Pedagogisch-psychometrische aspecten van centraal toetsen in Vlaanderen

Haalbaarheidsstudie – Perceel 1

Stichting Cito

Onderzoek, Kennis & Innovatie

Haalbaarheidsstudie – Perceel 1
Bas Hemker
Remco Feskens
Marieke van Onna
Iris Smits

Met bijdragen van
Anton Béguin
Hendrik Straat
Saskia Wools

Inhoudsopgave

Managementsamenvatting

In deze haalbaarheidsstudie naar de pedagogische-psychometrische aspecten van de introductie van gecentraliseerde proeven in Vlaanderen uitgewerkt. We gaan we in op de vraag wat er nodig is om centrale toetsen vorm te geven.

Voordat men een kwalitatief goede toets kan ontwikkelen moeten eerst de uitgangspunten van de toets duidelijk zijn. De uitgangspunten van een toets zijn gedefinieerd rondom een drietal vragen. Dat betreft de vragen *wat* we willen meten, *wie* we willen meten en *waarom* we willen meten. Wat we willen meten heeft betrekking op de vaardigheden wiskunde en Nederlands, met in ieder geval de onderdelen begrijpend lezen, schrijven en grammatica. Door bij de omschrijving van de toetsinhoud aan te sluiten bij de bestaande eindtermen en curricula, krijgen de toetsen en de daaropvolgende resultaten betekenis voor het onderwijs. Wie we willen meten heeft betrekking op de doelgroep en op de niveaus waarover men uitspraken wil doen. De doelgroep aan wie de toetsen worden voorgelegd zijn leerlingen in het 4de en 6de leerjaar van het lager onderwijs, en het 2de en 6de jaar secundair onderwijs (einde eerste en derde graad). De proeven worden opgezet om uitspraken te doen op leerling, klas, docent of schoolniveau of over het onderwijssysteem als geheel.

Waarom er getoetst wordt verwijst naar het doel van de toetsen. De centrale toetsen moeten kwaliteitsvol onderwijs in Vlaanderen helpen te beschermen en te ondersteunen, door beleid meer te kunnen onderbouwen op basis van objectieve informatie. Dit doel kent twee aspecten: kwaliteitsbewaking en kwaliteitsbevordering. Het eerste aspect is wat we noemen de summatieve functie van de toetsen: we willen na afloop van het gegeven onderwijs kijken welke van de onderwijsdoelen behaald zijn. Als hier belangrijke consequenties aan hangen voor de betrokkenen, dan wordt het voor hen high-stakes toetsen. Bij high-stakes toetsen willen deelnemers zo goed mogelijk presteren. Aan de ene kant levert dit gemotiveerde betrokkenen, aan de andere kant kan dit tot mogelijke fraude leiden. Het tweede aspect is wat we de formatieve functie van toetsen noemen: op

basis van de toetsgegevens willen we de betrokkenen helpen te verbeteren. Fouten worden vooral gezien als gelegenheid om van te leren. Merk op dat het belang voor verschillende betrokkenen kan verschillen: leerlingen kunnen een toets niet als belangrijk voor henzelf zien, terwijl de toetsresultaten voor een school wel belangrijk kunnen zijn.

Tussen deze aspecten moet een balans gevonden worden. Doordat motivatie zo'n impact heeft op de prestaties, en we de scholen willen kunnen vergelijken, moet de motivatie van de leerlingen zoveel mogelijk "gestandaardiseerd" worden. Dat betekent dat binnen een schooljaar de waarde van de toets voor iedere leerling hetzelfde moet zijn, terwijl ondertussen de toetsen, zoals afgesproken, niet een te zwaar gewicht mogen krijgen voor de individuele leerlingen. Daartoe worden in dit rapport enige suggesties gedaan. Merk ook op dat als in het begin van de invoering het formatieve aspect benadrukt wordt dit kosten scheelt die anders aan fraudebestrijding besteed moeten worden. Ook kan dat helpen de toetsen meer geaccepteerd te krijgen in het onderwijsveld. In een later stadium kunnen de toetsen alsnog aan summatief belang winnen. Dat is makkelijker bij toetsen waar het onderwijsveld aan gewend is.

Binnen het perceel zijn een aantal specifieke vragen gesteld rond de thema's leerwinst, toetsontwikkeling en rapportage. In deel II van de dit rapport (Hoofdstuk 4, 5 en 6) worden deze een voor een behandeld. In Hoofdstuk 7 worden de antwoorden nog eens vanuit een praktisch kader beschouwd, en is de kernvraag bij een aantal dilemma's: hoe kunnen we dit aanpakken bij de invoer van centrale proeven in Vlaanderen?

Leerwinst en toegevoegde waarde modellen kunnen zeer waardevolle informatie opleveren voor leerlingen, ouders, scholen en beleidsmakers om het onderwijsproces te ondersteunen. Zodra leerwinst en toegevoegde waarde uitkomsten echter ingezet worden als afreken- of beoordelingsinstrument, kan dat snel averechtse effecten oproepen. Bij leerwinst zijn de belangrijkste uitdagingen dat tussen meetmomenten de eindtermen en samenstelling van de groep leerlingen verandert. Voor beide zaken worden oplossingen gegeven. Een oplossing die beide uitdagingen aanpakt, is een verschuiving van focus van leerwinst naar de evaluatie van de school per afname. In de verschillende modellen spelen de eerder behaalde resultaten nog wel steeds een belangrijke rol, naast andere factoren die toegevoegde waarde en de kwaliteit van de school kunnen kenmerken. Afhankelijk van het doel, de doelgroep en de beschikbare achtergrondinformatie kan voor een model gekozen worden.

In dit rapport bespreken we diverse aspecten die van belang zijn rondom toetsontwikkeling. Er worden suggesties gegeven hoe om te gaan met toetsverversing en het bekend raken van opgaven. Daarnaast wordt ingegaan op het gebruik van verschillende toetsversies. Het verkrijgen van vergelijkbare resultaten ondanks het gebruik van meerdere toetsversies is bij het gebruik van IRT te realiseren, gebruik makend van slimme test designs. Als we hierbij kijken naar de rotatie van eindtermen, zal de optimale oplossing afhangen of de focus vooral ligt op het vergelijken van scholen of van leerlingen. Er worden suggesties gedaan voor brede afnamen en afwijkingen van de gestandaardiseerde afname. Dat kan ook papieren afnamen, die mogelijk in de eerste jaren van de afnamen nodig zou kunnen zijn om problemen met digitale afnamen te mitigeren. Het belangrijkste is dat bij afwijkingen van de standaard digitale afname, equivaleringsonderzoek noodzakelijk is, om te controleren of de resultaten verschillende typen afnamen wel vergelijkbaar zijn. Daarnaast gaan we in op adaptieve toetsen waar leerlingen items krijgen voorgelegd die

aansluiten bij het niveau dat ze op een eerder onderdeel van de toets hebben laten zien. Tot slot komen onderwerpen aan bod zoals de betrouwbaarheid van de toetsen, en hoe om te gaan met externe toetsinformatie. Bij dat laatste onderwerp wordt de meerwaarde benadrukt van externe toetsen voor moeilijk meetbare vaardigheden.

Het laatste thema betreft de rapportages. We bespreken de mogelijke vormen van rapportages en de impact van die vormen. Dit wordt geïllustreerd door middel van voorbeelden. Een belangrijk onderscheid is te maken tussen relatieve normen, met een sterke summatieve, high-stakes associatie, en absolute normen, met een formatieve associatie. Bij een relatieve normering wordt de prestatie van een leerling gerelateerd aan de prestaties van andere leerlingen. Bij een absolute normering, wordt op basis van expertonderzoek de prestatie van een leerling gerelateerd aan een inhoudelijke standaard, gebaseerd op het te verwachten niveau van de leerling. In dergelijke rapportages kan de vaardigheid ook gerelateerd worden aan de opgaven in de itembank, om zo een inhoudelijke betekenis te geven aan de vaardigheidsschaal.

Een kernboodschap is dat het belangrijk is het veld nauw en grootschalig te betrekken bij de ontwikkeling van rapportages. We onderstrepen het belang van relevante en begrijpelijke rapportages van toetssystemen. We stellen een werkwijze voor die ondersteunt dat de wijze waarop de resultaten gepresenteerd worden aansluit bij de gebruikers van de rapportages. Daartoe moet bijvoorbeeld bekend zijn wat de leerkrachten en scholen willen weten, en hoe ze de toetsresultaten gebruiken om de kwaliteit van het onderwijs verbeteren. De ontwikkeling van rapportages, in samenwerking met de doelgroep, is minstens zo belangrijk als de ontwikkeling van de toetsen, terwijl dit vaak effectief in besteedde tijd en geld achterblijft. Het voordeel van het grootschalig met het veld samen ontwikkelen van rapportages is dat deze dan aangepast zijn op het kennisniveau van de lezer, en op de beoogde vervolghandelingen. Dit betekent dat het minder moeite kost om de toetsen sneller geaccepteerd te krijgen en om leerkrachten vervolgens toetsen goed te laten gebruiken voor kwaliteitsverbetering van hun onderwijs. Deels gerelateerd hieraan worden ook suggesties gegeven om ranking tegen te gaan.

Tot slot, in deze haalbaarheidsstudie naar de pedagogische-psychometrische aspecten van de introductie van gecentraliseerde proeven in Vlaanderen trachten we een inzicht te geven in het brede palet van mogelijkheden dat te vinden is bij het ontwerp van toetsen en examens. We hopen met de voorbeelden die we geven te inspireren bij de keuze van een haalbare strategie.

1. Introductie

Dit rapport betreft de resultaten van het pedagogisch-psychometrisch perceel (perceel 1) van de haalbaarheidsstudie aangaande het introduceren en toepassen van gestandaardiseerde, gevalideerde, genormeerde netoverschrijdende proeven in Vlaanderen. In dit rapport gebruiken we hiervoor vaak de kortere term “centrale toetsen”, die het gecentraliseerde karakter van de productie en afname van de toetsen weerspiegelt. Dit perceel is een van de drie percelen aangaande de haalbaarheidsstudie. De overige twee percelen betreffen organisatorische aspecten (perceel 2) en technisch-juridische aspecten (perceel 3), met ieder een eigen rapport. De haalbaarheidsstudie moet antwoord geven op wat er nodig is om vanaf het schooljaar 2022-2023 te starten met het afnemen van deze toetsen waarbij uiteindelijk jaarlijks zo’n 300.000 leerlingen verschillende toetsen moeten afleggen.

1.1 Aanleiding en omschrijving van beoogde opzet van de centrale toetsen

De vraag naar een haalbaarheidsstudie aangaande centrale toetsen kan in een brede context geplaatst worden. In de afgelopen jaren is Vlaanderen geconfronteerd met een onverwachte en sterke achteruitgang in lezen, wiskunde en wetenschappen op basis van diverse nationale en internationale peilingen waaronder PIRLS 2016 en PISA 2018 (Beleidsnota, 2019)[1]. Deze bevindingen werden de aanleiding voor de Vlaamse regering om in het regeerakkoord (2019)[2] het doel te verankeren om te streven naar excellent onderwijs.

De vrijheid van onderwijs in Vlaanderen is een groot goed en biedt scholen de ruimte om onderwijsdoelen[3] op eigen wijze te realiseren en een eigen pedagogisch beleid te

[1] Minister van Onderwijs, Sport, Dierenwelzijn en Vlaamse Rand (2019). Beleidsnota Onderwijs 2019-2024 (`https://www.vlaanderen.be/publicaties/beleidsnota-2019-2024-onderwijs`) – zie ook `https://onderwijs.vlaanderen.be/nl/nieuwe-pisa-en-pirls-repeat-resultaten-voor-vlaanderen`.

[2] Departement Kanselarij en bestuur (2019). Vlaamse regering 2019-2024. Regeerakkoord. `https://www.vlaanderen.be/publicaties/regeerakkoord-van-de-vlaamse-regering-2019-2024`.

[3] `https://onderwijsdoelen.be`

voeren. Een voorwaarde blijft echter wel dat zij zo goed mogelijk, bij voorkeur excellent onderwijs, realiseren[4]. Iedere leerling moet de kansen krijgen om de beste versie van zichzelf te worden en daarom is een belangrijke pijler van het streven van de Vlaamse regering om talenten te ontwikkelen.

Om na een grondige hervorming[5] meer grip te krijgen op de onderwijskwaliteit is de Vlaamse regering van plan om een pakket aan maatregelen door te voeren waaronder het ambitieus en op grote schaal aanscherpen van de eindtermen[6] van het leerplichtonderwijs[7]. Het inzichtelijk maken van de beheersing van die eindtermen vindt dan plaats door middel van gestandaardiseerde net- en koepeloverstijgende proeven.

Het is de bedoeling om met deze centrale toetsen leerwinst van alle leerlingen objectief in kaart te brengen. Het is tevens de bedoeling dat de onderwijswereld hiermee in de spiegel kan kijken en dat het mogelijk wordt met deze informatie de kwaliteit van het Vlaamse onderwijs op een hoog peil te houden en gericht te verbeteren[8]. Deze objectieve informatie helpt de Vlaamse overheid om kwaliteit van het onderwijs te monitoren door middel van leerwinst, en daarmee beleid te onderbouwen en waar nodig bij te sturen. Het helpt de scholen de kwaliteit van het eigen onderwijs te monitoren via leerwinst van leerlingen. Tenslotte helpt het leerlingen om aan te tonen dat zij de passende kennis en competenties hebben, hetgeen hen kan helpen voor de doorstroom binnen het onderwijs. Om deze informatie te verkrijgen zijn de meetdoelen van de instrumenten[9]:

1. Het meten van het bereiken van de eindtermen
2. Het in kaart brengen van de leerwinst van leerlingen
3. Het in kaart brengen van de leerwinst van scholen

Naast deze proeven nemen Vlaamse onderwijsinstellingen deel aan internationaal vergelijkend onderzoek om de kwaliteit van het Vlaams onderwijs permanent te kunnen monitoren. Vlaanderen is voornemens deel te nemen aan PIRLS in 2021, het PIAAC onderzoek in 2021-2022 en de volgende rondes van PISA in 2022 en 2025. Het beleid is zo ingericht dat de centrale toeten in eerste instantie Nederlands (begrijpend lezen, schrijven, grammatica) en wiskunde meten. Alle scholen dienen de proeven af te nemen bij alle leerlingen op twee momenten in het lager onderwijs (het 4de en 6de jaar van het basisonderwijs) en op twee momenten in het secundair onderwijs (aan het einde van de eerste en de derde graad van het secundair onderwijs[10]; het 2de en 6de jaar). Tevens is gesteld dat de gestandaardiseerde proeven moeten worden afgenomen bij zowel alle netten zoals gemeenschapsonderwijs (GO), officieel gesubsidieerd onderwijs (OGO), en vrij

[4] In het regeerakkoord (zie noot 2) is dit op pagina 13 beschreven: “De vrijheid van onderwijs blijft een belangrijk uitgangspunt. De overheid bepaalt wat de leerlingen moeten kennen en kunnen, de scholen en leerkrachten bepalen hoe ze dit pedagogisch realiseren.”

[5] `https://onderwijs.vlaanderen.be/nl/van-29-studiegebieden-naar-8-studiedomeinen-in-2e-en-3e-graad-secundair-onderwijs`

[6] `https://onderwijs.vlaanderen.be/nl/onderwijsdoelen`; zie ook Beleidsnota Onderwijs 2019-2024, OD 1.2 “Eindtermen scherp, duidelijk en uitdagend formuleren”.

[7] In Nederland wordt hiervoor de term funderend onderwijs gebruikt.

[8] `https://www.benweyts.be/Alle-Vlaamse-leerlingen-krijgen-centrale-toetsen`

[9] Regeerakkoord (noot 2) - paragraaf 1.2.1; en Beleidsnota Onderwijs 2019-2024 (noot 1) – OD 1.3.

[10] Beleidsnota Onderwijs 2019-2024 – OD 1.3.

gesubsidieerd onderwijs (VGO) alsmede alle koepels binnen deze netten[11]. De resultaten van de toetsen moeten worden teruggekoppeld aan scholen, overheid en onderzoekers. Aan scholen moeten de resultaten op zowel leerling- als schoolniveau worden teruggekoppeld. Aan de overheid en voor onderzoekers moeten de resultaten in geanonimiseerde vorm op individueel niveau worden teruggekoppeld. Expliciet is aangegeven dat het niet het doel is om scholen te classificeren[12]. De resultaten van de proeven zullen worden gebruikt om scholen waarvan de leerlingen significant minder leerwinst genereren een vrij te kiezen (verplicht) begeleidingstraject aan te bieden om de kwaliteit van hun onderwijs te vergroten. In Hoofdstuk 2 wordt dieper ingegaan op het doel van de toetsen en het gebruik van de resultaten op leerling-, leraar-, school- en systeemniveau.

1.2 Opzet van het rapport

In dit rapport zullen we met de bovengenoemde gestelde kaders en wensen ten aanzien van centrale toetsen in Vlaanderen een antwoord trachten te geven op de vraag wat er nodig is om centrale toetsen vorm te geven. We starten in Hoofdstukken 2 en 3 met een algemeen kader. In Hoofdstuk 2 gaan we in op het gebruik van toetsen, i.e., verschillende doelen van toetsen die met elkaar in balans gebracht moeten worden. In Hoofdstuk 3 geven we een psychometrisch kader. We gaan dieper in op itemresponstheorie (IRT), wat het meest geschikt lijkt voor de aanpak van veel van de uitdagingen die de haalbaarheid van de centrale toetsen heeft, en waar in latere hoofdstukken naar gerefereerd zal worden.

De uitdagingen voor het beleid krijgen in het bestek dat vooraf ging aan dit onderzoek vorm door middel van de vragen. Binnen het pedagogisch-psychometrisch perceel van het onderzoek zijn deze geordend onder drie thema's: leerwinst, toetsontwikkeling en omgaan met resultaten. In dit rapport zijn de vragen uit het bestek in Hoofdstukken 4, 5 en 6 per thema beantwoord, zo nodig aangevuld met antwoorden op enkele andere relevante vragen. Bij het beantwoorden van de vragen wordt ingegaan op de voor- en de nadelen van de diverse scenario's die gerelateerd zijn aan een vraag. Bij de voor- en de nadelen worden ook de tijd en de kosten per optie meegenomen.

Bij de beantwoording zal de theoretische verhandeling gerelateerd worden aan de Vlaamse onderwijspraktijk. Daar waar dit aan de orde is, bespreken we ook de relatie met de twee andere percelen die onderdeel uitmaken van de haalbaarheidsstudie. We zien het eerste perceel als het pedagogisch-psychometrisch fundament voorwaardelijk voor de beantwoording van de meer toegepaste vragen op het organisatorische en technisch-juridische vlak. Merk op dat de twee andere percelen evenzeer richtinggevend zijn voor het eerste perceel omdat het pedagogisch-psychometrisch fundament optimaal toepasbaar moet zijn. Er is dus sprake van sterke synergie tussen de drie percelen.

Ook wordt het bestek aangaande de inrichting van het steunpunt "Ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde net- en koepeloverschrijdende toetsen in Vlaanderen" relevant geacht voor de beantwoording van de vragen. De vragen waar de haalbaarheidsstudie een antwoord op moet geven en de daaruit volgende scenario's hebben

[11] https://onderwijs.vlaanderen.be/nl/officieel-en-vrij-onderwijs-onderwijsnetten-en-koepels#:~:text=In%20het%20officieel%20onderwijs&text=Het%20gesubsidieerd%20officieel%20onderwijs%20omvat,zijn%20verenigd%20in%202%20koepels%3A&text=Provinciaal%20Onderwijs%20Vlaanderen%20(POV)

[12] Regeerakkoord (noot 2) - paragraaf 1.2.1; Beleidsnota Onderwijs 2019-2024 (noot 1) – OD 1.3.

kennelijk ook invloed op de werkzaamheden van dit steunpunt. Aan de andere kant, en mede daardoor, geldt ook dat wat het steunpunt dient uit te voeren volgens dat bestek relevant is voor deze haalbaarheidsstudie. Zodoende is ook het bestek van die opdracht meegenomen in de aanpak.

Na de drie hoofdstukken aangaande de drie thema's volgt Hoofdstuk 7 waarin de scenario's die benoemd worden onder de vragen bij de drie thema's in samenhang terugkomen.

Deze scenario's hebben tot doel de keuzes die te maken zijn wetenschappelijk te ondersteunen, en de gevolgen ervan in kaart te brengen. Op basis van de keuzen die volgen naar aanleiding van de drie rapporten aangaande de haalbaarheidsstudie, zal het steunpunt een uiteindelijke praktische invulling geven aan de centrale toetsen.

Bij de bespreking van de scenario's in Hoofdstuk 7 komen ook de voor- en nadelen opgesomd terug, evenals aspecten aangaande de kosten. Deze zijn mede afhankelijk van de randvoorwaarden die ook benoemd worden. De risico's en de kansen per scenario worden beschreven, inclusief de effecten van de scenariokeuze op elementen als de vertrouwelijkheid van het toetsmateriaal, (het voorkomen van) strategisch gedrag en toetsfraude. Ook wordt ingegaan op de tijdsinvestering en de doorlooptijd.

We starten het rapport zoals gesteld met het inleidende Hoofdstuk 2 over het gebruik van toetsen. De thema's leerwinst, toetsontwikkeling en omgaan met resultaten hebben onderling een sterke samenhang daar waar het gaat om het onderliggende doel van centraal toetsen. Vragen die over deze thema's beantwoord dienen te worden, zijn: Wat willen we meten, hoe doen we dat, en wat zegt dat op leerling-, school- en regioniveau? Wat zijn de doelstellingen rond het centraal meten? Hoe gaan we de meting en de daaruit volgende observaties over leerwinst gebruiken? Hoe zorgen we dat het beoogde gebruik ook het toegepaste gebruik is? Welk misbruik van de toets moet zeker vermeden worden en hoe doen we dat? Wat is de rol van de rapportage hierin, en wat zijn de gevolgen hiervan voor de toetsontwikkeling? Leidend in al deze vragen zijn de doelen van centraal toetsen. Deze doelen zijn divers en kunnen elkaar mogelijk tegenwerken of tot duurdere scenario's leiden. In Hoofdstuk 2 gaan we in op deze doelen en mogelijke tegenstrijdigheden. Hier wordt het kader geschetst waar de rest van het rapport op stoelt, omdat het toetsdoel gevolgen heeft voor zowel de beantwoording van de vragen per thema, als de scenario's die daaruit volgen. Daarom starten we dit rapport met een hoofdstuk over het gebruik van toetsen.

I Achtergrond

2. Gebruik van toetsen in balans

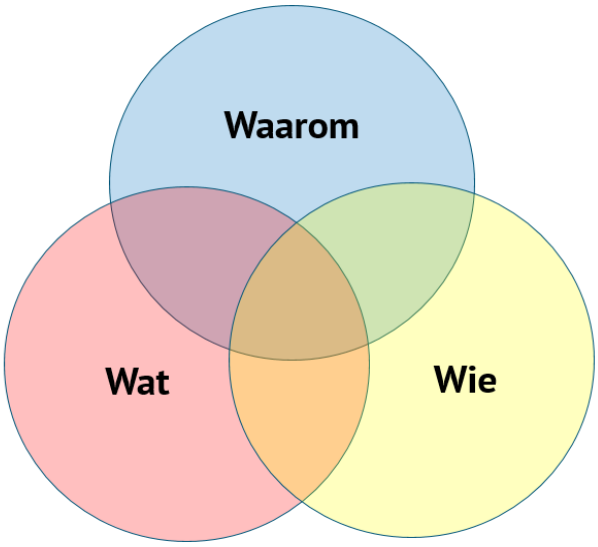
Voordat men een kwalitatief goede toets kan ontwikkelen, moeten eerst de uitgangspunten van de toets duidelijk zijn. Voor instanties die toetsen ontwikkelen is dit punt cruciaal, omdat de aard en de opzet van het ontwikkelingsproces van de toetsen voor een belangrijk deel volgen uit deze uitgangspunten. Het is evident dat ook voor het opzetten van een systeem met centrale toetsen de vragen die de kern vormen voor de uitgangspunten de basis zijn van de inrichting van de toetsen en de scenario's.

De vraag of uitgangspunten van de toets goed uiteen zijn gezet, is een zeer belangrijk criterium om de kwaliteit van toetsen te bepalen. Het is ook het eerste criterium waarnaar gekeken wordt in beide beoordelingssystemen die binnen Nederland formeel gebruikt worden om tests en toetsen te evalueren, met name het COTAN beoordelingssysteem[1] en het RCEC beoordelingssysteem[2]. Als het doel niet duidelijk is, dan wordt het ook moeilijk om andere kwaliteitscriteria zoals betrouwbaarheid en validiteit goed te kunnen beoordelen. De uitgangspunten van een toets zijn gedefinieerd rondom een drietal vragen. Dat betreft de vragen waarom we willen meten, wie we willen meten en wat we willen meten.

In eerste instantie zijn dergelijke vragen zeer simpel te beantwoorden, zo ook in het geval van de centrale proeven in Vlaanderen. De aanleiding om proeven te ontwikkelen – het waarom – is de wens om meer grip te krijgen op de onderwijskwaliteit. Centrale proeven hebben als doel om het behalen van nieuw ingevoerde eindtermen te kunnen

[1] Evers, A., Lucassen, Meijer, R. & Sijtsma, K. (2010). COTAN beoordelingssysteem voor de kwaliteit van tests. Amsterdam: NIP/COTAN. Dit beoordelingssysteem is ontwikkeld voor de evaluatie van de kwaliteit van psychologische tests in het Nederlandse taalgebied, dus ook voor gebruik in Vlaanderen. Het wordt ook toegepast voor formele evaluatie van onderwijskundige toetsen. Zie ook: `https://www.psynip.nl/uw-beroep/cotan/`.

[2] Sanders, P.F., Brouwer, A., Eggen, T. & Veldkamp, B. (2018). RCRC beoordelingssysteem voor de kwaliteit van studietoetsen en (praktijk) examens. Enschede: RCEC. Dit beoordelingssysteem is gebaseerd op het COTAN-systeem, maar richt zich specifiek op de beoordeling van toetsen in het onderwijs. Zie ook: `www.rcec.nl`.



Figuur 2.1: Uitgangspunten van een toets

controleren en om de leerwinst voor leerlingen en scholen in kaart te brengen. Wie we meten is aangegeven als alle leerlingen in het Vlaamse onderwijs die in het vierde en het zesde leerjaar zitten van het basisonderwijs, de leerlingen aan het einde van de eerste graad van het Secundair Onderwijs en de leerlingen aan het einde van het Secundair Onderwijs. Wat gemeten wordt, wordt reeds beschreven als de vakken Nederlands (begrijpend lezen, schrijven en grammatica) en wiskunde.

In dit hoofdstuk wordt nader ingegaan op overwegingen die een rol spelen bij de vragen waarom, wie en wat wij toetsen. In de aanbestedingstekst zijn de wie, wat en waarom in grote lijnen al benoemd en daarom kaderstellend. In dit inleidende hoofdstuk beginnen wij met het neerzetten van een bredere visie op meten. Daarbij komen achtereenvolgens de vragen over het waarom van het meten, wie het onderwerp van de meting is en wat er gemeten wordt aan bod. Deze drie vragen hebben een onderlinge samenhang. Als we het hebben over het niveau van de opgaven gaat het zowel over wat we meten, wie we meten maar ook waarom we meten. Als we kijken naar de rapportagevorm komen al deze drie zaken ook samen, hetgeen ook geldt voor de betrouwbaarheid en de validiteit van de toetsen. Zoals diverse beoordelingssystemen ook al doen, is het afzonderlijk benaderen van deze vraagstukken zinvol om er ook zeker van te zijn dat ieder van deze vragen beantwoord kan worden. Dat gebeurt in algemene zin in Sectie 2.1 en geeft het kader van overdenkingen rondom de toetsen. In de tweede paragraaf van dit hoofdstuk worden de drie vragen verder beschouwd met het zicht op de haalbaarheid van centrale proeven in Vlaanderen waarbij we ook verder ingaan op de kritiekpunten op (centraal) toetsen. Dit betreffen zaken waarmee rekening gehouden moet worden bij het ontwikkelen van de scenario's.

2.1 Een kader voor het vormgeven van toetsen

2.1.1 Het toetsdoel: Waarom willen wij meten?

Het introduceren van een toets op enig moment in de onderwijspraktijk vloeit voort uit een informatiebehoefte waarbij de informatie zodanig is dat er ook op gestuurd kan worden. De grootste uitdaging in het toetsontwerp is om te borgen dat de prestaties die de leerlingen neerzetten daadwerkelijk het construct reflecteren waarin wij geïnteresseerd zijn[3]. Dit heeft te maken met validiteit. De klassieke benadering maakt een onderscheid tussen diverse vormen van validiteit, zoals inhoudsvaliditeit, constructvaliditeit, *face validity* of criteriumvaliditeit. De modernere opvatting van dit kernbegrip in de wereld van het testen is dat validiteit meer als eenheid wordt beschouwd[4], met speciale aandacht voor de elementen die validiteit kunnen bedreigen[5]. Vervolgens worden in een validiteitsonderzoek bewijzen verzameld dat de toets aan het doel van validiteit voldoet[6].

Een zeer praktische definitie van validiteit die ook binnen deze huidige beschouwing past is "de mate waarin de test[7] aan zijn doel beantwoordt" (Drenth & Sijtsma, 2006)[8]. Het is dus belangrijk dat het doel van de toetsen goed omschreven is. Om dit validiteitsvraagstuk goed te kunnen beantwoorden, is de eerste stap in het opstellen van het toetsontwerp het stellen van de vraag waarom we willen meten: wat is het doel van het inwinnen van informatie?

In grote lijnen zijn er twee verschillende doelen om een toets af te nemen[9]. Het eerste doel is om informatie te verkrijgen die steun biedt aan het leerproces. Het tweede doel is om informatie te verkrijgen over waar de leerling staat na afronding van het leerproces. Hoewel deze toetsdoelen duidelijk verschillen van aard, bedienen toetsen en examens in de praktijk vaak een combinatie van beide functies. Het is echter altijd van belang om goed in het oog te houden dat de verschillende gebruiken van dezelfde toetsresultaten in balans zijn, zodanig dat voor beide doelstellingen betekenisvolle informatie uit de toetsresultaten onttrokken kan worden[10].

Toetsen die ons ter ondersteuning van het leerproces inzicht moeten geven in het niveau van leerlingen noemen wij formatief[11]. Deze toetsen zijn dan onderdeel van een formatief proces waarin informatie over het leren wordt verkregen met als doel om het leren verder te verbeteren. Gardner beschrijft dit als "het proces van het zoeken naar en interpreteren

[3] Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, 50, pp. 1-73.

[4] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

[5] Crooks, T.J. & Kane, M.T. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, 3, 3.

[6] Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.), *Educational Measurement 4th ed.*, pp. 17-64. Westport, CT: American Council on Education/Praeger.

[7] Lees ook: proef, toets, examen of assessment.

[8] Drenth, P. J.D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen (4e herziene druk) [Test theory. Introduction in the theory and application of psychological tests (4th revised ed.)]*. Houten, The Netherlands: Bohn Stafleu van Loghum. Zie met name pagina's 334-340.

[9] Zie bijvoorbeeld R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.) (1967). *The methodology of evaluation. Perspectives of curriculum evaluation*. Chicago, IL: Rand McNally. pp. 39–83.

[10] Koretz, D. (2008). Measuring Up: What Educational Testing Really Tells Us. Cambridge, MA: Harvard University Press, 2008.

[11] Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139–148.

van bewijsmateriaal voor gebruik door leerlingen en hun leraren, om vast te stellen waar de leerlingen zich bevinden in het leerproces, waar ze heen moeten en hoe ze daar het beste kunnen komen"[12]. Dergelijke informatie kan vaak al verkregen worden uit bijvoorbeeld een vraag die een leerling stelt, maar deze informatie kan ook systematisch worden verzameld door middel van een formatieve toets[13]. Kenmerken van deze formatieve toetsen zijn dat deze toetsen worden afgenomen om te ondersteunen tijdens het leren met als expliciet doel om het leren te verbeteren. Om dit te bewerkstelligen is een duidelijke link nodig tussen het toetsresultaat en het onderwijsmateriaal. Gemaakte fouten tijdens de toets moeten relevante informatie opleveren over welke onderdelen meer aandacht behoeven. Een ideale formatieve toets kan zelfs informatie geven over het type denkfout dat de leerling maakt tijdens de toets om vervolgens de didactiek verder te kunnen toespitsen op het remediëren van deze denkfout[14].

Toetsen die aan het eind van een leerproces worden afgenomen om een oordeel te vellen over het leerresultaat, noemen wij summatief. Op basis van het toetsresultaat, in veel gevallen in combinatie met andere gegevens, worden beslissingen over leerlingen genomen. De beslissing kan dan onder andere betrekking hebben op de overgang naar het volgende leerjaar, best passend vervolgonderwijs of diplomering of certificering. Kenmerkend voor dit type toetsen is dat er aan het eind van een leerproces wordt gekeken wat het resultaat van dit proces is. Het is daarbij de bedoeling dat de leerling zo veel mogelijk laat zien wat zij of hij allemaal kan, waarbij de leerling probeert zo min mogelijk fouten te maken.

Als startpunt is bij de genoemde doelen de nadruk gelegd op de leerlingen. Zij maken de toetsen en zij vergroten hun vaardigheid. De opdeling van de doelen in formatief en summatief kan ook gemaakt worden op een geaggregeerd niveau, zoals dat van de school of op systeemniveau. Ook hier is het formatieve doel het verbeteren van het leren. Op schoolniveau heeft dit betrekking op hoe het onderwijs op school verbeterd kan worden om het leerproces van de leerlingen te verbeteren daar waar dat kan. Dat betreft een echte kwaliteitsbevordering van het onderwijs. Anderzijds kunnen toetsen voor scholen ook summatief gebruikt worden om een oordeel te vellen over het leerresultaat van de leerlingen van de school. Dit is meer kwaliteitsbewaking, en heeft net als bij een summatieve toets bij leerlingen meer een connotatie van "afrekenen".

Hierbij kan ook opgemerkt worden dat of de toets als formatief of summatief gezien wordt, ook afhangt van hoe de school de consequenties van de toets ervaart. Als de school de consequentie van het verkregen resultaat als "afrekenen" ervaart, krijgt de toets meer kenmerken van een summatieve toets. Dat is bijvoorbeeld het geval als ouders de informatie gebruiken voor de selectie van de school voor hun kinderen of als het aangeboden (verplichte) begeleidingstraject niet als ondersteunend maar als onprettig wordt ervaren. Een eindoordeel voor een school is in die zin ook niet echt anders dan voor een leerling. Zowel op leerling- als op schoolniveau kunnen summatieve toetsresultaten consequenties hebben met een oordelend karakter. Daar waar bij een negatief eindoordeel op leerlingniveau de leerling mogelijk een jaar over moet doen, zal een school met een

[12] Gardner, J., (2012). *Assessment and Learning. 2nd ed.*, Los Angeles: Sage.

[13] Sluijsmans, D. & Kneyber, R. (2016). *Toetsrevolutie. Naar een feedbackcultuur in het voortgezet onderwijs*. Phronese, Uitgeverij.

[14] Zie bijvoorbeeld Bowen, L.,Kennedy, P.C., Seipel, B., Carlson, S.E., Biancarosa, G. & Davison, M. L. (2019). Can We Learn from Student Mistakes in a Formative, Reading Comprehension Assessment? *Journal of Educational Measurement*, v56 n4 p815-835 Win 2019.

negatief eindoordeel mogelijk een jaar extra toezicht of extra verantwoording moeten afleggen over hun onderwijskwaliteit.

Op systeemniveau worden zelden toetsresultaten gebruikt met een summatieve functie. Als een land of deelstaat onderzoek doet naar hun een eigen (onderwijs)systeem, is dat vooral om het systeem te verbeteren, dus met een formatieve functie. Kijken we naar internationale studies op systeemniveau, dan zien we wel dat toetsresultaten naast formatief ook gebruikt worden met een summatieve functie. In internationale studies is een onderdeel van de studie vaak het vergelijken van een land met andere landen – een relatieve beoordeling – hier komen dan elementen naar voren met een meer summatief karakter. Van elk land wordt dan het voorgaande onderwijsbeleid beoordeeld en afgezet tegen die van andere landen.

De opdeling van de doelen in formatief en summatief kan dus een spelen op zowel leerling-, school- of systeemniveau. In Tabel 2.1 staan de verschillen tussen formatief en summatief toetsen nogmaals op een rij gezet.

Type toetsdoelen	**Vaststellen van het leerresultaat**	**Ondersteuning van het leerproces**
Kernbegrip	Oordelen	Helpen
Doel	Een beslissing nemen	Het leren verbeteren
Benadering van fouten	Vermijden	Om van te leren
Wanneer (t.o.v. leeractiviteit)	Aan het einde	Tijdens
Toetsresultaat	Eindoordeel	Terugkoppeling naar lesmateriaal
Benaming	Summatief	Formatief

Tabel 2.1: Verschil tussen summatieve en formatieve toetsen op diverse criteria

Voorgaande laat zien wat voor verschillende informatie een toets op kan leveren en hoe die informatie gebruikt kan worden om verschillende doelen te behalen. Naast de informatiebehoefte kunnen er echter ook andere redenen spelen om te willen toetsen. Dit kan zijn om leerlingen te motiveren om de leerstof beter te bestuderen zodat de leerprestaties verbeteren. Het kan ook dienen om ervoor te zorgen dat scholen de lessen meer richten op de leerstof die gedekt wordt door de toetsen. Dat zijn zaken die niet direct met de meting te maken hebben, maar wel het gevolg zijn van het feit dat er gemeten wordt. Deze redenen vallen onder consequentiële validiteit[15]. Consequentiële validiteit heeft zowel betrekking op de mogelijke positieve gevolgen van meten (bijvoorbeeld motivatieverhoging) als de mogelijke negatieve gevolgen van meten (bijvoorbeeld curriculumvernauwing). In Sectie 2.2 gaan we daar dieper op in.

2.1.2 De doelgroep: Over wie moeten de toetsresultaten een uitspraak doen?

In het onderwijs worden toetsen aan de leerlingen voorgelegd. Het resultaat van het onderwijs wordt als het goed is weerspiegeld in wat de leerling doet of kan. In de vorige paragraaf zijn we er eerst vanuit gegaan dat het waarom van de toets impliciet betrekking

[15]In het Engels, *Consequential validity*. Zie: Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement, 3rd ed.*, pp. 13-103. New York: Macmillan.

heeft op de leerling. Als we kijken over wie uitspraken gedaan worden naar aanleiding van de toetsresultaten, dan zijn dat niet alleen de leerlingen. Op basis van de resultaten wordt ook gekeken naar de impact van handelingen en eigenschappen van de leerkracht, de school, of het schoolbestuur. De doelgroep kan ook op een nog meer geaggregeerd niveau: de koepel, de regio, het net of heel Vlaanderen.

Het kiezen van de doelgroep is doorgaans relatief eenvoudig: een groep leerlingen op een bepaald niveau wordt geselecteerd, waarna bepaald wordt tot op welk aggregatieniveau uitspraken gedaan worden. Het is ook van belang te weten of er uitsluitcriteria zijn die aangeven welke leerlingen de toets niet hoeven te maken. De keuze voor de doelgroep en de uitsluitcriteria (of juist het ontbreken ervan) hebben vaak praktische consequenties. Dat betreft bijvoorbeeld de productie van de opgaven. De omschrijving van de doelgroep helpt richting te geven aan de moeilijkheid. Doordat we weten welke groepen in staat moeten zijn deze toetsen te maken, weten we ook in welke vorm (varianten van) opgaven beschikbaar moeten zijn.

Naast het bestaan van uitsluitcriteria is er ook het recht van deelname. Al zou een leerling op basis van de uitsluitcriteria de toets niet hoeven te maken, is het goed mogelijk dat de leerling de toets wel wil maken[16]. Een dergelijke leerling moet de meest geschikte versie van de toets aangeboden krijgen. Vanuit het idee van een brede afname zullen er meer varianten mogelijk zijn[17]. Als een leerling een toets gemaakt heeft, hoort deze ook een rapportage te krijgen van het resultaat, in welke vorm dan ook[18]. Voor de rapportage op schoolniveau zouden de leerlingen die onder de uitsluitcriteria vallen niet mee hoeven (of moeten) tellen. Overigens is de verwachting dat het aantal leerlingen die vallen onder de uitsluitcriteria, en dus ook het aantal scholen met dergelijke leerlingen, zeer beperkt is, gezien de ambitie zoveel mogelijk leerlingen bij de centrale toetsen te betrekken.

2.1.3 De toetsinhoud: Wat willen we meten?

De toetsinhoud beschrijft datgene wat we willen meten. Het beschrijven van toetsinhoud is een belangrijke, maar geen eenvoudige taak[19]. Hoe duidelijker dit omschreven wordt, hoe makkelijker het wordt om later opgaven erbij te construeren die reflecteren wat de toets wil meten. Dus een toetsinhoud is meer dan een aanduiding van domeinen en subdomeinen; en meer dan een inhoudelijke omschrijving van een (sub-)domein aan de hand van de concepten en de relaties daartussen. Bovenal beschrijft de toetsinhoud welk gedrag een leerling met betrekking tot de leerstof zou moeten vertonen, en in het geval van misconcepties van de leerstof zou kunnen vertonen. Daarbij kan bijvoorbeeld onderscheid gemaakt worden tussen productieve en reproductieve vaardigheden of andere classificaties van vaardigheden zoals die van Bloom[20].

Het is ook een noodzaak dat de beschrijving van het 'wat' toetsbare vaardigheden

[16] Er kunnen diverse redenen zijn waarom een leerling dat zou willen, zoals dat de leerling zich niet uitgesloten wil voelen, of dat de ouders erop staan dat dat de leerling niet anders behandeld wordt.

[17] Voor een nadere uitwerking zie Sectie 5.6.

[18] Meer over de vormen van rapportage in Hoofdstuk 6.

[19] Zie voor een uitgebreide introductie ook Webb, N.L. (2006). Identifying content for student achievement tests. In: S.M. Downing & T.M. Haladyna, *Handbook of test development*. Mahwah, NJ: Erlbaum

[20] Bloom, B.S.; Engelhart, M.D., Furst, E. J., Hill, W.H. & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay Company.

omvat. Naast het gewenste of 'juiste' gedrag is het daarnaast aan te raden om gangbare denkfouten, misconcepties of incorrecte werkwijzen op te nemen. Dit alles scherpt namelijk aan wat je wilt meten. Concrete voorbeelden van gedrag kunnen eveneens helpen om de toetsinhoud zo scherp mogelijk te omschrijven. Dit is ook een element dat een belangrijke rol speelt in het gebruik van het Evidence-Centered Design (ECD)[21].

De voorwaardelijke vaardigheden die een leerling moet beheersen om zich de leerstof eigen te kunnen maken, krijgen eveneens een plaats in de beschrijving van de toetsinhoud. Daarnaast worden vaardigheden die voortbouwen op hetgeen je wilt meten ook opgenomen. Deze aspecten van de beschrijving helpen om de vergelijkbaarheid van de vaardigheid over meerdere meetmomenten in kaart te brengen. Hiermee komen we ook op een ander onderscheid in wat we meten: is dat het vaardigheidsniveau, is het de groei in vaardigheid, of beiden?

Als toetsen leerwinst in kaart moeten brengen, is het handig om te weten hoe leerwinst inhoudelijk omschreven kan worden. Wat kon een leerling eerst niet en nu wel? In de beschouwing van leerwinst moet het 'wat' dusdanig gedefinieerd worden dat het te meten construct van deze vaardigheid duidelijk is. Betreft de gemeten vaardigheid eigenlijk wel een enkele vaardigheid die serieel aangeleerd wordt in vaste stappen, of betreft het een vaardigheid die uit verschillende aspecten is opgebouwd, die niet noodzakelijk sequentieel geleerd hoeven te worden of cumulatief opgebouwd zijn. Als voorbeeld hiervan kunnen we rekenen/wiskunde beschouwen: men kan de onderdelen meetkunde en statistiek los van elkaar leren, maar beiden zullen vooraf gegaan moeten zijn door kennis in cijferen.

Bestaande curricula geven aan welke stof aangeleerd wordt in de diverse fasen van een schoolloopbaan. Door bij de omschrijving van de toetsinhoud aan te sluiten bij de bestaande curricula, krijgen de toetsen en de daaropvolgende resultaten betekenis voor het onderwijs. Als de toetsinhoud afwijkt van wat er daadwerkelijk op school aangeleerd wordt, dan zeggen de toetsen weinig over de kwaliteit van het onderwijs: het daadwerkelijke onderwijs wordt immers niet gemeten. Het kan zijn dat er diverse curricula gehanteerd worden binnen het onderwijs, die verschillen in de stof die aangeleerd wordt. Gelukkig kennen deze vaak overeenkomsten als het gaat om basisvaardigheden als taal en wiskunde. De toetsinhoud kan dan aansluiten bij het gemeenschappelijke deel van de curricula, om betekenisvol te zijn voor alle scholen.

Een dergelijke toetsomschrijving levert een breed scala aan gedragingen op. Om ervoor te zorgen dat toetsen bij deze omschrijvingen telkens dezelfde gewichten aan de diverse vaardigheden of subdomeinen geven, wordt vaak met een toetsmatrijs gewerkt[22]. In een toetsmatrijs wordt gespecificeerd hoeveel vragen er opgenomen worden over ieder subdomein of iedere subvaardigheid. Een dergelijke toetsmatrijs zorgt ervoor dat diverse toetsvarianten, zowel binnen één afnameperiode als tussen afnameperiodes, inhoudelijk vergelijkbaar blijven. Prestaties op de toetsen blijven dan vergelijkbaar, er wordt niet iets anders gemeten.

De inhoud van de toets refereert aan een ander aspect van validiteit, namelijk meet de

[21] Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1(5). Voor een korte introductie in ECD zie ook: Mislevy, R., Almond, R.G. & Lucas, J.F. (2003). A Brief Introduction to Evidence-centered Design, `https://www.ets.org/Media/Research/pdf/RR-03-16.pdf`.

