap en de Vlaamse Gemeenschap. Brussel: Koning Boudewijnstichting.

Jimerson, J. B. (2014). Thinking about data: Exploring the development of mental models for "data use" among teachers and school leaders. *Studies in Educational Evaluation*, *42*, 5–14. https://doi.org/10.1016/j.stueduc.2013.10.010

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112, 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th edition, pp. 17-64). Washington, DC: American Council on Education/Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.

Kavadias, D. (2008). *Syntheseverslag: Zoektocht naar de meest relevante indicatoren ter voorspelling van leerachterstand op een samengesteld bestand op basis van de attestendatabank.* (Intern Rapport in opdracht van het Ministerie van de Vlaamse Gemeenschap – Departement Onderwijs SEP-Rapport 2008/3; Identificatie van relevante kenmerken, constructie van indicatoren, ontwikkeling van meetinstrumenten & koppeling van databanken, p. 44). UA - IOIW.

Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to Promote Data Use for Instructional Improvement: Actions, Outcomes, and Lessons from Three Urban Districts. *American Journal of Education*, *112*(4), 496–520. https://doi.org/10.1086/505057

Ketelaar, E., Beijaard, D., Boshuizen, H. P. A., & Den Brok, P. J. (2012). Teachers' positioning towards an educational innovation in the light of ownership, sense-making and agency. *Teaching and Teacher Education*, *28*(2), 273–282. https://doi.org/10.1016/j.tate.2011.10.004

Kim, J. & K. Choi (2008): Closing the Gap: Modeling within-school Variance Heterogeneity in School Effect Studies. *Asia Pacific Education Review*, *9* (2): 206-220.

Knoester, M. & W. Au (2017): Standardized testing and school segregation: like tinder for fire? *Race Ethnicity and Education*, *20* (1): 1-14.

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

Koenker, R. (2017). Quantreg: Quantile regression. R package version 5.34. Retrieved from https://CRAN.R-project. org/package=quantreg

Konstantopoulos, S., Li, W., Miller, S., & van der Ploeg, A. (2019). Using Quantile Regression to Estimate Intervention Effects Beyond the Mean. *Educational and Psychological Measurement*, 79(5), 883–910. https://doi.org/10.1177/0013164419837321

Kyriakides, L., Creemers, B., & Panayiotou, A. (2020). Developing and Testing Theories of Educational Effectiveness Addressing the Dynamic Nature of Education. In J. Hall, A. Lindorff, & P. Sammons (Eds.), *International Perspectives in Educational Effectiveness Research* (pp. 33–69). Springer International Publishing. https://doi.org/10.1007/978-3-030-44810-3_3

Laurijssen, I. & Glorieux, I. (2020). *Sociale verschillen in vaardigheidsniveaus leerlingen die dezelfde studierichting volgden.* Steunpunt Onderwijsonderzoek, Gent.

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., R. French, C. Charlton & W. Browne (2014): Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, 39 (5): 307-332.

Lee, S., Kim, S. H., & Kwon, B. C. (2017). VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 551–560. https://doi.org/10.1109/TVCG.2016.2598920

Lee, Y. & J.A. Nelder (2006): Double hierarchical generalized linear models (with dis- cussion). *Applied Statistics*, *55*: 139-185.

Lester, H. F., Cullen-Lester, K. L., & Walters, R. W. (2019). From Nuisance to Novel Research Questions: Using Multilevel Models to Predict Heterogeneous Variances. *Organizational Research Methods*, 1094428119887434. https://doi.org/10.1177/1094428119887434

Lester, H.F., K.L. Cullen-Lester & R.W. Walters (2019): From Nuisance to Novel Research Questions: Using Multilevel Models to Predict Heterogeneous Variances. *Organization Research Methods*, Online First, DOI: 10.1177/1094428119887434.

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

Liu, H., J. Van Damme, S. Gielen & W. Van Den Noortgate (2015): School processes mediate school compositional effects: model specification and estimation. *British Educational Research Journal*, *41* (3): 423-447.

Logan, S. & R. Johnston (2010): Investigating gender differences in reading. *Educational Review*, *62* (2): 175-187.

Mandinach, E. (2012). A Perfect Time for Data Use: Using Data-Driven Decision Making to Inform Practice. Educational Psychologist, 47(2), 71-85.