[22] Zie bijvoorbeeld Lane, S., Raymond, M. R., & Haladyna T.M. (2015). *Handbook of test development, second edition*. New York, Routledge.

toets wat hij zou moeten meten. In meer klassieke termen gaat het dan om inhoudsvaliditeit. Daarbij gaat het niet alleen om een goede beschrijving van de toetsinhoud, maar ook over de formulering van opgaven, de omstandigheden waarin toetsen worden afgenomen, of het daadwerkelijke gebruik van de toetsresultaten. Al deze aspecten kunnen leerlingen namelijk belemmeren om hun vaardigheid te tonen[23] of een foutief beeld geven van de vaardigheid van de leerling. Het genoemde toepassen van het ECD bij de omschrijving van toetsinhoud en verdere toetsontwikkeling helpt om de validiteit van een toets hoog te houden.

Tot slot moet opgemerkt worden dat de wijze waarop gemeten wordt, moet passen bij de gemeten vaardigheid. Als we bijvoorbeeld de spreekvaardigheid van een leerling willen achterhalen, zullen schriftelijke meerkeuze opgaven geen echte valide meting opleveren. Wanneer vastligt dat de meting gestandaardiseerde centrale toetsen betreft die digitaal afgenomen moeten worden, dan legt dit enige beperkingen op daar waar het de mogelijk te meten vaardigheden betreft. Als daarnaast bij de centrale toetsing de eis gesteld zou worden dat er geen (externe) beoordelaars betrokken mogen zijn bij het evalueren van de leerlingprestaties, dan beperkt dit de mogelijk te meten vaardigheden nog verder. Een eis van automatisch scoorbare opgaven levert een beperking op in de mogelijke antwoordwijzen die een kandidaat kan geven. Een mogelijke, kostenefficiënte oplossing is de moeilijker te toetsen vaardigheden en competenties zoals schrijven en spreken, onder de verantwoordelijkheid van de scholen te laten vallen[24]. In Hoofdstuk 5, wanneer dieper ingegaan wordt op de toetsontwikkeling, zal verder ingegaan worden op de uitdagingen die de randvoorwaarden van de afname oplevert.

2.1.4 De rapportage: De terugkoppeling

Een belangrijk onderdeel van toetsing is de terugkoppeling van het toetsresultaat. Zonder een goede terugkoppeling is het uitvoeren van een toets niet zinvol. Het omvat het wie, het wat en het waarom van de toets, omdat het van belang is om te weten op welke manier prestaties van leerlingen gewaardeerd gaan worden. De vraag is of deze normering absoluut of relatief gaat zijn. Bij een absolute normering, wordt de prestatie van een leerling gerelateerd aan een bepaalde standaard. We meten of een leerling bepaalde stof beheerst of niet. Bij een relatieve normering wordt de prestatie van een leerling gerelateerd aan de prestaties van andere leerlingen. Met de toetsscore worden de leerlingen gerangschikt van zwak naar sterk, van lage vaardigheid naar hoge vaardigheid.

Bij een absolute normering ligt de focus van de toetsinhoud op een specifiek bereik binnen de vaardigheidsdimensie, namelijk dat deel dat overeenkomt met de standaard die getoetst wordt. Als er een specifieke minimale vaardigheid geëist wordt, of als het doel is te onderzoeken hoeveel leerlingen in een populatie een streefvaardigheid behaald hebben, hebben we te maken met een absolute normering. Inhoudelijke omschrijvingen

[23] Crooks, T.J. & Kane, M.T. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, 3, 3.

[24] Binnen het Nederlandse systeem worden de competentiegerichte vereisten om bijvoorbeeld aan de certificeringseisen van het examen te voldoen, die gemeten moeten worden met behulp van een praktische uitvoering, onder de verantwoordelijkheid van de school uitgevoerd. Er worden eisen gesteld waaraan voldaan moet worden, maar de meting zelf valt buiten de verantwoordelijkheid van het College voor Toetsen en Examens. Deze keuze is gemaakt omdat een centraliseerde uitvoering van een dergelijke meting technisch lastig uit te voeren is en wanneer dat wel gepoogd zou worden hoge kosten met zich mee zou dragen.

van standaarden moeten dan vertaald worden naar scores op de toets die aangeven of een leerling de stof wel of niet beheerst. Dergelijke standaardbepalingen zijn op verschillende manieren uit te voeren (zie bijvoorbeeld Cizek & Bunch, 2007)[25]. Pas na een standaardbepaling zijn de scores op een toets inhoudelijk te interpreteren. Bij de invoering van de referentieniveaus taal en rekenen[26] in het basis- en het secundair onderwijs in Nederland is gebruik gemaakt van standaardbepalingen op referentiesets van opgaven voor de terugkoppeling van de resultaten[27]. Meer informatie hoe een dergelijk onderzoek kan worden uitgevoerd, wordt gegeven in Hoofdstuk 6, waarin verder ingaan wordt op de rapportage waarbij gebruik gemaakt wordt van absolute normering.

In het geval van een relatieve normering zal een breder bereik omschreven worden om alle leerlingen goed te kunnen ordenen. Met name als de verschillen tussen leerlingen relatief groot zijn, maakt het voor de toetsinhoud uit of er absoluut of relatief getoetst wordt. Een relatieve wijze van normering kan op verschillende manieren plaatsvinden. Bekende voorbeelden zijn percentielscores en ranking, maar bij alle vormen waarbij de score van een persoon geïnterpreteerd wordt in relatie tot (relevante) anderen is sprake van een relatieve normering. Soms is een relatieve normering iets lastiger te herkennen, maar zo is de IQ-score ook een vorm van relatief normeren: met het gekende gemiddelde van 100 en standaardafwijking van 15 kan men de relatieve positie van ieder geteste persoon ook bepalen.

Relatieve vormen van normeren kunnen betrekking hebben op individuele leerlingen wanneer we leerlingscores bekijken, maar ook op scholen en op landen. Voor deze laatste twee heeft de relatieve normering vaak de vorm van een ranking: de scholen dan wel landen worden geordend op basis van hun geaggregeerde scores. In Hoofdstuk 6 over de rapportages wordt verder ingegaan op deze vorm van rapporteren.

Een leerlingrapportage per toets kan heel summier zijn (denk bijvoorbeeld aan een cijfer), of juist uitgebreid. Zo kan naast het rapporteren van een algemeen niveau ook op subdomeinen het niveau van de leerling aangegeven worden. Dit wordt vaak een profielanalyse genoemd. Daarnaast kan een rapportage het huidige niveau aangeven, maar ook de leerwinst sinds het vorige meetmoment. Een rapportage kan voorbeelden omvatten van wat er wel of niet wordt beheerst, of de gegevens in een rapportage kunnen met figuren worden geïllustreerd.

Rapportages worden afgestemd op het toetsdoel, de doelgroep en de toetsinhoud van de toetsing. Echter, het werkelijke gebruik van een rapportage kan afwijken van het doel van de toetsing. Rapportages kunnen een andere interpretatie krijgen dan bedoeld. Zo speelt vaak bij relatieve rapportages dat men graag beter wil scoren dan anderen, en binnen de normgroep beter dan het gemiddelde. Dat heeft vaak tot gevolg dat wanneer relatieve normen gebruikt worden dat fouten minder worden gezien als iets om van te leren, maar meer iets om te vermijden. Dat is een benadering die past bij summatief toetsen, en niet bij formatief toetsen. Wellicht niet door iedereen, maar er zullen belanghebbenden zijn die zodoende de toets ook een summatieve functie meegeven. Dus, ook als het doel formatief is, kan het gebruik door de vorm van de rapportage eerder summatief zijn. Het doel van de

[25] Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

[26] `https://www.slo.nl/thema/meer/taal-rekenen/`

[27] `http://www.toetsspecials.nl/html/referentiesets/publicaties_referentiesets.shtm`

toets moet weerspiegeld worden in de vorm van de rapportage, want het werkelijk gebruik van de toets is het gevolg van die rapportage.

In relatie hiermee heeft Daniel Koretz een belangrijke observatie: "Besluitvormers moeten bepalen welke doelen het belangrijkst zijn voor een test, en vervolgens accepteren dat het resultaat hen kost in termen van andere doelen."[28] Formatief bedoelde toetsen kunnen soms summatief gebruikt gaan worden, met consequenties voor het formatief gebruik, namelijk dat de kans op fraude toeneemt, en daarmee de kans op het optimale formatief gebruik afneemt. Veel kwantitatieve toetsresultaten kunnen uiteindelijk ook resulteren in rangordeningen van leerlingen of scholen. Als er kwantitatieve gegevens beschikbaar zijn die geordend kunnen worden, zullen er belanghebbenden zijn die dat gaan doen. Verder in het rapport worden manieren beschreven die kunnen helpen om de impact hiervan te beperken. Gecombineerd summatief en formatief gebruik komt ook aan de orde, zij het dat de observatie van Koretz wel een punt blijft om rekening mee te houden bij het bepalen van de balans.

De neiging van het publiek om scholen te willen rangordenen is niet alleen een uitdaging als er relatief wordt genormeerd, maar ook als er alleen absoluut wordt genormeerd op leerlingniveau. Wanneer bekend is hoeveel leerlingen op een school de standaard halen, dan kunnen scholen ook op basis van dit percentage geordend worden. Het is de wens binnen het bestek om het maken van rangordeningen tegen te gaan[29]. Daarnaast is het, hoewel moeilijker, wel mogelijk om eerlijke rangordeningen te publiceren. Bijvoorbeeld door normgroepen van scholen, of leerlingen, te formeren met vergelijkbare achtergrondkenmerken. Binnen deze normgroepen is de ordening eerlijker. Een andere optie betreft bewerkingen van de scores door middel van (lineaire) regressies of andere vormen van wegingen, waarbij scores gecorrigeerd worden voor achtergrondvariabelen.

Een wijze om de rangordening van scholen wellicht niet zozeer tegen te gaan maar duidelijk wel te relativeren, is door in de rapportage ook informatie te geven over de meetfout van het - al dan niet bewerkte - gemiddelde resultaat van de school. Het rapporten van de meetfout kan helpen om duidelijk te maken dat gemiddelden wellicht kunnen verschillen, maar dat deze gevonden verschillen mogelijk niet significant en door toeval gevonden zijn[30]. Het is dan wel van belang dat de interpretatie van de gerapporteerde meetfout dan wel effectgrootte correct gebeurt. Hierover meer in Hoofdstuk 6.

Daarnaast is het aan te raden om in rapportages niet alleen de resultaten van een enkel jaar te gebruiken. Bij rapportages over resultaten van meerdere jaren, zijn de resultaten gebaseerd op meer data waardoor de standaardmeetfout van de gemiddelden kleiner is

[28] "*Decision makers should determine what goals are most important for a test and then accept the fact that the result will cost them in terms of other goals*" – Daniel Koretz in *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press, 2008.

[29] Bij de beschrijving van de rapportage wordt verder ingegaan op mogelijkheden om dit ook te beperken.

[30] Statistische significatie wordt bepaald door in de evaluatie van de verschillen in de gemiddelden deze te relateren aan de standaardmeetfout van de schatting van de schoolgemiddelden.

en bovendien is dan zichtbaar of de resultaten enige stabiliteit vertonen. Hoe rapportages geïnterpreteerd en gebruikt worden heeft ook betrekking op de validiteitsvraag. Het betreft dan niet zozeer de validiteit van een meting, maar de validiteit (geldigheid) van de conclusies en besluiten die verschillende mogelijke belanghebbenden (ouders, de inspectie, of anderen die een oordeel over een school hebben) aan een meetuitslag verbinden[31].

Als we kijken naar het combineren van de doelen van toetsen voor de leerlingen en scholen, dan zijn er in principe een viertal mogelijkheden: de toets die voor zowel de school als de leerling summatief, dan wel formatief is, en de combinaties waarbij de toets voor de school formatief is en de leerling summatief of juist omgekeerd. Deze combinaties worden bij de beschrijving van de scenario's verder uitgewerkt, in Hoofdstuk 7. Het is wel belangrijk te realiseren dat om de resultaten over de scholen goed te kunnen interpreteren, de functies van de toetsen voor de leerlingen binnen alle scholen identiek moeten zijn. Dit heeft ook te maken met hoe omgegaan wordt met de ervaren mate van belang van de toets.

Toetsen met grote belangen en minder grote belangen

Toetsen verschillen in de mate waarin er belang aan gehecht wordt. Het belang dat aan een toetsresultaat gehecht wordt, dus of een toets als van groot of minder groot belang ervaren wordt, heeft effect op degene die de toets gaat maken. Toetsen en examens verzamelen altijd informatie over de leerling, maar soms slechts uitsluitend met het doel om deze te aggregeren op leerkracht-, school- of systeemniveau, om vragen te kunnen beantwoorden als:

- Is de leerkracht effectief in het geven van onderwijs?
- Zijn er scholen die significant minder progressie boeken?
- Hoe ontwikkelt het landelijk onderwijsniveau zich?

Om dergelijke vragen goed te kunnen beantwoorden, is het van belang ook de motivatie van de leerlingen mee te nemen: zijn de leerlingen voldoende gemotiveerd om de best mogelijke prestatie neer te zetten? Uit diverse ervaringen weten wij dat het ervaren belang door een leerling van invloed kan zijn op de toetsprestatie[32]. Wanneer leerlingen een groot belang toedichten aan de uitkomst van een toets, ervaren zij deze als "high-stakes", en bij vrijwel geen belang als "low-stakes". Het toegekende belang kan duidelijke invloed hebben op de voorbereiding van de toets, door bijvoorbeeld goed te studeren van tevoren, of op tijd naar bed te gaan. Tijdens het maken van de toets zal de leerling geconcentreerder werken en minder opgaven overslaan[33]. Het verschil tussen de prestatie in een high-stakes en een low-stakes situatie op een en dezelfde opgave kan aanzienlijk zijn, en voor groepen leerlingen verschillen[34]. Merk overigens op dat er tussenvormen zijn tussen high-stakes

[31] Verhelst, N., Staphortius, G., & Kleintjes, F. (2001). Scholen langs de meetlat. Arnhem: Cito. `https://www.researchgate.net/publication/237758068_Scholen_langs_de_meetlat`

[32] Keizer-Mittelhaëuser M.-A. (2014). Modeling the effect of differential motivation on linking educational tests. Doctoral dissertation. Tilburg University, Tilburg, the Netherlands.

[33] Hemker, B.T. (2012). The impact of motivation: modeling motivation in educational measurement. Presentation presented July 4, 2012 at the ITC conference, Amsterdam.

[34] Een verschil met een effectgrootte van 0,40 is gevonden in vertoonde vaardigheid tussen leerlingen in een low-stakes en een high-stakes situatie. Voor meisjes was het effect 0,30 terwijl het voor jongens 0,50 was. Dat laatste betekent dat jongens in een low-stakes situatie een vaardigheid vertoonden die een halve standaardafwijking verschilde van hun prestaties wanneer zij een groot belang aan de toets hechtten (Hemker,

en low-stakes, afhankelijk van de mate waarin het toetsresultaat impact heeft op een beslissing. Er zijn vele variaties mogelijk die liggen tussen 0% (zeer low-stakes) en 100% (zeer high-stakes).

Ook andere betrokkenen kunnen een groot belang toekennen aan de uitkomst van een toets. Dit hangt onder andere af van het aggregatieniveau van de toetsuitkomsten. Voor leerkrachten is dat bijvoorbeeld het geval wanneer zij op de resultaten van de toets worden afgerekend, bijvoorbeeld als de resultaten gaan om hun aanstelling – zeker als de aanstelling nog niet vast is–, hun salariëring of hun status. Voor een school wordt een toets van groot belang als de resultaten gevolgen hebben voor de financiering van de school, als er sprake is van een publieke ranking of als ouders de resultaten worden aangeboden om hen te helpen bij de keuze van de school voor hun kinderen[35].

Daarnaast kunnen toetsuitkomsten ook voor de overheid van groot belang zijn. Een groot deel van het overheidsbudget gaat naar onderwijs. Een overheid wil dan ook zien dat dit geld welbesteed is, en wil op basis van bewijs-gestuurde besluitvorming goed onderwijs stimuleren. Rapportages over de resultaten van het gevoerde overheidsbeleid zijn van belang voor de regerende partijen, omdat deze invloed hebben op hun electorale draagvlak. Een goed voorbeeld hiervan vinden we in de uitkomsten van peilingsonderzoek. Dit betreft nationale peilingen, maar zeker ook internationale peilingen waarvoor veel publieke aandacht is dankzij de ranking van landen. Als een land gestegen of gedaald is in vaardigheid of ranking kan dit politieke gevolgen hebben.

In al deze gevallen is de leerling degene die de toets maakt. Vaak zien we in de praktijk dat er een discrepantie is tussen de ervaren belangen door de leerling en het belang van het toetsresultaat voor een leerkracht, de school of het ministerie. Dit kan tot ingewikkelde tegenstrijdigheden leiden. Wanneer er voor de leerling belangen vanaf hangen zal deze proberen een goed resultaat te behalen. Maar wie de toets maakt, is niet altijd degene die wordt beoordeeld. Als voor de leerling de belangen juist laag zijn, maar voor betrokkenen op geaggregeerde niveaus hoog, dan kunnen resultaten lastiger te interpreteren worden. Dan worden leerkrachten, scholen en/of overheden afgerekend op 'typische resultaten' en niet op 'maximale resultaten'. Wanneer er daarnaast resultaten bekend zijn op vergelijkbare high-stakes toetsen, kunnen er paradoxale resultaten worden gevonden. Dit speelt bijvoorbeeld bij peilingen een rol.

In een Nederlands onderzoek[36] zijn dezelfde opgaven in een voor de leerlingen high-stakes situatie en een low-stakes situatie gebruikt. De eerste situatie betrof een belangrijke beslissing aan het einde van de basisschool voor leerlingen (toets in februari). De tweede situatie was een peiling in het laatste jaar van het basisonderwijs die geen impact had voor de voortgang van een leerling (toets in mei/juni). Hierbij was een duidelijk verschil in vaardigheid te zien. De getoonde vaardigheid die in de eerste situatie door twee derde van de leerlingen behaald werd, werd in de tweede situatie door slechts de helft van de leerlingen behaald. In internationaal onderzoek is gevonden dat dit motivatie-effect om de best mogelijk prestatie neer te zetten, kan verschillen tussen groepen. Zo is in verschillende

2012, zie vorige voetnoot).

[35]Vaak is het zo dat, als de ouders informatie kunnen verkrijgen en zo kunnen achterhalen welke school hoger scoort dan een andere school, een journalist, al dan niet met behulp van iemand die technisch begaafd is, dat ook kan achterhalen voor álle scholen. De journalist kan dan alsnog een ranking publiceren.

[36]Hemker, B.T. (2012). The impact of motivation: modeling motivation in educational measurement. Presentation presented July 4, 2012 at the ITC conference, Amsterdam.

landen (i.e., Nederland, Frankrijk[37] en Canada[38]) gevonden dat dit effect meer bij jongens dan bij meisjes speelt. Dit heeft ook effect op hoe de verschillende leerwinstresultaten en schoolresultaten geïnterpreteerd kunnen worden. Als gevolg daarvan mag het belang dat leerlingen aan de resultaten op de toetsen hechten niet sterk tussen scholen verschillen om een zinvolle interpretatie te geven aan de verschillen tussen prestaties op de verschillende scholen. Als op de ene school de toets geen impact heeft voor de leerling zal deze bij een schoolpopulatie met een gelijke vaardigheid toch aanzienlijk slechter presteren dan de school waarbij de toets wel meetelt voor de leerling. Dit is iets waar ook rekening mee gehouden moet worden bij de invoer van centraal toetsen.

Strategisch gedrag als gevolg van de ervaren belangen van de toets

Naarmate de ervaren belangen bij een toets groter zijn, neemt ook de kans op negatieve neveneffecten toe, en zodoende ook de moeite die gestoken moet worden om dit tegen te gaan. We onderscheiden diverse verschillende vormen van negatieve neveneffecten die bij toetsen met grote belangen kunnen voorkomen.

De eerste vorm is extra investering van de ouders om de toetsprestatie van hun eigen kinderen te verbeteren. Vaak gebeurt dit door middel van extra bijles. Hoewel dit de vaardigheid van enkele individuele leerlingen vergroot, vergroot dit ook de kansenongelijkheid van leerlingen. Immers, meer vermogende ouders kunnen meer investeren in extra les voor hun kinderen om daarmee de toetsprestatie – en mogelijke extra kansen die daaraan verbonden zijn – te vergroten.

Een tweede negatief neveneffect komt voort vanuit een extra belang dat door de scholen wordt gevoeld om leerlingen goed te laten presteren. Als een toets voor leerkrachten, de school, of beiden van groot belang is, dan zullen ook zij proberen zo hoog mogelijke scores te verkrijgen op die toets. Dat kan door *teaching to the test*: in de klas extra veel aandacht te besteden aan de onderwerpen op de toets, of het type vragen dat gesteld wordt. Op zich leidt dit neveneffect wel tot een vergroting van de vaardigheid, maar het kan ook leiden tot curriculumvernauwing, wat betekent dat vaardigheden waar scholen en leerlingen minder op afgerekend worden, ook minder aan bod komen in het onderwijs. Dat is met name problematisch wanneer de vaardigheden die minder onderwezen worden wel belangrijk zijn, maar alleen ondergesneeuwd raken omdat ze moeilijker grootschalig toetsbaar zijn. Dit fenomeen kan al plaatsvinden als de toets alleen voor de leerlingen van groot belang is, maar zal nog sterker zijn als dit ook voor leerkrachten en scholen geldt.

Een schadelijkere vorm van teaching to the test is het geval wanneer niet zozeer de gemeten vaardigheid onderwezen wordt, als wel hoe men vragen erover moet beantwoorden. Stel dat een test die spelling meet altijd bestaat uit meerkeuze opgaven waarbij de leerling een foutief gespeld woord moet herkennen, dan kan dat tot een passieve vorm van spellingsonderwijs leiden, waarbij het gaat om het herkennen van spelfouten wanneer men weet dat er een spelfout is, en niet om een actieve vorm van spelling. In sommige gevallen leidt het ook tot het aanleren van trucjes om een test beter te maken (“test wiseness”) die

[37] Keskpaik, S. & Rocher, T. (2015). La motivation des élèves français face à des évaluations à faibles enjeux, éducation & formations n° 86-87 (`http://cache.media.education.gouv.fr/file/revue_86-87/58/2/depp-2015-EF-86-87-motivation-eleves-francais-face-evaluations-faibles-enjeux_424582.pdf`)

[38] Sarwar, G.S., Zerpa, C., Hachey, K., Simon, M., & van Barneveld, C. (2012). Teaching Practices and Student Motivation That Influence Student Achievement on Large-Scale Assessments. *International Journal of Education*, Vol.4, No.3.

niet meer te maken hebben met de vaardigheid zelf. Bekendheid met de vraagvorm moet er alleen voor zorgen dat de leerling die de vaardigheid wel heeft er niet door verward wordt, maar niet dat de leerling die de vaardigheid niet heeft een grotere kans heeft een goede prestatie neer te zetten.

Het is belangrijk te realiseren dat het toetsresultaat alleen maar een afbeelding is van de vaardigheid, maar niet de vaardigheid zelf is. Hoe meer de opgaven in de toets een goede weerspiegeling zijn van de te meten vaardigheid, hoe minder het een probleem is dat leerlingen met dat soort opgaven oefenen. Als het construct dat gemeten wordt volledig gedekt wordt door alle mogelijke opgaven die een leerling kan krijgen (zoals bij de tafels met getallen onder de 10) dan is oefenen uiteraard prima. Als de vaardigheid meer divers is, legt dat een grote druk op instanties die de toetsen maken om de diversiteit van de vaardigheid te vangen in de opgaven. Dat kan lastig zijn binnen de randvoorwaarden waaraan de opgaven moeten voldoen. In veel gevallen zal dit dan ook tot een vernauwing van de interpretatie van de vaardigheid leiden.

Een derde negatief neveneffect is gelegen in het manipuleren van de resultaten op verschillende meetmomenten door leerkrachten of scholen. Voorafgaand aan een toets kan een school bijvoorbeeld zwakke leerlingen aanmoedigen om zich 'ziek' te melden. Of, in het geval dat metingen van leerwinst voor een school van belang zijn, kan bij de eerste meting de prestatie gedrukt worden door de school, bijvoorbeeld door een gebrekkige voorbereiding bij de eerste afname. Bij de tweede afname van hetzelfde cohort in een later leerjaar kan dan wel een optimale prestatie worden neergezet. Daardoor lijkt de leerwinst groter. Strategisch gedrag op scholen waarbij zij een bepaalde statistiek proberen te optimaliseren, kan zo onwenselijke effecten opleveren waar rekening mee gehouden moet worden.

Een bekend grootschalig voorbeeld van toetsen die niet alleen voor leerlingen, maar ook voor leerkrachten en scholen van belang waren, zijn de toetsen die in de Verenigde Staten afgenomen werden in het kader van de "No Child Left Behind (NCLB)"-wetgeving. Deze wet was in 2001 vol goede bedoelingen aangenomen. Het zou ertoe moeten leiden dat kansarme studenten aantoonbaar goed onderwijs zouden krijgen, wat moest leiden tot goede resultaten op gestandaardiseerde toetsen. Deze toetsen moesten door alle leerlingen gemaakt worden[39] en gingen uit van leerwinst: elk jaar zou de school beter moeten presteren op de toetsen dan het jaar ervoor. Er waren serieuze consequenties voor scholen die niet aan de eisen voldeden, tot aan opheffing toe. Het zou ook een betere betrokkenheid van leerkrachten moeten bewerkstelligen. Hoewel initieel positieve effecten gerapporteerd zijn, had het ook dusdanig veel nadelen, zoals de hier beschreven curriculumvernauwing en onderwijs alleen gericht op het goed maken van de toets, wat ook fraude tot gevolg had, waardoor NCLB als een mislukking wordt beschouwd[40] en deze wet in 2015 is afgeschaft. Er is veel geschreven over NCLB[41,42] waaruit lering te trekken valt bij het invoeren van

[39] In ieder geval moest minsten 95% van de leerlingen per school de toets maken.

[40] Eskelsen García, L. & Thornton, O. (2015). 'No Child Left Behind' has failed. Washington Post, February 13, 2015. `https://www.washingtonpost.com/opinions/no-child-has-failed/2015/02/13/8d619026-b2f8-11e4-827f-93f454140e2b_story.html`.

[41] Meier, D., & Wood, G., Editors (2004) Many Children Left Behind: How the No Child Left Behind Act Is Damaging Our Children and Our Schools. Boston: Beacon Press.

[42] McGuinn, P.J. (No Child Left Behind and the Transformation of Federal Education Policy, 1965-2005 (Studies in Government & Public Policy).

grootschalig verplicht centraal toetsen.

Een ander negatief neveneffect bij toetsen met grote belangen is mogelijk frauduleus gedrag. Wanneer de belangen heel groot zijn, is het extra verleidelijk om te proberen de opgaven al vooraf te kennen en daarmee de kans op een goed toetsresultaat aanzienlijk te vergroten. Daarom zal bij een dergelijke toets veel moeite moeten worden gestoken in het geheim houden van de opgaven. Wanneer opgaven iedere keer volledig nieuw zijn dan hoeft slechts het huidige toetsmateriaal beveiligd te worden. Vaak wil men echter vanuit kostenbesparing en voor toetstechnische kwaliteitsborging opgaven uit eerdere toetsen hergebruiken. In het geval van grote toetsbelangen vergt dit een extra zorgvuldige voorbereiding. Als er sprake is van hergebruik van opgaven, dan is het zinvol een grote opgavenbank te hebben. Bij een grote opgavenbank gaat het een kandidaat veel moeite kosten om al deze opgaven te kennen[43]. Als de kandidaat een paar opgaven zou kennen heeft dat bij een grote bank ook weinig impact op het eindresultaat omdat er zoveel andere opgaven zijn. Meer over itembanken volgt later.

Ten slotte kunnen er ook tijdens de toetsafname negatieve neveneffecten optreden. Een leerling kan door middel van spieken (spiekbriefje, bij andere leerlingen kijken, of zelfs via contact met mensen buiten de toetslocatie) of door middel van gebruik van oneigenlijke hulpmiddelen proberen een beter beeld van zichzelf te geven. Om dit tegen te gaan zijn aanvullende maatregelen vaak noodzakelijk. Er zijn methoden om dit via *data forensics* te achterhalen[44], maar in de meeste gevallen zal ook ter plekke bij de afname geobserveerd moeten zijn dat er gefraudeerd is.

Fraude door leerkrachten en scholen kan ook plaatsvinden rondom de toetsafname. Dit zal minder snel spelen als de toetsen alleen voor de leerlingen van belang zijn. Als de leerkracht of school geen belang heeft bij de resultaten, dan kunnen zij tijdens de afname als surveillant optreden om fraude door de leerlingen tegen te gaan, en kunnen zij de toetsen van hun eigen leerlingen, of wellicht van de leerlingen van een collega nakijken. Echter, als de toetsen voor leerkrachten en scholen van groot belang worden, dan wordt de verleiding groter om de resultaten positief te beïnvloeden tijdens en na de afname. Dat kan door de leerlingen net iets meer tijd te geven dan toegestaan, of niet in te grijpen als er door leerlingen gebruik gemaakt wordt van ongeoorloofde hulpmiddelen. In meer extreme gevallen kunnen leerkrachten de toetsresultaten beïnvloeden door tijdens de afname verduidelijking te geven van vragen, hints te geven, of zelfs juiste antwoorden te geven. Wat bij digitale afname minder geldt, maar in het geval van papieren afname wel, is dat leerkrachten incorrecte antwoorden kunnen aanpassen in correcte antwoorden, of bij open opgaven een milde beoordeling geven. Voor deze zaken zijn oplossingen mogelijk zoals het inzetten van externe toezichthouders bij afname op school of zelfs het verplaatsen van een toetsafname naar een externe locatie, maar daaraan zijn meestal extra kosten verbonden.

Naarmate de belangen minder worden voor betrokkenen zal de kans op ongewenst

[43]Bij een grote, goed opgezette opgavenbank is het kennen van de gehele bank ook een indicatie dat de leerling de beoogde vaardigheid heeft. Bij de theorie-examens rijvaardigheid kunnen de kandidaten alle opgaven in de bank kennen, omdat het kennen van de gehele bank betekent dat zij de theorie beheersen. Bij dergelijke min of meer openbare opgavenbanken, is de normering vaak ook vrij streng.

[44]Ommering, van, C.J., de Klerk, S. & Veldkamp, B.P. (2019). Getting to Grips with Exam Fraud: A Qualitative Study Towards Developing an Evidence Based Educational Data Forensics Protocol. In: *Technology Enhanced Assessment*.

gedrag ook kleiner zijn. Dat houdt ook in dat de maatregelen die genomen moeten worden om dit gedrag tegen te gaan beperkter kunnen zijn, hetgeen over het algemeen betekent dat er minder kosten gemaakt hoeven te worden. Door de mate van belangen te beïnvloeden, kunnen de kosten dus beïnvloed worden. Naarmate de geteste persoon de belangen als lager ervaart, zal de kans op strategisch gedrag verminderen maar zal de kandidaat die getest wordt minder vaak de optimale prestatie leveren. Dat kan gevolgen hebben voor de voorbereiding van de toets, maar kan ook tijdens de test zelf blijken. Daarom is het altijd noodzaak om te zoeken naar een gepaste balans tussen kosten, toegekende belangen en het prestatieniveau waarin men geïnteresseerd is. In Tabel 2.2 zijn deze aspecten van belangen overzichtelijk bij elkaar gezet.

	high-stakes	**low-stakes**
Niveau prestatie	Optimale prestatie	Typische prestatie
Strategisch gedrag	Grote kans	Weinig noodzaak
Toetsverversing	Vaak	Kan langer mee leren
Controle op fraude	Streng	Mild
Kosten van controle op fraude	Kostbaar in beheersing	Lager

Tabel 2.2: Belangen: hoog – laag (high-stakes vs low-stakes)

Verschillende belanghebbenden kunnen ook een verschillende perceptie hebben van de belangen. Leerlingen kunnen bijvoorbeeld de toetsen als (relatief) low-stakes zien, terwijl de scholen (leerkrachten, directie) ze als high-stakes ervaren[45]. Als de leerling de toets als niet belangrijk voor zichzelf acht, zal het niveau van de prestatie eerder naar die van de typische prestatie gaan. De school kan echter er van alles aan proberen te doen desalniettemin de prestatie van de leerling hoog te krijgen. Dat kan door alsnog te proberen de toetsen als belangrijk voor de leerling te maken, door de toets wel mee te laten tellen bij de beslissing en zo voor de leerlingen een summatieve functie aan de toets mee te geven. Op die manier kunnen scholen hun best doen toch een optimale prestatie uit te lokken bij hun leerlingen. Wanneer de scholen van elkaar verschillen in de mate waarin het toetsresultaat toch impact heeft voor de leerling levert dit verschillen tussen scholen op die niet noodzakelijk het gevolg zijn van de kwaliteit van het onderwijs.

Als de toets voor een van de groepen belanghebbenden van groot belang is, dan is er altijd kans op strategisch gedrag en fraude. De vorm van dat gedrag en mogelijke fraude verandert alleen wel afhankelijk van de belanghebbende. Daar waar leerlingen bijvoorbeeld spiekbriefjes of stiekem ongeoorloofde hulpmiddelen kunnen inzetten, kunnen scholen bij de afname de leerlingen helpen of ongeoorloofde hulpmiddelen aanbieden. Beide typen belanghebbenden kunnen proberen (delen van) de toetsen voor de afname te achterhalen. Over het algemeen kan gesteld worden dat als voor een van de belanghebbenden de toetsen als high-stakes ervaren worden, de kenmerken zoals in de kolom high-stakes genoemd staan dominant zijn.

[45] In principe is het omgekeerde uiteraard ook mogelijk (high-stakes voor de leerlingen en low-stakes voor de scholen), al is dat in het kader van het beoogde doel van de centrale toetsen niet heel waarschijnlijk.

2.2 Toetsdoelen in relatie tot centrale toetsen in Vlaanderen

Uit de voorgaande paragraaf volgt dat de doelen van de toetsen van groot belang zijn voor de inrichting van diverse aspecten van toetsen, en zeker voor de drie thema's die in dit perceel de nadruk hebben. De toetsdoelen –en vooral hoe de belanghebbenden die ervaren– hebben effect op de wijze waarop leerwinst gemeten en geïnterpreteerd kan worden. Maar ook hebben deze effect op de kans op fraude en daarmee op de kosten van de toetsontwikkeling. Tot slot heeft de rapportage vooral impact op de manier waarop belanghebbenden de toetsresultaten zullen gaan gebruiken en zodoende op hoe zij de belangen ervaren. Om die reden is het zeer noodzakelijk om op een rij te zetten wat exact de beoogde doelen zijn voor de verschillende belanghebbenden. Daar waar deze nog niet geheel vastomlijnd zijn zal dat tot verschillende scenario's leiden.

Bij de evaluatie van de uitgangspunten zoals die in het begin van dit hoofdstuk gegeven worden, kan het bestek samen met het regeerakkoord[46] en de beleidsnota onderwijs[47] als primaire bron beschouwd worden. Bij de vraag **wat** we meten blijkt duidelijk uit de genoemde stukken dat de toetsinhoud van de proeven de vaardigheden Nederlands en wiskunde gaan worden. Binnen de vaardigheid Nederlands zijn de onderdelen begrijpend lezen, schrijven, en grammatica specifiek benoemd. Binnen het vak wiskunde worden geen expliciete onderdelen genoemd die als afzonderlijke vaardigheid gezien worden.

Voor de toetsinhoud zal sterk naar de vernieuwde eindtermen gekeken worden. Het is uit de discussies rond deze eindtermen duidelijk dat dit nog een moeizaam proces is[48]. Zolang het niet evident is wat er precies gemeten moet worden, en bij wie, is dat problematisch voor het produceren van toetsen. Een min of meer algemene omschrijving die weinig concreet is, levert voor het steunpunt een serieuze uitdaging op om een goede toetsmatrijs te maken. Het gevaar bestaat dat zij door de keuzes die ze in de productie van de opgaven maken dan effectief de eindtermen bepalen, hetgeen voor alle betrokkenen een onwenselijke situatie is. Het steunpunt is gebaat bij een zo concreet mogelijke, door alle betrokkenen breed gedragen omschrijving van de eindtermen op ieder niveau. Zonder dat loopt de haalbaarheid van de centrale toetsen een serieus gevaar.

De eindtermen verschillen ook per niveau, en dus ook van **wie** er gemeten worden. Twee zaken zijn van groot belang bij de omschrijving van de doelgroep van de toetsen. Welke niveaus meten we, en welke uitsluitcriteria worden gehanteerd voor leerlingen om niet deel te hoeven nemen aan de toetsen[49]. Daarin zijn de richtlijnen zeer duidelijk. De beoogde te meten leerlingen zitten op vier verschillende niveaus: in het basisonderwijs zitten ze in het vierde en het zesde leerjaar, in het secundair onderwijs zitten ze in het tweede en het zesde leerjaar. De keuze van de leerjaren zijn allen scharnierpunten in het onderwijs, en geven alle vier een vorm van afsluiting aan. Het vierde jaar in het basisonderwijs is de afsluiting van de basis van het basisonderwijs waarbij de focus sterk

[46] Departement Kanselarij en bestuur (2019). Vlaamse regering 2019-2024. Regeerakkoord. https://www.vlaanderen.be/publicaties/regeerakkoord-van-de-vlaamse-regering-2019-2024.

[47] Minister van Onderwijs, Sport, Dierenwelzijn en Vlaamse Rand (2019). Beleidsnota Onderwijs 2019-2024 (https://www.vlaanderen.be/publicaties/beleidsnota-2019-2024-onderwijs).

[48] Zie bijvoorbeeld, https://www.leuvenpubliclaw.com/juridisch-getwist-over-de-nieuwe-eindtermen-secundair-onderwijs-hoorzitting-in-het-vlaams-parlement/, maar ook diverse nieuwsberichten over dit onderwerp.

[49] Zoals al gesteld staat daar ook het recht van deelname tegenover. In Hoofdstuk 5 wordt bij de uitwerking van de brede deelname verder ingegaan op de mogelijkheden om bij deze groepen toetsen af te nemen.

ligt op de basisvaardigheden (rekenen en Nederlands lezen en schrijven). Het zesde leerjaar is het afsluitende jaar van het basisonderwijs, waarna het secundair onderwijs volgt. Het tweede leerjaar van het secundair onderwijs is de afsluiting van de eerste graad. Dat is het moment waarop (in de meeste gevallen) de definitieve keuze gemaakt wordt tussen de vier onderwijsvormen (aso, kso, tso en bso) die daarna in de tweede en derde graad gevolgd zal worden. Het zesde jaar betreft formeel de afsluiting van de derde graad en daarmee het secundair onderwijs.

Wie mee moeten doen is ook duidelijk: alle leerlingen in Vlaanderen, ongeacht het net of de koepel, wat ook betekent dat alle scholen moeten deelnemen. De toetsen moeten inclusief zijn zodat iedereen die op school zit de toets moet kunnen maken, ongeacht mogelijke beperkingen. Als echt alle leerlingen in Vlaanderen meedoen aan de centrale toetsen betekent dat dat ook alle leerlingen buitengewoon onderwijs mogen dan wel moeten deelnemen. Dit kan grote consequenties voor de kosten hebben omdat het maken van een nieuwe versie van een toets die geschikt is voor afname van een groep leerlingen met een specifieke beperking vaak kostbaar is. Zeker als er meervoudige uitdagingen zijn, bijvoorbeeld met zowel met zien als met bewegen, dan zijn de groepen voor wie de aparte versies gemaakt worden ook vaak (zeer) klein. In hoofdstuk 5 wordt verder ingegaan op mogelijkheden voor een brede afname.

De doelen van de centrale toetsen zoals aangegeven in het regeerakkoord en de beleidsnota zijn, zoals eerder genoemd in de introductie, het meten van het bereiken van de eindtermen, het in kaart brengen van de leerwinst van leerlingen en het in kaart brengen van de leerwinst op schoolniveau. Het is duidelijk dat wat er vervolgens met die informatie gebeurt, dus hoe die informatie gedeeld wordt en welke beslissingen op basis van die informatie gemaakt worden, van groot belang zijn om het gebruiksdoel van de toetsen te doorgronden. Dat bepaalt namelijk voor een zeer groot deel welke impact de centrale toetsen hebben en hoe de scenario's uitgewerkt worden.

2.2.1 Beoogde wijze van terugkoppeling

In het regeerakkoord wordt gemeld dat het de bedoeling is de resultaten aan de scholen terug te koppelen op leerling- en schoolniveau. Wat de scholen met de informatie op leerlingniveau moeten doen, dus welke gevolgen de meting voor de leerling heeft, is echter niet gegeven. Bij de eerste meting zal de informatie vooral gaan over het al dan niet behalen van de eindtermen. Deze informatie is op zichzelf criteriumgericht omdat wat een leerling wel of niet kan een inhoudelijke relatie heeft met de gemeten vaardigheid. De toetsen zijn dan absoluut genormeerd. De aandacht gaat in zo'n geval uit naar **welke** (clusters van) eindtermen wel en niet gehaald zijn, om vervolgens het onderwijs daarop aan te passen. Daarmee kan de toetsing een formatieve rol hebben in het leerproces. Deze formatieve rol zal echter in de volgende fase van het onderwijs doorlopen, aangezien het beoogde afnamemoment aan het einde van een fase in het onderwijstraject is, in de overgang naar een nieuwe fase.

Het moment van de meting, aan het einde van een fase, is over het algemeen meer overeenkomstig met het summatief gebruik van toetsen. Zeker als we ook kijken naar het laatste meetmoment (zesde jaar van het secundair onderwijs). Heeft een leerling voldoende eindtermen behaald? Ongeacht waar de precieze grens tussen voldoende en onvoldoende ligt, krijgt het centrale toetsen dan een summatieve functie. Dat is zeker het geval wanneer

op basis daarvan een beslissing volgt over de voortgang van het leertraject van de leerling. Dat geldt ook wanneer een relatieve normering wordt toegepast. De leerling krijgt dan een terugkoppeling van de prestaties relatief ten opzicht van de andere vergelijkbare groep leerlingen. Ongeacht de definitie van de "vergelijkbare groep" wordt een indicatie gegeven hoeveel procent van de leerlingen beter dan wel slechter presteert.

Wanneer de leerling voor een tweede keer gemeten wordt met centrale toetsen, minstens twee jaar na de eerste meting, wordt het mogelijk individuele leerlingen ook een terugkoppeling te geven over de individuele leerwinst. Vaak is de formatieve functie van deze terugkoppeling van leerwinst voor de leerling beperkt, aangezien die vooral gebaseerd is op de inhoudelijke interpretatie van het behalen van een eindterm. Dit betreffen vaak relatieve interpretaties, waarin de leerwinst van een leerling afgezet wordt tegen de leerwinst die bij andere leerlingen plaatsgevonden heeft. Bij die interpretatie is bovendien het startpunt van de eerste meting van belang: is de leerwinst vergelijkbaar met leerlingen die bij die eerste meting dezelfde vaardigheid hadden? De beperkte formatieve functie van de terugkoppeling van leerwinst op leerlingniveau is in dat de leerling kan nagaan bij welke vaardigheden wellicht niet de groei is behaald die te verwachten was. Het is niet evident hoe beslissingen over de individuele leerling voortvloeien uit die informatie. Het is af te raden op basis van de informatie over de leerwinst enig (summatief) besluit over de leerling te nemen. Een dergelijk besluit zal eerder gebaseerd moeten worden op de op dat moment geleverde prestatie, ongeacht eerdere prestaties.

In de terugkoppeling naar scholen kan de informatie over de leerlingen op schoolniveau samengevat worden. Daarvoor zijn verschillende mogelijkheden, die in hoofdstuk 5 over rapportages nader uitgewerkt worden. Op basis van de formatieve informatie over welke eindtermen behaald zijn, kan de school de lessen aanpassen. Op basis van summatieve informatie waarbij de prestaties van de school afgezet worden tegen de prestaties van andere vergelijkbare scholen, kan een school zich spiegelen. Een school kan zo een indruk krijgen of ze het goed doen, gemiddeld of matig doen ten opzichte van andere scholen. Het is informatie die scholen vaak graag zelf wel willen hebben, om zo beter grip te hebben op de prestaties. Dat zal zowel gelden voor de informatie over het behalen van de eindtermen, als informatie over leerwinst na een volgende meting[50].

Een nadeel blijft echter bij summatieve informatie dat een school niet graag ziet dat ze bij de slechter presterende scholen behoort, en de prestaties mogelijk zal willen verhogen op een oneigenlijke manier en niet door beter onderwijs. Bij een terugkoppeling van de informatie die alleen zichtbaar is voor de betrokkenen op de school zal dit negatieve effect van summatief toetsen beperkt zijn, maar als ook anderen deze informatie ter beschikking krijgen, zal de druk op scholen toenemen. Het kan tot onderwijsvernauwing leiden: een sterk vergrote aandacht voor de inhoud van de toetsen ten koste van andere voor onderwijs ook zeer relevante zaken. Het zal hier van de onderwijskundige visie afhangen of dit als een groot probleem gezien wordt.

In het regeerakkoord en de beleidsnotitie staat dat de informatie geanonimiseerd wordt

[50]Merk op dat op schoolniveau een tweede meting al plaats kan vinden in het jaar na de eerste meting. Er wordt dan weliswaar geen cohort gevolgd, en dus niet de leerwinst van leerlingen gemeten, maar wel kan gezien worden of de school een betere (of slechtere) prestatie levert binnen het gemeten leerjaar. Of dat het gevolg is van een andere vorm van lesgeven of dat het de vaardigheid van de groep zelf betreft, is daarbij niet definitief te achterhalen.

op individueel niveau en aan de overheid (inclusief inspectie en onderwijsverstrekkers) en onderzoekers ter beschikking gesteld wordt. Op schoolniveau is die anonimiteit niet genoemd en (naar het zich laat aanzien) niet beoogd. Er wordt gemeld: "Aan de hand van de resultaten van de net- en koepeloverschrijdende proeven, kunnen we[51] bijsturen waar dat nodig is. Scholen waarvan de leerlingen significant minder leerwinst genereren op die proeven, moeten in een vrij te kiezen begeleidingstraject stappen om de kwaliteit van hun onderwijs te verhogen[52]." Dit impliceert een formatieve functie aangezien het doel lijkt te zijn het leerproces te verbeteren. Dat wordt verder benadrukt door de zeer duidelijke stellingname: "Het is absoluut niet mijn bedoeling om een rangschikking van scholen op te stellen, wel om de leerwinst te vergroten[53]."

Ondanks deze duidelijke bedoelingen voor de scholen bestaat er wel een risico dat de toetsresultaten als high-stakes ervaren zullen worden door de scholen. Dat zal onder meer afhangen van hoe zij een bijsturing beleven. Als zij dit als ongewenst zien, zullen scholen dit gevolg trachten te vermijden. De belangen zijn ook groot als de resultaten publiek ter beschikking komen en de ouders de informatie over schoolprestaties ook kunnen gebruiken in functie van schoolkeuze, zoals gesteld wordt in de aanvraag voor het steunpunt voor de ontwikkeling van de centrale toetsen[54]. In die aanvraag wordt erkend dat daarmee het risico van een rangschikking van scholen toeneemt en dat mogelijk de uitkomsten van Perceel 3 van de haalbaarheidsstudie, aangaande de technisch-juridische aspecten, uitkomst kan brengen[55]. Echter ook zonder die algemene ranking voor heel Vlaanderen maakt de beschikbaarheid van de informatie aan ouders voor de schoolkeuze de toets al van groot belang voor de school. Aangezien de locatie de schoolkeuze meestal beperkt hoeven zij niet de informatie te hebben van heel Vlaanderen, maar van een relatief kleine verzameling scholen die voor hen relatief makkelijk te bereizen is. Zij kunnen daarmee een lokale rangschikking maken die voor de school van groot belang is, ook als de betrokken onderwijsinstellingen de mogelijkheid krijgen om de informatie over hun instelling te contextualiseren. Hiermee verandert de formatieve functie van de toetsen in een summatieve functie met grote belangen, met de beschreven mogelijke nadelige gevolgen van dien. Dan maakt het de facto niet meer uit of er een algemene Vlaamse rangordening is of niet.

Die uitdaging is wellicht nog groter bij het rapporteren van de gegevens over leerwinst die na twee metingen kan plaatsvinden. Ook hierbij is de uitdaging dat de geobserveerde leerwinst beïnvloed wordt door zaken die niet direct aan de school toe te wijzen zijn. Het

[51] In de beleidsnotitie wordt vaker verwezen naar 'we', waarbij de interpretatie kan variëren van de inwoners van Vlaanderen als geheel ('het Vlaamse volk') naar de Vlaamse overheid, dan wel deze regering. In dit geval gaan we in de interpretatie ervan uit dat het de Vlaamse overheid betreft.

[52] Citaat uit Beleidsnota Onderwijs 2019-2014 (`https://www.vlaanderen.be/publicaties/beleidsnota-2019-2024-onderwijs`), te vinden op p.37.

[53] Citaat te vinden p. 65 van de Beleidsnota Onderwijs 2019-2014. Dit wordt net iets anders verwoord bij paragraaf 1.2.1 (p.23) van het 2019-2024 Regeerakkoord.

[54] Zie de aanvraag tot erkenning en toelating als steunpunt voor het thema "Ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde net- en koepeloverschrijdende toetsen in Vlaanderen", p.10 paragraaf "Rapportage van de resultaten op leerling-, klas-, school- en systeemniveau en feedback aan scholen, leraren en leerlingen" (`https://onderwijs.vlaanderen.be/nl/oproepen-voor-onderzoeksvoorstellen#Voorbij`)

[55] Een algemene ranking zal voor het steunpunt een stevige uitdaging worden. De wet openbaarheid van bestuur kan hier een rol spelen, maar ook de mogelijkheid om via de voor ouders beschikbare informatie door slim zoeken – ondersteund door digitale technieken – een rangordening te verkrijgen.

contextualiseren is dan van nog groter belang. In Hoofdstuk 4 over leerwinst laten we zien dat dit kan door toepassing van diverse modellen, waarbij een model de (lokale) rangordening kan beïnvloeden – het nadeel is echter dat er geen enkel model bestaat dat de werkelijkheid volledig omvat ("All models are wrong")[56] . Wat de keuze van het beste model is, zal deels ook afhangen van arbitraire aannames en subjectieve standpunten. Als er vervolgens zware consequenties uit gaan voortkomen, kan dat problematisch zijn. Hoe hier mee om te gaan is een onderwerp van Hoofdstuk 4. Dat onder welk model dan ook de resultaten, en daarmee de toetsafname van groot belang kunnen zijn voor scholen is het belangrijkste punt voor dit hoofdstuk.

Zoals eerder in het huidige hoofdstuk aangegeven is, bepaalt de vorm van de rapportage het gebruik van de toets, en daarmee ook het door de gebruikers ervaren doel van de toetsen. Zodoende is de informatie voor het steunpunt aangaande de rapportage van groot belang om het door de belanghebbenden ondervonden doel te kennen. Er wordt in die aanvraag dan ook gesteld dat het belangrijk is dat de feedback zo geformuleerd wordt dat deze te gebruiken is voor de beoogde doelstellingen. Voor een beoogd formatief gebruik betekent dat dat de rapportage inhoudelijk is. Ook wordt aangegeven dat de feedback op klas- en daarmee (deels) op leerkracht-niveau plaats moet vinden. Het is evident dat deze informatie vooral formatief moet zijn, en geenszins summatief[57]. Het beoordelen van individuele leerkrachten op basis van toetsresultaten van leerlingen is iets dat ten allen tijden vermeden dient te worden[58].

De in de aanvraag voor het steunpunt genoemde rapportage met de analyses op systeemniveau geeft ook duidelijk weer dat het doel niet alleen meten op leerling- of schoolniveau is, maar dat er ook een hoger beleidsdoel speelt. Die informatie is voor de overheid van groot belang en wordt formatief gebruikt om het onderwijs te verbeteren, niet zozeer om af te rekenen.

2.2.2 Kritieken op centraal toetsen in Vlaanderen

In bovenstaande alinea's is al aangegeven dat er enige uitdagingen zijn in het balanceren van de toetsdoelen. Naast kritiek op het gebruik van toetsen in Vlaanderen[59], was men zeer lange tijd dusdanig huiverig voor centrale toetsen dat deze niet zijn ingevoerd. Zeker na de aankondiging van de (mogelijke) invoer van centrale toetsen kunnen in de publieke opinie op diverse plekken kritieken gevonden worden[60,61]. In deze kritiek zien we onderwerpen en zorgen terug die hierboven ook aangestipt zijn. Willen de centrale toetsen een succes

[56]Box, G.E.P. (1976), Science and statistics, *Journal of the American Statistical Association*, 71 (356): 791–799.

[57]Rapportage met relatieve normen leidt snel tot een summatieve interpretatie, en leidt af van een formatieve interpretatie. Bij relatieve normering gaat de focus af van de inhoud van de schaal, en wat de leerling nu aan vaardigheid bezit, en richt de aandacht op anderen. Hoe anderen het doen, zegt inhoudelijk weinig over wat iemand zelf kan. Een inhoudelijke normering gericht op de relatie met de opgaven of eindtermen is meer informatief in een formatief kader. Meer over de vorm van de rapportage in Hoofdstuk 6.

[58]Wainer, H. (2011). *Uneducated Guesses: Using Evidence to Uncover Misguided Education Policies*. Princeton University Press.

[59]Standaert, R. (2014). De becijferde school - meetcultus en meetcultuur. Acco uitgeverij.

[60]Voorbeelden hiervan zijn Wenmackers, S. & Ginis, V. (2020). Kritiek: Te vroeg om een nieuwe toetsfabriek te bouwen. De Standaard, 30 juni 2020 (`https://tinyurl.com/y6al7op7`).

[61]Spruyt, B. & Van Houtte, M. (2020). Opinie: Maak van gestandaardiseerde toetsen geen centrale testen (`https://www.vub.be/TOR/opinie-maak-van-gestandaardiseerde-toetsen-geen-centrale-testen/`).

worden dan moet aan voldoende randvoorwaarden voldaan worden. De toetsen moeten ingezet kunnen worden om te helpen, en niet om de leerlingen en de scholen te beoordelen. De kwaliteitszorg moet de kwaliteit van het onderwijs bevorderen en ondersteunen, en niet alleen bewaken. Zeker niet als dat bewaken tot negatieve maatregelen voor leerlingen en scholen kan leiden.

Als de toetsen alleen zouden leiden tot het afrekenen van leerlingen en scholen dan is het lastig te stellen wanneer er voldoende rekening gehouden moet worden met de sociale context. Er is geen model dat alles verklaart en overal rekening mee kan houden, waarover meer wanneer we het hebben over toegevoegde waarde. Bij het afrekenen wordt het leren vaak uit het oog verloren, doordat fouten afgestraft worden, terwijl men juist van fouten kan leren. Zorgen over vernauwing van het curriculum en teaching to the test worden genoemd, evenals de kosten van de invoer. Die kosten zullen met name spelen naarmate de centrale testen meer high-stakes voor scholen en leerlingen worden, omdat veel maatregelen genomen moeten worden om mogelijke fraude tegen te gaan.

Ook de zorgen over de onduidelijkheid wat de inhoudelijke richting is waarin dergelijke centrale testen ontwikkeld worden, is genoemd. Het is noodzakelijk om extreem duidelijke –bij voorkeur zeer concrete- definitieve eindtermen te hebben. Het moet duidelijk zijn wat de eindtermen voor iedereen betekenen en in welke mate deze voor de verschillende onderwijsvormen gelijk aan elkaar zijn.

Tot slot moet het uiteindelijke doel strak gedefinieerd zijn. Is de focus kwaliteitszorg, kwaliteitsbevordering of kwaliteitsbewaking? Is de focus meer gericht op de formatieve functie van het toetsen, of de summatieve functie? Echt formatief toetsen is onderdeel van een cyclus. De cyclus start bij het communiceren van verwachtingen. Deze stap wordt gevolgd door het verzamelen van informatie, bijvoorbeeld met behulp van toetsen. De informatie die dat oplevert moet geïnterpreteerd worden, waarna de communicatie met de leerling volgt. Dat laatste bevat in ieder geval instructie op maat –relevant voor waar de leerling is in het leerproces– in combinatie met verwachtingen van wat er geleerd wordt, waarmee de eerste stap van de cyclus weer bereikt is. Het is duidelijk dat deze formatieve toetsen een directe positieve invloed (kunnen) hebben op wat de leerlingen kunnen en kennen. In die zin zou het mooi zijn als het merendeel van de toetsen in het onderwijs een formatieve functie heeft. Helpen om leerlingen (en leraren) nog vaardiger te maken, daar kan niemand op tegen zijn.

Er zijn hele boeken gewijd aan de voordelen van formatief gebruik, en die vooral ingaan op de nadelen van summatief toetsen[62]. De stelling dat we geheel af zouden kunnen van die summatieve toetsen, lijkt echter niet houdbaar. We kennen vast allemaal de vraag "moeten we dat kennen voor de toets?" En als dat dan niet het geval was, dan negeerde de meerderheid van de klas dat deel van de stof. Hoe interessant de stof verder ook was. Maar omgekeerd, als het juist wel deel van het proefwerk is, dan kan dat leerlingen stimuleren iets te leren waar ze anders uit zichzelf niet aan zouden beginnen. En in verrassend veel gevallen gaan leerlingen de stof daarna ook leuk vinden. Zo hebben zowel summatieve als formatieve toetsen een belangrijke rol in het onderwijs. Ook kan een goede summatieve toets een leerkracht ondersteunen in het bereiken van kansengelijkheid. In plaats van de centrale toetsen te zien als geïnstitutionaliseerd wantrouwen in leerkrachten kunnen ze in een summatief kader waarbij de vaardigheid van de leerling centraal staat ook gezien

[62] Stobart. G. (2008). *Testing Times: The Uses and Abuses of Assessment*. Routledge.

worden als hulpmiddel om goede beslissingen te nemen. Een leerkracht is ook maar een mens en kan daar ook vaak hulp bij gebruiken[63].

We hebben beide vormen van toetsen nodig. Er is wel een belangrijke waarschuwing: als je eenzelfde toetsmoment zowel formatief als summatief wil inzetten, kan dat behoorlijk wat problemen geven, vanwege de verschillen in de benadering van fouten: vermijden dan wel ervan leren. Daardoor is het wenselijk voor een dergelijk toetsmoment één vorm van toetsen te kiezen.

In alle gevallen spelen bij de invoer van centrale toetsen ook begrippen als gelijkheid en vergelijkbaarheid een rol. Dat wil zeggen, dat onafhankelijk van waar je school staat je als leerling gelijke kaders krijgt, gelijke centrale meetmomenten, of deze nu een formatieve of een summatieve functie hebben. De balans tussen de voor- en nadelen bij het summatief inzetten van de toetsen moet gekend zijn. Daarbij nemen we zeker in ogenschouw dat de haalbaarheid van de centrale toetsen ook over aanvaardbaarheid gaat. Zonder voldoende acceptatie in het veld, is de kans dat de centrale toetsen haalbaar en succesvol te handhaven zijn een enorme uitdaging. Daarom is het van belang draagvlak te creëren en verbinding te zoeken met het onderwijsveld bij de invoer van centrale toetsen.

[63] Zie onder andere `https://en.wikipedia.org/wiki/Decision-making` en `https://en.wikipedia.org/wiki/Bias`.

3. Itemresponstheorie

Veel onderwerpen in de volgende hoofdstukken hebben betrekking op modellen die afkomstig zijn uit de itemresponstheorie (IRT). Het grote voordeel van IRT ten opzichte van bijvoorbeeld klassieke testtheorie is dat het heel veel flexibiliteit biedt in het vergelijken van resultaten behaald op verschillende toetsen. Juist in een context waarin gezocht wordt naar mogelijkheden om te werken met verschillende varianten van toetsversies, biedt IRT dus aantrekkelijke voordelen. Om deze reden geven we in dit hoofdstuk eerst een korte introductie over deze modellen die een prominente rol spelen in het moderne onderwijskundig meten. De volgende hoofdstukken die de vragen uit het bestek adresseren, kunnen los van dit hoofdstuk gelezen worden. Dit hoofdstuk dient dan ook vooral gezien te worden als ondersteunend hoofdstuk dat achtergrondinformatie geeft over IRT. Het hoofdstuk gaat ook in op de voor- en de nadelen van de IRT, en de voorwaarden die moeten gelden voordat deze modellen toegepast kunnen worden, die zodoende ook impact hebben op de mogelijke scenario's.

3.1 IRT modellen

IRT omvat statistische meetmodellen waarin het antwoordgedrag van leerlingen gemodelleerd wordt als een functie van persoons- en itemkenmerken[1,2]. Meestal betreft het 'gedrag' de kans op een goed antwoord, het meest belangrijke persoonskenmerk is veelal de vaardigheid van een leerling, en het meest belangrijke kenmerk van een item de itemlocatie of itemmoeilijkheid. De itemlocaties en de persoonsvaardigheden worden veelal afgebeeld op één dimensie: de latente trek of vaardigheidsschaal.

Er bestaan vele modellen binnen de IRT. De belangrijkste modellen voor dichotoom

[1] Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Psychology Press.

[2] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

gescoorde items (dat wil zeggen items die hetzij goed of fout worden gescoord), zijn het 1-parameter logistisch model (ook wel bekend onder het Rasch-model)[3], het 2-parameter logistisch model[4] en het 3-parameter logistisch model[5]. De modellen verschillen in het aantal itemkenmerken (of itemparameters) dat wordt opgenomen in het model om het responsgedrag van leerlingen te verklaren. Een parameter die altijd opgenomen is in de modellen is de positie op de schaal. Als dit per item de enige parameter in het model is om dit item te beschrijven dan is deze parameter te interpreteren als de moeilijkheid van de opgave. In een 2-parameter logistisch model wordt een discriminatie-parameter toegevoegd in het model. Deze parameter geeft aan in welke mate het item een onderscheid maakt tussen minder en meer vaardige leerlingen. Een 3-parameter logistisch model tenslotte, voegt nog een derde itemkenmerk toe aan het model. Dit kenmerk modelleert dat leerlingen ook door te gokken een item goed kunnen maken. De belangrijkste modellen uit de IRT voor polytoom gescoorde items (items die een maximum score hoger dan 1 hebben) zijn het *partial credit model*[6] en het *generalized partial credit model*[7].

Met voldoende afnamegegevens kan een IRT-model geschat worden. Als een model op alle items in de itembank geschat is, dan zijn de statistische itemkenmerken, zoals de moeilijkheid, van alle items met elkaar te vergelijken. Het schatten van de itemkenmerken en het beoordelen of deze itemkenmerken ook passen bij de data (zie Secties 3.3 en 3.5) wordt kalibreren genoemd. Je kunt dan ook de moeilijkheid van sets van items, ofwel toetsvarianten die uit een verzameling items (zie ook Paragraaf 5.1.3 over toetsitemdatabanken) zijn samengesteld, voorspellen. Iedere meting die plaatsvindt met toetsvariant -oftewel een verzameling van opgaven- uit deze gekalibreerde itembank, kan worden gerelateerd aan de vaardigheidsschaal die geschat is met het IRT-model. Op basis van een score - op welke toetsvariant dan ook - kan de vaardigheid van de kandidaat bepaald worden. Zo kunnen kandidaten die verschillende toetsen maken alsnog direct met elkaar vergeleken worden. Dat is een groot verschil met klassieke testtheorie waarbij dat aanzienlijk moeilijker, zo niet onmogelijk is[8]. IRT biedt zo de mogelijkheid om veel varianten van een toets uit te brengen, en de resultaten van de leerlingen dan toch zeer goed met elkaar vergelijkbaar te maken.

3.2 Equivaleren

Verschillende toetsversies kunnen met behulp van IRT geëquivaleerd worden. Equivaleren betreft het statistisch proces om tot vergelijkbare scores te komen die behaald zijn op

[3]Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests.(Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

[4]Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

[5]Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

[6]Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174.

[7]Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm, *Applied Psychological Measurement*, 16(2), 159-177.

[8]Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer Science+Business Media, Inc.

verschillende toetsversies[9]. Ook bij een reguliere equivalering zijn er vijf voorwaarden die algemeen beschouwd worden als noodzakelijk voor equivalentie[10]. Deze vereisten zijn:

- Het vereiste van een gelijk construct: de twee test(versie)s moeten beiden metingen zijn van een en hetzelfde construct (latente eigenschap, vaardigheid).
- Het vereiste van gelijke betrouwbaarheid: de twee varianten van de tests moeten hetzelfde betrouwbaarheidsniveau hebben.
- De symmetrie-eis: de gelijkstellingstransformatie voor het in kaart brengen van de scores van Y tot die van X zou het omgekeerde moeten zijn van de vergelijkingstransformatie voor het in kaart brengen van de scores van X tot die van Y.
- De billijkheidseis: Het mag niet uitmaken voor een kandidaat welke van de twee test(versie)s de examinandus daadwerkelijk aflegt.
- De populatie-invariantie-eis: de transformatie van de scores van de ene naar de andere versie moet gelijk zijn ongeacht de keuze van (sub)populatie waarvan deze is afgeleid.

Het vergelijken van toetsen binnen eenzelfde niveau, in een zelfde afnamejaar, of over afnamejaren heen, wordt horizontaal equivaleren genoemd. Wanneer we te maken hebben met het vergelijkbaar maken van toetsen die aanzienlijk van niveau verschillen, zoals het geval is wanneer we leerjaren met elkaar vergelijken, dan is er sprake van verticaal equivaleren. Deze laatste vorm van equivaleren is lastiger omdat verschillen in het construct tussen verschillende leerjaren niet alleen kwantitatief van elkaar verschillen (leerlingen krijgen door de tijd een hogere vaardigheid), maar ook kwalitatief (leerlingen krijgen door de tijd een iets andere vaardigheid). In de volgende sectie over de voorwaarden voor het toepassen van IRT zal duidelijk worden waarom.

3.3 Voorwaarden voor het gebruik van IRT

Het werken met IRT biedt dus veel voordelen, maar voordat van deze voordelen genoten kan worden, moet een IRT model wél passen op de data. Dat betekent dat aan een aantal voorwaarden voldaan moet worden. De eerste voorwaarde is dat een opgave maar één vaardigheid meet, en dat de opgaven op de vaardigheidsschaal allen dezelfde vaardigheid meten. Dit is de aanname van unidimensionaliteit[11]. De tweede aanname is de aanname dat de kans op een specifiek antwoord op een vraag niet beïnvloed wordt door de vragen ervoor (of erna). Dat komt erop neer dat de vragen geen hints mogen bevatten voor andere vragen in de toets[12]. Tot slot is er de aanname dat de kans op een antwoord een specifieke wiskundige formule volgt, waarbij de vaardigheid van de kandidaat een rol speelt en de kenmerken van het item[13].

[9] Kolen, M.J., & Brennan, R.L. (1995). *Test Equating*. New York: Springer.

[10] Holland, P.W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement 4th ed.*, (pp. 187-220).Westport, CT: Praeger.

[11] Voor een introductie rondom multidimensionele IRT modellen verwijzen wij naar Ackerman„ T.A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests Are Measuring, *Applied Measurement in Education*, 7(4), 255-278.

[12] Deze aanname wordt lokaal stochastische onafhankelijkheid genoemd.

[13] Deze aanname wordt aangeduid met monotoniciteit.

Deze voorwaarden hebben consequenties voor de toepasbaarheid van het model. Met name de eis van unidimensionaliteit levert belangrijke beperkingen op. Het is moeilijk een schaal te vinden die unidimensioneel is in zeer strikte zin. Daartoe moeten de opgaven in sterke mate hetzelfde meten. Zo zal het mogelijk zijn een unidimensionele schaal te vinden voor de vaardigheid om te kunnen optellen met getallen onder de tien. De schaal met opgaven voor de vaardigheid voor zowel optellen als aftrekken zal al minder unidimensioneel zijn, maar die is wel weer meer unidimensioneel dan een schaal waar de vaardigheid "rekenen" gemeten wordt – zeker als daar "kale opgaven" en vraagstukken[14] gecombineerd worden. Het is duidelijk dat hoe meer unidimensioneel de schaal is, hoe meer gedetailleerd het construct beschreven wordt.

De vraag is dan: is de schaal unidimensioneel genoeg? In het geval de strikt unidimensionele schalen allen onderling (latent) zeer hoog samenhangen –zoals dat bij rekenen binnen een leerjaar meestal het geval is– is de samenstelling naar een schaal rekenen meestal geen probleem. Mocht dat niet het geval zijn, zoals mogelijk bij rekenen met of zonder context dan kan ervoor gekozen worden om binnen IRT verschillende schalen voor de twee te onderscheiden vaardigheden te maken. Op basis van de correlatie tussen deze vaardigheden valt er dan mogelijk weer een enkele score te maken.

Bij een unidimensionele meting kan gesteld worden dat de verschillen tussen leerlingen kwantitatief zijn: de ene leerling heeft een hogere vaardigheid dan een andere leerling. Zodra de verschillen tussen leerlingen niet kwantitatief maar kwalitatief worden, betekent dat dat de ene leerling iets anders kan dan de andere leerling. Dat maakt een vergelijking op de schaal moeilijk, omdat dit niet noodzakelijk unidimensionele verschillen zijn. De verschillen tussen metingen[15] zijn zogezegd niet op een lijn weer te geven. In Hoofdstuk 2 was hiervoor het verschil al gegeven tussen leerlingen die meer van meetkunde weten en anderen die meer van statistiek weten, wat beiden onder wiskunde valt, maar toch lastig te vergelijken is. Er kan ook sprake zijn van een kwantitatief én een kwalitatief verschil tussen leerlingen. Het kwalitatieve verschil maakt het dan moeilijk het kwantitatieve verschil goed op waarde te schatten, hetgeen tot een onder- en overschatting van het verschil kan leiden.

Zoals met de voorbeelden ook deels al is aangegeven, zijn verschillen in het curriculum een belangrijke oorzaak voor kwalitatieve verschillen tussen metingen. Als groepen leerlingen van elkaar verschillen wat betreft de onderwerpen die ze gehad hebben binnen het curriculum wordt de vergelijking bemoeilijkt. Dat kan een uitdaging zijn bij het volgen over leerjaren. Een voorbeeld kan dit wellicht duidelijk maken. Als we twee rekensommen vergelijken, namelijk de vraag wat 327 opgeteld bij 8.984 oplevert en de vraag wat de wortel van 4 is, dan geldt voor leerlingen op de basisschool dat de eerste som weliswaar moeilijk zal zijn, maar op te lossen, maar dat de tweede vraag aanzienlijk moeilijker zal zijn, omdat basisschoolleerlingen, op een enkeling na, nog nooit van een wortel gehoord zullen hebben. In het secundair onderwijs zal juist de vraag naar de wortel van 4 als gemakkelijker ervaren worden vergeleken de optelsom en minder fouten opleveren. Doordat opgaven over een grote afstand tussen leerjaren moeilijk uitwisselbaar zijn, kan dat een probleem

[14] Met kale opgaven worden sommen zonder context bedoelt $(1 + 1)$, daar vraagstukken in een verhaal zijn opgenomen ("Jan koopt een appel en Toby koopt een appel. Hoeveel appels hebben zij samen?").

[15] Dit kan dan gaan over metingen van de twee verschillende leerlingen die bijvoorbeeld een verschillend curriculum hebben gevolgd, dan wel twee metingen van dezelfde leerling met een tussenpoos van twee tot vier jaar.

opleveren voor de unidimensionaliteit.

Dit probleem van kwalitatieve verschillen kan ook spelen als een onderwerp op een andere manier behandeld wordt. Het curriculum van wiskunde in het bso is aanzienlijk minder theoretisch dan in het aso. Daardoor verschillen de leerlingen aan het einde van de derde graad tussen deze twee stromen niet alleen kwantitatief van elkaar maar ook in wat zij kunnen. Als een proef aan het einde van de derde graad bestaat uit vooral praktische opgaven, dan zal het geobserveerde verschil in vaardigheid tussen bso- en aso-leerlingen kleiner zijn dan als de proef vooral bestaat uit theoretische opgaven.

Een andere voorwaarde binnen IRT modellen is dat de geschatte parameters in de modellen constant zijn voor alle groepen[16]. We zien dat deze voorwaarde bijvoorbeeld geschonden wordt bij de twee voorgaande rekenopgaven. Als we nu de moeilijkheidsparameter van de optelsom (β_0) constant veronderstellen zal de moeilijkheidsparameter van de wortelopgave (β_w) binnen het basisonderwijs hoger zijn dan β_0 (dus, $\beta_w > \beta_0$), terwijl deze binnen het secundair onderwijs lager ligt dan β_0 (dus, $\beta_0 > \beta_w$). Het is duidelijk dat β_w dan niet constant is[17].

Verschillen tussen geschatte parameters bij verschillende groepen zijn ook te interpreteren als vraagpartijdigheid of *differential item functioning* (DIF)[18]. Een andere vorm van DIF kan plaatsvinden als de groepen van leerlingen zodanig in hun taalvaardigheid verschillen dat de gebrekkige leesvaardigheid een probleem oplevert voor het oplossen van bijvoorbeeld vraagstukken[19]. In dat geval zullen leerlingen in een laagtaalvaardige groep vraagstukken als relatief (veel) moeilijker beschouwen dan kale sommen in vergelijking met leerlingen in een meer taalvaardige groep. Als twee leerlingen even goed in rekenen zijn, dan heeft een leerling uit de eerste groep een lagere kans het item goed te beantwoorden dan een leerling uit de taalvaardige groep. Ook dit levert problemen bij het vergelijken van leerlingen op. Er zijn verschillende manieren om dit op te lossen. Een mogelijke manier is om het taalniveau van de vraagstukken zo laag te maken dat dit niveau geen belemmering meer vormt om de opgave goed te maken. Als dat niet mogelijk is, dan is het een optie om een aparte vaardigheidsschaal te maken voor de vaardigheid om vraagstukken te kunnen maken en de vaardigheid om kale sommen te maken. Het gevolg is dan wel dat er in dat geval twee rekenvaardigheidsschalen zijn[20].

Het is duidelijk dat de situatie rond de schendingen van unidimensionaliteit en niet-constante parameters ook uitdagingen opleveren voor het bepalen van leerwinst. Een

[16] Iets exacter is het om te stellen dat de afstand tussen parameters constant zijn, op een lineaire transformatie na.

[17] Uiteraard kan dit ook beschreven worden in de vorm waarbij β_w constant verondersteld wordt, en β_0 niet contant is.

[18] Holland, P.W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

[19] Hiermee worden rekenvragen bedoeld waarin door middel van een (korte) tekst een realistisch toegepast kader gegeven wordt waarin het maken van de rekensom de oplossing geeft. In Nederland worden dergelijke vraagstukken redactiesommen genoemd.

[20] De vergelijking tussen de twee typen rekenvaardigheid, horend bij de kale som dan wel een vraagstuk, kan bij een leerling interessante informatie opleveren in een profielanalyse. Dat er twee vaardigheden onderscheiden worden, kan wel een nadeel zijn als we het alleen over "rekenen" willen hebben. Er zijn echter diverse methoden beschikbaar om alsnog tot een samengestelde rekenschaal te komen. Als beiden vaardigheden geen perfecte (latente) correlatie hebben –en dat is het geval wanneer de analyses laten zien dat er gebruik gemaakt moet worden van twee aparte schalen– dan maakt het voor die samengestelde rekenscore wel uit hoe de beide typen gewogen worden. Zijn ze evenwaardig, of moet de ene vorm zwaarder wegen dan ander? Een psychometrisch antwoord op deze vraag valt hier niet op voorhand te geven.

belangrijke voorwaarde om de groei van leerlingen op de vaardigheidsschaal te kunnen volgen, is dat meetmomenten die nu minstens twee jaar uit elkaar liggen, met elkaar te verbinden moeten zijn. Om twee meetmomenten met elkaar te verbinden moeten er opgaven zijn die op beide meetmomenten afgenomen kunnen worden. Dat is met deze afstand lastig omdat de moeilijkste vraag voor het eerdere moment te makkelijk is voor het latere moment, en de makkelijke vraag van het latere moment te moeilijk is voor het eerdere moment. Dat is al het geval wanneer er twee jaar tussen zitten, laat staan vier jaar. De oplossing is ook een tussenliggend moment mee te nemen om de schalen met elkaar te verbinden waarin opgaven uit het hogere en het lagere jaar wel samen opgenomen kunnen worden. Dit hoeft overigens geen enorm grootschalige meting te zijn. Vanaf rond de 1000 waarnemingen zal het mogelijk zijn de twee relevante jaren te koppelen door middel van een meting in een tussenliggend jaar[21]. In dat geval gaat het bij die meting niet zozeer om de leerlingen te meten, maar om de opgaven van de verschillende metingen aan elkaar te kunnen relateren. Binnen de IRT is het namelijk voor de kalibratie noodzakelijk dat er een toetsontwerp is waarbij er een koppeling is tussen de verschillende leerjaren. Hoe dit eruit ziet wordt bij de uitwerking van de designs verder behandeld.

De assumptie van lokaal stochastische onafhankelijkheid kan geschonden worden als de relatie tussen opgaven groter is dan door het IRT-model voorspeld: gegeven de vaardigheid van een leerling zijn de opgaven niet gerelateerd, en de enige correlatie tussen de opgaven wordt verklaard door de vaardigheid van de leerling. Dat betekent bijvoorbeeld dat vragen geen hints mogen bevatten voor andere vragen in de toets. Ook mag het geven van een juist antwoord op één vraag niet een voorwaarde zijn voor het juist beantwoorden van een andere vraag. Als samenhang tussen opgaven moeilijk te vermijden is, dan kan de verzameling opgaven als een enkele opgave beschouwd worden, waarbij de som van de losse opgaven de itemscore wordt. Dit kan bijvoorbeeld bij leesvaardigheid een oplossing zijn als een aantal opgaven over een enkele tekst gaan, waardoor deze opgaven onderling meer samenhangen dan opgaven over andere teksten. Een nadeel van deze oplossing is dat deze opgaven dan ook altijd in dezelfde samenstelling meegenomen moeten worden als het overkoepelende item in verschillende toetsen terecht komt.

Schendingen van de wiskundige relatie tussen parameters en kans op een antwoord kunnen worden opgevangen door een ander IRT-model te kiezen. Het Rasch-model vooronderstelt bijvoorbeeld dat alle items op gelijke wijze onderscheid maken tussen laag- en hoog-vaardige leerlingen. Als dit niet het geval is, kan voor een 2-parameter logistisch model gekozen worden.

3.4 Voordelen van IRT

De mogelijkheid tot equivaleren biedt een heleboel flexibiliteit, met alle voordelen van dien. Doordat niet iedereen dezelfde toets hoeft te maken om de resultaten met elkaar vergelijkbaar te maken, is het mogelijk binnen een afnameperiode verschillende toetsen af te nemen. Dit is het eerste grote voordeel van IRT.

Een tweede voordeel is dat er ook opgaven uit andere toetsen aan gerelateerd kunnen worden. Als alle leerlingen altijd een centrale toets maken, zijn gegevens van een andere

[21] Het exacte aantal hangt ook af van hoe de steekproef getrokken wordt, en bij een steekproef van scholen waarin hele klassen meedoen van de intraklasse-correlatie-coëfficiënt (ICC).

toets daaraan te relateren. Voor zo'n koppeling is geen bijkomende dataverzameling meer nodig, anders dan de afname van die andere toets. Deze andere toets kan een toets met leerplan-specifieke opgaven zijn, een toets die vanuit de koepel wordt uitgegeven, of een toets die gerelateerd is aan een internationale peiling. Dit thema zal behandeld worden in Secties 5.2 en 5.3.

Een derde voordeel is dat er een brede meting van de toetsinhoud per school kan plaatsvinden als niet alle leerlingen dezelfde toets hoeven te maken. De inhoud van de toets wordt niet alleen gedefinieerd met de opgaven in een enkele toets, maar met alle opgaven in de itembank. Als op een school verschillende versies van een toets afgenomen worden, is de meting op die school divers en dekt dit op schoolniveau het curriculum beter. Dit voordeel wordt bij systeemmetingen zoals peilingen vaak benut (zie Secties 5.2 en 5.3).

Een vierde voordeel is dat toetsvarianten toegespitst kunnen worden op de vaardigheid van leerlingen. Alleen opgaven die geschikt zijn voor een leerling, dus niet te moeilijk of te makkelijk, worden dan in zijn of haar variant opgenomen. Met behulp van adaptieve toetsen –waar verschillende mogelijkheden voor zijn– kan ook bij iedere leerling een geschikte toets voorgelegd worden (zie Sectie 5.5).

Met een itemverzameling die gekalibreerd is onder een IRT model is het ook mogelijk om leerwinst te meten. Met behulp van IRT is het mogelijk de resultaten op verschillende toetsen[22] op verschillende meetmomenten aan elkaar te relateren. Op beide momenten vindt dan een goede meting van zijn of haar vaardigheid plaats, en de groei in vaardigheid daartussen kan dan goed worden vastgesteld (zie Hoofdstuk 4).

Een laatste voordeel van de mogelijkheid tot equivaleren is het reduceren van fraude, doordat er per afname verschillende toetsvarianten kunnen worden voorgelegd. Dit kan op meerdere manieren. Tijdens de afname kan er moeilijker overleg plaatsvinden over de opgaven, want leerlingen krijgen andere opgaven, of dezelfde opgaven in een andere volgorde voorgelegd. Maar ook: als er grote hoeveelheden opgaven gebruikt worden in de diverse varianten, is het lastiger om deze allen te verzamelen en openbaar te maken. Er zijn ook andere varianten denkbaar om de geheimhouding te maximaliseren, bijvoorbeeld: leerlingen maken op één moment alle dezelfde opgaven, maar er is een set ankeropgaven die verspreid over de leerlingen en weinig frequent wordt afgenomen, naast de reguliere opgaven. Dit is dan in principe voldoende voor de vergelijking tussen jaren. Het is dan wel van belang dat al deze ankeropgaven van (zeer) goede kwaliteit zijn, hetgeen voor een deel een empirische vraag is. De reguliere opgaven worden na de afnameperiode niet meer hergebruikt, de ankeropgaven worden bij een volgende afnameperiode wel hergebruikt, maar dan in combinatie met een nieuwe set reguliere opgaven.

3.5 Nadelen van IRT

Een belangrijke uitdaging bij het gebruik van IRT kan zijn dat de rapportage er anders uitziet dan wat leerkrachten gewoon zijn. Leerkrachten zijn in hun klaspraktijk ermee vertrouwd te werken met toetsscores – vaak de somscores of aantal goede antwoorden. Daarmee kunnen zij leerlingen direct vergelijken. Wanneer leerlingen verschillende versies met verschillende opgaven maken, is dat niet meer mogelijk. De toetsscore kan voor de ene

[22] Hierbij is de inhoud van de toets aangepast op de te verwachten vaardigheid die op de verschillende meetmomenten zou moeten verschillen. Dat kan al dan niet met adaptieve toetsen.

leerling lager liggen vanwege de moeilijkheid van de gemaakte opgaven (of toetsversies) in vergelijking met andere leerlingen die andere items of versies gemaakt hebben. Dat is zeker het geval wanneer er sprake is van adaptieve afnames waarbij de moeilijkheid van de aangeboden (verzameling) items aangepast wordt aan de ingeschatte vaardigheid van de leerling.

Met behulp van IRT zijn de verschillende versies vergelijkbaar te maken, maar deze vergelijking moet ook aan de betrokkenen gecommuniceerd worden op een manier die zij als logisch en acceptabel zien. In Vlaanderen is daar al ervaring mee met rapportages vanuit het peilingsonderzoek die gebruik maken van IRT. Hoe groter de belangen zijn, hoe groter de noodzaak van de overtuigingskracht van deze rapportages. Meer over de rapportages is te vinden in Hoofdstuk 6.

Tot slot is een uitdaging dat het correct gebruik van IRT-modellen afhangt van de modelpassing. Als een model niet past, is de vraag of de conclusies die vanuit het model getrokken worden over de juistheid van de vergelijkbaarheid van scores van verschillende versies wel gewettigd zijn. De passing is een empirische vraag waar uiteindelijk pas echt antwoord op te geven is na de afname, wanneer de data verzameld zijn. Het vertrouwen op de bruikbaarheid van het model is daarmee altijd een risico dat zeker bij belangrijke beslissingen serieus genomen moet worden.

Wat betreft dit laatste punt zijn er gelukkig wel mogelijkheden de risico's te beperken. De statistische passing van het model voor alle items –zeker wanneer de aantallen waarnemingen per opgave in de tienduizenden kunnen vallen– zal bijna gegarandeerd geschonden worden. Zelfs een kleine schending van het model kan een misfit opleveren. Statistische passing is in deze situatie minder van belang dan de robuustheid van het model: is de grootte van de schendingen van het model zodanig dat het voor de uiteindelijke conclusies effectief niets uitmaakt? Dat valt goed te onderzoeken door de schattingen op basis van het model te vergelijken met de geobserveerde waarden[23] en die verschillen te evalueren.

Het risico is ook deels op te vangen door te proeftoetsen om te zien of bepaalde veronderstellingen kloppen. Dat betreft kleinere aantallen leerlingen, maar dat betekent vaak dat als daar een significant probleem gevonden wordt, dat ook een betekenisvol probleem is. Als er bij de werkelijke afname alsnog blijkt dat op zeer beperkte schaal *misfit* plaatsvindt, kan besloten worden om niet correct functionerende opgaven niet mee te nemen in de uitkomsten. Dit heeft echter ook wel nadelen aangezien de leerlingen wel tijd en moeite hebben gestoken in het maken van de opgave[24]. Het verwijderen van de opgaven zal als noodoplossing gezien moeten worden.

[23]Vanuit het model kan een schatting gedaan worden wat betreft de verdelingseigenschappen van een toets, en deze kunnen rechtstreeks vergeleken worden met de geobserveerde verdeling van de scores bij deze toets. Dat kan voor de gehele populatie en voor subgroepen binnen de populatie. Daar waar het de opgaven betreft kunnen geschatte p-, rit- en rir-waarden vergeleken worden met de geobserveerde waarden.

[24]Het goed rekenen van de opgave lost daarbij niet veel op omdat deze opgave dan als extreem makkelijke opgave gezien wordt en de normering daar dan weer voor corrigeert. Het weglaten bij iedereen en het goed rekenen bij iedereen zijn door de aangepaste normering effectief gelijk, als er niet een vaste cesuur op de scoreschaal ligt.

3.6 Conclusie

Ondanks de mogelijke risico's en uitdagingen die met de toepassing van IRT bestaan is het aan te raden deze modellen te gebruiken binnen het centrale toetsen in Vlaanderen. Het belangrijkste gevaar is dat de risico's niet onderkend worden, en men van een altijd werkend 'magisch model' uit zou gaan. Voor alle risico's en uitdagingen zijn oplossingen te vinden, sommige nu al te voorzien, anderen wanneer ze zich aandienen. We hebben de volste overtuiging dat het steunpunt voor de centrale toetsen ruimschoots geëquipeerd is om de uitdagingen te onderkennen en te ondervangen. Dan blijven de belangrijke voordelen van IRT over die het gebruik ervan rechtvaardigen:

- Door het gebruik van een vaste meetschaal kan een cesuur gemakkelijk geëquivaleerd worden, wat betekent dat alle leerlingen, over scholen en door de tijd heen eerlijk met elkaar vergeleken worden;
- Leerwinst valt ook inhoudelijk te interpreteren door middel van de (geïllustreerde) absolute voortgang op de vaardigheidsschaal (zie Hoofdstuk 4);
- Itembanken gebaseerd op IRT zijn flexibel in het gebruik en bieden de mogelijkheid tot stapsgewijze aanvulling met meer opgaven zodat zowel vernieuwing als continuïteit geborgd is (zie Hoofdstuk 5);
- Gerichte verzamelingen items (toetsversies) kunnen aangeboden worden, passend bij de ingeschatte vaardigheid van de leerling zodat de toets ook echt relevant is voor de leerling (zie Hoofdstuk 5);
- De vaardigheidsschaal is zeer goed aanschouwelijk te maken door een grote hoeveelheid opgaven zodat een bepaalde vaardigheidsscore niet alleen een abstract begrip wordt, maar werkelijk een betekenis kan krijgen in termen van wat de leerling kan (zie Hoofdstuk 6).

II Onderzoeksvragen

4. Leerwinst en toegevoegde waarde

In dit hoofdstuk wordt ingegaan op verschillende definities die bestaan rondom leerwinst en toegevoegde waarde. Vervolgens gaan we in op verschillende operationalisaties van deze begrippen. Hierbij besteden we speciaal aandacht aan leerwinst in het perspectief van leerlingen, scholen en beleidsmakers. Vervolgens beschrijven we verschillende modellen die leerwinst en toegevoegde waarde in kaart brengen en welke overwegingen gemaakt dienen te worden in de keuze voor een bepaalde benadering. Tenslotte gaan we in dit hoofdstuk in op welke manier leerwinst gemeten kan worden aan de hand van de graadspecifieke eindtermen.

4.1 Definities van leerwinst

Een in het Nederlands taalgebied gebruikelijke definitie van leerwinst is de toename van vaardigheden of kennis van leerlingen gedurende een bepaalde periode[1]. Deze winst weerspiegelt dus het verschil in kennis van leerlingen op twee verschillende momenten. Het is daarbij natuurlijk van belang dat het verschil in kennis –uitgedrukt in toetsprestaties– betrekking heeft op toetsen die hetzelfde inhoudelijke domein meten[2]. Op het moment dat deze toetsen door middel van een model uit de itemresponstheorie (zie Hoofdstuk 3) op één schaal gekalibreerd zijn,[3] wordt de interpretatie van leerwinst en toegevoegde waarde modellen eenvoudiger. Leerwinst is dan eenvoudig te berekenen, het is dan simpelweg het

[1] Bosker, R. (2012). De toegevoegde waarde van een school: Begripsbepaling, meting en causale attributie. In A.B. Dijkstra & F.J.G. Janssens (red.) *Om de kwaliteit van het onderwijs*, (pp. 93-104). Den Haag: Boom Lemma.

[2] Hamilton, L.S., McCaffrey, D.F. & Koretz, D.M. (2006). Validating achievement gains in cohort-to-cohort and individual growth-based modeling contexts. In R.W. Lissitz (Ed.) *Longitudinal and value added models of student performance*, (pp. 407-435). Maple Grove, MN: JAM Press.

[3] Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer Science+Business Media, Inc.

verschil tussen twee schaalscores. De periode waarover de leerwinst bepaald is, staat niet vast; deze kan kort (weken) of lang zijn (jaren), en alles daartussen.

Een voor de hand liggende operationalisering van leerwinst is om deze te zien als het verschil in scores tussen een toets aan het einde en een toets aan het begin van de periode. Dat zien we ook terug in de onderzoeksrapporten over PIRLS-repeat[4], waar leerwinst berekend wordt op basis van het verschil tussen de score op PIRLS2016 (4e leerjaar) en de Vlaamse PIRLS2018 (6e leerjaar). Ook daarbij wordt leerwinst gezien als het verschil in scores tussen twee meetmomenten en daarmee als de groei die leerlingen doormaken. Dit verschil kan ook geaggregeerd worden over groepen, bijvoorbeeld over klassen, scholen, regio's, of gewesten.

Dit betekent dat leerwinst ook op een geaggregeerd niveau te beschouwen valt. De gemiddelde leerwinst van een groep leerlingen kan berekend worden voor bijvoorbeeld een school of scholengemeenschap. Niet elk deel van deze geaggregeerde leerwinst is echter aan de school toe te schrijven. Leerlingen kunnen immers ook een toename in leerwinst laten zien, omdat ze in een niet-schoolse omgeving in vaardigheid zijn gegroeid. Dit kan bijvoorbeeld plaatsvinden in de thuisomgeving of op een vereniging. Juist omdat de thuissituatie en achtergrondsituatie in bredere zin van leerlingen per school kan verschillen en daarmee potentieel de gemiddelde leerwinst per school verschillend kan zijn, is het van belang om school- en leerlingeffecten van elkaar te scheiden. Daarom wordt ook vaak gebruik gemaakt van de aan leerwinst gerelateerde term "toegevoegde waarde", ook bekend als "*value-added*" in de Engelstalige literatuur. De toegevoegde waarde van de school is de bijdrage van de instelling aan de leerwinst van haar leerlingen.

4.2 Operationalisaties van leerwinst

4.2.1 Leerwinst in het perspectief van leerlingen, scholen en beleidsmakers

Door de introductie van centrale toetsen in Vlaanderen kan de leerwinst voor (vrijwel) alle leerlingen worden berekend. Een goede interpretatie van leerwinst vereist echter het plaatsen van de leerwinst in een context waarin deze behaald is. Dit geldt voor leerwinst geoperationaliseerd op het niveau van individuele leerlingen, het niveau van de school en het onderwijssysteem. Het meenemen van de omgeving in de analyse van leerwinst is geen sinecure en een behoorlijke methodologische uitdaging, zoals ook Van Landeghem en collega's (2019) al opmerkten[5].