Mandinach, E. B., & Gummer, E. S. (2012). Navigating the landscape of data literacy: It IS complex. San Francisco, CA: WestEd and Education Northwest.

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Mandinach, E. B., & Schildkamp, K. (2020). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*, 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. American Journal of Education, 112(4), 496-520.

Marx, D.M., S. J. Ko & R.A. Friedman (2009): The "Obama effect": How a salient role model reduces race-based performance differences. *Journal of Experimental Social Psychology*, *45* (4): 953-956.

März, V., & Kelchtermans, G. (2013). Sense-making and structure in teachers' reception of educational reform. A case study on statistics in the mathematics curriculum. *Teaching and Teacher Education*, *29*(1), 13–24. https://doi.org/10.1016/j.tate.2012.08.004

Miller, M.E. (2015, December 14). *How many service designers does it take to define service design* [Blogpost]. Retrieved from: https://blog.practicalservicedesign.com.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning, School Effectiveness and School Improvement, 25 (2), 231-256.

Nicols, S.L., & Berliner, D.C. (2008). Testing the joy out of learning. *Educational Leadership*, March 2008, 14-18.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? Educational Evaluation and Policy Analysis, 26(3), 237–257.

OECD. (2016a). *PISA 2015 Results (Volume 1)*. Paris: OECD Publishing. https://doi.org/10.1787/9789264266490-en.

OECD. (2016b). *Low-Performing Students: Why They Fall Behind and How To Help Them Succeed* Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264250246-en

Ogbu, J.U. (1987): Variability in minority school performance: A problem in search of an explanation. *Anthropology & Education Quarterly*, *18* (4): 312-334.

Opdenakker, M.-C. (2020). Three Decades of Educational Effectiveness Research in Belgium and the Netherlands: Key Studies, Main Research Topics and Findings. In J. Hall, A. Lindorff, & P.

Sammons (Eds.), *International Perspectives in Educational Effectiveness Research* (pp. 231–286). Springer International Publishing. https://doi.org/10.1007/978-3-030-44810-3_10

Pierce, R., & Chick, H. (2011). Teachers' intentions to use national literacy and numeracy assessment data: a pilot study. *The Australian Educational Researcher*, *38*(4), 433–447. https://doi.org/10.1007/s13384-011-0040-x

Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: a psychological perspective on the role of the teacher. *Educational Psychology*, *38*(6), 734–752. https://doi.org/10.1080/01443410.2018.1426834

Raudenbush, S.W. & A.S. Bryk (1987): Examining correlates of diversity. *Journal of Educational and Behavioral Statistics*, *12*: 241-269.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). Evaluation: A systematic approach. Thousand Oaks: Sage Publications.

Salvi del Pero, A. & A. Bytchkova (2013): *A Bird's Eye View of Gender Differences in Education in OECD Countries*, OECD Social, Employment and Migration Working Papers, No. 149, Paris: OECD Publishing.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement*, *24*(1), 1–38. https://doi.org/10.1080/09243453.2012.691100

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Schildkamp, K., & Lai, M. K. (2013). Conclusions and a Data Use Framework. In *Data-based Decision Making in Education* (pp. 177–191). Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_10

Schildkamp, K., & Poortman, C. L. (2015). Factors influencing the functioning of data teams. Teachers College Record, 117(4), 1-30.

Schildkamp, K., & Teddlie, C. (2008). School performance feedback systems in the USA and in The Netherlands: a comparison. *Educational Research and Evaluation*, *14*(3), 255–282. https://doi.org/10.1080/13803610802048874

Schildkamp, K., Lai, M.K., & Earl, L. (2012). Data-based decision making in education: challenges and opportunities. Dordrecht: Springer.

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2016). Factors promoting and hindering data-based decision making in schools. School Effectiveness and School Improvement.

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement*, *28*(2), 242–258. https://doi.org/10.1080/09243453.2016.1256901

Schleicher, A. (2018). *World Class: How to Build a 21st-Century School System, Strong Performers and Successful Reformers in Education*. OECD Publishing, Paris, https://doi.org/10.1787/9789264300002-en.

Sirin, S. R. (2005): Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, *75* (3): 417-453.

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE.