In de literatuur wordt ook vaak gesproken over "type A" en "type B" effecten in de context van toegevoegde waarde en schooleffectiviteit[6]. Het type A effect is vooral van belang voor ouders die een school voor hun kind willen selecteren. Het type B effect is meer van toepassing voor beleidsmakers die willen weten welk deel van de leerprestaties toe te

[4]Van Landeghem, G., Dockx, J., Aesaert, K., Van Damme, J. & De Fraine, B. (2019). *PIRLS, de peilingen begrijpend lezen en loopbanen doorheen het lager onderwijs. De impact van alternatieve trajecten op de interpretatie van de prestatiemetingen*. Leuven: Centrum voor Onderwijseffectiviteit en -evaluatie KU Leuven, 2019.

[5]Van Landeghem, G., Dockx, J., Aesaert, K., Van Damme, J. & De Fraine, B. (2019). *PIRLS, de peilingen begrijpend lezen en loopbanen doorheen het lager onderwijs. De impact van alternatieve trajecten op de interpretatie van de prestatiemetingen*. Leuven: Centrum voor Onderwijseffectiviteit en -evaluatie KU Leuven, 2019.

[6]Raudenbush, S., & Willms. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307–335.

schrijven zijn aan de door de school te beïnvloeden factoren[7]. Raudenbush & Willms (1995) beargumenteren dat de leerprestaties beïnvloed kunnen worden door 1) leerlingeigenschappen (bijvoorbeeld intelligentie of doorzettingsvermogen), 2) een toevallige component, 3) de omgeving van een school, waar een school zelf niet of nauwelijks invloed op heeft en uiteindelijk ook 4) het schoolbeleid. Alleen het schoolbeleid valt onder de invloedssfeer van een school en dat is dan ook het gedeelte dat de toegevoegde waarde van een school zou moeten uitdrukken. Het type A effect wordt gezien als het gecombineerde effect van de omgeving waarin de school zich bevindt en de kwaliteit van het schoolbeleid. Het type B effect heeft alleen betrekking op de kwaliteit van het schoolbeleid. Terwijl ouders niet per se geïnteresseerd zullen zijn in het kunnen maken van het onderscheid tussen de verbetering van de leerprestaties van hun kinderen door de omgeving versus door het beleid van de school, en dus voldoende hebben aan een inschatting van een type A effect, is het voor beleidsmakers wel van belang om vast te kunnen stellen welk van de leerprestaties nu bepaald zijn door het schoolbeleid. Type B effecten zijn daardoor bij uitstek geschikt om tot een eerlijke beoordeling te komen van de toegevoegde waarde, door de scholen zelf, of door de beleidsmakers. Juist omdat omgevingseffecten en schoolbeleid gecorreleerd kunnen zijn, is het niet eenvoudig een type B effect goed vast te stellen (Raudenbush, 2004).

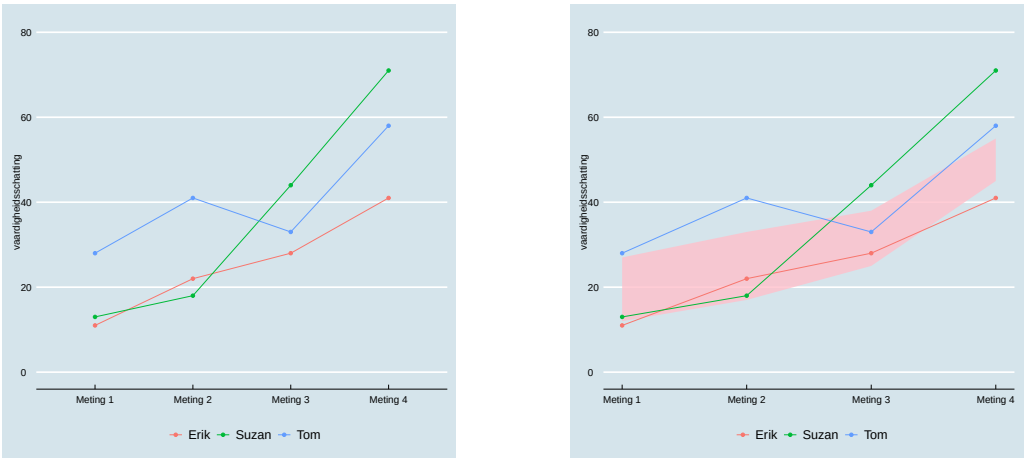
Afhankelijk van het niveau waarop men uitspraken wil doen rondom leerwinst en toegevoegde waarde, spelen er verschillende uitdagingen. Op het niveau van leerlingen is het belangrijkste vraagstuk te bepalen of een leerling voldoende gegroeid is. Op het niveau van de school is wellicht het belangrijkste vraagstuk hoe vast te stellen welk didactisch handelingsperspectief gebruikt kan worden om leerlingen verder te helpen. Beleidsmakers, tenslotte, zullen vooral geïnteresseerd zijn of de bijdrage van scholen aan leerwinst voldoende is. In de volgende paragrafen gaan we in op deze verschillende perspectieven.

Leerlingen

Voor een leerling is het van belang om vast te stellen of hij of zij voldoende gegroeid is. Dit is te doen door te beoordelen of de leerling al dan niet bepaalde vaststaande standaarden heeft bereikt. Dit thema zal verder behandeld worden in Sectie 4 van dit hoofdstuk, waarin leerwinst in de context van graadspecifieke eindtermen besproken wordt. Daarnaast kan een relatieve vergelijking plaatsvinden met de prestatie van anderen. Ten slotte is ook een vergelijking mogelijk met een eerdere prestatie van de leerling zelf. In dit geval is een longitudinale dataverzameling noodzakelijk. Alleen dan kan de groei van leerlingen ten opzichte van hun eerdere prestaties in kaart gebracht worden. Een manier om een indruk van de groei van een leerling te krijgen, is het werken met groeicurves. In Figuur 4.1 is een illustratief voorbeeld te vinden van de groeicurves van drie fictieve leerlingen op vier meetmomenten. De vier toetsen zijn op één schaal gekalibreerd, zodat de uiteindelijke toetsresultaten onderling vergelijkbaar zijn en leerlingen dus ook in de tijd gevolgd kunnen worden.

In Figuur 4.1a zijn de vaardigheidsscores van Erik, Suzan en Tom afgebeeld. Te zien is dat alle drie de leerlingen leerwinst hebben behaald wanneer de resultaten van Meting

[7] Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioural Statistics*, 29(1), 121–129.



(a) Groeicurves

(b) Groeicurves met normering

Figuur 4.1: Groeicurves

4 vergeleken worden met de resultaten behaald op Meting 1. Ook is te zien dat Tom een knikje in "zijn" leerwinst laat zien tussen Meting 2 en Meting 3. Deze daling is op verschillende manieren te verklaren, maar kan in ieder geval aanleiding geven om nader te onderzoeken of de ontwikkeling van deze leerling nog voorspoedig loopt. Een manier om de interpretatie van groeicurves te faciliteren, is het werken met zogenoemde normeringsgegevens. In dat geval is er een combinatie van een vergelijking met de leerling met zijn of haar eerdere prestatie en wordt zijn of haar prestatie vergeleken met de prestatie van anderen. Een voorbeeld hiervan is te vinden in Figuur 4.1b. Het roze gearceerde gedeelte van deze figuur zijn de normeringsgegevens. Leerlingen waarbij blijkt dat zij onder deze roze arcering presteren, kunnen aanleiding geven om extra aandacht aan hen te besteden. Dat geldt ook voor leerlingen die boven deze roze arcering presteren. Zij presteren bovengemiddeld en zijn wellicht gebaat bij additioneel uitdagend lesmateriaal.

Scholen

Leerwinst en toegevoegde waarden zijn instrumenten die ook in te zetten zijn door scholen om de kwaliteit van het gegeven onderwijs te verhogen. Gegevens rondom leerwinst zijn te gebruiken om inzicht te krijgen in de ontwikkeling van leerlingen op een bepaald vakgebied zoals in bovenstaande paragraaf betoogd is. Gegevens rondom toegevoegde waarde zijn ook door de scholen te gebruiken om de eigen onderwijspraktijk te evalueren. In een publicatie van de Nederlandse Onderwijsraad wordt geadviseerd om het eigenaarschap van dit instrumentarium bij de scholen zelf te laten liggen[8]. Dit wordt niet alleen geadviseerd om strategisch gedrag van scholen te voorkomen, maar ook omdat een beoordeling van scholen op alleen meetbare cognitieve vaardigheden als te smal wordt gezien. Gegevens over toegevoegde waarde kunnen wel een meer genuanceerd beeld geven van het succes van de onderwijspraktijk dan alleen separate eindscores. Een school waar vooral leerlingen zitten met een bovengemiddelde intelligentie zal minder inspanningen hoeven te leveren om een hoge gemiddelde eindscore te bereiken dan een school met juist veel leerlingen

[8] Onderwijsraad (2014). *Toegevoegde waarde: een instrument voor onderwijsverbetering - niet voor beoordeling*. Den Haag, Onderwijsraad, 2014.

met een benedengemiddelde intelligentie. Omgekeerd kan de laatste school echter wel veel "waarde" hebben toegevoegd aan de leeruitkomsten van haar leerlingen. Een hoge toegevoegde waarde en hoge eindopbrengsten kunnen dus wel samengaan, maar dat hoeft lang niet altijd het geval te zijn. Het in context in kaart brengen van leerwinst betekent dan ook dat traditioneel benedengemiddelde presterende leerlingen en scholen in absolute termen bij leerwinst analyses juist positief naar voren kunnen komen.

Beleidsmakers

Het is niet verwonderlijk dat beleidsmakers ook geïnteresseerd zijn in gegevens rondom leerwinst en toegevoegde waarde. Dezelfde gegevens die scholen gebruiken om te reflecteren op hun eigen onderwijsbeleid zijn ook op een landelijk niveau te aggregeren om een effectief onderwijsbeleid te kunnen vormgeven. Tegelijkertijd is ook al aangegeven dat er een aantal zaken spelen die tot voorzichtigheid manen om leerwinst en toegevoegde waarde gegevens te gebruiken op het niveau van het onderwijssysteem. Ten eerste zouden – zoals in Sectie 4.1 aangegeven – beleidsmakers op zoek moeten gaan naar zogenaamde type B effecten. Om een goed beeld te krijgen welke scholen een effectief onderwijsbeleid voeren, moet je namelijk een onderscheid kunnen maken tussen de effecten op leerwinst die door de kenmerken van de leerlingen, dan wel de omgeving van de school veroorzaakt zijn én dat deel dat wel toe te schrijven valt aan het schoolbeleid. Op het moment dat je dit onderscheid niet maakt, bestaat het risico dat er beleidsaanbevelingen worden gedaan voor alle scholen gebaseerd op scholen die in een bepaalde context opereren. Juist het goed onderscheid kunnen maken tussen omgevingseffecten en schoolpraktijk vraagt veel kennis over de context waarin de school zich bevindt. Het is complex om al deze informatie te verzamelen op het moment dat je je niet in de omgeving van de school bevindt. Daarnaast wordt het risico dat scholen proberen op een gekunstelde manier gunstig uit een toegevoegde waarde berekening te komen groter op het moment dat de consequenties van deze uitkomsten zwaarwegender worden. De consequenties zullen toenemen op het moment dat toegevoegde waarden niet alleen een rol spelen bij het verbeteren van het onderwijs, maar ook voor het verantwoorden van het onderwijs (zie ook Sectie 4.3.1).

4.2.2 Modellen voor het in kaart brengen van leerwinst en toegevoegde waarde

In de literatuur zijn verschillende benaderingen te vinden voor het in kaart brengen van leerwinst en toegevoegde waarde. Deze kunnen grofweg ingedeeld worden in regressie-,[9] stratificatie-,[10,11,12] en modelleringsbenaderingen[13]. Deze laatst genoemde methode

[9]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, 518-537, 2019.

[10]Rubin, D.B., E.A. Stuart & Zanutto, E.L. (2015) A Potential Outcomes View of Value- Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, Value-Added Assessment Special Issue, Spring, pp. 103-116. 2004.

[11]Janssens, F.J.G., Rekers- Mombarg, L. & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs*. Den Haag, Ministerie van OCW, 2014.

[12]Feskens. R.C.W.(2015). Measurement of value-added. Report of the project Development of standardized tools for the assessment and self- assessment of pupil achievement in schools of general education, stage II. Arnhem, Cito. 2015.

[13]Willms, J.D. (2011) An analysis plan on educational equality and equity: Recommendations for the oecd education at a glance. paper prepared for the OECD NESLI INES network for the collection and adjudication of

wordt ook (gedeeltelijk) toegepast in PISA 2018[14]. Deze verschillende benaderingen geven verschillende antwoorden op de vraag welke kenmerken deel zouden moeten uitmaken van een leerwinstanalyse. Afhankelijk van de gebruiker, het doel en de beschikbaarheid van achtergrondinformatie kan voor een van deze benaderingen gekozen worden.

Juist omdat er niet één manier is om leerwinst en toegevoegde waarde te berekenen, vergelijken we hier de verschillende mogelijkheden met elkaar op basis van de methodologische voor- en nadelen. Deze methoden zijn ook naast elkaar te gebruiken om scholen, en de overheid, zo goed mogelijk te informeren en te helpen het onderwijs nog beter te maken, en zo de nadelen van iedere specifieke methode te ondervangen. Bij de verschillende methoden kan op verschillende manieren rekening gehouden worden met achtergrondvariabelen, en factoren als zittenblijven, schoolveranderingen, en veranderingen van studierichting. Bijvoorbeeld door deze zaken in modellering mee te nemen, is het mogelijk de scholen te vergelijken, inzicht te krijgen in de oorzaken van de veranderingen en de toegevoegde waarde van de scholen.

Regressiemodellen

Er bestaan verschillende regressiemodellen om leerwinst en toegevoegde waarde te berekenen. We bespreken drie regressiemodellen in oplopende complexiteit. Dat zijn lineaire regressiemodellen, meerniveau-modellen[15], en groeicurve-modellen. Al deze modellen hebben met elkaar gemeen dat ze door middel van het toevoegen van achtergrondinformatie van leerlingen en scholen een genuanceerder beeld proberen te verkrijgen van leerwinst en toegevoegde waarde. Deze achtergrondvariabelen worden ook wel fairness-kenmerken genoemd en kunnen als onafhankelijke variabelen of covariaten in de verschillende modellen worden opgenomen. Het gebruik van deze variabelen dient om een betere interpretatie te kunnen geven aan de uitkomsten op het vlak van leerwinst en toegevoegde waarde. Het gebruik van deze variabelen moet er als het ware voor zorgen dat de effecten waar een leerling of school geen invloed op heeft ook geen rol spelen in de uitkomsten van het model.

Er is veel debat over de keuze welke achtergrondinformatie relevant is. Deze discussie komt verder aan de orde in Paragraaf 4.3.2. Op het moment dat een fairness-kenmerk deel uitmaakt van een model, wordt het effect van dit kenmerk op leerwinst weggenomen. Als bijvoorbeeld de variabele etniciteit deel uitmaakt van een model, corrigeert het model voor eventuele verschillen in leerwinst tussen de onderscheiden etnische groepen en wordt de groep leerlingen als een etnisch homogene groep beschouwd.

Bij een lineair regressiemodel wordt de score op de eindmeting voorspeld op basis van de score op de beginmeting. Ook andere (achtergrond)variabelen kunnen een rol spelen bij het voorspellen van de score op de eindmeting . Deze variabelen worden dus gebruikt om de relatie tussen de begin- en eindmeting te “corrigeren” voor de relevante verschillen tussen leerlingen en leerlingpopulaties.

---

system- level descriptive information on educational structures, policies and practices (NESLI). UNB-CRISP, Fredericton, 2011.

[14]OECD (2018). PISA for Development Assessment and Analytical Framework: Reading, Mathematics and Science, OECD Publishing, Paris.

[15]Meerniveau-modellen staan ook bekend als multilevel-modellen. Zie onder andere Hox, J.J. (2010). *Multilevel Analysis: Techniques and Applications*. New York: Routledge; Raudenbush, S.W & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.)*. Thousand Oaks, CA: Sage.

Een nadeel van lineaire regressiemodellen is dat deze modellen geen rekening houden met de structuur van onderwijskundige data. Leerlingen die op dezelfde school zitten, lijken namelijk meer op elkaar dan je op basis van toeval zou kunnen verwachten. De leerlinggegevens zijn met andere woorden niet volledig onafhankelijk van elkaar. De uitkomsten van een lineair regressiemodel zijn om deze reden licht vertekend. Meerniveau-modellen houden wel rekening met deze zogenaamde hiërarchische datastructuur[16]. De gegevens zijn in een meerniveau-model in twee lagen op te delen (leerlingen die genest zijn binnen scholen), maar ook in drie lagen (leerlingen die genest zijn binnen klassen die wederom genest zijn binnen scholen). Een ander belangrijk voordeel van meerniveau-modellen ten opzichte van regressiemodellen is dat er binnen meerniveau-modellen meer flexibiliteit bestaat rondom het in kaart brengen van leerling- en groepseffecten.

De meeste complexe regressiemodellen die we hier bespreken zijn zogenaamde groeicurve-modellen. In deze modellen modelleren we de ontwikkeling in leerprestaties met behulp van wederom een meerniveau-model waarbij hier de metingen genest zijn binnen leerlingen. Waar bij een standaard meerniveau-model leerlingen genest zijn binnen scholen, en daarmee leerling- en schooleffecten goed van elkaar kunnen worden gescheiden, gaat een groeicurve-model nog een stapje verder. Binnen deze modellen zijn ook de groeipatronen tussen leerlingen te onderscheiden. Dit model schat dus voor iedere leerling een afzonderlijke aanvangsniveau en ontwikkelingssnelheid. Een groeicurve-model is nog uit te breiden met een derde schoolniveau. Het belangrijkste voordeel van een groeicurvemodel ten opzichte van de andere regressiemodellen is dat het een nog gedetailleerder beeld geeft van zowel de leerontwikkeling van leerlingen als de toegevoegde waarden van scholen. De nadelen zijn vooral te vinden in de complexiteit van de modellen en de eisen waar de data aan moeten voldoen om deze modellen in te kunnen zetten. De voorwaarden waar de data aan moeten voldoen, zullen verder behandeld worden in Paragrafen 4.3.3 en 4.3.4.

Stratificatiemodellen

Stratificatiemodellen hebben als doel om alleen leerlingen met een vergelijkbare achtergrondpositie met elkaar te vergelijken. Bij de toepassing van deze strategie wordt de totale leerlingpopulatie verdeeld in subgroepen met een vergelijkbare predispositie[17]. Daarbij is het gebruikelijk om de toetsscore op de eerste meting te gebruiken om verschillende subgroepen te construeren[18]. Het centrale idee daarbij is dat alle relevante informatie is opgenomen in deze ene variabele. De analyse wordt vervolgens voor elk van de subgroepen apart uitgevoerd. Leerlingen worden in dit model dan ook alleen vergeleken met leerlingen met dezelfde beginscore. Voor iedereen met dezelfde beginscore wordt de verwachte score op het tweede meetmoment en de bijbehorende standaarddeviatie berekend. Een voorbeeld van een zogenaamde scoretabel is te vinden in Tabel 4.1.

In Tabel 4.1 is te zien dat acht leerlingen op de eerste meting geen enkele opgave goed hebben gemaakt. Een toetsscore nul op de eerste meting komt in dit geval overeen met een vaardigheidsscore 30. Deze acht leerlingen hebben een gemiddelde vaardigheidsscore van

[16] Hox, J., Moerbeek, M. & van de Schoot, R. (2018). *Multilevel Analysis: Techniques and Applications 3rd edn.*, Routledge, London.

[17] Deeks, J.J., Dinnes J, D'Amico, R., Sowden, A.J., Sakarovitch, C., Song, F., Petticrew, M., & Altman, D.G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7, 27.

[18] Keuning, J. & Feskens, R. (2013). Meten van leerwinst en toegevoegde waarde op basis van niveaugestandaardiseerde groeiscores. Paper gepresenteerd tijdens Onderwijs Research Dagen, Brussel.

Toetsscore meting 1	Vaardigheid meting 1	n meting 1	Gemiddelde vaardigheid meting 2	Standaard - deviatie meting 2
0	30	8	86	9.5
1	32	12	89	8.5
2	34	22	94	8.1
...	...	...	...	...
48	120	121	151	8.2
49	125	115	154	8.1
50	133	60	155	7.1

Tabel 4.1: Voorbeeld scoretabel stratificatiemodel

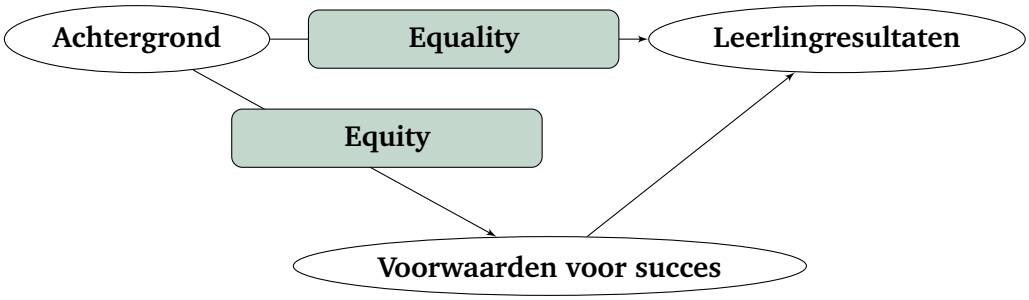
86 behaald op het tweede meetmoment. De bijbehorende standaarddeviatie is 9,5. In een stratificatiemodel worden deze leerlingen met elkaar vergeleken en hun prestatie op de tweede meting wordt dus afgezet tegen de prestatie op de tweede meting van leerlingen met dezelfde score op meting 1. Op het moment dat leerlingen met een toetsscore nul op meting 1 een hogere vaardigheidsscore dan 86 hebben op de tweede meting, is de interpretatie dat zij een positieve "leerwinst" laten zien. Om ook de leerwinstuitkomsten van leerlingen met verschillende scores op de eerste meting met elkaar te kunnen vergelijken, worden de leerwinst uitkomsten uitgedrukt in Z-scores:

$$Z_i = \frac{\Theta_{2i} - \mu_{\Theta_2|r_1}}{\sigma_{\Theta_2|r_1}} \tag{4.1}$$

De Z-scores (Z_i) worden dus per leerling berekend door de gemiddelde vaardigheidsscore op het tweede moment van iedereen met dezelfde toetsscore op het eerste meetmoment ($\mu_{\Theta_2|r_1}$) af te trekken van de vaardigheidsscore van een leerling op het tweede meetmoment (Θ_{2i}). Dit resultaat wordt vervolgens gedeeld door de standaarddeviatie van de vaardigheidsscores op het tweede moment van leerlingen met dezelfde score op het eerste meetmoment ($\sigma_{\Theta_2|r_1}$). Doordat de resultaten nu vergelijkbaar zijn voor alle leerlingen (onafhankelijk van hun prestatie op de eerste meting), kunnen de Z-scores ook geaggregeerd worden op bijvoorbeeld schoolniveau. Op het moment dat de aantallen leerlingen per toetsscores relatief klein zijn, zoals in dit voorbeeld ook wel het geval is, kan men ervoor kiezen om verschillende scorepunten op de eerste meting samen te voegen. Deze, in vergelijking met verschillende (meerniveau)-regressie technieken, relatief eenvoudige procedure maakt de stratificatiebenadering ook relatief transparant met betrekking tot dit soort keuzes. Verder speelt het feit dat er achtergrondvariabelen ter beschikking staan geen rol in dit soort analyses. Een belangrijk voordeel van deze benadering is dat je niet van elke leerling hetzelfde groeipatroon meer verwacht en dat ook traditioneel benedengemiddeld presterende leerlingen en scholen eerder een positief geluid terug gerapporteerd krijgen op het moment dat ze beter presteren dan andere leerlingen met dezelfde predispositie. Een nadeel van deze methode is dat deze leunt op de aanname dat alle relevante informatie rondom de achtergrond van leerlingen in het resultaat op de eerste meting terug te vinden is.

Modelleringbenaderingen

De kwaliteit van een school valt dus te berekenen door te proberen het deel van de leerwinst dat toe te schrijven is aan scholen te isoleren. De kwaliteit van scholen is echter ook in kaart te brengen door expliciet kenmerken van goed onderwijs in een model op te nemen. Dit wordt onder andere gedaan door een model ontwikkeld door Willms (2001)[19]. Dit model is schematisch weergeven in Figuur 4.2.



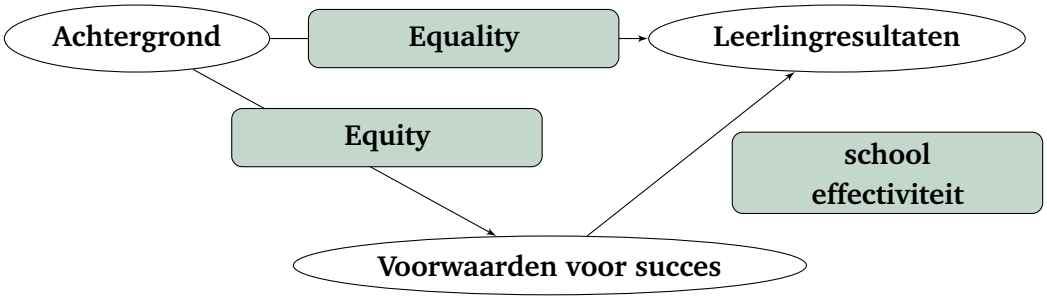
Figuur 4.2: Het model van Willms (2011)

Dit model veronderstelt dat leerlingresultaten tot stand komen door een combinatie van achtergrondkenmerken van leerlingen (Achtergrond) en structurele- en proceskenmerken van de institutionele context (Voorwaarden voor succes). De achtergrondkenmerken van leerlingen kunnen een direct effect hebben op leerlingen hun resultaten, maar ook een indirect effect via de institutionele voorwaarden voor succes. Het model maakt daarbij een onderscheid tussen *equality* (gelijkheid) en *equity* (kansengelijkheid, cf: Willms, 2011). Gelijkheid drukt in dit model uit in hoeverre schoolpopulaties vergelijkbaar zijn in de verdeling van hun onderwijsuitkomsten. Kansengelijkheid is een normatief concept dat ook een beoordeling en interpretatie vereist van de waargenomen verschillen tussen schoolpopulaties en hun toegang tot middelen en schoolprocessen die een effect op onderwijsuitkomsten kunnen hebben. In onderzoek naar onderwijseffectiviteit wordt het effect van achtergrond "uitgepartieerd" om een zuiverdere indruk te krijgen van de invloed van de veronderstelde voorwaarden voor succes. Een uitbreiding van het model van Willms is dan ook te vinden in Figuur 4.3. In dit model wordt de toegevoegde waarde of schooleffectiviteit van scholen weergegeven door de relatie tussen voorwaarden voor succes en leeruitkomsten.

Zo kan op basis van dit model gericht gezocht worden naar verschillen tussen scholen na inachtneming van meerdere achtergrondkenmerken (*Equality* of gelijkheid). De vraag wordt dan: maakt het voor de resultaten van leerlingen van gelijke achtergrond (intelligentie, sociaal-economische status, interesse, leertempo, et cetera) uit op welke school zij onderwijs volgen? Ook kunnen voorwaarden voor succes meegewogen worden, waarmee de vraag wordt: hebben leerlingen van verschillende achtergrond gelijke toegang tot voorwaarden voor succes op verschillende scholen?

Het model van Willms (2011) laat zien dat verschillen tussen scholen groot kunnen zijn, ofwel door achtergrondkenmerken van de leerlingen zelf, ofwel doordat zij niet dezelfde toegang hebben tot voorwaarden voor succes. Eerlijkheid heeft betrekking op

[19]Willms, J.D. (2011) An analysis plan on educational equality and equity: Recommendations for the oecd education at a glance. paper prepared for the OECD NESLI INES network for the collection and adjudication of system-level descriptive information on educational structures, policies and practices (NESLI). UNB-CRISP, Fredericton, 2011



Figuur 4.3: Het model van Willms (2011) uitgebreid

hoe goed landen erin slagen leerlingresultaten onafhankelijk van achtergrondkenmerken van leerlingen te maken en dus een gelijke toegang tot de voorwaarden voor succes te realiseren voor alle leerlingen, ongeacht hun achtergrond.

Deze benadering geeft de mogelijkheid om voorwaarden voor succes expliciet te modelleren. Zo is het bijvoorbeeld mogelijk te onderzoeken of het al dan niet aanbieden van adaptief onderwijs of het geven van feedback binnen de klas bijdraagt aan de voorwaarden voor succes. De hiervoor in te zetten statistische modellen zijn zogenoemde *structural equation models* of structurele vergelijkingsmodellen. Dit zijn modellen die regressie- met meetmodellen combineren[20]. Een van de belangrijkste voordelen van structurele vergelijkingsmodellen is dat een variabele gelijktijdig te modelleren is als een onafhankelijke en als een afhankelijke variabele. Dit maakt het mogelijk om het construct “voorwaarden voor succes” zowel als voorspeller op de “onderwijsuitkomsten” en als voorspelde variabele van “achtergrondkenmerken” te gebruiken binnen hetzelfde model. Hiermee volgt het analysemodel de theoretische veronderstellingen van het model van Willms. Deze benadering is minder geschikt om de toegevoegde waarde van individuele scholen te bepalen. Ze is echter zeer geschikt om scholen duidelijk te maken welke factoren bij kunnen dragen aan het verhogen van de toegevoegde waarde.

4.3 Overwegingen bij de keuze voor een leerwinst benadering

4.3.1 Strategisch gedrag

Op het moment dat de uitkomsten van berekeningen van leerwinst en toegevoegde waarde van groter belang worden voor leerlingen of scholen, zal ook de kans op ongewenst strategisch gedrag van de betrokkenen toenemen. Ongewenst strategisch gedrag met als doel op een gunstige wijze in de uitkomsten te verschijnen, kan zich op verschillende manieren voordoen. Zo kunnen scholen bijvoorbeeld besluiten om (zwak presterende) leerlingen uit te sluiten van toetsen of scholen kunnen proberen expres laag te presteren op de beginmeting om het verschil met de nameting bij een berekening van toegevoegde waarde[21] te maximaliseren. Daarnaast kunnen nog andere ongewenste effecten optreden op het moment dat het belang van een prestatie op een toets zeer groot wordt. Scholen kunnen zich hun schoolpraktijk buitenproportioneel gaan richten op het onderwijzen van de gevraagde kennis en vaardigheden die in de toets aan de orde komen ten koste van

[20] Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley.

[21] Janssens, F.J.G., Rekers-Mombarg, L. & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs*. Eindrapportage.Groningen: GION.

andere ook belangrijke onderwijsgebieden. Deze zogenaamde teaching to the test effecten kunnen een verenging van het onderwijs tot gevolg hebben.

Uitkomsten op het vlak van leerwinst en toegevoegde waarde die gebruikt worden met het doel om het onderwijs te verbeteren, zullen minder gevoelig zijn voor strategisch gedrag dan uitkomsten die gebruikt worden voor afrekendoeleinden. In die zin sluit de zogenaamde Campbell's law[22] hierbij aan: *"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor*"[23].

Het gebruik van diverse methoden om leerwinst en toegevoegde waarde te bepalen, heeft een aantal voordelen. Een ervan is dat mogelijke tegenstrijdigheden die kunnen opduiken beter tegen elkaar af te zetten zijn om zo een volledig beeld te krijgen in de sterke en zwakke punten van de school. Een ander voordeel is dat, doordat er sprake is van meer dan één indicator, de scholen ook niet over kunnen gaan tot het optimaliseren van één enkele waarde, maar dat alle aspecten van het onderwijs op school relevant blijven.

4.3.2 Achtergrondvariabelen

Op dit moment bestaat er in de literatuur geen consensus over het aantal variabelen en de specifieke variabelen die meegenomen zouden moeten worden om toegevoegde waarde van scholen succesvol te bepalen[24,25]. Er is wel brede consensus dat prestaties op een eerder tijdstip cruciaal zijn voor het bepalen van schooleffecten. Variabelen die vaak als *fairness* kenmerken worden opgenomen in de modellen zijn daarnaast leerlingkenmerken zoals geslacht, migratie-achtergrond en de sociaaleconomische status. In Engeland - waar een brede traditie bestaat van het gebruik van toegevoegde waarde modellen - wordt vaak een indicator opgenomen die aangeeft of leerlingen in aanmerking komen voor een gratis schoolmaaltijd[26].

Al deze variabelen kunnen op een geaggregeerd niveau ook op schoolniveau in een meerniveau-model worden opgenomen. De gedachte daarachter is dat het lastiger is voor een school toegevoegde waarde te creëren voor een groep leerlingen met een lastige uitgangspositie, niet alleen doordat individuele leerlingen eventuele leerwinst belemmeren, maar er ook een gezamenlijk negatief groepseffect op leerwinst kan bestaan. Op het moment dat er een fairness-kenmerk wordt opgenomen in een model, betekent het dus ook dat er andere verwachtingen voor deze groep leerlingen wordt gecreëerd. Of dit gewenst of ongewenst is, valt ook per vakgebied te besluiten. Zo kan bijvoorbeeld het hebben van een taalachterstand bij leerlingen een andere rol spelen bij de resultaten voor wiskunde dan voor begrijpend lezen. Het is cruciaal dat alle relevante stakeholders consensus bereiken

[22] Campbell, D. T. (1976). *Assessing the Impact of Planned Social Change*. Hanover New Hampshire: The Public Affairs Center, Dartmouth College.

[23] Zie ook: Janssens, F.J.G., Rekers-Mombarg, L. & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs*. Eindrapportage. Groningen: GION.

[24] Manzi, J., San Martín, E., & Van Bellegem, S. (2011). School System Evaluation by Value-Added Analysis Under Endogeneity. *Psychometrika*, 79(1),130- 153.

[25] McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67–101.

[26] 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), pp. 518-537.

over welke set achtergrondvariabelen meegenomen moeten worden in de correctie van leerwinst en toegevoegde waarde berekeningen.

4.3.3 Stabiliteit van schattingen

Leerwinst en toegevoegde waarden berekeningen zijn gebaseerd op statistische modellen die variëren in complexiteit. Hoe complexer de modellen zijn, hoe meer gegevens nodig zijn om het model goed te kunnen schatten. Op het moment dat men uitspraken wil doen op het niveau van scholen, geldt dat voor kleine scholen de onzekerheid rondom de schatting over de toegevoegde waarde van deze school groot kan zijn[27]. Bolhaar en Scheer[28] adviseren om drie cohorten leerlingen samen te voegen om zo schooleffecten te berekenen over meerdere jaren. Dit voorkomt dat uitschieters in een specifiek jaar een te grote rol spelen. Het opleveren van stabiele schattingen die niet te veel van jaar tot jaar fluctueren, heeft ook als voordeel dat gebruikers niet het vertrouwen in de toegevoegde waarde modellen verliezen[29].

4.3.4 Ontbrekende data

Voor analyses om leerwinst en toegevoegde waarde te bepalen zijn veel gegevens nodig. Dit betreft niet alleen toetsgegevens afkomstig uit longitudinale dataverzamelingen, maar ook maken veel modellen gebruik van achtergrondinformatie van leerlingen, klassen en scholen. De gezochte gegevens zijn lang niet altijd aanwezig. Dan kan veroorzaakt zijn doordat gegevens eenvoudig niet bestaan of opgeslagen zijn. Voorbeelden hiervan zijn leerlingen die een toets niet hebben gemaakt of naar een andere school zijn gegaan. Ook leerlingen die een jaar overdoen, “zittenblijvers” of “doubleurs”- leiden tot ontbrekende data. Deze vorm van ontbrekende data heeft misschien wel de meeste impact op de uitkomsten van toegevoegde waarde modellen, omdat hier de kans op selectiviteit in de data het grootst is. Leerlingen die blijven zitten, zijn immers vaak ook de zwakkere leerlingen[30]. Het niet meenemen van de gegevens van zittenblijvers, kan dus leiden tot een overschatting van de uitkomsten van leerwinst modellen[31]. Het moge duidelijk zijn dat juist deze vorm van ontbrekende data meegenomen moet worden in de analyses. Dit kan door te werken met beslisregels, waarop bij een bepaald percentage zittenblijvers naast de toegevoegde waarde uitkomsten een waarschuwingsvlag verschijnt die aanleiding geeft om het gesprek met de desbetreffende school aan te gaan. Ook kan ervoor gekozen worden bij een te hoog percentage zittenblijvers de toegevoegde waarde cijfers helemaal niet meer te rapporteren vanwege te veel twijfel over de geldigheid van de uitkomsten. Naast het werken met beslisregels, kan ervoor gekozen worden om de ontbrekende gegevens met een model te

[27] McCaffrey, D. & Lockwood, J.R. (2008). Value-added models: Analytic issues. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14.

[28] Bolhaar, J. & Scheer, B. (2019). *Verschillen in leerresultaten basisscholen*. CPB notitie. Den Haag: Centraal Planbureau.

[29] Janssens, F.J.G., Rekers-Mombarg, L. & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs*. Eindrapportage. Groningen: GION.

[30] Timmermans, A.C. (2012). Value added in Educational Accountability: Possible Fair and Useful? Groningen: University of Groningen.

[31] Janssens, F.J.G., Rekers-Mombarg, L. & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs*. Eindrapportage. Groningen: GION.

schatten. Zogenaamde imputatie modellen, kunnen ontbrekende gegevens met behulp van andere wel beschikbare gegevens als het ware invullen[32]. Een nadeel van het imputeren van data is dat het ten koste gaat van de transparantie en daarmee de uitlegbaarheid van het model. Terwijl ontbrekende gegevens door zittenblijvers of schoolverlaters bij toegevoegde waarde modellen problematisch zijn, is dit voor het berekenen van individuele leerwinst veel minder het geval. De gegevens van twee meetmomenten kunnen nog steeds betekenisvol met elkaar in verband worden gebracht. Het grootste praktische probleem is dat bij leerlingen die van school wisselen er voor een systeem moet worden gezorgd dat de oude leerlinggegevens op de nieuwe school van de leerling terecht komen.

Daarnaast is niet altijd de informatie geregistreerd die bij voorkeur in het model was opgenomen. Dit geldt vooral voor contextuele kenmerken waarop men graag zou willen corrigeren om de toegevoegde waarde te berekenen of voor kenmerken die de onderwijspraktijk binnen een klas of school in kaart zouden moeten brengen. Tenslotte mogen gegevens vanwege privacywetgeving soms niet gedeeld of voor een langere periode bewaard blijven. De beschikbaarheid van gegevens zou daarom een factor moeten zijn bij de keuze voor het model dat men graag wil gebruiken om leerwinst en toegevoegde waarde te berekenen. Dat geldt ook voor de bijbehorende interpretatie.

Wat analyses en de interpretatie ook kan compliceren is dat door de tijd heen variabelen kunnen verdwijnen of niet beschikbaar kunnen zijn, toegevoegd of relevant kunnen worden, of van betekenis of definitie kunnen veranderen. Een voorbeeld van dat laatste kan zijn als het beleid rond zittenblijven aangepast wordt. De zittenblijvers uit twee metingen zijn daardoor lastiger met elkaar te vergelijken. Als deze variabelen een rol spelen in het model kan dit ook het gevonden resultaat beïnvloeden.

4.4 Graadspecifieke eindtermen

Het gebruik van leerwinst en toegevoegde waarde modellen in combinatie met graadspecifieke eindtermen is goed mogelijk. Dit kan enerzijds door gebruik te maken van eerdere metingen en deze op te nemen in het leerwinst en toegevoegde waarde model. Ook als deze resultaten niet op een gekalibreerde schaal liggen, kan de eindmeting opgenomen worden in de modellen. Het is daarbij uiteraard wel van belang dat de eindmeting dezelfde vaardigheid meet als de metingen die gebruikt worden als beginmeting (zie ook Sectie 4.1). Het is anderzijds ook mogelijk voor een criteriumgerichte beoordeling te kiezen. Hier worden de prestaties van leerlingen bekeken in vergelijking met vaststaande standaarden, zoals de eindtermen en de aan de eindtermen gekoppelde bouwstenen. In deze leerwinstmethodologieën staat de groei van leerlingen centraal, in directe koppeling met de eindtermen: hoeveel procent van de leerlingen behaalt een niveau in de eerste, en hoeveel in de volgende meting? Ook hier kan gecorrigeerd worden voor diverse achtergrondkenmerken.

Een belangrijk aandachtspunt bij het berekenen van toegevoegde waarden is het tijdvenster tussen twee metingen waarop de toegevoegde waarde berekeningen gebaseerd zijn. Dit lijkt met name een punt van aandacht bij het meenemen van de 3de graad eindresultaten in het secundair onderwijs. Op het moment dat deze resultaten voor een leerwinstmeting worden gebruikt, zouden toetsresultaten van vier jaar ervoor meegenomen moeten worden

[32] Little, R.J.A. & Rubin, D.B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

als beginmeting in een berekening van de toegevoegde waarde. Dit is voor een dergelijke berekening een vrij lange periode. Het is hierbij zeer lastig om causale uitspraken te doen over in welke mate het resultaat van de toegevoegde waarde daadwerkelijk toe te schrijven is aan een effectief schoolbeleid. Dit is met name ook het geval omdat in het derde en vierde jaar van het secundair onderwijs nog relatief veel wisselingen van leerlingen naar andere scholen plaatsvinden. Op het moment dat er een beginmeting zou kunnen gebruikt worden die afgenomen wordt aan het einde van de tweede graad, zou - ook gegeven dat er minder wisselingen plaats vinden in het vijfde en zesde jaar - weer wel een toegevoegde waarde berekening kunnen worden uitgevoerd.

Anders is de situatie bij individuele leerwinst berekeningen: hier is het grotere tijdvenster tussen begin- en eindmeting minder problematisch. De beginmeting zou opgenomen kunnen worden als achtergrondvariabele in een regressiemodel of als een stratificatievariabele in een stratificatiemodel (zie Paragraaf 4.2.2). De enige uitdaging ligt hier in het gemeten construct: is wiskunde en Nederlands aan het einder van de eerste graad voldoende vergelijkbaar met die vaardigheden aan het einde van de derde graad. Oplossingen hoe daar achter te komen worden gegeven in Hoofdstuk 7.

4.5 Conclusie

Uiteindelijk is het lastigste van leerwinst de interpretatie van de uitkomst. Vragen die bij de interpretatie van leerwinst opkomen zijn bijvoorbeeld: Wanneer is er voldoende leerwinst behaald? Is dat voor alle leerlingen hetzelfde, of kan het over leerlingen verschillen wat voldoende leerwinst is? Of kan het over scholen verschillen? Maar ook: hoe interpreteren we de verschillen in leerwinst verkregen met de ene methode met medeneming van een serie variabelen met de leerwinst zoals bepaald met een andere, eveneens wetenschappelijk onderbouwde methode met een andere verzameling relevante variabelen?

De beantwoording van deze – en vergelijkbare – vragen omtrent de operationalisatie op niveau van de individuele leerlingen, op niveau van de school en op niveau van het onderwijssysteem, is tevens contextgevoelig. De definitieve beantwoording moet dus plaatsvinden binnen, maar ook samen met het Vlaamse onderwijsveld. Het samen met het onderwijsveld optrekken kan enerzijds bewerkstelligen dat scholen in eerste instantie meer gebruik maken van data om hun beleid te evalueren en te verbeteren. En anderzijds worden door het gezamenlijk optrekken met het onderwijsveld, eventuele negatieve aspecten van leerwinstberekeningen, zoals eenzijdige publieke rangordeningen (*rankings*), verenging van het onderwijs en teaching to the test, voorkomen.

De resultaten van de verschillende manieren van analyse kunnen ook naast elkaar bestaan. Zoals de Type A en de Type B leerwinst voor verschillende belanghebbenden een verschillende waarde kan hebben, geldt dat ook voor de andere manieren om naar leerwinst te kijken. Er is niet één model dat het gehele verhaal vertelt. Juist door de verschillende indicatoren van leerwinst en toegevoegde waarde naast elkaar te beschouwen krijgt men het gehele beeld. Ook voorkomt dat een focus op slechts een enkele maatschappelijke indicator en kan zo – in ieder geval deels – Campbell's Law vermeden worden.

5. Toetsontwikkeling

In dit hoofdstuk worden verschillende aspecten en vragen rondom toetsontwikkeling besproken. Aspecten die onder andere aan bod zullen komen, hebben betrekking op de toetsinhoud (Secties 5.1, 5.2 en 5.3), de toetsvorm (Secties 5.1, 5.5 en 5.8), de toetsdoelgroep (Sectie 5.6) en de toetskwaliteit (Secties 5.4 en 5.7). Een groot deel van de definitieve antwoorden op deze vragen zal ook afhangen van de keuzes die in Hoofdstuk 2 aangehaald zijn, namelijk in welke mate aan de centrale toetsen belangrijke consequenties verbonden zijn voor de verschillende betrokkenen. Het uitgangspunt dat vaststaat bij alle vragen rond de toetsontwikkeling is dat alle leerlingen binnen goed omschreven leerjaren[1] jaarlijks worden getoetst voor één of meerdere leerdomeinen. Deze proeven zullen in eerste instantie focussen op Nederlands (begrijpend lezen, schrijven, grammatica) en wiskunde[2]. Het voornemen is dat al deze toetsen digitaal worden afgenomen.

5.1 Toetsverversing

Toetsen kunnen op verschillende momenten geheel of gedeeltelijk vervangen worden. Een belangrijke reden om toetsen te vervangen is het voorkomen dat leerlingen een oneigenlijk voordeel hebben op het moment dat items voor toetsafname bekend zijn. Daarnaast is een belangrijke reden om een toets te vervangen om aan te sluiten bij een veranderende onderwijsinhoud. We bespreken in deze sectie varianten waarin met vaste toetsen wordt gewerkt (Paragraaf 5.1.1) en een variant waarin toetsen jaarlijks vervangen worden (Paragraaf 5.1.2). Het gebruik van een toetsitemdatabank wordt behandeld in Paragraaf 5.1.3. Deze variant biedt de mogelijkheid om de toets telkens gedeeltelijk te vervangen. Tenslotte wordt het gebruik van een gekalibreerde itembank in Paragraaf 5.1.4

[1] Dit betreft het vierde en het zesde leerjaar in het lager onderwijs en aan het einde van de eerste en de derde graad van het secundair onderwijs.

[2] `https://www.vlaanderen.be/publicaties/beleidsnota-2019-2024-onderwijs`, p.37 en p.86.

behandeld.

5.1.1 Vaste toetsen

Wanneer toetsen een puur formatieve functie hebben, dan zijn vaste toetsen die langer meegaan goed toepasbaar. In het geval de formatieve functie consequent wordt toegepast, is het voor een leerling zinloos –en wellicht zelfs schadelijk– om te frauderen. Wanneer de leerling de toets oefent, in plaats van de vaardigheid, of wanneer de leerling afkijkt bij een medeleerling, dan kan de leerkracht de leerling niet goed helpen om de volgende stap te nemen in het onderwijsproces. Als dankzij fraude de vaardigheid van de leerling aanzienlijk hoger lijkt dan dat deze werkelijk is, dan krijgt de leerling te moeilijke stof aangeboden en mist deze de vaardigheden die voorwaardelijk zijn om die hogere stap te nemen. Dan is de nieuwe stof onbegrijpelijk voor de leerling, en worden ondertussen de voorwaardelijke vaardigheden niet getraind. Dat betekent dat fraude hier ook niet heel waarschijnlijk[3] is en de noodzaak voor toetsverversing vanuit dit perspectief minder aanwezig.

Het scenario waarin toetsen een puur formatieve functie hebben en vaste toetsen toepasbaar zijn, heeft een aantal voordelen. Het eerste is dat in dit scenario waarin de belangen zo laag mogelijk zijn voor de leerling en de leerkracht, de typische prestatie van de leerling[4] wordt gemeten. Dat is wat de leerling laat zien in de meest dagelijkse situatie zonder voor de testafname extra te studeren, of op een oneigenlijke wijze probeert de prestatie te verhogen. Het voordeel is dat het mogelijk is de meest gepaste vorm van reactie op het resultaat te geven. Het is dan wel noodzakelijk dat er na de toetsafname nog sprake is van substantieel onderwijs. In dat geval mag de afname niet geheel aan het einde van het leerjaar plaatsvinden. Als het doel niet zozeer op de leerling gericht is, maar vooral het ondersteunen van de school is, is een afname op het eind van het leerjaar minder een probleem. Op basis van het feedbackrapport kan de school zien waar de sterke en zwakke punten van het lesgeven op de basisvakken zit en kan deze voor de aanvang van het nieuwe schooljaar aanpassingen doorvoeren door accenten in het onderwijs te veranderen. Een ander groot voordeel van het scenario waarin toetsen een puur formatieve functie hebben en vaste toetsen mogelijk zijn, betreft de kosten. Als jaarlijks opnieuw ontwikkelen niet nodig is, scheelt dat uiteraard kosten. Vernieuwing zou dan bijvoorbeeld maar eens in de 4 of 5 jaar nodig zijn. Het is dan bijvoorbeeld mogelijk geleidelijk deze toetsen te vernieuwen door ieder jaar 20% of 25% van de opgaven te vervangen. Dit kunnen dan bijvoorbeeld de net iets minder goed functionerende opgaven zijn. Ook kan zo een nieuw itemtype ingevoerd worden. In het geval van digitale afnamen hoeven daartoe ook geen nieuwe toetsboekjes of nieuwe antwoordbladen gedrukt te worden. Doordat er geen additionele fraudepreventiemaatregelen getroffen hoeven te worden, is een externe controle niet vereist. Ook ontbreekt de noodzaak om alle kandidaten tegelijkertijd de toets te laten maken, aangezien het uitwisselen van de opgaven geen zin heeft.

Een belangrijk nadeel is dat als de toetsen in tweede instantie wél een belangrijke rol gaan spelen bij beslissingen, de kans op fraude aanzienlijk toeneemt. Dan is de geheimhouding van items niet te waarborgen. Er zal dan een prikkel zijn voor diverse

[3] Zeker als er geen relatieve normering gegeven wordt, is deze noodzaak er niet.

[4] De oorspronkelijke term komt uit de arbeidspsychologie: Sackett, P.R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, 73(3), 482–486.

partijen (leerlingen, ouders, trainingsbureaus, leerkrachten, schoolbesturen) om een zo goed mogelijke prestatie voor te bereiden. Zeker in het geval de toetsen bekend zijn, is de kans groot dat die training op de toetsinhoud gericht zal zijn waardoor de uiteindelijk verkregen toetsscore geen goede indicator meer zal zijn van de werkelijke vaardigheid. Als voor een dergelijk scenario gekozen wordt, is het dus cruciaal dat de toetsen nooit een high-stakes summatieve toepassing krijgen.

5.1.2 Jaarlijkse produceren van toetsen

Toetsen kunnen ook bij elke afname helemaal ververst worden. Dit is vooral relevant op het moment dat er veel kans op fraude bestaat. Dit is een scenario dat in Nederland wordt toegepast bij de Centrale Examens. Deze examens worden op papier afgenomen ter afsluiting van het voortgezet onderwijs. Voor ieder vak waar een centraal examen is, bepaalt dit centraal schriftelijk examen (cse) voor 50% het eindcijfer. Op basis van deze eindcijfers wordt bepaald of een kandidaat geslaagd is voor de gevolgde opleiding. Dit zijn duidelijk toetsen met grote belangen voor de leerlingen. Ook voor de scholen zijn er belangen omdat hun slagingspercentages centraal geregistreerd worden. Voor ieder vak en voor alle niveaus van de centrale examens worden ieder jaar nieuwe varianten ontwikkeld. Dat betreft een nieuwe variant voor de eerste afname (het eerste tijdvak - het eigenlijke eindexamen) en een nieuwe variant voor de herkansing (het tweede tijdvak, met beperktere deelname). Ook is er nog een variant voor de extra late herkansingsmogelijkheid voor kandidaten die bij één van de eerste twee tijdvakken ziek waren of een staatsexamen afleggen (het derde tijdvak, met zeer beperkte deelname). Bij een wisselende beperkte set vakken wordt deze variant jaarlijks niet vernieuwd. Als reserve, voor als de geheimhouding van het examen ernstig geschonden blijkt, is er voor ieder regulier examen ook een vierde alternatief examen beschikbaar. Aangezien deze versie zelden of nooit gebruikt wordt, hoeft deze niet regelmatig ververst te worden.

Het voordeel van deze examens is dat ze met zoveel veiligheidsmaatregelen omgeven zijn, dat de opgaven zelf met aan zekerheid grenzende waarschijnlijkheid voor het afnemen ervan onbekend blijven. Iedereen kan wel van te voren zeer veel informatie inwinnen over de examens en de regels daaromtrent via een publieke website[5] waar voor ieder jaar is aangeven wat voor vaardigheden er bij ieder van de examens gevraagd wordt van de leerling. Eveneens is exact de dag en het tijdstip (start- en stoptijd) van het examen aangegeven[6]. Door de gelijktijdige afname is een uitwisseling van opgaven tussen leerlingen niet mogelijk. Zeer kort na de afname wordt het examen zelf publiek gemaakt via dezelfde website, samen met eventuele bijlagen en correctievoorschriften. Mogelijke aanvullingen op correctievoorschriften kunnen volgen op basis van commentaar van betrokkenen. Als laatste document verschijnt ongeveer een maand na de afname van het examen, de omzettingstabel waarmee te bepalen is welke toetsscore welk oordeel krijgt[7]. Deze examens worden in de voorbereiding op nieuwe examens ook veelvuldig gebruikt om te oefenen, zodat leerlingen gewend raken aan het type vragen en dankzij de correctievoorschriften en omzettingstabellen ook een inschatting kunnen maken van hoe ze ervoor staan.

[5] `https://www.examenblad.nl/`

[6] `https://www.examenblad.nl/examenrooster`

[7] Cijfers tussen 1 en 10, waarbij de grens tussen voldoende en onvoldoende ligt bij 5,5.

Het voordeel van deze informatie is dat de examenvorm zelf geen verrassing hoeft te zijn en de leerlingen kunnen oefenen met het optimaal indelen van hun tijd tijdens het examen. Het nadeel van teaching to the test is zeer beperkt. Bij het examen moeten leerlingen welomschreven vaardigheden hebben, waarbij de examens een goede dekking van die vaardigheden moeten hebben. Doordat de specifieke vragen niet bekend zijn, moet de leerling bekwaam zijn in de gehele vaardigheid voor een optimaal resultaat. Daar waar de gemeten vaardigheid bij het centrale examen een deelvaardigheid betreft van het gehele vak, zoals bij moderne vreemde talen alleen lezen wordt bevraagd, moeten de overige relevante vaardigheden bij die talen binnen de schoolexamens – die de overige 50% van het eindoordeel per vak uitmaken – gemeten worden.

Een nadeel van deze vorm van afname betreft de kosten. Doordat ieder examen –in drievoud– vernieuwd moet worden, waarbij de ontwikkeling met veel veiligheidsmaatregelen plaatsvindt, is de ontwikkeling kostbaar. Een ander nadeel is dat een directe, rechtstreekse vergelijking tussen overeenkomstige examens zodoende niet mogelijk is – noch over jaren, noch over tijdvakken binnen een jaar. Dat is het gevolg van het feit dat ieder examen geheel nieuw is, en dus overlap in opgaven tussen examens geheel ontbreekt. Het is echter wel belangrijk dat deze vergelijking te maken is. Ten eerste is dat eerlijk naar de leerlingen en scholen toe, omdat de zak-slaag beslissing voor een vak niet van de toevallige versie moet afhangen (tijdvak, afnamejaar), maar van de getoonde vaardigheid. Ten tweede geeft die vergelijking ook inzicht in de voortgang over de tijd. Om nu deze vergelijking toch te kunnen maken, worden diverse pre- en post-test procedures toegepast, in combinatie met andere methoden. De eindexamens in Nederland worden op papier afgenomen, waarbij de verdeling over de scholen met (kostbare) veiligheidsmaatregelen omringd zijn. Wanneer de examens op eenzelfde wijze binnen een vastgesteld kort tijdsbestek digitaal moeten worden afgenomen, is het een uitdaging dat de infrastructuur, zowel op alle scholen als centraal, daar geschikt voor moet zijn. Doordat de afnamevorm gestandaardiseerd moet zijn, is het noodzakelijk eenheid in een aantal toetsafnamekenmerken na te streven, met name:

- Afname device (b.v., desktop, laptop, tablet, of zelfde mobiele telefoon);
- Instellingen op het device (o.a., weergave, besturing, beveiliging);
- Examenopstelling voor digitale afname.

In het geval het niet mogelijk is dat alle scholen exact dezelfde digitale afname aan kunnen bieden, dan moet in ieder geval aangetoond worden dat deze equivalent zijn, om ervoor te zorgen dat de leerlingen en scholen goed met elkaar te vergelijken zijn. Dat geldt ook als er sprake is van een storing die de digitale afname bemoeilijkt. Als het eenmalige high-stakes afname betreft, zijn de kosten hier aanzienlijk hoger, om iedere onvoorziene omstandigheid te dekken.

5.1.3 Toetsitemdatabank

Een toetsitemdatabank of itembank is een gestructureerde verzameling van toetsvragen die een bepaald concept of inhoudsdomein meten. Dit impliceert ook dat er per inhoudsdomein een itembank zou moeten worden ontwikkeld. Hoe meer relevante informatie per item opgeslagen is, hoe meer de itembank aan kracht wint. Naast de inhoud van de vraag en

het antwoordmodel kan heel veel meer informatie opgeslagen worden. Hierbij kan gedacht worden aan:

- Praktische informatie bij het maken van de items:
  - Item-identificatie (itemlabel)
  - Gebruiksgegevens (item constructeur);
  - Status van het item (is het item goedgekeurd om in te zetten voor gebruik in een toets);
  - Gegevens over de gebruikte bronnen bij het maken van het item.

- Informatie relevant voor de toetsmatrijs:
  - Inhoudskenmerken – zoals relatie met eindterm (algemeen en gedetailleerde niveaus);
  - Beoogd niveau van de vaardigheid (voor wie is het item bedoeld);
  - Bevragingsniveau (bijvoorbeeld kennis, begrip, toepassing of denken);
  - Gebruik van de vraag (zoals, examen, formatief, proeftoets).

- Psychometrische informatie (basis):
  - Informatie over de toetsen waarin het item al dan niet is afgenomen;
  - Soort vraag (meerkeuze, kort antwoord, lang antwoord, essay, hotspot, etc.);
  - Aantal antwoord categorieën (bij meerkeuze vragen);
  - Sleutel (bij meerkeuze vragen);
  - Het aantal opties dat gekozen mag worden (voor meerkeuze vragen en andere keuze response items);
  - Beoordelingsvoorschrift (bij open vragen);
  - Maximum score.

In principe is een itembank dus niets meer dan een verzameling opgaven, die allen tot dezelfde toetsinhoud behoren. Een itembank wordt vaak voor administratieve doelen opgesteld. Binnen een itembank kun je namelijk eenvoudig tellen hoeveel opgaven met bepaalde kenmerken er beschikbaar zijn. Zeker als er opgaven hergebruikt gaan worden, of als meerdere toetsvarianten voor één afnameperiode samengesteld worden, is een eenduidige administratie van itemkenmerken noodzakelijk. Wellicht is het allerbelangrijkste element bij een itembank het correct gebruik van itemlabels. Deze dienen *uniek* en *stabiel* te zijn. Een unieke identificatie betekent dat verschillende items nooit een zelfde itemlabel mogen hebben binnen en tussen verschillende bestanden. Een stabiele identificatie betekent dat hetzelfde item altijd hetzelfde itemlabel zou moeten hebben binnen en tussen verschillende bestanden. We spreken hier expliciet over verschillende databestanden, omdat bijvoorbeeld bij de analyse van toetsafname data ook gerefereerd wordt naar itemlabels. Daarbij is het noodzakelijk om ook een eenduidige koppeling naar de itembank te kunnen maken om bijvoorbeeld analyses te kunnen uitvoeren voor verschillende inhoudskenmerken die geadministreerd zijn in de itembank.

De onderwijsnota en het bestek spreken over net- en koepeloverschrijdende toetsen, die in eerste instantie gericht zijn op Nederlands (begrijpend lezen, schrijven, grammatica) en

Wiskunde. Wat betreft de itembanken betekent dat, dat er een itembank voor Nederlands en een itembank voor Wiskunde nodig is. De bank van Nederlands kan opgedeeld worden in drie compartimenten, gebaseerd op de drie specifiek genoemde vaardigheden begrijpend lezen, schrijven, en grammatica. Dit kan leiden tot in principe vier[8] verschillende itembanken die niet direct aan elkaar gerelateerd zijn, maar wel dezelfde vorm kunnen hebben.

Aangezien het de bedoeling is dat uit de itembanken toetsen komen die inhoudsvalide zijn, moet er binnen de betreffende itembank een goede afbakening zijn van het toetsdomein. Dat houdt in dat in principe alle eindtermen en onderwijsdoelen die onder de te meten vaardigheden vallen met opgaven gerepresenteerd zijn. Dat is nodig om ervoor te zorgen dat we uit de itembank toetsen kunnen maken die de kennis of vaardigheden meten die beoogd worden om te meten. Merk op dat deze eindtermen per niveau verschillen en deze voor het vierde leerjaar van het lager onderwijs anders zijn dan die aan het einde van de derde graad van het secundair onderwijs. Het is ook evident dat de operationalisatie van de eindtermen in opgaven fundamenteel verschillende opgaven oplevert voor de verschillende niveaus. Bij het maken van de itembank moet daarbij gelet worden op ten eerste een goede dekking en operationalisatie van de eindtermen, en dus de vaardigheid, binnen ieder van de vier te meten (leer)jaren. Daarnaast moet gekeken worden naar hoe deze vier meetmomenten met elkaar verbonden kunnen worden om leerwinst te kunnen definiëren. Daartoe is het aan te raden ook te kijken naar (de operationalisatie in termen van opgaven van) de eindtermen voor de tussenliggende leerjaren.

Het gebruik van een itembank waaruit toetsen worden samengesteld, heeft een aantal aantrekkelijke voordelen. Deze voordelen hebben betrekking op een financieel voordeel op langere termijn door hergebruik van items en een eenvoudiger te organiseren proces van itemconstructie. Deze voordelen scharen we onder efficiency voordelen. Een belangrijk psychometrisch voordeel is dat de vergelijkbaarheid van resultaten, behaald op toetsen afgenomen in verschillende jaren, kwalitatief beter te realiseren is. Een evident nadeel van het niet gebruiken van itembanken en daardoor niet jaarlijks verversen van de gehele toets, is de kans op het bekend geraken van de items voordat deze worden afgenomen in de toetsen. Het psychometrisch voordeel van itembanken én hoe om te gaan met de geheimhouding van opgaven bespreken we in de volgende sectie rondom “gekalibreerde itembanken”.

5.1.4 Gekalibreerde itembanken

In de vorige drie secties zijn respectievelijk vaste toetsen, jaarlijks vernieuwde toetsen en het gebruik van een itembank besproken als varianten die gebruikt kunnen worden bij toetsverversing. Een goede onderliggende itembank wordt aanbevolen bij alle varianten. Een gekalibreerde itembank is nog een uitbreiding van bovenstaande scenario’s die in deze

[8] De itembank Nederlands kan het best opgesplitst worden naar de drie vaardigheden die genoemd worden omdat –zo suggereert de onderwijsnota– het doel is om over deze vaardigheden apart te rapporteren. Ook als het doel uiteindelijk is om een vaardigheidsniveau voor Nederlands weer te geven, samengesteld uit de drie genoemde vaardigheden, zijn deze vaardigheden dusdanig divers dat deze lastig als unidimensionele schalen gezien kunnen worden. Daarvoor verschillen de vaardigheden te veel en zijn de correlaties tussen de vaardigheden te laag om ze als inwisselbaar te zien. De correlaties voor de vaardigheden binnen Wiskunde –wellicht met uitzondering van einde van graad 3– zijn over het algemeen hoog genoeg om te werken met een enkele schaal of itembank voor Wiskunde.

sectie besproken worden.

- Psychometrisch kenmerken (statistisch)
  - Basisinformatie: hoeveel personen afgenomen, welke leerjaren;
  - Klassieke informatie: p-waarde (moeilijkheid), Rit- en Rir-waarden[9];
  - Item-respons-theorie-kenmerken: itemparameters per vaardigheidsschaal, standaardfouten van de geschatte parameters;
  - Bij beoordeelde vragen: beoordelingskenmerken.

Als de statistische itemkenmerken uit een IRT model opgenomen worden in de itembank, én een rol spelen bij de samenstelling van toetsvarianten, dan spreken we van een gekalibreerde itembank. Als een IRT model geldt voor een verzameling opgaven in een itembank, dan zijn alle opgaven op dezelfde vaardigheidsschaal geplaatst. Voor de opgaven in die bank is het mogelijk om iedere deelverzameling van opgaven –dus voor iedere toets die samengesteld wordt uit opgaven uit de bank– een meting te relateren aan die vaardigheidsschaal. Voor die toetsen geldt dat op basis van een score -op welke toets dan ook- de vaardigheid van de kandidaat bepaald kan worden. Zo kunnen kandidaten die verschillende toetsen maken alsnog direct met elkaar vergeleken worden. Dat is een groot verschil met klassieke testtheorie waarin dat aanzienlijk moeilijker, zo niet onmogelijk is[10].

Het werken met een itembank waarvoor een IRT model geldt, biedt veel flexibiliteit, met alle voordelen van dien. Doordat niet iedereen dezelfde toets hoeft te maken om de resultaten met elkaar vergelijkbaar te maken, is het mogelijk binnen een afname verschillende toetsen af te nemen. Ook biedt IRT de mogelijkheid om trends door de jaren heen waar te nemen zonder dezelfde toetsen te hoeven gebruiken. Hergebruik van een deel van de opgaven uit het voorgaande jaar maakt het mogelijk door middel van IRT de uitkomsten van verschillende jaren met elkaar te vergelijken. Ook het volgen van leerlingen over leerjaren heen is mogelijk met behulp van opgaven uit een IRT-gekalibreerde itembank. Het is duidelijk dat aan leerlingen die twee, of zelfs vier jaar verder zijn in hun onderwijs niet dezelfde opgaven voorgelegd kunnen worden als bij de meting ervoor. Het is zodoende bij leerwinstmetingen van individuele leerlingen evident dat zij door de tijd verschillende toetsen voorgelegd krijgen. Desalniettemin willen we de resultaten vergelijkbaar maken om de groei van de leerling op een vaardigheid inzichtelijk te maken.

Het tweede voordeel van een gekalibreerde itembank is dat daar ook veel meer opgaven aan te relateren zijn van andere afnamen. Door de afnameverplichting dat de centrale toetsen door alle leerlingen in de beoogde leerjaren worden afgenomen, is het afnamedesign relatief simpel: een kandidaat heeft altijd een centrale toets gemaakt, die daarmee te relateren valt aan de gegevens van een andere toets. Voor zo'n koppeling is geen bijkomende dataverzameling meer nodig, anders dan de afname van die andere toets. Deze andere toets kan een toets met leerplan-specifieke opgaven zijn die op school wordt afgenomen of

[9] Deze informatie kan ook over de gehele populatie of over deelgroepen gegeven worden; aan valt te raden hier gebruik te maken van de met een IRT model geschatte klassieke waarden in plaats van de direct geobserveerde waarden. Het doet er dan namelijk niet meer toe wie de opgave nu gemaakt heeft, en levert zo ook bij adaptief gebruik van de bank waarden op die over opgaven onderling vergelijkbaar zijn.

[10] Voor de verschillen in equivaleren tussen klassieke testtheorie en IRT verwijzen we naar Kolen, M.J., & Brennan, R.L. (1995). *Test Equating*. New York: Springer.

vanuit de koepel wordt uitgegeven of een toets die gerelateerd is aan een internationale peiling. Merk wel op dat het psychometrisch op zich niet cruciaal is wat de oorsprong van de toets is, maar voor de praktische uitvoering uiteraard wel. Het gaat dan om de noodzakelijke technologische aanpassingen, zowel wat betreft de digitale afname, als ook in de dataverzameling. Ook moeten de verantwoordelijkheden voor wat betreft het beheer van de gegevens en het geven van labels aan items uitgewerkt worden. Vanuit scholen[11] is dat lastiger dan wanneer de uitvoer van de internationale peiling (deels) in dezelfde handen ligt als het beheer van itembank van de Vlaamse centrale toetsen[12] .

Praktische uitdagingen

Er zijn een aantal praktische zaken waar rekening mee gehouden moet worden als er met een itembank gewerkt wordt. Opgaven kunnen verouderen omdat de gebruikte contexten niet meer voorkomen in de dagelijkse praktijk en niet meer herkenbaar zijn voor nieuwe generaties leerlingen. Ook kunnen opgaven verouderen omdat ze te publiek zijn geworden, bijvoorbeeld als ze in voorbeeldtoetsen zitten. In dat geval meet een item niet meer de gewenste vaardigheid van leerlingen. Deze verouderde items verdwijnen uit de itembank. Om de itembank in omvang voldoende groot te houden, is verversing van de bank nodig. Dat wil zeggen dat er jaarlijks nieuwe items toegevoegd moeten worden. De mate waarin dit moet gebeuren, hangt niet alleen af van de technische of inhoudelijke houdbaarheid van items, maar ook van de mate waarin items geheim moeten blijven, en de mate waarin dat lukt.

Een tweede uitdaging zit in de omvang van de itembank. Een uitgebreide itembank is een kostbare zaak. Als een variant van een itembank gekozen wordt waarbij tijdens afname adaptief een toetsvariant wordt samengesteld, dan zou het aantal opgaven in de bank ongeveer tien keer zo groot moeten zijn als de lengte van de toets. Daarnaast moeten bij een adaptieve toets reeds veel gegevens over de opgaven verzameld worden voordat de eerste afname plaats kan vinden. Deze gegevens stammen vaak uit proeftoetsing. Dit zorgt voor een extra kostenpost, naast zorgen omtrent de geschiktheid van de afnamegegevens, of omtrent het bekend worden van opgaven. Ook moet het gehele domein of de gehele vaardigheid gerepresenteerd zijn in een itembank. Een kleine itembank kan niet de hele vaardigheid dekken. Maar ook een grote itembank kan te éénzijdig gevuld zijn. 'Kale' rekensommen zonder context (bijv. 2 x 4=) kunnen bijvoorbeeld snel geconstrueerd worden. Een wiskunde-bank kan snel in omvang toenemen door veel kale rekensommen toe te voegen. Echter, de vraag is of een dergelijke itembank de wiskunde-vaardigheid afdoende dekt.

Andere nadelen van het gebruik van een gekalibreerde itembank ten opzichte van een "normale" itembank zijn te vinden in het gebruik van IRT. Deze nadelen zijn eerder besproken in Hoofdstuk 3.

[11] Dit zijn stappen die wellicht genomen kunnen worden na een aantal succesvolle jaren van afnamen van centrale toetsen in Vlaanderen. Meer hierover meer in Sectie 5.3 die verder ingaat op de uitwerking van vraag 8 en 9 van de aanvraag.

[12] Een voorbeeld waarbij schalen van Vlaamse meting en een internationale meting gecombineerd worden is het PIRLS-repeat project. Hierin werden de Vlaamse peilingsresultaten begrijpend lezen in het basisonderwijs weergegeven op de meetschaal van PIRLS. Zie: Dockx, J., Van Landeghem, G., Aesaert, K., Van Damme, J., & De Fraine, B. (2019). Begrijpend lezen van het vierde naar het zesde leerjaar. Herhalingsmeting van PIRLS in 2018 vergeleken met PIRLS 2016. Leuven: CO&E. (`https://onderwijs.vlaanderen.be/nl/progress-in-international-reading-literacy-study-pirls`).

Geheimhouding van de opgaven

Een belangrijke factor bij een itembank is de vraag hoe geheim de opgaven moeten zijn. Dat betreft de openbaarheid van de itembank, maar ook de flexibiliteit van de afnames. Met dat laatste wordt bedoeld dat het voor te stellen is dat niet iedere leerling op hetzelfde moment de examens maakt, wat ervoor zorgt dat de opgaven bekend kunnen raken voor latere afnamen. Dat gevaar is groter als er flexibiliteit is binnen een school omdat de leerlingen dan eerder met elkaar in contact staan en de inhoud van opgaven kunnen uitwisselen. De reden om alsnog voor flexibiliteit te kiezen, is dat het een minder grote druk op de IT-infrastructuur is omdat er minder computers tegelijkertijd nodig zijn[13], en de piekbelasting in het centrale systeem lager is[14].

Als het niet van groot belang is dat de opgaven geheim zijn, kan er voor een openbare itembank gekozen worden. Dit heeft als voordeel dat een duidelijke verantwoording naar het publiek mogelijk is over het totaal aan gebruikte opgaven. Dit kan uiteraard leiden tot oefenen van opgaven, maar dat kan juist tot een gewenste verbetering van de vaardigheid of kennis leiden. Er zijn wel een aantal voorwaarden waaraan voldaan moet zijn bij een dergelijke opgavenbank. Ten eerste moet de omvang van de benodigde (openbare) itembank groot genoeg zijn om dit positieve effect van training te hebben. Een kleine itembank kan niet de hele vaardigheid dekken. Daarnaast is ook het type vaardigheid dat gemeten wordt van belang, en daarmee ook het type opgaven. Als de omvang van het kennisdomein beperkt is, zoals sommen onder tien of alle hoofdsteden van Europa, dan is het trainen met alle mogelijke opgaven geen probleem omdat dit de vaardigheid dekt. Is het construct breder, zoals woordenschat van een 12-jarige, dan is dat lastiger. Dan moet de opgavenbank ook aanzienlijk groter zijn om het te meten construct voldoende te kunnen dekken, omdat anders het kennen van de opgaven een slechte representatie is van de woordenschat. De trainingsduur moet ook overeenkomen met de tijd die er normaal staat voor het aanleren van een leerdoel.

Het is voor te stellen om met een relatief kleine bank te werken, in combinatie met een vaste toets voor belangrijke beslissingen. Een beperkte uitbreiding van de bank met ankeropgaven voor de vergelijking over jaren is dan in principe voldoende. Het voordeel hiervan is dat de productie van de opgaven dan relatief goedkoop is omdat het aantal opgaven beperkt is. Het is dan wel van belang dat al deze opgaven (zeer) goed van kwaliteit zijn, hetgeen voor een deel een empirische vraag is. Dat betekent dat het vooraf uittesten van de opgaven van groter belang is. Echter, wanneer de belangen hoog zijn dan is het uittesten een precaire situatie, want bij het uittesten kunnen de opgaven ook bekend raken. Bij het beheer van een kleine bank moet complete geheimhouding geborgd zijn, wanneer de uitkomsten van de toetsen uit de bank van groot belang zijn. Dat betekent weinig flexibiliteit in de afname, en voldoende veiligheidsmaatregelen. Alleen in het geval de toetsen puur formatief ingezet gaan worden, zijn die zaken minder problematisch.

Een grotere itembank is om een aantal redenen aan te raden. Zoals al genoemd, is het bij een grote itembank mogelijk om een te meten construct of domein breder te dekken.

[13] Dit zal waarschijnlijk meer spelen in het basisonderwijs dan in het secundair onderwijs. In het geval van een examen opstelling van de computers voor een gehele jaargroep zou dat ook voor (een deel van) het secundair een uitdaging kunnen zijn.

[14] In hedendaagse toepassingen lijkt dit steeds minder een probleem te zijn. Dit valt verder onder de expertise van Perceel 3.

Niet een enkele opgave maar een grotere hoeveelheid geeft een meer valide dekking van de eindterm. Daar waar een enkele leerling nog steeds uiteindelijk maar een beperkt aantal opgaven maakt tijdens de afname, is de dekking op het geaggregeerd niveau veel beter als een grote diversiteit aan opgaven ingezet kan worden.

Een ander belangrijk voordeel van een grote itembank is dat het mogelijk is de bank op te delen in een openbaar deel en een geheim deel. Het is het gemakkelijkst het openbare deel pas na afname openbaar te maken. Maar het is zelfs mogelijk dit vooraf te doen, bijvoorbeeld als oefenmateriaal. Openbaar maken kan ook in fases gebeuren. Stel dat het vanwege de behoefte aan flexibiliteit gewenst is een toets op verschillende momenten af te kunnen nemen, bijvoorbeeld op drie. Dan kan op ieder van die momenten een unieke toets beschikbaar gesteld worden. Deze drie toetsen kunnen dankzij een geheim deel met elkaar verbonden worden en de toetsscores zijn dan met behulp van IRT te equivaleren. Een manier om een deel geheim te houden is door dit bij slechts een relatief beperkt aantal leerlingen af te nemen, en dit geheime deel ook in kleinere delen op te knippen. Zeker als de opgaven bij een digitale afname niet in een vaste volgorde staan, is het relatief gemakkelijk ze geheim te houden.

Ook als er maar een toets is die op een enkel moment wordt afgenomen is een geheim deel zinvol. Een geheim deel is dan in te zetten om over de tijd een jaarvergelijking te verkrijgen. De opgaven in het geheime deel komen niet in het publieke domein terecht en zijn daardoor na een jaar opnieuw te gebruiken. Na een jaar kunnen dan opgaven uit het geheime deel onderdeel gaan uitmaken van een regulier deel dat weer publiek gemaakt wordt als de “reguliere toets”. Ook zou het geheime deel te gebruiken zijn voor de toetsen in “tussenjaren” die het mogelijk maken om metingen die twee leerjaren van elkaar verschillen met elkaar te verbinden.

Een itembank is makkelijk groter te maken doordat er met de jaren steeds meer opgaven aan toegevoegd kunnen worden. Er kan voor gekozen worden de nieuwe opgaven niet mee te nemen bij het bepalen van het toetsresultaat totdat de standaardfout van de parameters klein genoeg is. Daarna kan een dergelijke opgave functioneren als een gewone bankopgave[15].

Ook kan deze bank continu aangevuld worden met resultaten van proefafnames, al moeten die proefafnames dan wel zoveel mogelijk onder dezelfde condities plaatsvinden als de werkelijke toets. Leerlingen kunnen bijvoorbeeld minder gemotiveerd zijn bij een proefafname en dit kan tot afwijkende schattingen leiden. Het is zelfs voor te stellen dat de itembank deels uitgebreid wordt met opgaven uit externe bronnen, zoals van de koepels of scholen zelf. Het is echter aan te raden om deze stap pas te nemen wanneer er een goed werkende itembank bestaat.

5.2 Selectie te toetsen onderwijsdoelen

De vrees bestaat dat bij selectie van een beperkt aantal onderwijsdoelen om te toetsen, in het onderwijs de aandacht voor die onderwerpen ten koste gaat van de andere onderwerpen

[15] Dergelijke opgaven worden ook wel zaai-opgaven genoemd. Deze worden in de echte toetsafnamen gezaaid en groeien als ze voldoen uit tot volwaardige opgaven. Afhankelijk van de eisen die gesteld worden rond de acceptabele standaardfout van de parameters, wat ook weer afhankelijk is van het gebruikte model kan dit variëren van 400 tot 1.600 afnamen waarbij de score niet meetelt, totdat dit item voor de vaardigheidsbepaling gebruikt kan worden.

die minstens zo belangrijk zijn, maar die niet aan de orde komen in de centrale toetsen. Hoe groter de belangen van de toetsen zijn voor de betrokkenen, hoe groter dat gevaar. Bij verschillende toetsafnames kunnen verschillende onderwijsdoelen opgenomen worden. Dit kan gerealiseerd worden met behulp van een zogenaamd rotatieprincipe. Redenen om dit te doen, kunnen naast het voorkomen van curriculumvernauwing, ook het voorkomen van een te lange toetstijd voor leerlingen zijn. Zo kan er ook gerapporteerd worden op het niveau van clusters van eindtermen.

Wanneer er sprake is van een digitale itembank zijn er verschillende afnamevormen mogelijk. Het is mogelijk om uit de itembank een toets te maken, met een vaste volgorde van opgaven. Een alternatief is dat er verschillende toetsversies gebruikt worden. Een belangrijk onderscheid dat bij het gebruik van meerdere toetsversies moet worden gemaakt, is òf dat deze weliswaar uit verschillende opgaven bestaan, maar deze wel dezelfde vaardigheid meten òf dat er meerdere toetsversies bestaan die verschillende vaardigheden meten. De principes voor het adequaat samenstellen van verschillende toetsversies zijn in beide opties hetzelfde. De implicaties voor het gebruik zijn verschillend tussen beide opties. We bespreken eerst in Paragraaf 5.2.1 de principes van rotaties en selectie. In Paragrafen 5.2.2 en 5.2.3 komen de implicaties op het meten van leerwinst op het niveau van respectievelijk de individuele leerlingen en van de school aan de orde.

5.2.1 Rotatie

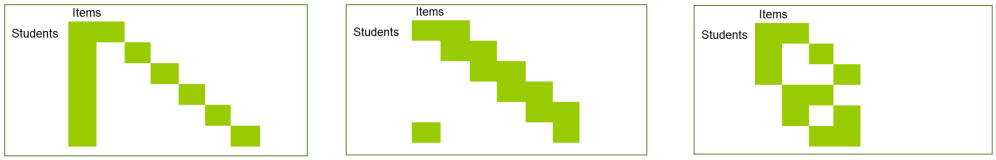
Rotatie kan plaatsvinden *binnen* en tussen toetsafnames. Rotatie die binnen een toetsafname plaatsvindt, leidt tot het gebruik van verschillende toetsversies voor leerlingen. Rotatie van verschillende onderwijsdoelen die plaatsvindt tussen toetsafnames kan –maar hoeft niet – leiden tot eenzelfde toets voor de groep leerlingen op een toetsafnamemoment, maar met verschillende toetsinhoud tussen de afnamemomenten.

Er zijn meerdere redenen om gebruik te maken van verschillende toetsversies binnen een toets. Een eerste reden kan zijn om te faciliteren dat niet iedereen op hetzelfde moment een toets kan maken. Het kan gewenst zijn dat leerlingen die bijvoorbeeld door ziekte niet in staat zijn geweest deel te nemen aan een toets, een andere toetsversie voorgelegd krijgen bij de herkansing. Dat geldt zeker voor leerlingen die een herkansingsmogelijkheid krijgen aangeboden, wanneer ze een toets de eerste keer niet voldoende hebben gemaakt. Een tweede reden om te werken met verschillende toetsversies is dat het mogelijkheid geeft om de items te matchen aan de vaardigheid van de leerlingen. Vormen van deze zogenaamde adaptiviteit worden behandeld in Sectie 5.5. Een derde reden om met meerdere toetsversies te werken is dat het gelegenheid biedt om een compleet inhoudsdomein te meten op geaggregeerd niveau, terwijl gelijktijdig de toetstijd op individueel niveau beperkt blijft. Dit kenmerk zal verder behandeld worden in Paragraaf 5.2.2.

Een meer algemeen voordeel van gebruik te maken van verschillende toetsversies is dat de opgaven uit een itembank minder snel bekend raken. Zeker als er meer momenten zijn waarop de toets te maken valt. Een aanvullend middel om tegen te gaan dat opgaven bekend worden, is om de opgaven in een willekeurige volgorde aan te bieden. Dit kan ook afkijken tegengaan, als het niet mogelijk is om een examenopstelling te gebruiken. Ook is het bij een willekeurige volgorde –mogelijk als er een versie is of verschillende versies– een voordeel dat geheime opgaven minder snel opvallen. Bij een vaste volgorde kunnen leerlingen aan elkaar vragen wat de eerste, tweede of derde (etc.) opgave was,

en als leerlingen dan merken dat in hun toetsen opgaven in een verschillende volgorde stonden, kan dat verwarring opleveren, of wordt bekend dat er geheime opgaven zijn. Als de volgorde willekeurig is dan is het moeilijker te achterhalen welke opgaven nu net anders zijn.

Het verkrijgen van vergelijkbare resultaten ondanks het gebruik van meerdere toetsversies is bij het gebruik van IRT te realiseren, maar een cruciale voorwaarde is het creëren van *overlap* in het rotatiedesign. Dit betekent dat binnen de verschillende toetsversies dezelfde items terugkomen (zogenaamde ankeritems) die ervoor zorgen dat alle toetsversies met elkaar verbonden zijn. Voorbeelden van dit soort design zijn het *anchor item design*, het *block interlaced design* en een *balanced block design*. Illustraties van deze designs zijn te vinden in de Figuren 5.1a - 5.1c.



(a) Anchor item design (b) Block interlaced design (c) Balanced block design

Figuur 5.1: Designs

De Figuren 5.1a - 5.1c zijn schematische weergaves van verschillende designs. In de rijen zijn (groepen) leerlingen te vinden. De kolommen representeren (verzamelingen) items. Te zien is verschillende groepen leerlingen die verschillende verzamelingen items maken. Alle design zijn echter voorbeelden van verbonden designs. Er kan een link worden gemaakt van elke verzameling items naar een andere verzameling items via een deel van de items dat door beide groepen leerlingen wordt gemaakt. Deze verbondenheid maakt het eenvoudiger om een gemeenschappelijke gekalibreerde IRT schaal te realiseren.

Het anchor item design (Figuur 5.1a) kenmerkt zich door één gemeenschappelijk deel van de toets dat door alle leerlingen wordt gemaakt. Dit is een design dat overzichtelijk is, maar in statistische termen niet het meest efficiënte design is. De gemeenschappelijke of anker-items zullen namelijk met veel meer precisie worden geschat dan de andere items, omdat deze door veel meer groepen studenten worden afgenomen. In Figuur 5.1b wordt een block interlaced design schematisch weergegeven. Dit design heeft als voordeel dat alle items door ongeveer dezelfde aantallen leerlingen worden gemaakt. Alle items kunnen dus met dezelfde statistische precisie geschat worden. Het is bovendien een overzichtelijk en daardoor praktisch uitvoerbaar design. Het aantal toetsversies blijft vaak beperkt en het is eenvoudig om controle te houden over de inhoudsrestricties die aan de samenstelling van toetsversies verbonden zijn. In Figuur 5.1c tenslotte, wordt een balanced bock design gepresenteerd. Dit is het meest complexe design en lastiger praktisch uitvoerbaar. Het statistische voordeel dat met dit design behaald wordt, is dat alle item paren geobserveerd worden en zo goed geschikt zijn voor het detecteren van item misfit.

Als er meerdere toetsversies gebruikt worden binnen een toetsafname, kan het toewijzen van een versie aan een leerling zowel random plaatsvinden als door middel van toewijzing vooraf. Dat laatste is aan te raden doordat er dan meer grip is op welke versies op een school terechtkomen. Ook zijn vormen van toewijzing van toetsen te bedenken die een meer adaptief karakter hebben. Hierover meer in Sectie 5.5 die verder ingaat op adaptiviteit.

Een extreme vorm van willekeurigheid is dat het aanbieden van opgaven volledig aan het toeval wordt overgelaten en in principe iedere leerling een eigen toets kan krijgen. Hierin valt wat meer ordening aan te brengen door de opgaven van te voren op basis van eindtermen en inhoud op te delen in verschillende groepen en de willekeur alleen binnen die groepen te laten plaatsvinden, zodat niet een leerling bijvoorbeeld alleen optelsommen krijgt, maar de toets inhoudelijk gevarieerd blijft. Deze variant is waarschijnlijk niet aan te raden. De vergelijking is toch net iets lastiger dan bij de andere voorstellen en de meerwaarde is gering.

5.2.2 Gevolgen van rotatie op individueel niveau

Het is mogelijk om uit de itembank een toets te maken, met een vaste volgorde van opgaven. De koppeling met andere leerjaren en andere afnamejaren kan dan plaatsvinden door middel van externe toetsen waarvan de uitkomsten via de itembank op een IRT-schaal te projecteren zijn, zodat de leerwinst en trends te ontdekken zijn. Het voordeel van deze toets met de vaste vorm is dat de leerlingen allen op exact dezelfde items, in exact dezelfde volgorde vergeleken worden. Als toetsen zeer high-stakes zijn, en iedere leerling zonder additionele statistische modellen (zoals IRT) met elkaar vergeleken moeten worden, is dit een optie. Dit is de toetsvorm die bij de Nederlandse centrale eindexamens wordt gebruikt. Een andere mogelijkheid is om na een aantal jaar de centrale toetsen aan te vullen met een focustoets, die dieper op een onderwerp ingaat om daar in meer detail de vaardigheid in Vlaanderen te meten, zoals bij (internationale) peilingen met enige regelmaat gebeurt. Dit laatste is echter een scenario voor als het centraal toetsen meer ingeburgerd is.

Het is ook goed mogelijk om uit een itembank te werken met verschillende toetsversies waarin gevarieerd wordt met items die dezelfde vaardigheid meten. Dat betekent dat leerlingen een terugrapportage krijgen op het niveau van de drie vaardigheden begrijpend lezen, schrijven, en grammatica voor het vak Nederlands en voor de vaardigheid op het gebied van Wiskunde. De dekking van de toetsen betreft wel de eindtermen, maar het aantal opgaven per eindterm is te klein om daar bij een leerling zeer nauwkeurig over te meten. Bij Wiskunde betekent dat dat de opgaven wel de vaardigheden getallen, verhoudingen, meten en meetkunde, en verbanden omvatten, maar bij de terugrapportage op individueel niveau er niet afzonderlijk onderscheid gemaakt zal worden[16]. Deze vorm van rotatie heeft geen gevolgen voor het volgen van leerlingen. Het roteren van onderwijsdoelen op individueel niveau heeft gevolgen voor het meten van de leerwinst op individueel niveau, hetgeen dan lastiger wordt. Individuele leerwinst en ontwikkeling is dan niet meer goed mogelijk omdat de unidimensionaliteit geschonden is.

5.2.3 Gevolgen van rotatie op geaggregeerd niveau

Binnen (internationale) peilingsonderzoeken is het heel gebruikelijk om gebruik te maken van complexe rotatiedesigns. Dit komt mede doordat in dit soort onderzoeken de tijd die beschikbaar is voor het maken van items vrij beperkt is. Deze beperkte tijd is vooral ingegeven door het inspelen op de verwachte lagere motivatie van leerlingen om de toets

[16]Merk op dat de terugrapportage wel gebruik kan maken van iets als een profielanalyse [Verhelst, N.D. (2012). Profile Analysis: A Closer Look at the PISA 2000 Reading Data. *Scandinavian Journal of Educational Research*, 56(3): 315–32] waardoor er wel iets gezegd kan worden op een meer gedetailleerd niveau, maar dat is niet een vaardigheidsschatting op detailniveau.

te maken, maar ook om de totale toetsbelasting op het reguliere lesgeven te beperken. Dit betekent dat er van verschillende toetsversies gebruik gemaakt wordt, waarbij individuele leerlingen slechts van een beperkt aantal (deel)domeinen items voorgelegd krijgen. Desondanks kunnen, door gebruik te maken van een verbonden design, op geaggregeerd systeemniveau uitspraken gedaan worden voor alle domeinen die in de toets zijn opgenomen. Daarmee kunnen de eindtermen breder gedefinieerd worden en dat is vooral interessant als de interesse meer uitgaat naar de meting op een geaggregeerd niveau (klas, school of nog verder geaggregeerd). Als de meting van het individu en de individuele leerwinst niet het belangrijkste meetdoel is, dan is het zelfs mogelijk om de versies niet alleen uit verschillende opgaven bij dezelfde vaardigheid te laten bestaan maar ook uit opgaven voor verschillende vaardigheden.

Als er sprake is van verschillende toetsversies, waarbij voor alle versies samengenomen het aantal gebruikte opgaven aanzienlijk groter is dan in een enkele toets past, is op een geaggregeerd systeemniveau wel op een meer gedetailleerd niveau leerwinst te meten. Als op een school verschillende versies van een toets afgenomen worden, kan zo een domein in volle breedte afgenomen worden. Op die manier kan de gehele vaardigheid ook gedetailleerd op een school gemeten worden. Wanneer aangenomen wordt dat iedere observatie beschouwd kan worden als een random trekking uit de leerlingpopulatie (al dan niet gecorrigeerd voor achtergrondvariabelen), dan kan zo ook de leerwinst op de school gedetailleerd bijgehouden worden.