Strickdorn, M., Hormess, M., Lawrence, A., & Schneider, J. (2018). *This is service design doing. Applying service design thinking in the real world: A practitioners' handbook.* Sebastopol: O'Reilly Media.

Tashakkori, A., & Teddlie, C. (2002). Handbook of Mixed Methods in Social & Behavioral Research. Thousand Oaks, CA: Sage Publications.

Timmermans, A. & S. Thomas (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.

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

Van Droogenbroeck, F., Spruyt, B., Lemblé, H., Bongaerts, B., Siongers, J., & Kavadias, D. (2020). *TALIS 2018 Vlaanderen - Volume II*. Brussel: VUB.

Van Ewijk, R. & P. Sleegers (2010a): The effect of peer socioeconomic status on achievement: A meta-analysis. *Educational Research Review*, *5* (2): 134-150.

Van Gasse, R., Vanhoof, J., Mahieu, P., & Van Petegem, P. (2015). Informatiegebruik door schoolleiders en docenten. Antwerpen: Garant.

Van Gasse, R., Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2017). The impact of collaboration on teachers' individual data use. *School Effectiveness and School Improvement*, *28*(3), 489–504. https://doi.org/10.1080/09243453.2017.1321555

Vandenbroeck, M., Vanlaar, G., Bellens, K., Van Damme, J., & De Fraine, B. (2016). *Het Vlaams lager onderwijs in TIMSS 2015: Wiskunde en wetenschappen in internationaal perspectief en in vergelijking met vorige deelnames*. Leuven: KU Leuven, Centrum voor Onderwijseffectiviteit en -evaluatie.

Vanhoof, J., De Maeyer, S., Van Petegem, P., Penninckx, M., & Quintelier, A. (2016). Scenario's voor leer(winst)monitoring in Vlaanderen: Een ontwerponderzoek naar haalbaarheid en wenselijkheid. Universiteit Antwerpen, eindrapport OBPWO 13.03.

Vanlommel, K., & Schildkamp, K. (2019). How Do Teachers Make Sense of Data in the Context of High-Stakes Decision Making? *American Educational Research Journal*, *56*(3), 792–821. https://doi.org/10.3102/0002831218803891

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*(March 1994), 75–83. https://doi.org/10.1016/j.ijer.2017.02.013

Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2017). Data use by teachers: the impact of motivation, decision-making style, supportive relationships and reflective capacity. Educational studies, 42(1), 36-53.

Vanthournout, G., Donche, V., Speltincx, G., Gijbels, D., & Van Petegem, P. (2012). One size fits all? Feedback op leer- en motivatiekenmerken bij de instroom van de lerarenopleiding. *VELON: Tijdschrift voor Lerarenopleiders, 33*(4), 35-43.

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: perceptions of primary school principals. School Effectiveness and School Improvement, 21(2), 167-188.

Visscher, A. J. (2002). A Framework for Studying School Performance Feedback Systems. In A. J. Visscher & R. Coe (Eds.), *School Improvement through Performance Feedback* (pp. 41–72). Swets & Zeitinger.

Visscher, A. J., & Coe, R. (Eds.). (2002). School improvement through performance feedback. Lisse, Nederland: Swets & Zeitlinger.

Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the Data-Informed District. School Effectiveness and School Improvement. An International Journal of Research, Policy and Practice, 23(2), 159-178.

Wenger, M., H. Gärtner & M. Brunner (2020): To what extent are characteristics of a school's student body, instructional quality, school quality, and school achievement interrelated? *School Effectiveness and School Improvement*. Online First, DOI: 10.1080/09243453.2020.1754243.

Wu, A. D., & Zumbo, B. D. (2007). Understanding and Using Mediators and Moderators. *Social Indicators Research*, *87*(3), 367. https://doi.org/10.1007/s11205-007-9143-1

Xie, G., & Zhang, Y. (2020). School of golden touch? A study of school effectiveness in improving student academic performance. *The Journal of Chinese Sociology*, *7*(1), 7. https://doi.org/10.1186/s40711-020-00118-7

Zehner, F., Kroehne, U., Hahnel, C. & Goldhammer, F. (2020). PISA reading: Mode effects unveiled in short text responses. *Psychological Test and Assessment Modeling, 62* (1), 85-105.