Samenvattend kan gesteld worden dat het simpelste scenario voor de vergelijking van de leerlingen, scholen en verandering door de tijd heen (leerwinst) is dat scholen of leerlingen bij iedere meting op alle onderwijsdoelen getoetst worden. Op detailniveau kan de rapportage verschillen per meetniveau (leerling, school, systeem). Bij slimme rotatie van onderwijsdoelen levert het geen problemen op voor systeemmetingen, maar kan het wel problematisch zijn op het individuele niveau (vergelijking tussen individuele leerlingen onderling en door de tijd).

5.2.4 Moeilijk te toetsen vaardigheden

De vorige paragrafen over rotatie betroffen de rotatie van de opgaven. Als het gaat om de selectie van de te toetsen onderwijsdoelen is het ook van belang in te gaan op de moeilijker te toetsen vaardigheden. Veel van de genoemde vaardigheden zijn goed te meten met behulp van meerkeuze-opgaven en andere opgaven waarbij de antwoorden automatisch te scoren zijn[17]. Dergelijke opgaven zijn echter niet voor metingen van alle vaardigheden op alle niveaus te gebruiken.

Voor de vaardigheden Nederlands lezen en Nederlands luisteren, en delen van Nederlandse taalbeschouwing zal het mogelijk kunnen zijn om op alle beoogde niveaus met automatisch scoorbare opgaven te werken. Er is hier veel ervaring mee. Als we kijken naar de meer actieve vaardigheden zoals schrijven en spreken in het Nederlands is dat lastiger.

[17]Binnen afnamen op de computer zijn er talloze vormen van opgaven te ontwikkelen die automatisch te scoren zijn. Dit kan korte-antwoord opgaven betreffen, waarbij de kandidaat een enkel woord of getal opgeeft, of zelfs een paar woorden. Ook typisch *drag-and-drop* opgaven zijn mogelijk, of opgaven waarbij een kandidaat in opdracht een aantal handelingen moet verrichten die gevolgd en beoordeeld worden door de computer. Het is evident dat hoe complexer de opgave wordt, en hoe complexer de automatische beoordeling, hoe duurder het wordt om een dergelijke opgave te ontwikkelen. Over het algemeen zijn de meerkeuze opgaven het goedkoopst om te ontwikkelen, en om te scoren. Meer over andere vormen van opgaven in Sectie 5.8

Voor metingen binnen het basisonderwijs is het wellicht nog voor te stellen dat opgaven die zeer gestructureerde, korte antwoorden uitlokken nog automatisch te scoren zouden kunnen zijn. Op een hoger vaardigheidsniveau is het, gezien de gestelde eindtermen, niet goed mogelijk automatisch gescoorde opgaven te maken die deze vaardigheden valide kunnen meten. Wellicht dat later er technieken zullen komen die dat wel mogelijk zouden kunnen maken, maar de verwachting is niet dat deze op korte termijn binnen de centrale toetsen in Vlaanderen ingezet kunnen worden tegen aanvaardbare kosten[18].

Als we kijken naar schrijfvaardigheid dan zou een centrale afname eventueel mogelijk zijn, zij het zonder de automatisch gescoorde opgaven. Dat betekent dat deze opgaven anders gescoord moeten worden, en zullen er beoordelaars nodig zijn. Dit levert een aantal uitdagingen op die opgelost moeten worden. Het begint met de keuze wie deze beoordelaars moeten zijn. Wanneer dat bekend is, moet bepaald worden hoe de leerling-antwoorden bij deze beoordelaars terecht komen. Tot slot moet bij menselijke beoordelingen bepaald worden wat de beoordelaarsovereenstemming is, omdat een verschil in beoordelaars ook een impact kan hebben op de gerapporteerde scores. Een gerelateerde vraag is hierbij ook hoeveel beoordelaars per schrijfproduct nodig zijn.

Hier zijn verschillende scenario's mogelijk. Een eerste is dat de schrijfproducten van de leerlingen door leerkrachten beoordeeld worden. Dit kunnen de eigen leerkrachten zijn, maar wanneer de antwoorden digitaal opgeslagen zijn dan is het ook mogelijk dat deze willekeurig over verschillende leerkrachten in Vlaanderen verdeeld worden. Leerkrachten krijgen dan producten te beoordelen van leerlingen van verschillende scholen zodat niet alle leerlingen van een school door één leerkracht beoordeeld worden. Zo wordt de impact van een individuele beoordelaar voor de school beperkt. Als de toets uit verschillende subtaken bestaat, zijn er ook technieken mogelijk[19] waarbij het werk van een enkele leerling per subtaak verdeeld wordt over verschillende leerkrachten, zodat het niet een enkele beoordelaar is waardoor de leerling beoordeeld wordt en zo de impact van de enkele beoordelaar beperkt wordt voor een individu. Uiteraard kan het verdelen van de toetsen of van subtaken over beoordelaars ook plaatsvinden wanneer niet leerkrachten de beoordeling uitvoeren, maar externe beoordelaars gebruikt worden. Daar waar bij leerkrachten het beoordelen van schrijftaken nog als onderdeel van hun reguliere werk gezien kan worden – ze zouden anders toch ook schrijftaken van (hun eigen) leerlingen moeten beoordelen – is dat bij externe beoordelaars niet het geval. Wanneer van externe beoordelaars gebruik gemaakt wordt, zal dat extra kosten met zich meebrengen[20]. Ter

[18]Hoe hoog de kosten zullen zijn is uiteraard nog niet duidelijk aangezien de techniek nog niet zover is. De ervaring met dergelijke technieken is dat de eerste jaren dat dergelijke nieuwe technieken beschikbaar zijn de kosten hoog liggen. Zeker wanneer de belangen hoog liggen is het gebruik van dergelijke technieken problematisch omdat bij grote belangen deze technieken (vrijwel) foutloos dienen te werken.

[19]Deze technieken worden in Nederland gebruikt bij de beoordeling van de Staatexamens Nederlands als tweede taal: `https://www.staatsexamensnt2.nl/item/over-staatsexamens-nt2`.

[20]De additionele kosten door het gebruik van externe beoordelaars is van een aantal zaken afhankelijk. Ten eerste van hoeveel leerlingen deze schrijftaken maken. Ten tweede hoe lang het beoordelen van een leerlingwerk duurt. Ten derde hoeveel deze beoordelaars betaald krijgen. Als we uitgaan van 75.000 leerlingen per leerjaar, de tijd per leerling in het lager onderwijs 6 minuten gemiddeld is en in het secundair onderwijs 12 minuten (of dit realistisch is, is geheel afhankelijk van de uitgegeven taken, maar lijken eerder onderschattingen dan overschattingen), dan betreft dit 45.000 uur werk. Als de beoordelaars dit werk tegen €10 per uur zouden willen doen (wat ongeveer overeenkomt met minimumloon) dan kost het beoordelen van deze werken, met één beoordelaar per toets ongeveer €450.000. Het is aannemelijk dat dit bedrag hoger uit zou vallen omdat

controle van de beoordelaarsovereenstemming – wat nodig is om te bepalen of de metingen voldoende betrouwbaar zijn – zal een deel van de werken twee maal beoordeeld moeten worden[21].

Het gebruik van menselijke beoordelaars levert ook altijd een aantal andere zaken op waar rekening mee gehouden moet worden. Ten eerste is dat de additionele tijd die het kost ten opzichte van de automatische beoordelingen. Ten tweede moet er een systeem gebouwd of gekocht worden dat de leerlingwerken verdeeld over de beoordelaars, en de scoring door de beoordelaars goed bijhoudt. Dit laatste zal ook additionele kosten opleveren[22].

Voorgaande betrof het meten van schrijfvaardigheid. Wanneer spreektaken meegenomen worden in de centrale toetsing spelen dezelfde overwegingen, aangevuld met enige andere uitdagingen. Ten eerste is een centrale afname van spreektaken lastiger uit te voeren. Daar waar bij schrijftaken deze door een grote groep tegelijkertijd in stilte kan worden uitgevoerd, is dat bij spreektaken per definitie niet mogelijk. Dit levert een aantal aanvullende voorwaarden bij de afnamecondities.

Het tweede punt betreft de beoordeling bij spreektaken. Er is er de keuze om de beoordeling ter plekke tijdens de afname uit te voeren, achteraf op basis van een opname, of een combinatie van de twee, waarbij bijvoorbeeld bij een live-beoordeling een (online) video-opname gemaakt wordt die gebruikt kan worden door een tweede beoordelaar. Ook hierbij zal het ervan afhangen hoe de spreekvaardigheid geoperationaliseerd wordt wat de mogelijkheden zijn. Als deze opnamen opgeslagen moeten worden dan zijn hier ook diverse technische vereisten aangaande de opslag van de opnamen, die qua volume meer opslagruimte kosten dan geschreven teksten. Een punt waardoor beoordelingen van spreektaken tijdrovender zijn – en bij de externe beoordelaars dus ook kostbaarder – is dat de beoordelingstijd even lang is als de afnametijd. Daar waarbij schrijftaken de leerlingen meer tijd nodig hebben om te schrijven dan de beoordelaar nodig heeft om te beoordelen zal bij spreektaken de beoordelaar de gehele uitvoering van de spreektaak moeten beluisteren[23,24].

Bij zaken als rekenen en wiskunde spelen deze zaken iets minder. Op het niveau van het basisonderwijs is het goed mogelijk om gebruik te maken van meerkeuze opgaven of korte

---

deze personen ook getraind moeten worden, en het de vraag is of beoordelaars met voldoende kwalificaties gevonden kunnen worden die voor €10 per uur dit werk willen doen.

[21] Er zijn geen richtlijnen hoeveel werken precies dubbel beoordeeld moeten worden. Als het alleen gaat om te zien wat de beoordelaarsoverstemming is, kan dit beperkt zijn tot een paar 100. Als het gaat om het evalueren van de kwaliteit van de beoordelaars, dan zouden per beoordelaar minstens 10 dubbele beoordelingen nodig zijn. Als vervolgens blijkt dat de overeenstemming of de kwaliteit van (enkele) beoordelaars tegenvalt dan moeten voor een voldoende betrouwbare meting meer, zo niet alle werken dubbel beoordeeld worden, met alle kosten van dien.

[22] In Nederland wordt een digitale beoordelingsmodule gebruikt bij de beoordeling van de Staatexamens Nederlands als tweede taal: `https://www.staatsexamensnt2.nl/item/over-staatsexamens-nt2`. Hoe hoog de kosten precies zijn voor een dergelijk systeem is lastig te bepalen omdat er verschillende keuzes te maken zijn in hoe je zo'n systeem neerzet.

[23] Naast luisteren zal een beoordeling mogelijk ook vereisen dat de spreker gezien wordt tijdens de taak, bijvoorbeeld als het gaat om het beoordelen van een eindterm aangaande het gepast inzetten van lichaamstaal.

[24] In Nederland wordt voor zowel schrijven als spreken een digitale beoordelingsmodule gebruikt bij de beoordeling van de Staatexamens Nederlands als tweede taal: `https://www.staatsexamensnt2.nl/item/over-staatsexamens-nt2`. Voor de vaardigheid spreken beluisteren beoordelaars audiofragmenten, ingesproken door deelnemers, en beoordelen zij die op verschillende aspecten.

antwoordopgaven waarbij de leerling bijvoorbeeld alleen een getal in hoeft te vullen. Het is ook mogelijk creatievere automatisch scoorbare opgaven in te zetten waarbij de leerlingen handelingen op de computer moeten uitvoeren, bijvoorbeeld met een virtuele liniaal of het trekken van lijnen. Over het algemeen geldt hierbij wel dat hoe creatiever de opgave is, hoe kostbaarder de ontwikkeling ervan. Voor het hoogste niveau in het secundair onderwijs valt ook nog voor te stellen dat voor de moeilijkere meetbare wiskundige vaardigheden er ook open opgaven ingezet kunnen worden die een menselijke beoordelaar nodig hebben.

Bij menselijke beoordelingen voor wiskunde-opgaven kan een vergelijkbaar systeem opgezet worden als bij de schrijftaken[25]. Wat betreft het beoordelen van dergelijke taken is het over het algemeen wel zo dat de beoordelingsvoorschriften makkelijker te maken zijn, en de beoordelaarsovereenstemming ook makkelijker te verkrijgen is. Als daarnaast ook slechts een deel van de opgaven door een menselijke beoordelaar beoordeeld hoeven te worden, en niet de hele toets, zal dat ook tijd schelen en bij externe beoordelaars dus ook kosten schelen. Het is bij dergelijke opgaven dan wel van belang dat de leerlingen gewend moeten zijn om dergelijke opgaven op de computer te maken. Daar waar leerlingen, zeker in het secundair onderwijs, vaak wel gewend zijn om teksten op de computer te schrijven[26], is het maken van wiskunde opgaven op de computer een minder gebruikelijke handeling.

Het is uit bovenstaande duidelijk dat het opnemen van de niet-automatisch scoorbare opgaven in centrale toetsing eigen uitdagingen kent. Zoals aangegeven zijn er oplossingen om menselijke beoordelingen mee te nemen, maar deze kunnen kostbaar zijn. Deze uitdagingen geheel uit de weg gaan door centrale metingen van actieve en creatieve vaardigheden, zoals schrijven en spreken[27], geheel uit de weg te gaan, zal tot een ongewenste curriculumvernauwing leiden.

Voor de moeilijk te toetsen vaardigheden is het scenario van net- of koepel-specifieke toetsen naast de centrale toetsen een mogelijke oplossingsrichting. In Nederland is dit de aanpak die gevolgd wordt bij het onderscheid in het centraal schriftelijk en het schoolexamen. Het schoolexamen dekt de moeilijker centraal te meten vaardigheden zoals de spreekvaardigheid, literatuur en een deel van schrijfvaardigheid. Een variatie hierop is dat vanuit het steunpunt toetsontwikkeling een of meer mogelijke opdrachten wordt aangeboden aan de scholen, samen met de beoordelingsschema's en de scoring. Deze scores moeten dan in een centraal beheerd administratiesysteem worden ingevoerd voor

[25] Als er voor schrijftaken geen systeem is opgezet om de menselijke beoordelingen te ondersteunen, zal het waarschijnlijk kostbaar zijn om dit apart op te zetten voor alleen de beoordeling van de wat ingewikkeldere wiskunde opgaven. Dan is het financieel waarschijnlijk verstandiger om bij wiskunde alleen automatisch scoorbare opgaven op te nemen.

[26] Dit is niet alleen door het gebruik van de computer bij het maken van werkstukken en andere geleverde teksten, maar ook in het dagelijks gebruik in communicatie (zo schijnen 9 op de 10 Vlaamse jongeren een actief Facebookaccount te hebben: `https://datanews.knack.be/ict/nieuws/10-vaststellingen-over-het-digitale-mediagebruik-van-vlaamse-jongeren/article-normal-297281.html`). Er moet wel opgelet worden dat dit geen additionele ongelijkheid oplevert voor de leerlingen die niet over veel computervaardigheden beschikken.

[27] Bij het onderdeel literatuur speelt dezelfde problematiek als bij lezen en schrijven, daar waar het verwoorden van leerlingen hun gedachten, gevoelens en beleving bij het lezen, beluisteren en bekijken van literaire teksten lastiger te vangen is met automatisch scoorbare vragen, maar ook de keuze van de literaire teksten zelf bij een centrale meting enige beperkingen oplegt. Er is in dit stuk vanuit gegaan dat deze eindterm niet centraal gemeten zou moeten worden, maar als dat wel het geval is dan wordt aangeraden dit op schoolniveau te organiseren.

verdere analyses[28]. De (eind)beoordelingen van de lastiger meetbare vaardigheden op schoolniveau kunnen afgezet worden tegen de resultaten op de onderdelen die centraal gemeten worden[29]. Een dergelijke aanpak kan ook gevoerd worden wanneer de kerntoets aangevuld wordt met materiaal van onderwijsverstrekkers. In de volgende paragraaf gaan we verder in op het combineren van net- of koepel-specifieke toetsen naast de centrale toetsen.

5.3 Kerntoets aangevuld met materiaal onderwijsverstrekkers

Een van de manieren om curriculumvernauwing tegen te gaan, is door de centrale toetsing aan te vullen met school-eigen toetsmateriaal. Er zijn drie hoofdscenario's voor de aanvulling van kerntoetsen met net- of koepel-specifiek toetsmateriaal: ten eerste door de net- of koepel-specifieke toetsen naast de kerntoetsen af te nemen (5.3.2); ten tweede door delen van de centrale toetsen door net- of koepel-specifieke toetsen te vervangen (5.3.2); en ten derde door delen van de net- of koepel-specifieke toetsen te integreren in de ontwikkeling van de centrale toetsen (5.3.3). In Paragraaf 5.3.4 worden de gevolgen van de drie op de verwerking van de toetsresultaten bij het gebruik van de verschillende scenario's behandeld. Voordat we ingaan op het aanvullen van kerntoetsen met additioneel materiaal van onderwijsverstrekkers zullen we eerst iets dieper ingaan op het meten van complexer te toetsen vaardigheden zoals schrijfvaardigheid of spreekvaardigheid. Dit zijn vaardigheden die vaak met niet-automatisch scoorbare opgaven gemeten worden en daarmee enkele uitdagingen kennen bij het opnemen in de centrale toetsing.

5.3.1 Net- of koepelspecifieke toetsen naast de centrale toetsen

Het scenario van net- of koepel-specifieke toetsen naast de centrale toetsen is aan te raden bij het meten van vaardigheden die niet makkelijk met gesloten of korte antwoordvragen kunnen worden geoperationaliseerd. Dit kan bijvoorbeeld gaan om spreekvaardigheid of schrijfvaardigheid, zoals hierboven toegelicht. Die laatste vaardigheid is genoemd als te meten vaardigheid bij de centrale toetsen, maar als het gaat om het meten van de gehele breedte van de vaardigheid, waaronder het (door de leerling) zelf schrijven van een stuk tekst, kan dat vooralsnog niet in een automatisch scoorbare vorm[30].

Zolang de identificatie van leerlingen en scholen op orde is, kunnen de metingen op scholen en de resultaten op de koepel-specifieke toetsen gecombineerd en vergeleken worden met de resultaten op de kerntoetsen. Het combineren van koepel-specifieke toetsen met de kerntoetsen heeft een aantal belangrijke voordelen.

[28] Als de afgenomen proeven door de school zelf ontwikkeld worden dan is het weergeven van het eindresultaat voldoende. Als gewerkt wordt met centraal ontwikkelde opdrachten is het handig om ook op het niveau van de deelopdrachten, of wellicht zelfs beoordelingspunten, scores in te voeren. Als dit handmatig moet gebeuren dan is dat wel veel werk. Uit psychometrisch, statistische overwegingen zijn scores op itemniveau te ambiëren, maar als dat veel tijd vraagt voor de leerkrachten, dan kan het de haalbaarheid in de weg staan. In hoeverre dat het geval is, zal afhangen hoe makkelijk die invoer werkt.

[29] In Nederland is er een systeem waarbij de inspectie van het onderwijs de resultaten van de schoolexamens afzet tegen die van de centraal schriftelijke examens. Wanneer een school systematisch (meerdere vakken, meerdere jaren) afwijkende scores op de schoolexamens oplevert, zal dit tot verder onderzoek leiden.

[30] Via kunstmatige intelligentie en *machine learning* technieken wordt wel gewerkt aan het automatisch beoordelen van teksten, maar deze zijn nog niet zodanig ontwikkeld dat deze menselijke beoordelaars kunnen vervangen voor het beoordelen van individuele leerlingwerken.

Een eerste voordeel is dat het kan helpen om curriculumvernauwing tegen te gaan. Als deze moeilijk automatisch scoorbare vaardigheden, die niet gemakkelijk centraal te meten zijn, alsnog een centrale rol krijgen binnen een centraal datacentrum, is het voor het publiek en voor de scholen ook een duidelijk signaal dat deze vaardigheden ook van belang zijn. Pas dan kan ook een volledig valide meting van de vaardigheden Nederlands en Wiskunde plaatsvinden, omdat de inhoud zo beter gedekt wordt.

Het tweede voordeel is dat de combinatie van koepel-specifieke toetsen en kerntoetsen de scholen kan helpen bij het inschatten van het niveau van de niet-centraal gemeten vaardigheden. Wanneer een terugkoppeling plaatsvindt op de metingen binnen de school, gerelateerd aan de centraal gemeten vaardigheden, is in te schatten of de beoordeling van de leerlingen bij de eigen metingen wellicht te hoog of te laag is, gegeven de resultaten op de centrale toetsen. Uiteraard is het niet noodzakelijk dat deze centrale meting en de eigen meting geheel identiek zijn –er worden immers enigszins verschillende zaken gemeten– maar de scholen moeten wel kunnen aangeven hoe deze verschillen tot stand komen. Hierbij kan het ook helpen als vanuit het steunpunt toetsontwikkeling richtlijnen of suggesties komen hoe de niet-centraal gemeten vaardigheden, zoals opstellen of spreekbeurten, te beoordelen zijn.

Een derde voordeel is dat het werken met koepel-specifieke toetsen recht doet aan de eigenheid van de school. Een school kan eigen accenten leggen daar waar zij zich binnen het curriculum extra op willen richten. Als dit enige afwijkingen oplevert in de scores op de centrale toetsen kan dat ook onderdeel van de discussie zijn met bijvoorbeeld de onderwijsinspectie. In de terugkoppeling over de kwaliteit van de school en het onderwijs kunnen zo ook deze toetsen een belangrijke rol krijgen. In Nederland wordt een dergelijke aanpak gebruikt rond de eindexamens. Daar worden zaken centraal gemeten waar dat mogelijk is, en wordt het aan de school overgelaten waar dat niet mogelijk is. De hierboven genoemde voordelen worden ook als zodanig ervaren.

5.3.2 Vervangen van centrale toetsendelen door net- of koepel-specifieke toetsdelen

Het integreren van beide toetsen, in de zin dat delen van de kerntoets vervangen worden door koepel-specifieke toetsen die geacht worden hetzelfde te meten, is complexer dan het hierboven genoemde scenario van net- of koepel-specifieke toetsen naast de centrale toetsen. Een vaardigheid als leesvaardigheid is goed meetbaar met een centrale toets. Het vervangen van een deel hiervan door een eigen toets lijkt geen meerwaarde te hebben. Uiteraard is het wel mogelijk dat de school ook een eigen leestoets heeft binnen een afnamejaar –een enkele meting voor leesvaardigheid zou ook mager zijn- maar deze zou dan naast de centrale toets bestaan. Bij vervanging van de centrale toets door een eigen toets is het voor het handhaven van de vergelijkbaarheid noodzakelijk dat een deel van de leerlingen zowel de specifieke als de volledige kerntoets maakt. Dit scenario is om praktische reden beter niet te volgen bij de invoering van het centrale toetsen in Vlaanderen.

Voor de toekomst is er een scenario voor te stellen waarin er een Vlaamse itembank gerealiseerd wordt. Dat zou een bank kunnen zijn waarbij de opgaven die gemaakt zijn in het kader van de centrale toetsen aangevuld worden met opgaven die vanuit de netten, koepels of scholen gemaakt worden. Daartoe moet er eerst een goede centrale bank gecreëerd worden, ervaring opgedaan worden met de eerste afnames van de centrale toetsen en moet de digitale infrastructuur, inclusief de bijbehorende verantwoordelijkheden,

duidelijk ingericht zijn. Het zal dus een aantal jaar[31] duren voordat een dergelijk systeem voor het vervangen van een deel van de centrale toetsen door koepel-specifieke toetsen ingericht is. De vorm van de inrichting zal ook afhangen van de uiteindelijke vorm van de centrale toetsen. Het is voor nu nog te vroeg daar specifieke scenario's voor op te stellen.

5.3.3 Integratie van net- of koepel-specifieke toetsdelen in de centrale toetsen

Een derde manier is het integreren van de net- of koepel-specifieke toetsen in de kerntoetsen, door koepel-specifieke delen in te zetten als kerntoets-deel. In dat geval is afstemming tussen de diverse koepels en de centrale organisatie noodzakelijk. De koepels kunnen bijvoorbeeld deel uitmaken maken van de constructiegroep van de centrale toetsen. Dit zou hen de gelegenheid geven de eigen toetsen meer te relateren aan de centrale toetsen. Het zou ook betekenen dat leerlingen uit verschillende netten opgaven kunnen krijgen die oorspronkelijk ontwikkeld waren voor toetsen uit andere netten. Bij dit scenario is het van belang dat er voldoende overeenstemming is tussen de ontwikkelaars van de verschillende net-eigen toetsen.

Het scenario dat er net-eigen varianten van de centrale toetsen bestaan, is voor te stellen, maar levert voor de vergelijkbaarheid van de toetsen en resultaten wel de nodige uitdagingen. In theorie is het mogelijk, maar er bestaat de kans dat er sprake is van een niet waar te nemen serieuze schending van unidimensionaliteit, omdat geen sprake is van uitwisseling van opgaven. Om dit scenario te realiseren, is meer onderzoek nodig, en dat vindt plaats via de afname van de centrale toetsen. Om vanaf de aanvang van de centrale toetsen te starten met net-eigen centrale toetsen is ook vanuit een PR-standpunt niet aan te raden, aangezien dit lijnrecht lijkt in te gaan tegen het doel van de centrale toetsen. Kortom, voor dit moment is net-eigen varianten van de centrale toetsen geen optie, maar valt dit na het opdoen van voldoende ervaring met de centrale toetsen mogelijk in de toekomst te overwegen.

Het betrekken van de ontwikkelaar van de net-eigen toetsen heeft zeker een ander mogelijk voordeel. Als zij ook eigen leestoetsen en rekentoetsen hebben, kunnen de gegevens van deze toetsen ook toegevoegd worden aan de itembank, als voorloper van de Vlaamse Centrale Itembank. Ook kunnen zij zo wellicht een rol spelen in het aan elkaar koppelen van de schalen door de afname in de tussenliggende jaren. Dat kan als de afnames van de net-eigen toetsen deze verbinding vormen, maar dan moet er wel een link met de centrale toetsen zijn, hetgeen mogelijk is als zij mede deelnemen aan de constructie van de centrale toetsen.

5.3.4 Gevolgen voor de verwerking van de toetsresultaten

Ieder van de drie mogelijkheden heeft eigen gevolgen voor de verwerking van de toetsresultaten. Wanneer net- of koepel-specifieke toetsen naast de centrale toetsen bestaan, moeten de resultaten op de net- of koepel-specifieke toetsen ook opgeslagen worden, om ze optimaal te kunnen gebruiken. De minimumvariant hiervan is dat voor iedere leerling

[31] Hoeveel jaar dat zal kosten is moeilijk van te voren in te schatten. Daar spelen een aantal factoren een rol. Het tempo waarin het mogelijk is om opgaven te maken, wat weer afhankelijk is van het soort opgaven dat gemaakt wordt of kan worden. Ook heeft het met prioritering te maken. Als meer middelen besteed moeten worden aan fraudebestrijding, anders dan via uitbreiding van de itembank, of andere zaken, is er ook minder budget om de itembank snel te vernieuwen en uit te breiden.

de eindscore op de aanvullende toets aan een centraal datapunt wordt doorgegeven. Dat betekent dat er een goede identificatie van de leerling moet zijn. Iets wat rond het centrale toetsen ook al een kernopdracht is, waar onder andere bij Perceel 3 gekeken wordt naar mogelijkheden.

Eveneens moet er een goede identificatie zijn van de door de leerling gemaakte aanvullende toets. Dat laatste is nodig om te weten welke leerlingen dezelfde aanvullende toets hebben gemaakt, en welke leerlingen juist een andere aanvullende toets hebben gemaakt. De informatie van de aanvullende toetsen moet worden aangevuld met informatie over welke eindterm(en) deze toets beoogt te meten, en bij voorkeur worden aangevuld met de mogelijke score-range, de toets zelf, en het bijbehorende beoordelingsmodel. De beheerder van deze data is bij voorkeur de beheerder die ook de data van de centrale toetsen onder zijn hoede heeft. Het is bij de net- of koepel-specifieke toetsen ook mogelijk om gegevens op opgaveniveau door te geven, op voorwaarde dat sprake is van een duidelijke identificatie van de opgaven. Dit levert echter zo veel additionele administratie op, zowel aan de kant van de school als aan de kant van het databeheer, dat dit in het aanvangsstadium niet aan te raden is.

Het vervangen van centrale toetsdelen door net- of koepel-specifieke toetsdelen op afzonderlijk netniveau, zoals beschreven als tweede optie, is om die reden ook af te raden. Als het gaat om de integratie van net- of koepel-specifieke toetsdelen in de centrale toetsen (optie drie) dan is het beheer gelijk aan dat van het opstellen van een centrale toets. De invloed van de netten speelt dan vooral in de ontwikkeling van de centrale toets maar niet in de afname. Als de net-eigen-toetsen ooit een rol gaan spelen in de tussenliggende jaren, dan is aan te raden hetzelfde afname- en verwerkingssysteem te gebruiken als de centrale toetsen, omdat dan zowel de vorm van de toetsen als de verwerking van de data het meest in lijn is met de centrale toetsen.

5.4 De relatie tussen toetstijd, nauwkeurigheid, en betrouwbaarheid

5.4.1 Betrouwbaarheid

De betrouwbaarheid van een test geeft een indicatie of er veel of weinig toevallige meetfout te verwachten is en kan dan ook gezien worden als een maat voor consistentie van een meting[32]. Betrouwbaarheid kan worden beschouwd als de hypothetisch verwachte samenhang van twee toetsscores gegeven dat de condities hetzelfde zijn. Op het moment dat een meetinstrument een hoge betrouwbaarheid heeft, verwachten we dat de samenhang van de scores bij een (hypothetische) herhaalde meting ook hoog zal zijn, gegeven dat de vaardigheid van leerlingen niet veranderd is. Omgekeerd geldt bij een lage betrouwbaarheid van een meetinstrument dat de samenhang tussen de twee scores ook laag is.

Uiteraard is het wenselijk dat een meetinstrument een hoge betrouwbaarheid heeft en daarom is de betrouwbaarheid van een toets ook een criterium dat binnen alle beoordelingssystemen van toetsen een rol speelt. Betrouwbaarheid komt als apart criterium voor binnen het COTAN beoordelingssysteem, het RCEC beoordelingssysteem, het EFPA beoordelingssysteem[33] en de Amerikaanse *Standards for Educational and Psychological*

[32]Traub, R.E. & Rowley, G.L. (1991). Undertanding Reliability. *Educational Measurement, Issues and Practice*, 10(1), 37-45.

[33]EFPA (European Federation of Psychologists Associations) heeft ook een international gebruikt beoorde-

Testing.

Als toetsen een rol spelen bij beslissingen op individueel niveau (zoals voortgangscontrole), zijn er richtlijnen gegeven over de vereiste betrouwbaarheid van een toets. Verschillende beoordelingssystemen[34] hanteren hiervoor vuistregels. Een betrouwbaarheid[35] van onder de 0,70 wordt bij beslissingen over individuen als onvoldoende beschouwd, vanaf 0,70 maar onder 0,80 als voldoende, van 0,80 tot 0,90 als goed en vanaf 0,90 als uitstekend. De COTAN stelt dat voor metingen voor belangrijke beslissingen de betrouwbaarheid minstens goed dient te zijn. De eisen aan de betrouwbaarheid van een toets hangen dus ook af van de implicaties van de toetsresultaten voor het individu.

Verschillende factoren zijn gerelateerd aan de betrouwbaarheid van de toets. Deze factoren zijn onder te verdelen in kenmerken gerelateerd aan de toets, de afnamecondities en de groep leerlingen die de toets afnemen. Het eerste kenmerk van de toets dat van invloed is, is de lengte van de toets. De relatie met de toetslengte en betrouwbaarheid is eenduidig: hoe meer opgaven, hoe betrouwbaarder de meting. Hierbij wordt er wel vanuit gegaan dat al deze opgaven één en hetzelfde construct meten. De voorspelde relatie tussen toetslengte en betrouwbaarheid is vastgelegd in de zogenaamde *Spearman–Brown prophecy formule*[36]. De Spearman–Brown prophecy formule is als volgt gedefinieerd:

$$\rho^*_{xx'} = \frac{K\rho_{xx'}}{K\rho_{xx'} + 1 - \rho_{xx'}} \tag{5.1}$$

De betrouwbaarheid van een bestaande toets wordt gepresenteerd door ρ en k is de verlengingsfactor van de toets. De verwachte of voorspelde betrouwbaarheid van de nieuwe toets wordt weergegeven door ρ.

De relatie tussen betrouwbaarheid en toetsverlenging is grafisch weergegeven in Figuur 5.2 voor toetsen met een verschillende initiële betrouwbaarheid. In deze figuur is te zien dat bijvoorbeeld een toets met een zeer lage betrouwbaarheid van 0,1 (de zwarte lijn) met een factor 20 verlengd moet worden om een betrouwbaarheid van ongeveer 0,6 te bereiken.

Het aantal opgaven per gemeten vaardigheid moet voldoende groot zijn, waarbij voldoende bepaald wordt door de functie van de toets en de bijhorende eerder genoemde richtlijnen. Als we vier vaardigheden meten -wiskunde en drie vaardigheden bij Nederlands (begrijpend lezen, schrijven, grammatica)- dan gaat het om een voldoende aantal opgaven bij deze vaardigheden. Willen we echter op een gedetailleerder niveau uitspraken doen over een leerling, dan moeten de metingen per vaardigheid ieder voor zich ook betrouwbaar zijn. Dat betekent dat de hoeveelheid te gebruiken opgaven toeneemt naarmate men meer gedetailleerde uitkomsten wil hebben. De wens om een betrouwbare meting te hebben en om gedetailleerd te meten–en dus veel opgaven te gebruiken– moet echter in balans zijn met de beschikbare toetstijd voor een leerling. We kunnen niet te veel tijd van de leerlingen vragen.

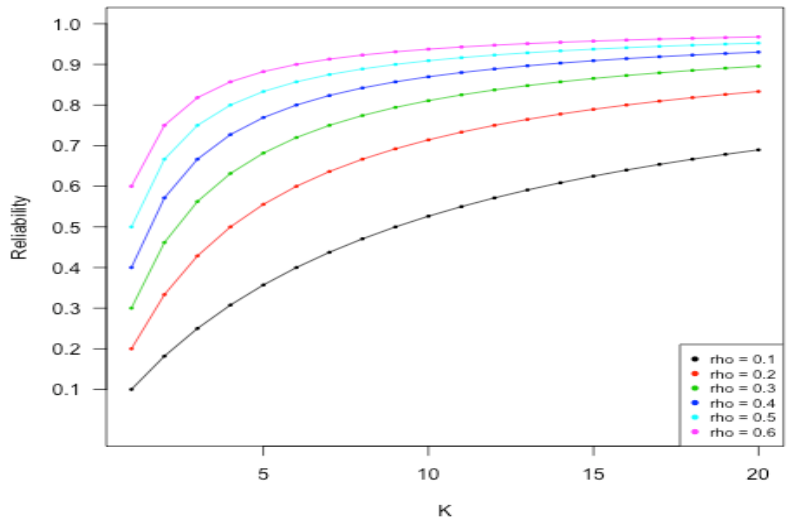
---

lingssysteem waar een Nederlandse vertaling van bestaat (`http://assessment.efpa.eu/download/9ccf47b3ac60e9bc9909c720e64e4195`).

[34]Hier worden de richtlijnen van het EFPA systeem en het COTAN systeem weergegeven.

[35]We hebben het hier over betrouwbaarheden over de gehele test op basis van inter-item-relaties zowel klassiek als met IRT bepaald (rho).

[36]Zie ook Stanley, J. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational Measurement. Second edition*. Washington, DC: American Council on Education.



Figuur 5.2: De relatie tussen betrouwbaarheid en toetsverlenging

Het verhogen van het aantal opgaven in een toets is niet de enige manier om de betrouwbaarheid te verhogen, er zijn ook andere manieren om dat te doen. De eerste manier betreft de kwaliteit van de opgaven. Hoe beter elk van de opgaven de te meten vaardigheid meet, hoe betrouwbaarder de toets. Bij heterogene vaardigheden –zoals grammatica, dat een aantal aspecten bevat– meet het ene item het ene aspect beter, en een andere opgave een ander aspect. Maar zolang ieder van de opgaven op zichzelf een goede meting is, zal de betrouwbaarheid toenemen. Het maken van goede opgaven is een kunst op zich. Het kan hierbij helpen om proeftoetsen uit te zetten met als doel om de beste opgaven te selecteren voor de uiteindelijke toets. Naast de toetslengte verlengen en de kwaliteit van het item materiaal verhogen, is een derde manier om de betrouwbaarheid te verhogen het vergroten van het onderscheid per opgave. Dat kan door opgaven op te nemen die niet alleen goed of fout gescoord worden, maar waarbij een verdere gradatie in scoring mogelijk is, zodat per opgave meer scorepunten te verdienen zijn. Met automatisch scoorbare opgaven behoort dit zeker tot de mogelijkheden.

Naast toetskenmerken spelen de afnamecondities een rol bij de hoogte van betrouwbaarheid van een toets. Een toetsafname die gestoord wordt door omgevingslawaai, zal meer behept zijn met meetfout. Duidelijke instructies voor het afnamen van een toets dragen bij aan betere afnamecondities.

De groep leerlingen ten slotte is de derde factor die invloed heeft op de hoogte van de betrouwbaarheid. Een toets kan betrouwbaarder zijn bij een grotere spreiding van de te meten vaardigheid in de populatie. Meer variatie in vaardigheid in de populatie levert voor de toets meer mogelijkheden op om deze waar te nemen, en dus om betrouwbaarder te meten. Dit is echter niet direct een kwaliteitskenmerk van de toets, maar van de te meten populatie. Het betekent wel dat bij het bepalen van de betrouwbaarheid van een toets het relevant is voor welke groep deze berekend wordt.

De betrouwbaarheid van een test is op verschillende manieren te bepalen, afhankelijk van welke bron van meetfout er verwacht wordt. Als het werken met (meerdere) beoorde-

laars mogelijk een bron van meetfouten is, is beoordelaarsbetrouwbaarheid van belang. Bij de geautomatiseerde beoordelingen die hier voorzien zijn, is dat geen thema. Daar waar een zekere mate van continuïteit verwacht wordt van de te meten eigenschap, zoals bij intelligentie, is test-hertest betrouwbaarheid van belang. Bij onderwijskundig meten is dat minder van belang. Het uitgangspunt daarbij is juist dat er sprake van ontwikkeling is en de gemeten vaardigheid dus toeneemt.

Het type meetfout dat bij onderwijskundig meten vooral aandacht krijgt, is meetfout die voortkomt uit de min of meer toevallige selectie van opgaven binnen een toets. De vraag die beantwoord moet worden is: zou de leerling dezelfde score of vaardigheidsschatting krijgen als hij of zij andere opgaven had gekregen? De maten die binnen de context van de centrale toetsen relevant zijn, zijn betrouwbaarheid op basis van inter-itemrelaties die vooral uit de klassieke testtheorie komen, en de methoden op basis van IRT. Bekende maten voor het eerste type betrouwbaarheid zijn Cronbach's alpha en Guttman's lambda-2[37]. De veronderstelde meetfout is bij iedere score even groot[38]. Hoewel het met de IRT methoden mogelijk is om ook voor een gehele toets de betrouwbaarheid te bepalen, wordt er binnen de IRT vanuit gegaan dat de mate van meetfout afhankelijk is van de positie van de leerling op de vaardigheidsschaal en de opgave of verzameling van opgaven die gemaakt zijn. Deze aanname is de belangrijke aanjager van het adaptief toetsen, waarover later meer.

Tot nu toe hebben we met name gesproken over de betrouwbaarheid van een enkele meting. Betrouwbaarheid van verschillen tussen twee metingen zijn over het algemeen lager dan die van de metingen op zich. De sterkte van die daling is voor een groot deel afhankelijk van de correlatie tussen de metingen. Om de leerwinst betrouwbaar te meten, is het dus een noodzakelijke voorwaarde dat de afzonderlijke metingen in ieder geval voldoende betrouwbaar zijn[39].

5.4.2 Nauwkeurigheid

Behalve de betrouwbaarheden kan ook gekeken worden naar de kwaliteit van de beslissing. Dat is vooral relevant als er sprake is van zak-slaag beslissingen. In een dergelijke situatie is het mogelijk de onterechte beslissingen te beschouwen, waarbij voor ten onrechte gezakte en ten onrechte geslaagde personen, indien ze een vergelijkbare (parallelle) toets hadden afgelegd, de beslissing anders geweest had kunnen zijn (dit worden ook wel niet-consistente beslissingen genoemd). Het percentage niet-consistente beslissingen is afhankelijk van de betrouwbaarheid en het percentage leerlingen dat zakt bij een specifieke cesuur. Tabel 5.1 laat het percentage niet-consistente beslissingen zien als functie van het percentage gezakte leerlingen en de betrouwbaarheid van een toets. Hieruit kunnen we afleiden dat hoe lager het percentage gezakte leerlingen, en hoe hoger de betrouwbaarheid, hoe lager het percentage niet-consistente beslissingen. Gebruik van deze tabel is overigens alleen

[37] Voor een toegankelijke uiteenzetting over de voor- en nadelen van de verschillende maten zie: Sijtsma, K. (2009): 'Over misverstanden rond Cronbachs alfa en de wenselijkheid van alternatieven'. *De Psycholoog*, Volume 44, p. 561 – 567.

[38] Bij de veronderstelde relatie tussen de meetfout per score (SEM) en de betrouwbaarheid (ρ) speelt ook de standaardafwijking van de scoreverdeling een rol (SD): $SEM = SD * \sqrt{(1-\rho)}$.

[39] Zie onder andere pagina 82 in Van den Brink, W.P. & Mellenbergh, G.J. (1998). *Testleer en toetsconstructie*. Boom Lemma Uitgevers, Amsterdam.

zinvol, wanneer de toetsscores ongeveer normaal verdeeld zijn[40].

	betrouwbaarheid						
% gezakt	**0,00**	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**1,00**
5	10	8	7	6	5	4	0
10	18	14	12	11	9	6	0
15	26	18	17	14	12	8	0
20	32	23	20	17	14	10	0
25	38	26	23	20	16	11	0
30	42	29	25	22	18	12	0
35	46	31	27	23	19	13	0
40	48	32	29	24	20	14	0
45	50	33	29	25	20	14	0
50	50	33	30	25	20	14	0

Tabel 5.1: Percentages niet-consistente beslissingen als functie van het percentage gezakten en de betrouwbaarheid

Wat nu een acceptabel percentage inconsistente beslissingen is, is lastig te bepalen. Vanuit de optiek van optimale eerlijkheid is dat uiteraard 0, alleen is dat in de praktijk niet haalbaar. Neem de leerling met een vaardigheid gelijk aan of zeer nabij de vaardigheid die nodig is om te slagen: dan is de kans dat deze persoon slaagt ook bij zeer hoge betrouwbaarheid aan kans onderhavig. Kortom, geheel zonder fouten kan niet. Maar hoeveel fouten acceptabel is, is een maatschappelijke vraag. Een belangrijke manier om deze vraag te benaderen is door de gevolgen van het niet halen van de cesuur te beperken. Dat betekent dat als er een foute beslissing genomen wordt de consequenties niet ernstig zijn. Dat kan door herkansingen in te voeren, maar ook –en dat lijkt in het geval van centrale toetsen in Vlaanderen niet onverstandig– de belangen voor de leerlingen niet te groot te maken.

Als we kijken naar de op schoolniveau benodigde betrouwbaarheid van een toets dan mag die lager zijn dan de betrouwbaarheid van een toets gebruikt wordt voor beslissingen op individueel niveau. Daar waar de betrouwbaarheid van de meting van het individu afhangt van het aantal items, hangt dat bij een school ook af van het aantal leerlingen. Als de meting van de individuele leerling niet optimaal betrouwbaar is –en dat is de betrouwbaarheid die klassiek dan wel met IRT bepaald wordt– dan is die meting op geaggregeerd schoolniveau een stuk nauwkeuriger. Dit valt intuïtief te begrijpen door de relatie tussen het aantal te verdelen scorepunten, en de betrouwbaarheid. Op individueel niveau is het aantal scorepunten alleen afhankelijk van de toets, terwijl dat op schoolniveau afhangt van het product van het aantal leerlingen en het aantal scorepunten van de toets. Daar geldt dus ook: hoe meer leerlingen, hoe betrouwbaarder de meting.

Het heeft een positief effect op de betrouwbaarheid (en de validiteit) van de meting, als meer uiteenlopende toetsversies, en dus meer verschillende opgaven, op de school worden afgenomen. Daar waar het op individueel niveau niet mogelijk is alle relevante opgaven uit de bank af te nemen, is dat op schoolniveau een optie. In dat geval maakt het al dan

[40]Tabel is afkomstig uit Veldhuijzen, N.H., Goldebeld, P., & Sanders, P.F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie. In T.J.H.M. Eggen & P.F. Sanders (Red.), *Psychometrie in de praktijk*, pp. 33-82.

niet opnemen van een enkele opgave in een toets ook niet uit, daar waar dat wel een rol speelt als slechts één versie van een toets wordt afgenomen op school.

Wat betreft de kwaliteit van de opgaven zijn er ook richtlijnen beschikbaar. Uiteraard zijn de inhoudelijke kwaliteiten van een item cruciaal, maar in het kader van de meetprecisie en de betrouwbaarheid zijn er ook psychometrische richtlijnen beschikbaar. Hier speelt de item-test-correlatie (de rit-waarde) een rol, oftewel de correlatie tussen de itemscore en de totale testcore. Veldhuijzen, Goldebeld en Sanders citeren de richtlijnen van Ebel en Frisbie waarbij een rit-waarde voor een item onder de 0,20 als slecht wordt gezien, vanaf 0,20 tot 0,30 als twijfelachtig, en 0,30 en groter als goed, dan wel zeer goed bij een waarde groter dan 0,40. Bij de COTAN richtlijnen wordt het onderscheid tussen goed en zeer goed niet gemaakt en wordt het 'twijfelachtig' van Ebel en Frisbie nog als voldoende gezien. Bij de evaluatie van de kwaliteit van de opgaven gaan we uit van een minimum rit-waarde van 0,20. De rit-waarde is overigens ook afhankelijk van de p-waarde van een opgave. Bij extreme p-waarden zal zelfs bij op zich goede opgaven de rit-waarde lager liggen. Het is echter vanuit psychometrisch perspectief in het algemeen niet aan te raden om opgaven met extreme p-waarden (zeer hoog of zeer laag) in een toets op te nemen.

Naast de betrouwbaarheid van een toets en de kwaliteit van opgaven, is ook de nauwkeurigheid van een meting van belang. Deze nauwkeurigheid hangt af van de variantie van de scores, en de betrouwbaarheid van een toets. Wanneer een toets vooral betrouwbaar is doordat de leerlingen sterk van elkaar verschillen, en niet zozeer door de kwaliteit van een toets, zal dat door de nauwkeurigheid van de toets in termen van de standaardmeetfout gesignaleerd worden. De nauwkeurigheid van een toets hangt in sterkere mate af van de kwaliteit van de opgaven, dan van de lengte van de toets.

Bij het beoordelen van de betrouwbaarheid in het kader van IRT wordt vooral gekeken naar de standaardfout van een meting en niet zozeer naar de betrouwbaarheid van de toets zelf. Zoals gezegd, is de betrouwbaarheid in de IRT niet voor iedere score gelijk. Voor iedere leerling valt zo een gedetailleerde voorspelling te maken over de meetnauwkeurigheid. Deze werkwijze is betekenisvoller dan een algemene betrouwbaarheidsmaat voor een gehele test. Het past ook in ons advies om in de aanpak van centrale toetsen IRT toe te passen. Vanuit de IRT aanpak is ook desgewenst gemakkelijk een klassieke schatting van de betrouwbaarheid te bepalen.

5.5 Veranderende adaptiviteit

Bij een adaptieve toets krijgen leerlingen items voorgelegd die aansluiten bij het niveau dat ze op een eerder onderdeel van de toets hebben laten zien. Adaptief toetsen heeft een aantal voordelen ten opzichte van het gebruik van lineaire toetsen waarin alle leerlingen dezelfde items maken. Ten eerste wordt een leerling niet of in ieder geval minder geconfronteerd met veel te moeilijke of veel te makkelijke opgaven. Dergelijke opgaven kunnen de kandidaat demotiveren en tot een minder zuivere meting leiden. Ten tweede kan een adaptieve toets tot een nauwkeurigere schatting leiden van de vaardigheid van de leerling. Dit voordeel wordt verder besproken in de onderstaande tekst box. Een derde voordeel van adaptieve toetsing is dat het automatisch tot het gebruik van verschillende toetsversies leidt, wat bijdraagt aan een veilige toetsafname en risico's van toetsfraude reduceert. Er bestaan twee belangrijke varianten van adaptief toetsen, computer adaptief toetsen (CAT) en multi-

stage-toetsen (MST). Deze beide varianten zullen worden besproken in de volgende twee secties.

Nauwkeurigheid en adaptief toetsen

Het doel van adaptief toetsen is het zo gericht mogelijk aanbieden van items aan leerlingen, waarbij gericht betekent dat de moeilijkheid van de items aansluit bij de vaardigheid van leerlingen. Juist deze items leveren het meeste informatie bij het vaststellen van de vaardigheid van leerlingen. Deze winst kan gebruikt worden voor het nauwkeuriger schatten van de vaardigheid van leerlingen vergeleken met het gebruik van een vergelijkbare set items in een lineaire toetsafname. Omgekeerd kan er ook voor gekozen worden om het aantal items in de toets te reduceren en dezelfde nauwkeurigheid van het schatten van de vaardigheid te behouden. Adaptieve toetsing maakt gebruik van twee kenmerken binnen IRT modellen, namelijk dat de vaardigheid van leerlingen makkelijk te vergelijken is, ook al maken ze verschillende opgaven, en dat de betrouwbaarheid afhangt van de locatie op de vaardigheidsschaal. Dat laatste betreft zowel de locatie van de leerling op de vaardigheidsschaal als de locatie van de opgave op de vaardigheidsschaal. Het principe is dat als de locatie van de leerling (vaardigheid) en de locatie van de opgave (moeilijkheid) aan elkaar gelijk zijn, de meetfout van de vaardigheidsschatting met die opgave minimaal is. Een opgave die niet te moeilijk en niet te makkelijk is, levert het meeste informatie op en dan meet je met deze opgave de vaardigheid het nauwkeurigst. Bij adaptief toetsen worden dergelijke opgaven aan de leerling voorgelegd. Open-antwoordvragen waarbij de leerling met een specifieke vaardigheid een kans heeft van ongeveer 0,5 om de opgave goed te beantwoorden zijn het meest informatief, en zijn vooral bij korte open-antwoordvragen zeer efficiënt. Bij meerkeuze-opgaven zijn iets gemakkelijkere opgaven, dus met een iets grotere kans om het item goed te beantwoorden, het meest informatief. Welk item het beste past, zal verschillen per kandidaat, omdat de kans op het goed beantwoorden van de opgave afhankelijk is van de vaardigheid van de leerling zelf.

5.5.1 Computer adaptief toetsen

Computer adaptief toetsen (CAT) is de meest verfijnde vorm van adaptief toetsen. Bij een CAT wordt na ieder afgenomen item de vaardigheid van de leerling opnieuw geschat en op basis van die schatting het meest geschikte volgende item gekozen[41]. De eerder genoemde voordelen van adaptief toetsen zijn bij uitstek van toepassing bij CAT. Onderzoek laat ook zien dat bij eenzelfde doelbetrouwbaarheid (of maximaal toegestane meetfout) een toets gemiddeld 25% a 30% korter kan zijn dan een lineaire toets[42]. In sommige literatuur wordt gesproken over een toets-verkorting van 40%[43], maar ervaring met realistische parameters,

[41] Weiss, D.J. (1982) Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.

[42] Verschoor, A., Eggen, T. (2014) Optimizing the test assembly and routing for multistage testing. In Yan, D., von Davier, A., Lewis, C. (eds) *Computerized Multistage Testing Theory and Applications*. CRC Press.

[43] Wainer, H., Dorans, N.J., Eignor, D. & Flaugher, R. (2000). *Computerized Adaptive Testing: a primer (2nd ed.)*. Lawrence Erlbaum.

ook in simulatiestudies, leert dat een dergelijk percentage niet gehaald wordt[44]. Een alternatief voor deze variatie in afnametijd is een CAT met een vaste lengte. Dan wordt de meerwaarde van de CAT vooral gehaald door de verbeterde betrouwbaarheid. Over het algemeen leidt adaptieve toetsing vooral tot een verbetering van de nauwkeurigheid indien de spreiding binnen de afnamepopulatie groot genoeg is. CAT heeft ook een aantal nadelen. Deze worden hieronder puntsgewijs besproken.

IT infrastructuur

CAT kan alleen met behulp van de computer plaatsvinden, omdat complexe berekeningen vereist zijn. CAT vergt vanwege achterliggende algoritmes veel rekenkracht, wat voor centrale toetsen uitdagingen zouden kunnen betekenen. Voor afname op computers op school moeten de computers snel genoeg zijn, en voor afname op een centrale server vereist dat dat deze ook een dergelijke rekenkracht aankan. In het verleden was het een belemmering als duizenden leerlingen tegelijkertijd de toets af moesten nemen, maar met de huidige rekenkracht en inrichting van servers geldt dat probleem steeds minder. Rekenkracht verschilt ook tussen varianten van CAT, waarbij bijvoorbeeld *shadow testing*[45] een variant is die veel rekenkracht vergt.

Omvang itembank

Voor het uitvoeren van een CAT is een voldoende grote itembank noodzakelijk. Een vuistregel gebaseerd op ervaringen is dat de bank voor een low-stakes test zes keer zo veel, en voor een high-stakes test tien keer zo veel items moet bevatten als dat de lineaire toets lang zou moeten zijn. Over het algemeen geldt in ieder geval dat hoe groter de bank, hoe beter CAT werkt. De bank moet ook zo gevuld zijn dat de toewijzing van de opgaven niet alleen van de moeilijkheid van de opgave afhangt, maar ook van de inhoud. De onderliggende toetsmatrijs moet ook binnen CAT gedekt zijn: een rekentoets kan niet alleen bestaan uit optelsommen of percentagesommen. Er zijn goede principes binnen CAT waarbij met inhoudelijke compartimenten van de bank gewerkt wordt om dit voor elkaar te krijgen, waarbij de opdeling van de bank de toetsmatrijs volgt. Als de matrijs vier deelvaardigheden veronderstelt, zal dat dus vier compartimenten van een itembank opleveren. Binnen ieder van deze compartimenten moet ook een voldoende groot aantal opgaven beschikbaar zijn met genoeg spreiding in moeilijkheid. Als er met dergelijke compartimenten gewerkt wordt, moet de itembank nóg groter zijn, of moeten er meer opgaven per leerling worden afgenomen. Dat is zeker het geval als er een relatie is tussen de moeilijkheid van de deelvaardigheid en de overkoepelende vaardigheid. Als rekenen moet bestaan uit optelsommen en percentagesommen, zal het voor leerlingen die niet heel goed zijn in rekenen lastiger zijn om makkelijke percentage-opgaven te vinden. Ook zal het lastiger zijn moeilijke optelsommen te vinden voor leerlingen die juist goed in rekenen zijn. Binnen het optelsommen-deel van de bank zullen de moeilijke opgaven snel uitgeput

[44] Dan kan het leerlingen wellicht 10 minuten toetstijd schelen per toets, maar dat voor behoorlijk hoge kosten om een dergelijk systeem op te zetten. Daarnaast wordt dat gemiddelde ook niet door iedereen gehaald. Er zullen ook leerlingen zijn die evenveel opgaven moeten maken als bij een lineaire toets. In de praktijk levert een dergelijke verkorting daardoor niet eens echt veel tijdswinst op omdat er toch een blok afgestemd was. Een deel van de leerlingen zal in de helft van de tijd klaar zijn, maar een deel ook niet. Dat levert mogelijk eerder onrust bij de afname op dan winst.

[45] Van der Linden, W.J. & Veldkamp, B.P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29(3): 273–291.

raken, en binnen het percentagesommen-deel juist de makkelijke. De ervaring leert dat dit betekent dat er extra opgaven nodig zijn, of dat leerlingen alsnog geconfronteerd worden met relatief (te) moeilijke of (te) makkelijke opgaven. Een andere oplossing is meer flexibiliteit toe te staan in de dekking van de toetsmatrijs, en dat er verschillen binnen redelijke bandbreedte mogen zijn tussen leerlingen voor wat betreft het aantal opgaven per deelvaardigheid. Of dat mogelijk is zal afhangen van de unidimensionaliteit van de hoofdvaardigheid over de deelvaardigheden heen. In ons voorbeeld: of de optelsommen en percentagesommen voldoende op een enkele rekenschaal liggen. Zoals het voorgaande al aangeeft heeft een te kleine itembank een aantal nadelen. De meerwaarde van CAT boven lineair toetsen is dan aanzienlijk minder groot. Een ander nadeel is dat de bank sneller uitgeput is: wanneer er weinig opgaven zijn, kan men sneller de route kennen naar de optimale score. Het is mogelijk om na te gaan of een item te vaak geselecteerd wordt door het algoritme (*exposure control*); het algoritme kan bij de toekenning van items daar rekening mee houden om dat iets tegen te gaan, maar na verloop van tijd leidt dit weer tot suboptimale toetsen en mogelijk zelfs minder informatieve toetsen dan lineaire toetsen omdat de items die gekozen mogen worden dan niet bij de leerling passen.

Steekproefomvang

Een uitdaging bij itembanken is dat de parameterschattingen van de opgaven zeer goed moeten zijn. Dat betekent dat we zeer zeker moeten weten dat de IRT-parameters in de itembank ook echt correct zijn. Volgens de COTAN voorschriften[46] zijn de richtlijnen dat er voor een 1-parameter logistisch model minstens 200 observaties moeten zijn, bij een 2- parameter logistisch model (met een parameter voor onderscheidend vermogen van de opgave erbij) er minimaal 400 observaties moeten zijn, en bij een 3- parameter logistisch model er 700 observaties nodig zijn . Eigen ervaringen betreffen vooral (een variant van) het 2- parameter logistisch model waarbij gevonden wordt dat voor lineaire toetsen deze aantallen observaties zeker kunnen voldoen, dat wil zeggen dat de schattingen van de toevoeging van de a-parameter voldoende is om met deze parameters te werken bij de equivalering van toetsen. Aangezien de moeilijkheidsparameter bij de selectie van opgaven in een CAT een zeer belangrijke rol speelt, is het van groot belang deze goed te schatten. Daarbij worden hogere aantallen observaties gehanteerd dan 400, met 700 tot 1.000 observaties per opgave. Een richtlijn die ook nog wel wordt toegepast is dat de standaardschattingsfout van de moeilijkheidsparameter onder een derde van de standaardafwijking hoort te liggen: se(β)<SD/3. Dit levert vaak ook wat hogere aantallen observaties op dan de genoemde 200 á 400 voor respectievelijk het 1- en het 2- parameter logistisch model. Met de aantallen leerlingen die deze centrale toetsen gaan maken zijn deze aantallen observaties goed te behalen, ook als een itembank gevuld moet worden met tien keer zo veel opgaven als er in een toets zitten. Het nadeel is echter dat deze gegevens bekend moeten zijn voor een eerste gebruik, en er dus heel veel voorwerk nodig is, met daarbij de nodige risico's, onder andere met het bekend worden van de opgaven. Er kan wellicht voor gekozen worden om pas nadat de centrale toets al een aantal jaar afgenomen is te werken met de ondertussen verkregen opgavenbank. Er moet dan wel

[46] De COTAN voorschriften baseren zich op Parshall, Davey, Spray en Kalohn (1998), maar de referentie naar de publicatie ontbreekt. Een goede referentie is hier Parshall. C.G, Spray, J.A., Kalohn, J.C. & Davey, T. (2002). *Practical Considerations in Computer-Based Testing*. ISBN : 978-0-387-98731-6.

sterk gelet worden op of er geen sprake is van een parametershift. Een andere werkwijze is een graduele transformatie van lineaire afnamen naar adaptieve afnamen tijdens het proces[47,48]. Het toepassen van de verschillende typen afnamen binnen een afnamejaar heeft praktische nadelen (een andere toets-ervaring binnen een afnamejaar) alsook nog psychometrische nadelen (verandering van parameters, en daardoor mogelijk andere vaardigheidsschattingen bij een zelfde antwoordpatroon).

5.5.2 Multi-stage-testing

Multi-stage-testing (MST)[49] is een vorm van adaptief toetsen die niet op item-niveau plaatsvindt, maar op een verzameling items. Zo'n verzameling items wordt een module genoemd. Leerlingen beginnen allen met dezelfde module (routing test genoemd) en op basis daarvan krijgen ze een moeilijkere of juist een makkelijkere module voorgelegd in het tweede deel van de toets. Dit wordt de tweede stage van de MST genoemd. Binnen een MST design kan gevarieerd worden met het aantal stages en het aantal verschillende modules dat binnen een stage wordt aangeboden. Uiteraard geldt dat hoe meer stages en verschillende modules worden ingezet, hoe meer adaptiviteit in het systeem wordt geïntroduceerd.

Een voordeel van MST ten opzichte van CAT is dat de itembank minder snel uitgeput raakt. Een tweede voordeel van deze vorm van adaptief toetsen is dat de toetsmatrijs beter beheerst kan worden omdat vooraf de varianten daarop ingesteld kunnen worden. Ten derde is de benodigde rekenkracht beperkter en is het makkelijker deze optimaal in te zetten. Ten vierde is het bij deze manier mogelijk om de items zonder excessief pretesten in te zetten. De schattingsprocedures zijn iets ingewikkelder dan bij lineair toetsen[50], maar aanzienlijk minder problematisch dan bij een directe schatting van persoonsparameters in een CAT. Ten vijfde is het ook makkelijker centraal te werken met toetsblokken op school. De winst van relevante opgaven voor de leerlingen wordt verkregen, met aanzienlijk lagere kosten dan bij een CAT. De efficiëntie van de simpelste vorm van MST – na de starttoets slecht één stap naar twee niveaus (een makkelijke en een moeilijke variant) – ten opzichte van een lineaire toets is al 12% tot 15%[51], wat bij meer niveaus en meer stappen oploopt naar hogere percentages waarmee de net iets grotere efficiëntie van een CAT ten opzichte van een MST bij grootschalige afnames beperkt is.

[47] Fink, A., Born, S., Spoden, C. & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, Volume 60, 2018 (3), 327-346.

[48] Verschoor, A., Berger, S., Moser, U. & Kleintjes, F. (2019) On-the-fly calibration in computerized adaptive testing. In Veldkamp, B., Sluijter, C (Eds.) *Theoretical and Practical Advances in Computer-based Educational Measurement*, Springer.

[49] Yan, D., von Davier, A., & Lewis, C. (2014). *Computerized Multistage Testing: Theory and Applications*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences; For practical applications see: Magis, D., Yan, D., & von Davier, A. (2017). *Computerized Adaptive and Multistage Testing with R*. Springer.

[50] Bechger, T., Koops, J., Zwitser, R., Partchev, I. & Maris, G. (2018). dexterMST: dexter for Multi-Stage Tests. `https://dexterities.netlify.app/2018/07/13/dextermst/`.

[51] Verschoor, A., Eggen, T. (2014) Optimizing the test assembly and routing for multistage testing. In Yan, D., von Davier, A., Lewis, C. (eds) *Computerized Multistage Testing Theory and Applications*. CRC Press.

5.6 Brede afname

Een brede afname geeft alle leerlingen de mogelijkheid om deel te kunnen nemen aan de toets, en hun vaardigheid of kennis aan te tonen. Als een aanpassing aan een reguliere variant wordt gedaan voor een leerling met beperkingen, dan biedt dit de leerling de kans om zijn vaardigheid te laten zien op hetzelfde vlak als een reguliere kandidaat. Dat wil zeggen dat belemmeringen worden weggenomen om de vaardigheid te tonen. Om een brede of inclusieve afname te kunnen realiseren, zijn er soms verschillende aanpassingen aan de toets nodig. De implicaties van aanpassingen, die nodig zijn voor het includeren van verschillende groepen, worden in deze sectie besproken. Het tweede gedeelte van deze sectie gaat in op het effect van deze veranderingen op de vergelijkbaarheid van de resultaten.

Merk op dat “inclusive assessment” in de literatuur twee benaderingen kent. Enerzijds is er een benadering die kijkt naar de behoeften van leerlingen met een beperking. Anderzijds is er een benadering waarbij gekeken wordt naar een eerlijke toetsing van leerlingen uit sociale achterstandssituaties of minderheden. In de eerste benadering betreffen de aanpassingen veelal de toetsvorm, in de laatste benadering meestal de inhoud, zoals het vermijden van overbodige taligheid van opgaven of van onbekende contexten.

5.6.1 Implicaties van aanpassingen

Aanpassingen aan de inhoud van de toets, zodat leerlingen uit minderheden of achterstandssituaties, geen onnodige hindernissen ervaren bij het maken van de toets, kunnen generiek worden doorgevoerd. Bij het schijven van de opgaven is het verstandig om van tevoren rekening te houden met mogelijke beperkingen, zeker als die aanpassingen relatief weinig moeite kosten en geen impact hebben op de validiteit van de opgaven. Denk hierbij aan relatief makkelijk taalgebruik –zeker bij de instructies rond de opgaven–, of kleurgebruik waarbij rekening gehouden wordt met kleurenblinden.

Het kan ook kostbaar zijn om aangepaste varianten voor leerlingen met beperkingen te produceren. Sommige aanpassingen (denk aan grootte van het lettertype, ondertiteling of verklanking) kunnen nog wel generiek worden meegenomen in de productie, in de zin dat alle leerlingen deze aanpassing kunnen gebruiken. Degenen die de ondersteuning niet nodig hebben, hebben er geen baat bij, maar ook geen last van. Andere aanpassingen zijn niet generiek mogelijk, of onwenselijk om generiek door te voeren.

Bij leerlingen met beperkingen kun je een onderscheid maken tussen fysieke beperkingen en mentale beperkingen. Bij fysieke beperkingen gaat het om beperkingen die niet gerelateerd zijn aan de vaardigheid die getoetst wordt. De aanpassingen zitten dan vaak in de condities, zoals andere presentatievormen of aangepaste temporisering. Daarnaast is een screening noodzakelijk van opgaven die onmogelijk te beantwoorden zijn voor deze kandidaten. Deze opgaven kunnen dan verwijderd worden uit de bijzondere variant, die dan iets wordt ingekort. Een dergelijke screening dient uitgevoerd te worden door experts die ervaring hebben met toetsing van het betreffende type beperkte leerlingen.

Bij mentale beperkingen[52], waarbij de beperking direct gerelateerd is aan de te toetsen vaardigheid, wordt meestal de moeilijkheid van de toets aangepast. Dit kunnen bijvoorbeeld

[52]Bij het gebruik van de Eindtoets was de wettelijke bepaling dat als kon worden aangetoond dat een leerling een IQ-score heeft onder de 70, bevestigd door een intelligentietest die twee jaar voor de eindtoets is afgenomen, de leerling de toets niet hoefde te maken. Als de leerling (de makkelijke versie van) de toets

leerlingen zijn die in het speciaal onderwijs zitten. Bij de Eindtoets Basisonderwijs heeft Cito jarenlang een variant gemaakt voor leerlingen die volgens de verwachting bij de 20% zwakst presterende kandidaten zouden horen. Dit zijn overigens niet allen leerlingen met mentale beperkingen.

Aan deze leerlingen is een gemakkelijkere versie van de toets voorgelegd, door gebruik te maken van de eenvoudigste vragen uit de reguliere toets, en deze aan te vullen met meer eenvoudige vragen. Dit heeft twee voordelen. Ten eerste zou een toets waarbij de meerderheid van de opgaven (veel) te moeilijk zou zijn, demotiverend zijn voor deze leerlingen. Door nu een toets te geven waar de leerling ook succeservaringen voelt, zal de leerling eerder proberen een zo goed mogelijke prestatie neer te zetten, in plaats van uit pure frustratie de toets op te geven. Een ander voordeel is dat als de opgaven op het niveau van de leerlingen is, de meting betrouwbaarder is. Zeker in het geval er meerkeuze opgaven gebruikt worden geven de te moeilijke opgaven geen goede weergave van de vaardigheid van deze leerlingen, en zal er sprake zijn van een relatief grote meetfout. Deze voordelen zijn niet alleen positief voor de meting van de individuele leerling, maar maakt ook de meting op schoolniveau beter. Door de aangebrachte overlap van opgaven tussen de speciale versie van de toets en de reguliere versie zijn de leerlingen die deze verschillende versies maken goed met elkaar te vergelijken en zijn ook uitspraken op geaggregeerd niveau mogelijk. Het maakt dan niet uit welke versie de leerling gemaakt heeft. Dergelijke aanpassingen in de afname zijn vergelijkbaar met adaptief toetsen waarbij de leerkracht bepaalt welke versie de leerling zou moeten krijgen. Als dit minder door de leerkracht gestuurd zou moeten worden, komen we uit bij een MST, welke geschikt is om toe te passen bij gemengde doelgroepen.

Bij beperkingen die direct samenhangen met de te meten vaardigheid is het dus aan te raden een makkelijkere variant van de toets aan te reiken, het inzetten van hulpmiddelen is vaak niet aan te raden. Als bijvoorbeeld leerlingen dusdanig kort in Vlaanderen wonen[53] dat hun vaardigheid van het Nederlands zeer laag is, is het af te raden om een hulpmiddel in te zetten dat deze leerlingen helpt een hogere score voor de vaardigheden Nederlands te verkrijgen. Die score zou dan een foutieve representatie van die vaardigheid zijn[54].

Als de beperkingen niet direct samenhangen met de te meten vaardigheid, bijvoorbeeld beperkingen op het gebied van de Nederlandse taal terwijl men de vaardigheid rekenen wil meten, zou er nog voor gekozen kunnen worden om bijvoorbeeld een woordenboek toe te staan, omdat het daarbij niet specifiek over de Nederlandse taal gaat. Het alternatief om de rekentoets in diverse andere talen uit te brengen zou kostbaar zijn, waarbij de equivalentie van die versies ook niet gegarandeerd is. Dit is zodoende af te raden. Beter is het om bij het taalgebruik bij de rekentaken uit te gaan van een basaal niveau, door relatief korte

wel maakt dan telde dit resultaat niet mee in de evaluatie van de school. Een andere wijze waarmee deze ontheffingsgrond aangetoond kon worden was wanneer gegevens uit het leerling- en onderwijsvolgsysteem bevestigen dat de ontwikkeling van de leerling zeer duidelijk achterloopt.

[53]Voor leerlingen in het laatste jaar van het basisonderwijs in Nederland was bij het maken van de Eindtoets een exclusiecriterium van leerlingen die minder dan vier jaar in Nederland woonden en onvoldoende Nederlands spraken. Deze leerlingen mochten de eindtoets maken, waarbij zij vaker de gemakkelijkere variant kregen, maar bij de evaluatie van de school hoefde de resultaten van deze leerlingen niet gebruikt te worden.

[54]Als het nu bijvoorbeeld om de vaardigheid “lezen” in het algemeen gaat, zou nog nagedacht kunnen worden of de toets in de taal gegeven kan worden waarin de leerling het meest vaardig is. Het is echter zo dat in de eindtermen het expliciet om de vaardigheid Nederlands lezen gaat. Het aanbieden van een anderstalige variant geeft daar geen goed beeld van.

zinnen te gebruiken en moeilijke woorden[55] te vermijden[56].

Afhankelijk van het aantal leerlingen met specifieke beperkingen, waarvoor speciale varianten ontwikkeld moeten worden, kan overwogen worden om de toetsen één op één mondeling af te nemen in plaats van afzonderlijke varianten te maken. Het maken van een aparte versie voor deze leerlingen is vaak zeer kostbaar, en ook hier is equivalentie[57] moeilijk aan te tonen[58]. Ook kunnen speciale toetsvarianten meerdere jaren meegaan en hoeven deze minder vaak ververst te worden als er maar weinig leerlingen zijn die er gebruik van maken.

5.6.2 Effecten op vergelijkbaarheid van de aanpassingen

Als het nodig is verschillende varianten van opgaven te gebruiken om een brede afname te realiseren, moet er onderzoek naar vraagpartijdigheid (DIF) worden uitgevoerd. De vraag is dan namelijk of de aangepaste opgave functioneert zoals de vergelijkbare reguliere opgave. Als dat niet het geval is, moet deze opgave als een geheel nieuwe opgave beschouwd worden. Het is dan wel nodig te onderzoeken of de aangepaste opgave nog wel hetzelfde meet. Dat betekent dus een controle op de aanname van unidimensionaliteit.

Als er opgaven verwijderd zijn voor beperkte leerlingen, kan de score op een verkorte variant via equivalering vertaald worden naar de reguliere toetsscore. IRT is hier heel geschikt voor. Ook het equivaleren van eenvoudigere varianten, waarin relatief makkelijke reguliere opgaven aangevuld zijn met eenvoudige vragen aan reguliere varianten kan met IRT. Hierbij speelt ook weer het risico dat er vraagpartijdigheid (DIF) optreed tussen de eenvoudige en reguliere variant.

In het geval er een toetsversie is waarbij alle items aangepast zijn, is er geen overlap tussen toetsen mogelijk. De toepassing van IRT-gerelateerde technieken worden dan wat lastiger[59]. Dat is bijvoorbeeld het geval bij een braille-variant van de toets, waarbij de toets ook niet in grote aantallen wordt afgenomen. De statische mogelijkheden zijn dan beperkt. Hier zal men vooral van aannames uit moeten gaan, aangaande de moeilijkheid van de toets (bijvoorbeeld de aanname dat de toets even moeilijk is), de vaardigheid van de groep (bijvoorbeeld: deze groep is even vaardig als de reguliere groep), of beiden.

Ook door middel van standaardbepalingsprocedures zouden de toetsen en aangepaste toetsen vergelijkbaar gemaakt kunnen worden. Wanneer de toetsen geheel verschillen dan kan voor beide toetsen een standaardbepalingsprocedure plaatsvinden[60]. De standaard die

[55] Het betreft hier moeilijke woorden die dan niet te maken hebben met de te meten vaardigheid. Lastige wiskunde termen die samenhangen met het beoogde vaardigheidsniveau kunnen uiteraard wel gebruikt worden.

[56] Achteraf kan de invloed van de taal bij rekenvaardigheden achterhaald worden met partijdigheidsonderzoek. Als gecorrigeerd voor andere verklarende vaardigheden (bijvoorbeeld aantal doublures, of opleiding van de ouders) aangetoond kan worden dat de thuistaal geen invloed heeft op het resultaat bij rekenen of wiskunde, dan draagt dat bij aan de evidentie dat de taal geen rol speelt bij het maken van de opgaven voor die vaardigheid.

[57] Meer over het aantonen van equivalentie in Sectie 3.2.

[58] Voor het statistisch aantonen van equivalentie zijn voldoende waarnemingen nodig. Een vuistregel is dat 100 leerlingen of minder in ieder geval te weinig zijn, en vanaf 400 leerlingen het goed mogelijk wordt. Over de aantallen daartussen kan verder gekeken worden naar de beoogde *power* om equivalentie aan te kunnen tonen.

[59] Dat geldt voor CML-schattingen maar ook voor MML-schattingen, omdat bij de laatste variant de aanname van een gelijke vaardigheid de populatie problematisch kan zijn.

[60] Cizek, G., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance*

gehanteerd wordt bij de procedure is dan gelijk in beide gevallen, waarmee de cesuurpunten op de twee toetsen vergelijkbaar zijn. De equivalentie van de andere scores dan het cesuurpunt is via aanvullende aannames, bijvoorbeeld een lineaire relatie, te bepalen.

Het kan zijn dat in plaats van een alternatieve toets de condities van de afname voor verschillende soorten leerlingen aangepast moeten worden. De aanname is dat bij een reguliere kandidaat het weghalen van deze belemmering geen effect zou hebben. Dit wordt dan een keuze tussen standaardisatie en eerlijke aanpassingen voor leerlingen. In die gevallen speelt de vraag wat de impact is van het loslaten van die standaardisatie[61]. Het loslaten van standaardisatie hoeft niet automatisch te betekenen dat de resultaten niet meer vergelijkbaar zijn. Empirisch kan onderzocht worden welk effect aanpassingen hebben op reguliere kandidaten, door reguliere kandidaten deels een reguliere variant en deels een aangepaste variant te laten maken. Daarnaast is het in het kader van vergelijkbaarheid lastiger als er geen resultaten zijn voor een subgroep van leerlingen[62], dan als er resultaten zijn die niet exact vergelijkbaar zijn, maar wel redelijk informatief.

Als hulpmiddelen ingezet worden, moeten deze wel aan een aantal voorwaarden voldoen. Zoals eerder al is aangegeven moeten deze hulpmiddelen niet direct samenhangen met de te meten vaardigheid. Dit heeft ook te maken met het punt dat bij een reguliere kandidaat het weghalen van deze belemmering geen effect zou hebben, oftewel het toevoegen van deze hulpmiddelen of aanpassingen zou niet tot een verbetering in prestaties moeten leiden. Het toevoegen van mogelijke braille afnamen, of bij kleurgebruik rekening houden met kleurenblinden zijn daar goede voorbeelden van: die zijn zinvol voor mensen met een belemmering, maar niet voor mensen zonder die belemmering. Wanneer het gaat over het voorlezen van de teksten bij een vaardigheid leesvaardigheid, of een alternatieve versie van een luistertest voor mensen met gehoorbelemmering, ligt dit wat complexer. Het zal waarschijnlijk niet zo zijn dat reguliere leerlingen baat hebben bij een aangepaste versie, maar de vraag is dan eerder of wel gemeten wordt wat er beoogd gemeten te worden. Wat hier acceptabel is, zal echt afhangen van de gekozen definitie en daaruit volgende operationalisatie van Nederlandse lees- dan wel luistervaardigheid. Daarmee kan bijvoorbeeld bepaald worden of er bij voorgelezen teksten wel of geen sprake is van een valide meting van leesvaardigheid[63]. Datzelfde geldt voor de definitie en de operationalisatie van rekenen/wiskunde en het mogelijke gebruik van rekenmachines voor leerlingen met dyscalculie[64]. Bij rekenvaardigheid kan het voorlezen van teksten echter

standards on tests. Thousand Oaks, CA: Sage Publications.

[61] In de opdracht is in paragraaf 2.1.2.6, van hoofdstuk IV. Technische Voorschriften, de vraag gesteld als (13): Welk effect heef het verminderen van de standaardisatie van de afname op de vergelijkbaarheid van de resultaten.

[62] Dat zou het geval zijn als er met exclusie gewerkt wordt.

[63] Of het ontbreken van verklankte leesteksten ten onrechte leerlingen met dyslexie benadelen bij een meting van leesvaardigheid, of dat dan juist bij een dergelijke meting correct wordt waargenomen dat deze leerlingen minder vaardig in lezen zijn, is een die hier niet opgelost gaat worden. Hierbij moet eerst een keuze gemaakt worden wat nu echt verstaan wordt onder “Nederlandse leesvaardigheid” en of daar bij verklanking spraken van kan zijn.

[64] Afhankelijk van die definitie is het al dan niet logisch om rekenmachines toe te staan als hulpmiddel. Dergelijke discussies blijken in de praktijk vaak lastiger dan wellicht van te voren te voorzien is, zoals ook in Nederland blijkt uit de heftige discussies tussen personen die het realistisch en het functioneel rekenen aanhangen (zie ook: Koninklijke Nederlandse Akademie van Wetenschappen (2008). Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering. Alkmaar: Bejo druk & print. ISBN 9789069846002;

wel makkelijker als oplossing voor een belemmering dienen.

Of het hulpmiddel ook echt een hulpmiddel is, zal ook sterk afhangen van de ervaring met het beoogde hulpmiddel. Als slechthorende leerlingen een kijk- en luistertoets krijgen waarbij een doventolk naast de spreker staat, maar waarbij de helft van de leerlingen geen ervaring heeft met een dergelijke tolk, dan neemt zo'n aanpassing natuurlijk de belemmering niet goed weg. Stel dat bij sommige rekenopgaven een in de afname opgenomen rekenmachine op het computerscherm gebruikt mag worden, maar dat leerlingen daar geen ervaring mee hebben, dan helpt dat niet. Daar waar bij reguliere afnamen onbekendheid met opgavenvormen een valide meting in de weg kan staan omdat de leerlingen verward raken door de afnamevorm, waardoor zij niet kunnen tonen wat zij kunnen, geldt dat ook voor de bekendheid met de hulpmiddelen.

5.7 Leereffecten in toetsontwikkeling

5.7.1 Psychometrische evaluatie en verbetering

Voor de continue verbetering van de toetsen is het belangrijk om het psychometrisch functioneren van items en toetsen te onderzoeken. Ook is een terugkoppeling en bespreking hiervan met de constructeurs noodzakelijk, zodat zij leren om betere items te maken.

Een onderdeel van de verbetering van de toetsen is de psychometrische terugkoppeling. Dat kan via klassieke indices als de p-waarde en de rit-waarden[65]. Dat zijn dan de geobserveerde waarden binnen de groep aan wie de opgaven is aangeboden. Als er gewerkt wordt met adaptieve toetsen, zou het opmerkelijk zijn als p-waarden dan niet liggen tussen 0,40 en 0,75. De rit-waarde kan ook aan de relatief lage kant liggen omdat binnen een groep de spreiding in vaardigheid niet zo groot is als in de populatie. Om die reden is het ook handig om zeker bij adaptieve toetsen ook de met behulp van IRT geschatte p- en rit-waarden weer te geven voor de gehele populatie. Desgewenst zijn ook IRT-modelparameters te rapporteren, maar die hebben pas betekenis als er enige bekendheid is met de vaardigheidsschaal. Het is daarbij goed mogelijk om door middel van automatisering duidelijke pasklare rapporten te produceren die itemschrijvers helpen hun opgaven goed te kunnen beschouwen.

5.7.2 Haalbaarheid en draagvlak

De toetsresultaten kunnen geëvalueerd en gevalideerd worden aan de hand van achtergrondkenmerken van leerlingen en scholen. Zowel het gebruik van de item- als de toetsresultaten zijn daarmee onderdeel van de plan-do-check-act-(PDCA-)cyclus, die tot een continue verbetering van de kwaliteit van de toetsen kan zorgen. Het toepassen van de PCDA-cyclus helpt zowel bij het doorvoeren van verandering en continue verbetering van de centrale toetsen als wel het verbeteren van hoe deze toetsen gemaakt worden.

De PDCA-cyclus begint met een fase die uitgaat van het gekozen scenario. In deze fase worden ook de haalbaarheid onderzocht en de middelen gepland, evenals de criteria voor

`https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/advies-rekenonderwijs-op-de-basisschool`).

[65] De rit-waarde was al geïntroduceerd. Een alternatieve maat is de rir-waarde, oftewel de item-rest-correlatie. Dat is de correlatie van de itemscore met de totaalscore waarbij de score op dat item zelf niet wordt meegenomen. In het geval van de rit-waarde is er sprake van een auto-correlatie omdat de score van het item zelf onderdeel is van de totaalscore. Dat is bij de rir-waarde niet meer het geval.

succes. In de actie-fase (Do) wordt het gekozen scenario uitgevoerd. Het is verstandig dit in kleine stapjes te doen door het scenario in de uitvoering op te breken in kleinere scenario's met ieder eigen achtereenvolgende of parallelle PDCA-cycli. In de controle-fase (Check) worden de (tussentijdse) uitkomsten afgezet tegen de succescriteria. Daar waar nodig moeten er aanpassingen plaatsvinden om tot een beter resultaat te komen. In de volgende fase (Act) worden de aanpassingen geïmplementeerd, waarna de plan-fase (Plan) weer toegepast wordt. In ieder geval zal er voor ieder jaar een nieuwe PCDA-cyclus uitgevoerd worden, maar zoals gezegd ook binnen een jaar diverse kleine cycli.

Het betekent in het geval van centrale toetsen in Vlaanderen dat de haalbaarheid toeneemt als er gestart wordt met een basisvariant van deze toetsen. In dit hoofdstuk is al enkele malen aangegeven wanneer iets toekomstmuziek is, dat later ingepast kan worden. Zo is wellicht aan te raden niet vanaf het eerste moment te starten met adaptieve afnamen, zeker niet met CAT maar waarschijnlijk ook niet met MST. Ook het toepassen van fraaie interactieve opgaven, en het gebruik van filmpjes, of het gebruik van open opgaven kan een stap te ver zijn, afhankelijk van de eerdere ervaringen. Het uitrollen van de meest basale vormen van centraal toetsen waarbij niets mis gaat, zal al een uitdaging zijn.

Het is wel verstandig om bij de ICT-component van het werk rekening te houden met zaken die verderop op de *roadmap* staan zodat als het nodig is daar alvast voorzieningen voor worden getroffen. Het uitvoeren ervan kan later in de tijd en kan zo ook bijgesteld worden op basis van de ervaringen bij de eerste afname. Bij de meest efficiënte uitwerking hiervan is het in ieder geval van belang dat de datastromen goed ingericht zijn. Dit is daarmee ook mede afhankelijk van de scenario's zoals deze beschreven zijn bij Perceel 3.

Het toepassen van de PDCA-cyclus heeft een aantal voordelen. Deze is ten eerste op verschillende manieren toe te passen in het ontwikkelproces. Heel veel stappen daarin kunnen beschreven worden als PDCA-cyclus. Het principe is intuïtief krachtig en eenvoudig te begrijpen. Het toepassen van de PDCA-cyclus maakt ook stappen en verwachtingen inzichtelijk voor de ontwikkelaars. Let wel, een PDCA-cyclus is geen doel op zich. Het toepassen ervan vereist de nodige inzet, maar mag er niet toe leiden dat activiteiten verzanden. Goed projectmanagement is nodig om ervoor te zorgen dat een in eerste instantie fraai plan vervolgens ook systematisch geëvalueerd wordt en dat de planning en opzet aangepast worden op basis van de evaluatie.

De in het bestek opgenomen vragen hebben betrekking op belangrijke onderwerpen die meer in detail ingaan op het hiervoor beschreven kader. De concrete beantwoording ervan vindt plaats in het kader van de uitwerking van de verschillende scenario's. Er is immers een sterke samenhang met de gekozen uitgangspunten, de overige onderdelen van de pedagogisch-psychometrische beschouwing en de elementen die aan bod komen bij de organisatorische en de technisch-juridische aspecten.

5.8 Toetsontwikkeling met papieren en digitale toetsvarianten

Het is de ambitie om alle centrale toetsen in Vlaanderen digitaal af te nemen. Dit biedt een aantal belangrijke voordelen. De grootste voordelen zijn te vinden in de schaalbaarheid in verdeling en verwerking van de toetsen. De verwerking van de resultaten zijn aanzienlijk gemakkelijker als er sprake is van geautomatiseerde scoring van de opgaven. Meerkeuze opgaven en kort-antwoordopgaven zijn hiervoor goed te gebruiken. Essay-opgaven waarbij

de leerlingen een langer antwoord moeten geven zijn lastiger in het geval van geautomatiseerde scoring. Als we een vaardigheid als schrijven –maar naar alle waarschijnlijkheid ook Wiskunde- op een hoger niveau zoals het einde van de derde graad willen meten dan kan het echter lastig zijn de beoogde eindtermen valide te meten met meerkeuze opgaven of kort-antwoordopgaven. Als het niet lukt om dat voor elkaar te krijgen moeten alternatieven verzonnen worden, hier wordt later dieper op ingegaan.

De digitale afname biedt ook de mogelijkheid om nieuw soort opgaven te ontwikkelen die op papier niet goed werken. Voorbeelden van dergelijke opgaven die in een digitale afname gebruikt kunnen worden, zijn zogenaamde *drag-and-drop*-opgaven, waarbij gesleept kan worden met objecten, of opgaven waar sprake is van interactie. Als echter bij de eerste afname geen zekerheid is dat alle scholen in alle leerjaren de toetsen digitaal af kunnen nemen, en er een noodoplossing kan zijn met papieren toetsen, is het verstandig om bij de eerste centrale toetsen niet vol op digitaal toetsen in te zetten. De nieuwe soort opgaven die met name geschikt zijn voor digitaal toetsen zijn daarnaast vaak opgaven die net iets meer van de techniek kunnen vergen. Het risico dat een dergelijke opgave bij de eerste afnames niet werkt op alle computers op alle deelnemende scholen maakt dat het verstandig is dit type opgaven te vermijden. Als er meer ervaring is met de digitale afnames geeft dit wel meer mogelijkheden om automatisch te scoren opgaven te maken voor het meten van vaardigheden die een hoger abstractieniveau behoeven (zoals de zogenaamde *higher-order thinking skills*).

Het zal echter een belangrijke vraag bij de haalbaarheidsstudie zijn of dit ook werkelijk mogelijk zal zijn in 2022-2023. De ervaring leert dat een grootschalige centrale afname waarbij tienduizenden leerlingen tegelijkertijd dezelfde toets digitaal moeten afnemen, eigen uitdagingen kent. Op alle scholen moet de infrastructuur dusdanig afgestemd zijn dat dit lukt, evenals dat, om tot een eerlijke vergelijking te komen, alle leerlingen enige ervaring met digitaal toetsen zouden moeten hebben. Om die reden kan het zeer zinvol zijn ook een scenario te ontwikkelen waarbij bij de centrale toetsing er –in ieder geval deels– op papieren varianten teruggevallen kan worden. Die papieren toets kan ingezet worden wanneer van te voren bekend is dat de school niet gereed is voor de digitale afname door een beperkte infrastructuur. Ook kan het voorkomen dat tijdens de afname er zich technische problemen voordoen waardoor een afname via de computer niet mogelijk is. Tot slot zou een papieren afname gebruikt kunnen worden bij leerlingen waarvan bekend is dat ze niet voldoende computervaardig zijn. Dat kan het geval zijn wanneer in hun huis geen computer aanwezig is, of in het geval van relatief jongere kinderen[66].

[66]De digitale generatie heeft meestal op jonge leeftijd al ervaring met computers, maar dat geldt niet voor iedere leerling, hetgeen ook door ongelijke kansen vanuit thuis kan komen. In het kader van kansengelijkheid moet ook met die leerlingen rekening gehouden worden.

6. Omgaan met de resultaten

Toetsscores en resultaten worden veelal gepresenteerd in rapportages. Deze rapportages zijn het middel waarmee toetsresultaten vertaald worden naar betekenisvolle acties[1]. Bij het beantwoorden van de vragen over hoe te rapporteren, is het van groot belang wat gerapporteerd moet worden, wie er ingelicht moet worden, en wat de ontvanger van de rapportage ermee moet doen, oftewel wat het doel van de rapportage is. Deze zaken moeten op elkaar afgestemd zijn omdat anders ongewenste neveneffecten op de loer liggen.

Bij de vraag 'wat' er gemeten wordt, gaat het niet alleen om de omschrijving van het construct (de toetsinhoud) en of deze in lijn is met de eindterm. De mate van detail waarop gerapporteerd wordt, is ook van belang. Is het bijvoorbeeld alleen gewenst de vier hoofdvaardigheden te meten (rekenen/wiskunde en Nederlands lezen, schrijven en grammatica), of zijn meer gedetailleerde uitspraken op onderliggende leerdoelen of domeinen gewenst? Daarnaast is het van belang of er gerapporteerd moet worden hóe goed een leerling in een bepaalde vaardigheid is, of dat het wellicht voldoende is te weten dát de leerling de eindterm behaald heeft. Tevens is het relevant om onderscheid te maken in rapportages waar de nadruk ligt op de vaardigheid op één moment in de tijd, versus rapportages waar de nadruk juist ligt op veranderingen in de tijd : bijvoorbeeld op de groei die een leerling heeft doorgemaakt.

Naast een keuze over 'wat' er gerapporteerd wordt, is het ook belangrijk over 'wie' de rapportage gaat: over een individuele leerling, over een klas, over alle leerlingen van één leerkracht, over alle leerlingen in een school of zelfs over het Vlaams onderwijsstelsel als geheel. Afhankelijk van het beoogde doel kan een verschillend aggregatieniveau gekozen worden om de toetsresultaten te presenteren.

Het 'wat' en het 'wie' worden in samenhang met elkaar gewogen. Het streven is om een rapportage te ontwikkelen waarbij afhankelijk van het doel van de toets, en het beoogde

[1] Hopster-den Otter, D., Wools, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2017). Formative use of test results: A user's perspective. *Studies in Educational Evaluation*, 52, 12-23.

gedrag van gebruikers van de rapportages, keuzes gemaakt worden met betrekking tot deze twee aspecten. Aangezien het belangrijk is dat gebruikers de rapportages kunnen inzetten en als relevant beschouwen, is het dan ook belangrijk om per doelgroep vast te stellen wat voor gebruikers het beoogde doel of de informatiebehoefte is.

Afhankelijk van het doel wordt bepaald welke informatie het best gepresenteerd kan worden. Als de leerling ingelicht wordt dan is andere informatie relevant dan als de leerkrachten of de scholen ingelicht worden. Dat geldt ook als de centrale toetsen dienen voor een systeemmeting, waarbij de centrale toetsen een deel van de functie van peilingsonderzoek hebben. Het is niet ongebruikelijk om in de praktijk verschillende vormen van rapportages te geven geschikt voor verschillende doelgroepen (leerling, ouders, leerkracht, school, bestuurder) en afgestemd op de informatiebehoefte.

Dit hoofdstuk start met een toelichting op het belang van relevante en begrijpelijke rapportages van toetssystemen. Vervolgens komen het wat, voor wie en het doel van rapportages aan bod. Daarnaast worden een aantal voorbeelden gegeven van mogelijke rapportagevormen. Het hoofdstuk sluit af met een aantal randvoorwaarden die vervuld moeten zijn opdat scholen en leerkrachten de resultaten op een goede manier kunnen gebruiken.

6.1 Het terugkoppelen van resultaten aan leerkrachten en scholen

Valide toetsscores zijn scores die geschikt zijn voor de beoogde interpretatie en het gebruik van een toets. Om te bepalen of toetsscores valide zijn, is het noodzakelijk dat goed gedefinieerd is wat de score betekent en hoe men deze score zou willen gebruiken: het doel van de toets. Het is dan ook van belang dat aangetoond wordt dat de toetsscores geschikt zijn voor de beoogde interpretatie en het beoogde gebruik[2]. Bij het beoogde gebruik speelt de terugrapportage een zeer belangrijke rol, want een juiste interpretatie van testresultaten is een noodzakelijke voorwaarde voor een adequaat gebruik[3]. Het ondersteunen van de gebruikers bij het interpreteren van de toetsresultaten is zodoende een belangrijk aspect van validiteit.

Het is dus duidelijk dat we de rapportage beschouwen als belangrijk element van de validiteit. Dit is ook in lijn met het geschetste kader van Hoofdstuk 2 waarbij duidelijk gemaakt is dat het doel van de toets en de vorm van de rapportage sterk aan elkaar gerelateerd zijn. Het doel van de toets en het daadwerkelijke gebruik worden met elkaar verbonden door de rapportagevorm: daar waar de rapportage het gevolg is van het doel, zal het de oorzaak van het uiteindelijk gebruik zijn. Dat betekent dat de vraag of de toets wordt gebruikt volgens het doel grotendeels zal afhangen van de rapportage. Bij het ontwikkelen van de rapportage moet zodoende "achteruit gedacht" worden: levert deze rapportage wel het beoogde doel op? En nog breder: past deze vorm van de rapportage bij de uitgangspunten van de toets?

Naast de gegevens die de ontwikkelaar van de terugrapportage wil overbrengen is het van belang om aan te sluiten bij de informatiebehoefte van de gebruiker. De rapportage

[2]American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

[3]Van der Kleij, F.M., Eggen, T.J.H.M, & Engelen, R.J.H (2014). Towards valid score reports in the Computer Program LOVS: A redesign study. *Studies in Educational Evaluation*, 43 (2014) 24–39.

kan gezien worden als vorm van communicatie: met een afzender, een boodschap en een ontvanger. De afzender van een toetsrapportage is meestal de toetsontwikkelaar die de resultaten wil presenteren. De boodschap heeft betrekking op de inhoud van de rapportage en het publiek bestaat uit mensen die de toetsresultaten willen gebruiken[4].

6.2 Wat wordt gerapporteerd?

Zoals eerder vermeld, is het belangrijk dat de rapportages op het juiste detailniveau resultaten van toetsen weergeven. Dit kan bijvoorbeeld door een totaalscore voor één van de hoofdvaardigheden te presenteren, of fijnmaziger, deelscores voor deelvaardigheden of domeinen. Afhankelijk van het toetsontwerp en het gekozen afnamedesign is het ook mogelijk om op eindtermniveau te rapporteren.

6.2.1 Betekenis van scores

In principe geldt dat een toetsscore alleen betekenis krijgt middels een normering. Middels een normering wordt een gerapporteerde vaardigheid gerelateerd aan verschillende elementen die interpretatie mogelijk maken. De vaardigheid kan geïnterpreteerd worden door:

a) Relatie met de inhoud
b) Relatie met een grenspunt of drempel
c) Relatie met een extern criterium
d) Relatie met andere leerlingen

Wanneer we het hebben over de relatie ten opzichte van de inhoud, dan hebben we het over de relatie tussen de vaardigheid en de opgaven. Het gaat er dan niet alleen om hoeveel opgaven een leerling goed maakt maar ook welke. Zoals bij de uiteenzetting van de IRT al genoemd is, zijn de opgaven op de vaardigheidsschaal te plaatsen: als de vaardigheid van de leerling bekend is, dan is ook bekend welke type opgaven gemakkelijk voor de leerling zijn, welke opgaven veel te moeilijk, maar ook welke opgaven precies bij de vaardigheid passen. Die opgaven representeren die inhoud en dat deel van het curriculum waar de leerling of de klas aan toe is. Op basis van deze resultaten kan vervolgens een rapportage ontworpen worden die richtlijnen of suggesties geeft over wat de leerling nodig heeft. Dit kan toegepast worden op individuele leerlingen voor individuele leerpaden of op (delen van) klassen als blijkt dat meerdere leerlingen op een bepaald niveau zitten.

Als de relatie met een grenspunt of standaard op de vaardigheidsschaal gelegd wordt, dan kan bepaald worden of een leerling een vaardigheid heeft die boven dat grenspunt ligt. Deze standaard correspondeert bijvoorbeeld met een vastgesteld niveau voor beheersing van een vaardigheid. Wanneer een standaard vastgesteld is, wordt deze vertaald naar een cesuur: een score die gehanteerd wordt om onderscheid te maken tussen beheerst/niet beheerst of wel/niet geslaagd. De standaard wordt vastgelegd middels een systematische procedure: een standaardbepaling. Bekende voorbeelden van deze procedures zijn de (aangepaste) Angoff en de bookmark-methode[5]. Er is ook een methode van standaardbepalen

[4]Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S.M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710).Mahwah, NJ: Lawrence Erlbaum.

[5]Zie bijvoorbeeld: *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*

ontwikkeld die zeer goed werkt op het bepalen van de standaard op IRT schalen: de *data-driven direct consensus-*, oftewel de 3DC-methode[6]. Een Nederlands praktijkvoorbeeld van een vaste inhoudelijke standaard die gebruikt wordt in verschillende toetsen en examens zijn de Referentieniveaus Taal en Rekenen[7,8].

Referentieniveaus Taal en Rekenen

In Nederland is in de wet vastgelegd dat alle leerlingen op verschillende momenten in hun schoolloopbaan laten zien of zij voldoen aan de 'referentieniveaus'. Deze niveaus zijn voor Nederlandse taal en Rekenen inhoudelijk beschreven. In de periode 2011 – 2014 is voor deze inhoudelijke niveaus een standaard vastgesteld waarmee ongeacht het toetsinstrument, dezelfde 'lat' kan worden gehanteerd voor iedere leerling.

Voor Nederlandse Taal zijn 3 niveaus beschreven: 1F, 2F en 3F (F staat voor fundamenteel). Voor Rekenen zijn 4 niveaus bepaald: 1F, 1S (streefniveau), 2F en 3F. Dit zijn in beide gevallen oplopende niveaus waarbij 1F het laatste niveau is en 3F het hoogste. Het 2F niveau wordt zowel gebruikt aan het eind van het basisonderwijs (11 jaar), als aan het eind van het voortgezet onderwijs (16 jaar) én aan het eind van het middelbaar beroepsonderwijs (18 jaar en ouder).

Om de niveaus vast te leggen is per niveau een expertpanel samengesteld met daarin 8 tot 10 leden. Deze leden waren evenredig verdeeld over de verschillende relevante onderwijssectoren, waren afkomstig uit de wetenschap, uit de praktijk en waren alleen inhoudelijk expert. De expertpanels werden ingezet voor twee activiteiten: het vaststellen van een 'referentieset' en het vaststellen van een inhoudelijke standaard.

Een referentieset bevat een groot aantal opgaven (in dit voorbeeld 130) die gezamenlijk een goede representatie zijn van het inhoudelijke domein. De opgaven zijn afkomstig uit de verschillende onderwijssectoren: basisonderwijs, voortgezet onderwijs én beroepsonderwijs. En dekken gezamenlijk alle inhoudelijke aspecten van het domein af. Het expertpanel had een rol in de selectie en het vaststellen van deze referentieset.

uit 2007 van G.J. Cizek en M.J. Bunch.

[6] `https://www.cito.nl/kennis-en-innovatie/tools-voor-toetsontwikkelaars/tools-voor-toetsconstructie/3dc`

[7] `https://www.rijksoverheid.nl/onderwerpen/taal-en-rekenen/referentiekader-taal-en-rekenen`

[8] `http://www.toetsspecials.nl/html/referentiesets/default.shtm`

Referentieniveaus Taal en Rekenen vervolg

Vervolgens zijn alle opgaven eenmalig afgenomen in alle onderwijssectoren. De data die hiermee beschikbaar kwamen zijn onder andere gebruikt in de standaardbepalingsprocedure. Voor de standaardbepalingsprocedure is het expertpanel aangevuld met 10 leraren uit verschillende sectoren. Het panel heeft vervolgens via een extended Angoff (rekenen) en de 3DC methode (taal) bepaald wat inhoudelijk het niveau moet zijn wat een leerling die een bepaald referentieniveau heeft moet halen. In de volgende fase is op basis van de verzamelde data gepresenteerd wat de impact van de gezette standaard was: gegeven deze standaard, hoeveel % van de leerlingen behaalt dan op dit moment het niveau? Dit werd gebruikt om de standaard eventueel nog iets bij te stellen.

In de praktijk worden de referentiesets gebruikt om in verschillende lopende toets- en examenprogramma's te bepalen of leerlingen een bepaald referentieniveau behaald hebben. Hiervoor is bij elk programma de standaard eenmalig overgebracht en dit wordt in iedere toets omgezet naar een cesuur: hoeveel opgaven moet ik goed maken om niveau 2F te behalen. Belangrijk is dat voor elke toets, in elk onderwijsniveau, dit niveau altijd te vertalen is naar de oorspronkelijke nationale standaard die gezet is via de referentiesets.

Een andere manier om betekenis aan een vaardigheidsschaal te geven is door de punten op die schaal te relateren aan de relatie met een extern criterium. Een voorbeeld van een dergelijk extern criterium is het succesvol afronden van een opleiding in het aso, tso, kso of bso. Hiervoor moet de schoolcarrière van leerlingen met doorstroomonderzoek door de tijd gevolgd worden. Als de vaardigheidsverdelingen van de onderscheiden groepen voldoende uit elkaar liggen, kan bepaald worden wat de vaardigheidsgrens is die het beste onderscheid maakt tussen de verschillende niveaus. Dit kan een bruikbare vorm zijn in een terugrapportage. Zo kan bijvoorbeeld worden gerapporteerd wat de kans is van een bepaalde leerling om een specifieke vervolgopleiding succesvol af te ronden. Het is evident dat een dergelijk systeem tijd kost. Leerlingen moeten immers meerdere jaren gevolgd worden. Het is ook een systeem dat door de jaren heen onderhouden moet worden, omdat er ook acties volgen naar aanleiding van de voorspellingen op basis van de modellering, die impact hebben op de voorspellende waarde van het model zelf.

Het is ook mogelijk dat het criterium van de normering niet in de toekomst ligt, maar op het moment van de afname zelf. Dat zou bijvoorbeeld de inschatting van de leerkracht over de leerling kunnen zijn met een driedeling naar vaardigheidsniveau: onder verwacht niveau, op of rond verwacht niveau, (duidelijk) boven verwacht niveau. De instructies moeten duidelijk zijn dat de leerkracht niet noodzakelijk drie even grote groepen hoeft te maken. Als deze gegevens over heel Vlaanderen verzameld worden en de vaardigheidsverdelingen van de drie groepen met elkaar vergeleken worden, ontstaan er grenspunten tussen deze drie groepen op de vaardigheidsschaal, gebaseerd op alle afnamen[9] die een in Vlaanderen gedeelde betekenis van de waarden op de schaal opleveren.

[9]Uiteraard is dit ook mogelijk met een steekproefopzet, mits er voldoende scholen en leerkrachten geselecteerd worden die hun mening geven. In dit geval zou het zeker bij een eerste versie minstens 400 leerkrachten per leerjaar en niveau moeten betreffen. Later zouden wellicht bij updates van deze grenzen lagere aantallen mogelijk zijn.

Vaardigheidsscores kunnen ook betekenis krijgen als de leerlingen onderling met elkaar vergeleken worden. De vaardigheidsverdeling van alle leerlingen binnen een leerjaar kan in verschillende groepen verdeeld worden op basis van de resultaten. In veruit de meeste gevallen gebeurt dit op basis van de percentielverdeling. De percentielscore geeft weer hoeveel procent van de leerlingen in dat leerjaar een gelijke of lagere vaardigheid hebben. Een percentielscore van 10 betekent dat 10% van de leerlingen eenzelfde of lagere vaardigheid heeft. Dat betekent dat 90% van de leerlingen een hogere vaardigheid laat zien. Logischerwijs betekent een percentielscore van 90 dus dat 90% vergelijkbaar of slechter presteert en 10% beter. Het is mogelijk van de percentielindeling af te wijken en de percentielscores te clusteren in betekenisvolle groepen zodat er niet 100 groepen zijn, maar bijvoorbeeld vijf betekenisvolle groepen[10]. De waarden op de vaardigheidsschaal die deze grenspunten tussen de vaardigheidsgroepen markeren, worden daarmee meer betekenisvol (sterk benedengemiddeld, onder gemiddeld, gemiddeld, bovengemiddeld, sterk bovengemiddeld).

6.2.2 Vergelijkbaarheid

Als we relatieve verdelingen maken, is de vraag “wie meten we nu?” (wat is de toetsdoelgroep?) zeer relevant want deze vraag bepaalt het referentiekader voor de vergelijking. Zeker wanneer deze referentie gebruikt wordt in de rapportage is het van groot belang dat duidelijk is wie de vergelijkingsgroep is. Bij leerlingen is nog te verantwoorden om het gehele leerjaar als referentiekader te nemen, om leerlingen onderling te vergelijken. Het nadeel van deze rapportagevorm is dat er geen directe link is met de inhoud van de toets en relatieve normering zich daarmee minder leent om aan te geven wat leerlingen nodig hebben om zichzelf te ontwikkelen.

Omdat een relatieve normering weinig handvatten biedt om het onderwijs voor een leerling in te richten, is het sterk geassocieerd met summatief gebruik van de rapportage. De rapportage vertelt dat er meer of minder leerlingen zijn die hoger scoren op een vaardigheid, maar dat zegt niet wat er gedaan moet worden om dat te verhogen. Een uitzondering hierop vormt de profielanalyse. Hierbij wordt de leerling vooral met zichzelf vergeleken. Als de leerling op een vaardigheid of over vaardigheden heen gemiddeld percentiel X haalt dan is op basis van de profielanalyse te bepalen op welke (sub)vaardigheden de leerling relatief goed presteert (meer dan percentiel X) of wat minder (lager dan percentiel X). De leerkracht kan er dan voor kiezen om met de leerling te werken aan de (sub)vaardigheid die gezien zijn of haar algemene vaardigheidsniveau de meeste aandacht verdient. Het aardige hiervan is dat dit ook leerlingen die goed presteren helpt om nog beter te presteren.

[10]Binnen het Nederlands leerlingvolgsysteem is een ABCDE-opdeling gebruikt waarbij de A-groep de 25% best presterende leerlingen betrof, de B-groep percentiel 51 tot en met 75, de C-groep percentiel 26 tot en met 50. De % laagste presterende leerlingen werd in twee groepen verdeeld waarbij de E-groep de 10% minst presterende leerlingen betrof en de D groep percentiel 11 tot en met 25. Omdat de C-groep, doordat deze in het midden lag vaak als de “gemiddelde” groep gezien werd – met name door personen die niet vaker met dit systeem werkten, zoals ouders die terugrapportage terug kregen–, terwijl deze leerlingen benedengemiddeld presteerden, is voor een nieuwe opdeling gekozen van vijf groepen, ieder met een breedte van 20 percentiel punten. Dat is de I-V verdeling, met de I de 20% best presterende leerlingen, II de 20% die daar net onder zit, III de groep rond het gemiddelde (percentiel 40-60), IV de 20% daar net onder, en V de 20% slechtst presterende leerlingen. Binnen de groep V is ook de categorie V* die de 10% slechtst presterende leerlingen weergeeft, en binnen I de groep I* voor de 10% best presterende leerlingen.

Over relatief normeren

Relatieve normeringen zijn zeer populair, en vrijwel iedere commercieel uitgebrachte test of toets heeft een dergelijke normering, of heeft een normering die afgeleid is van een relatieve normering, zoals de IQ-score. Ook bij toetsen in het onderwijs is de relatieve normering zeer vaak toegepast. Er is hier een aantal redenen voor aan te wijzen.

Een eerste reden is dat veel gebruikers het prettig vinden een resultaat te duiden door het te vergelijken met anderen. Of een bepaald aantal goede antwoorden goed of slecht is, is voor een leerkracht vaak moeilijk te bepalen, maar als deze ziet dat relatief veel mensen een hogere score, dan wel een lagere score hebben, dan geeft dat een leerkracht vaak meer houvast. Veel onderwijsprofessionals vragen ook om relatieve normeringen omdat ze gewend zijn ermee te werken. Ze willen simpelweg weten hoe goed of slecht hun leerlingen het doen in vergelijking met een vergelijkbare groep. De enige manier waarop ze daar achter kunnen komen is via een dergelijke landelijke vergelijking (via een steekproef, of zelfs op populatieniveau, mogelijk opgedeeld naar subpopulaties). Daar maken veel onderwijsprofessionals graag gebruik van.

Een tweede reden voor de populariteit is dat een dergelijke vorm van normering relatief makkelijk, en met een relatieve grote mate van consensus te bepalen is. Het is van belang dat er sprake is van een representatieve steekproef van voldoende grootte , maar in het geval van onderzoek naar het functioneren van de test of toets is het verzamelen van dergelijke gegevens toch al noodzakelijk om de kwaliteit aan te tonen. Het geven van een relatieve normering betreft dan alleen het toepassen van simpele analyses op gegevens die toch al verzameld zijn. Doordat het voor toetsontwikkelaars relatief makkelijk is om een dergelijke normering te geven wordt deze ook vaak gegeven, waardoor toets-gebruikers deze vorm van normering ook goed kennen, en ook vaak verwachten.

Er is echter ook een aantal belangrijke nadelen aan deze vorm van normeren. Zoals al in de lopende tekst is aangegeven, is het van belang met wie de leerling vergeleken wordt. Afhankelijk van de vergelijkingsgroep kan een prestatie dus als relatief (heel) goed of (heel) slecht geïnterpreteerd worden, terwijl als de vergelijking met een andere groep plaatsvindt de interpretatie duidelijk anders kan zijn. Het is hierbij niet evident, of van tevoren vastgelegd wat de juiste vergelijkingsgroep is, maar de keuze bepaalt wel de waardering van het behaalde resultaat.

Een ander nadeel is dat, ook als wel heel duidelijk van tevoren bepaald is wat de referentiegroep is, dat de bruikbaarheid van deze rapportage beperkt is. Wanneer bijvoorbeeld bekend is dat de prestatie van een leerling ten opzichte van alle leerlingen in hetzelfde leerjaar in hetzelfde afnamejaar lager of juist hoger dan gemiddeld is, dan geeft dat niet aan of het een voldoende prestatie is. Veel meer dan de observatie dat het (veel) beter of slechter kan dan wat er gepresteerd was, levert dat niet op. Een bijkomend nadeel bij een dergelijke normering is dat ook als de leerling hard werkt en vaardiger wordt door de tijd heen, de leerling nog steeds als benedengemiddeld gekenmerkt kan worden. Dit kan zeer demotiverend werken.

Over relatief normeren vervolg

Een ander belangrijk nadelig gevolg van deze vorm van normering is dat bij relatief minder prestaties, de normering kan leiden tot stigmatisering. De leerling wordt als “minder dan de rest” gezien, door anderen en in het ongunstigst geval ook door zichzelf. Deze vorm van rapportage garandeert dat er -per definitie- een groep is voor wie geldt dat zij als relatief zeer slecht presterend uit de bus komt (zeg de 10% slechtst presterende leerlingen). Er is dan dus per definitie een groep kinderen die dan als belangrijke feedback van hun toetsprestatie terugkrijgt dat zij als “minder dan de rest van Vlaanderen” gekwalificeerd worden. Dit is pedagogisch een twijfelachtige praktijk. Zeker als na hard werken de leerling vaardiger wordt, maar nog steeds benedengemiddeld presteert, dan kan de leerling bij een dergelijke vorm van rapportage ook het gevoel krijgen dat de negatieve kwalificatie ook niet te veranderen is. Dit zou een goede reden zijn, om zeker in het basisonderwijs, een dergelijke vorm van terugrapportage naar leerlingen achterwege te laten.

Op individueel niveau kan de vaardigheid van een leerling op enig moment ook vergeleken worden met zijn of haar vaardigheid op een eerder moment. In *Toetsen op school*[11] worden de volgende functies toegekend aan een dergelijke rapportage:

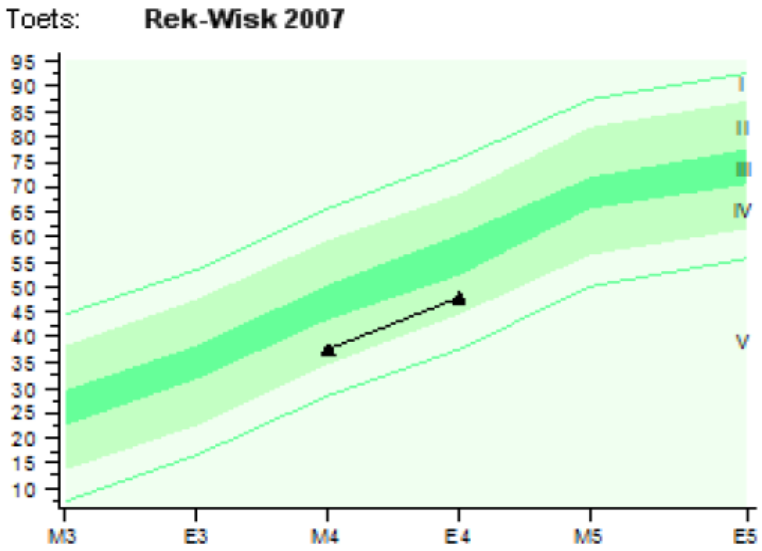
1. De ontwikkeling van de leerling ten opzichte van het voorgaande toetsmoment of de voorgaande toetsmomenten.
2. De ontwikkeling van een leerling in vergelijking met leeftijds- of groepsgenoten.
3. De voortgang van individuele leerlingen ten opzichte van de (tussen)doelen die men wil realiseren.

Bij de eerste functie gaat het over de vergelijking van de leerling met zichzelf, namelijk of de mate van vooruitgang van de leerling in de afgelopen periode te vergelijken is met de mate van vooruitgang in voorafgaande periodes, en of de ontwikkeling van de leerling nog steeds een opgaande lijn laat zien of dat een leerling stagneert in zijn of haar ontwikkeling op dat punt en weinig vorderingen vertoont. De tweede functie richt zich op de vergelijkbaarheid van de vorderingen in vergelijking met die van andere leerlingen uit het land over een langere periode en of een eventuele achterstand ten opzichte van andere leerlingen steeds groter wordt of dat de leerling deze juist inloopt. De derde en laatste functie maakt een vergelijking met de eindtermen en gaat in op de vraag of de leerling, als die zich in het getoonde tempo blijft ontwikkelen, de gestelde tussendoelen of einddoelen zal halen.

In principe gaan al deze vragen over hetzelfde: de vraag of wat de leerling presteert in overeenstemming is met wat je mag verwachten, waarbij de verwachting gevormd wordt door de eerdere prestaties van de leerling zelf, de prestaties van jaargenoten en de verwachting over eindtermen. Het rapporteren van de leerwinst van een individuele leerling is op verschillende manieren grafisch weer te geven. Dat kan bijvoorbeeld met behulp van de vaardigheidsschaal, waarbij de leerwinst op deze intervalschaal gegeven

[11] Sanders, P.F. (2017; herziene uitgave). *Toetsen op School*. Cito, Arnhem. `https://www.cito.nl/-/media/files/kennis-en-innovatie-onderzoek/toetsen-op-school/cito_toetsen_op_school.pdf?la=nl-nl`

wordt. Deze informatie kan meer betekenis krijgen door een vergelijking te maken met het verwachte traject, gegeven de eerdere resultaten, de vergelijking met anderen, en de relatie tot de einddoelen. Dit komt ook overeen met de drie functies van het rapporteren van de leerwinst.



Figuur 6.1: Voorbeeld voortgang op vaardigheidsschaal

In Figuur 6.1 is de toename in vaardigheid weergegeven tussen twee meetmomenten (M4 en E4). Daarnaast laten de balken zien dat de behaalde vaardigheidsscore valt in het vierde vaardigheidsniveau: percentiel 20-40. Dit is voor beide meetmomenten gelijk gebleven, hoewel de vaardigheidsscore van de leerling wel omhoog is gegaan. Hieruit kunnen we afleiden dat andere leerlingen in de vergelijkingsgroep die gebruikt is om de percentielen vast te stellen ook gegroeid zijn, en daarmee de relatieve positie van deze leerling ten opzichte van anderen niet veranderd is. Behalve de rapportage in termen van de vaardigheidsschaal is het ook mogelijk de bovenstaande figuur weer te geven als het percentage van de bank dat de leerling beheerst. Dat betreft dan de gehele bank waarbij bij de eerste meting 25% een buitengewone prestatie is, bij de tweede meting 50%, enzovoorts.

Op schoolniveau zien we vaak een grotere behoefte aan een relatieve normering waarbij de prestaties van de leerlingen van één school vergeleken worden met de prestaties van een andere school of andere scholen. Hoewel ook daarvoor geldt dat weten **hoeveel** procent van de andere scholen slechter of beter presteren dan je eigen school – ongeacht met welke scholen je eigen school vergeleken wordt – op zich weinig vertelt over **hoe** je je onderwijs kan verbeteren. Het gaat er dan vooral over **dat** er wel (of juist niets) moet veranderen, maar niet zozeer **wat**. Daarmee is de formatieve kracht van deze informatie ook beperkt. Zeker als de vergelijkingsgroepen heel gedetailleerd zijn, gebaseerd op alle relevante schoolvariabelen, kan dat tegenstrijdigheden opleveren. Als een goede school vergeleken wordt met andere goede scholen kan deze als relatief slecht naar voren komen, terwijl het onderwijs goed is. Evenzo kan een school met slecht onderwijs als relatief goed uit de bus komen als deze met slechter presterende scholen wordt vergeleken. Omgekeerd

werkt het ook zo dat het niet corrigeren voor een aantal relevante variabelen ook duidelijke nadelen heeft omdat niet iedere school dezelfde leerlingpopulatie heeft.

6.3 Wie is de gebruiker van rapportages?

Zoals eerder vermeld kunnen rapportages gezien worden als communicatiemiddel: met een afzender, een boodschap en een ontvanger. In het onderwijsleerproces heeft iedere ontvanger echter zijn of haar eigen unieke beslissingen te nemen op basis van de toetsresultaten[12]. Wanneer ontwerpers van rapportages de verschillende behoeftes onderzoeken van het doelpubliek, kunnen er rapportages worden ontwikkeld die enerzijds aansluiten bij de behoefte van de gebruiker, anderzijds bij het kennisniveau. Dit laatste punt wordt nog verder uitgewerkt als randvoorwaarde voor goed gebruik van de rapportage.

In het geval van Vlaamse centrale toetsen kunnen er verschillende doelgroepen onderscheiden worden: leerlingen, leerkrachten, scholen, (regionale) overheden. Afhankelijk van de specifieke behoefte van de doelgroep kunnen specifieke keuzes gemaakt worden over wat er gerapporteerd wordt, op welk detailniveau en hoe dit gevisualiseerd wordt. Het is mogelijk dat verschillende doelgroepen verschillende vragen hebben, of dat er voor verschillende groepen verschillende toetsdoelen gelden. Bijvoorbeeld, dat er voor de leerlingen een belang is om goed te presteren op de toets omdat het resultaat voor hen summatief wordt ingezet, terwijl de resultaten door leerkrachten ook gebruikt worden om hun onderwijs en curriculum te evalueren.

Het is van belang om in het ontwerp van de rapportages zowel inhoudelijk als visueel aan te sluiten bij de toekomstige gebruikers. Het is daarom aan te raden om de rapportage in samenspraak met gebruikers te ontwikkelen. Een ontwikkelproces voor rapportages start veelal met een behoefteanalyse (*needs assessment*)[13]. Doel van een behoefteanalyse is om in kaart te brengen waar de gemeenschappelijke basis is tussen het doel van de toetsontwikkelaar, de informatie die de toets zal opleveren en acties of handelingen die de toetsgebruiker zal uitvoeren op basis van de informatie. Een dergelijke behoefteanalyse is de eerste fase in het ontwerpen van een rapportage[14].

6.4 Het doel van de rapportages

Inmiddels is al meerdere keren het belang van een toetsdoel in relatie tot rapportages genoemd. Al eerder zijn verschillende doelen genoemd met als belangrijkste onderscheid formatief en summatief. Formatief wanneer toetsen ingezet worden ter ondersteuning van een leerproces, summatief wanneer toetsen ingezet worden om een oordeel te vellen met civiel effect. Deze doelen kunnen zowel op leerlingniveau als op een hoger aggregatieniveau (school, regio, landelijk) ingevuld worden.

[12] Zapata-Rivera, J. D., & Katz, I.R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports Assessment in Education: Principles. *Policy & Practice*, 21(4), 442–463.

[13] Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). Applying score design principles in the design of score reports for CBALTM teachers (ETS RM-12-20). Princeton, NJ.

[14] Hambleton, R.K., & Zenisky, A.L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. E. Geisinger (Ed.), *Handbook of testing and assessment in psychology* (pp. 479-494). Washington, DC: APA.

In het bestek is aangegeven dat de toetsen zowel op leerkracht- als op schoolniveau gebruikt moeten kunnen worden voor interne kwaliteitszorg[15]. Wanneer dit ingezet wordt met als doel om (zelf)evaluatie van scholen mogelijk te maken zodat leerkrachten en scholen hun eigen onderwijs kunnen verbeteren, is sprake van een formatieve doelstelling. De toetsresultaten hebben een informerende functie om de kwaliteit van het lesgeven te verhogen: de school realiseert zich door de resultaten dat een specifiek deelaspect of een specifieke eindterm wellicht te weinig aandacht heeft gekregen.

Bij formatief gebruik van de toetsen moeten de resultaten handvatten geven die tot het juiste onderwijskundige handelen leiden. Dat betekent dat de vorm van terugrapportage gericht moet zijn op de inhoud van de test. Hierbij is het van belang om verschillende ontvangers te onderscheiden. De theorie rondom formatief toetsen beschrijft nadrukkelijk een rol voor de leerling: hij/zij moet zelf eigenaar zijn van het leerproces en formatief toetsen speelt daarin een rol. Door zichtbaar te maken waar een leerling staat en waar een leerling naartoe moet, kan een leerling (in overleg met een leraar) vervolgens vaststellen hoe daar te geraken. Een kwantitatieve score is in dit geval niet afdoende, de leerling zal inhoudelijk inzicht moeten krijgen in zijn voortgang en snappen wat er wel en niet geleerd is. Vaak wordt in formatieve toetsing veel tijd besteed aan het creëren van een gedeeld begrip van de eindtermen zodat voor iedereen duidelijk is wat het einddoel is. Een rapportage zou bij deze eindtermen moeten aansluiten en ook voortgang hierop zichtbaar kunnen maken. Bijvoorbeeld door inhoudelijke domeinen verder op te splitsen in leerdoelen en te werken met rubrics: een inhoudelijke omschrijving van niveaus van functioneren die herkenbaar zijn voor een leerling.

Wanneer de formatieve rapportage bedoeld is voor leraren zal het hen informatie moeten verschaffen over de huidige vaardigheden van de leerling. En hulp moeten bieden bij het stellen van de diagnose. Bijvoorbeeld door misconcepties te rapporteren of zichtbaar te maken waar een leerling zich bevindt op een leerlijn. Ook hier kunnen rubrics helpen: de eerstvolgende stap is altijd om de leerling naar het volgende cognitieve functioneringsniveau te brengen. In alle gevallen geldt dat als we met complexe statistische modellen als IRT werken en resultaten rapporteren op een vaardigheidsschaal, dan moet ondanks de complexiteit van de analyses, de rapportage in verstaanbare taal en weergave gepresenteerd worden zodat deze helder en begrijpelijk is.

Interne kwaliteitszorg kan echter ook betrekking hebben op kwaliteitshandhaving, waarbij vooral (achteraf) gekeken wordt of de kwaliteit naar verwachting is: voldoende leerwinst, voldoende toegevoegde waarde, leerdoelen behaald, absoluut en in vergelijking met (vergelijkbare) andere scholen. Toetsresultaten kunnen in dit geval dienen om aan derden duidelijk te maken hoe het gesteld is met de kwaliteit van de school. Dan is sprake van een summatieve doelstelling. Aangezien in een summatieve context tegenvallende resultaten vooral een negatieve impact hebben, en fouten minder gezien worden als middel om van te leren, zal dit tot ander gedrag rond de toetsing leiden. Daarmee neemt de noodzaak om voorzieningen te treffen tegen fraude en andere ongewenste neveneffecten toe.

Wanneer resultaten inzichtelijk zijn voor ouders en deze een rol kunnen spelen bij de keuze van ouders voor een bepaalde school, worden de toetsen in dit geval high-

[15]Dit wordt in de aanvraag genoemd bij vraag 16 over de rapportage (perceel 1) en bij vraag 35 over de verplichte afname (perceel 2).

stakes. Van belang is dat duidelijk wordt dat centrale toetsen slechts een beperkt deel van onderwijskwaliteit betreft en daarmee in principe niet toereikend is als meting van volledige schoolkwaliteit. Als één van de indicatoren voor kwaliteit kunnen deze resultaten een rol spelen, wellicht zelfs als onderdeel van een andere indicator zoals 'onderwijsresultaten', of 'data gebaseerd kwaliteitsbeleid'.

6.4.1 Leerwinst

Bij het bepalen van leerwinst gaat de aandacht met name uit naar de rapportage op schoolniveau, ook in verband met de toegevoegde waarde van de school (zie ook Hoofdstuk 4). Voor de rapportage van de leerwinst van individuele leerlingen zijn er, net als bij de rapportage op schoolniveau, ook meerdere mogelijkheden. Bij het rapporteren van leerwinst kan de rapportage van vaardigheid op de inhoudelijke schaal laten zien hoeveel meer opgaven een leerling vergeleken met het eerdere meetmoment beheerst en wat voor opgaven dat zijn. De rapportage met grenspunten kan duidelijk maken hoeveel meer eindtermen een leerling door de tijd beheerst. Als de vaardigheidsschaal gebruikt wordt, dan valt ook informatie te geven in termen van een toegenomen vaardigheidsscore (al is dat, als deze niet gerelateerd worden aan opgaven of eindtermen minder betekenisvol). Leerwinst is ook te rapporteren als een verschuiving van percentielniveau. Een leerling kan op een hoger percentielpunt, of in een hogere percentielgroep, uitkomen dan voorheen, zeg van percentiel 50 naar percentiel 60. Het nadeel daarvan is dat dit een "nul-som-spel" is: als mensen vooruitgaan met de percentielen, gaan er per definitie ook mensen achteruit, ook als ze groeien. Ze groeien alleen iets minder hard dan de rest. Dat kan demotiverend werken, en als dit de enige vorm van rapportage zou zijn over leerwinst, dan geeft deze niet weer dat deze leerling ondertussen wel vaardiger geworden is dan dat hij of zij bij de eerdere meting was.

De summatieve functie van leerwinst is duidelijk. Het geeft weer wat over een bepaalde tijdsspanne (hier, twee of vier jaar) de toename in vaardigheid is. Daarmee is het voor een belangrijk deel terugkijken naar een leerproces. Bij de formatieve functie gaat het meer over het verbeteren van het leerproces nu en in de nabije toekomst, en het is een interessante vraag hoe de rapportage van leerwinst daar aan kan bijdragen.

Als we kijken naar de rapportage van leerwinst van het individu kunnen we zien of deze volgens verwachting verloopt. Is dat niet het geval, dan kan een leerkracht bij een positieve afwijking vooral doorgaan met wat zo goed werkt. Bij een negatieve afwijking kan een leerkracht aanpassingen doen in het leertraject om de trend te verbeteren. Merk op dat een periode van minstens twee jaar wel erg groot is, en het daarmee lastig kan zijn de oorzaak van een afwijking te duiden. Voor dergelijke aanpassingen is de leerling meer gebaat bij een kortere periode waarover feedback gegeven wordt, en het is dan ook zo dat de ervaring leert dat binnen goede scholen hier eerder op ingegrepen wordt. Voor de school is de rapportage van leerwinst waarschijnlijk relevanter. Dit gaat namelijk meer over de toegevoegde waarde van de school en van het onderwijssysteem. Die kan verder bijgestuurd worden als er een minder grote leerwinst gevonden wordt dan verwacht. Deze aanpassingen hebben meestal een langere cyclus dan de aanpassingen op leerlingniveau.

Leerwinst is zeer belangrijk voor de terugrapportage op schoolniveau. Een school moet bijvoorbeeld weten hoe hun leerlingen het doen in vergelijking met eerdere afnames. Laten ze nu een grotere leerwinst zien dan voorheen? Merk op dat hier de relatieve waarde

minder relevant is dan de absolute. Stel dat alle scholen een stap extra zetten en in absolute waarde 20% **meer** leerwinst zouden behalen dan voorheen, dan zie je dat niet terug in de relatieve cijfers: ieder blijft op hetzelfde percentielpunt, terwijl ze allen zo goed werk hebben verricht. Omgekeerd werkt het overigens ook: als alle scholen 20% **minder** leerwinst halen dan voorheen wordt dat ook niet gesignaleerd met een relatieve normering.

Wat op schoolniveau ook mogelijk zal zijn, is op een meer gedetailleerde vaardigheidsschaal te rapporteren. Voor leerlingen zal de betrouwbaarheid van de meting op de vier hoofdvaardigheden in ieder geval groot genoeg zijn. Naar alle waarschijnlijkheid zal er echter te veel toetstijd nodig zijn om met voldoende nauwkeurigheid een uitspraak te doen over de vaardigheid van individuele leerlingen op een meer gedetailleerd niveau. Op schoolniveau zal dat wel mogelijk zijn, omdat de hoeveelheid waarnemingen op schoolniveau de metingen betrouwbaar genoeg maakt (zie Hoofdstuk 5).

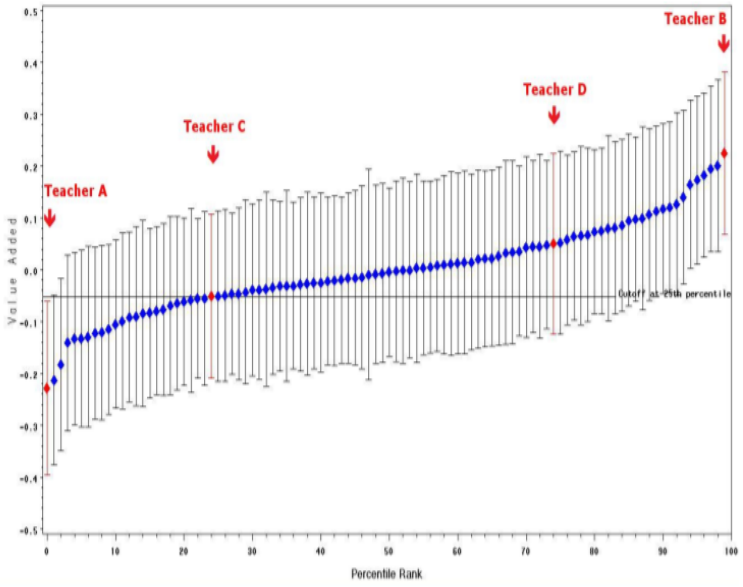
Leerwinst en toegevoegde waarde uitkomsten zijn behept met meetonzekerheid. Het is aan te raden deze onzekerheid rondom de uitkomsten op te nemen in de rapportages, ook om te voorkomen dat aan de uitkomsten in de vorm van puntschattingen in rankings te veel waarde gehecht wordt. Een voorbeeld hoe te voorkomen dat er een onterecht onderscheid gemaakt wordt tussen scholen, is door gebruik te maken van betrouwbaarheidsintervallen binnen de rapportages. Een voorbeeld daarvan is te vinden in Tabel 6.1.

Puntschatting	**Ondergrens**	**Bovengrens**	**Groep**
4	2.5	5.5	1
4.5	3	6	1
6	5	7	1
7	6	8	2
...	...	...	...
10	8.5	11.5	3
11	9.5	12.5	3
12	10.5	13.5	3

Tabel 6.1: Leerwinst rapportages en meetonzekerheid

Voor Tabel 6.1 maken we gebruik van een fictief voorbeeld. In de tabel is te zien dat er op basis van puntschattingen van leerwinst een onderscheid gemaakt kan worden tussen alle twaalf de scholen die zijn opgenomen in dit fictieve voorbeeld. Op het moment dat er echter ook rekening gehouden wordt met de onzekerheid rondom deze puntschattingen, worden de scholen alleen onderscheiden op het moment dat de ondergrens van een leerwinstschatting hoger is dan de bovengrens van de puntschatting van een andere school. In dit voorbeeld betekent het dat er nog maar drie groepen scholen van elkaar te onderscheiden zijn. Dit is overigens ook vaak in de praktijk het geval: Op het moment dat de onzekerheid rondom leerwinstschattingen meegenomen worden in de uitkomsten, levert dit vaak een grote middengroep van scholen op en twee kleine groepen scholen aan de uiteinden. Dit wordt verder geïllustreerd door Figuur 6.2 afkomstig uit Raudenbusch en Jean[16].

[16]Raudenbush, S.W., & Jean, M. (2012). *How should educators interpret value-added scores?* Palo Alto, CA: Carnegie Knowledge Network.



Figuur 6.2: Leerwinst rapportages en meetonzekerheid

In Figuur 6.2 wordt een voorbeeld in de Amerikaanse context gegeven om informatie te geven over de toegevoegde waarde van 100 verschillende leerkrachten. De puntschatting is weergegeven door de (blauwe en rode) diamanten. De verticale grijze lijnen weerspiegelen het betrouwbaarheidsinterval rondom de puntschattingen. Raudenbusch en Jean beargumenteren dat de betrouwbaarheidsintervallen helpen bij de interpretatie van de scores. In dit voorbeeld is wel een onderscheid te maken tussen leerkracht A en leerkracht B. Dit kan geconcludeerd worden, omdat de betrouwbaarheidsintervallen van beide leerkrachten niet overlappen. Dit geldt niet voor de uitkomsten van leerkrachten C en D. De betrouwbaarheidsintervallen van de uitkomsten van deze leerkrachten overlappen aanzienlijk. Voor een groot deel van de leerkrachten geldt dan ook dat er niet al te veel interpretatie aan de verschillen gegeven kan worden. Merk op dat dit voor het merendeel van de leerkrachten geldt. Er zou zelfs gekozen kunnen worden om de puntschatters niet weer te geven om te voorkomen dat daar alsnog mee gewerkt wordt en de meetfout genegeerd wordt.

6.4.2 Signaalfunctie

Naast het inzichtelijk maken van de situatie op scholen en in de klas, kunnen rapportages ook ontworpen worden met een bepaalde signaalfunctie als doel. Bijvoorbeeld om afwijkende scholen of klassen te identificeren. Er zijn twee redenen om extra aandacht te besteden aan mogelijkheden rondom het identificeren van afwijkingen.

Ten eerste, wanneer het doel van de meting is om bijvoorbeeld de 10 beste scholen te identificeren. De signaalfunctie wordt dan ingevuld in de vorm van een top 10. Het is een voorbeeld waarbij het doel van de meting expliciet gekoppeld wordt aan een uitsnede van data die een antwoord geeft op deze specifieke vraag. De rapportage vervult dan zelf een signaalfunctie.

Ten tweede, wanneer er afwijkingen worden geconstateerd binnen een rapportage. Denk bijvoorbeeld aan een school waar relatief weinig waarnemingen zijn. Het is dan mogelijk om via visuele waarschuwingen (kleuren, symbolen, et cetera) de aandacht te trekken

op de beperkte mogelijkheden ten aanzien van de interpretatie van deze specifieke data. Deze mogelijkheden nemen toe wanneer sprake is van een dashboard om rapportages weer te geven. Met name wanneer gekozen wordt voor dynamische rapportages: scholen kunnen zelf rapportages genereren, uitsnedes maken en bijvoorbeeld zelf vergelijkingsgroepen samenstellen. Wanneer gekozen wordt voor dynamische rapportages is een signaalfunctie voor afwijkingen des te belangrijk: wanneer bijvoorbeeld de vergelijkingsgroep te klein wordt om degelijke uitspraken te kunnen doen.

6.5 Voorbeelden van rapportages

In deze paragraaf presenteren we een aantal voorbeelden van rapportages. Deze voorbeelden zijn uiteraard niet uitputtend. Daarnaast dienen de rapportages specifiek voor dit programma ontwikkeld te worden. Deze voorbeelden dienen slechts als illustratie van een aantal hiervoor benoemde aspecten en van de mogelijkheden.

6.5.1 Inhoudelijke betekenis vaardigheidssschaal

Door IRT te gebruiken is het mogelijk om naast de vaardigheidsschaal gebruik te maken van een "itembankscore". Deze score is de verwachte score als de leerling alle opgaven van de itembank gemaakt zou hebben, in plaats van het deel van de itembank in de toets die de leerling gemaakt heeft. Deze itembankscore kan de gehele itembank betreffen als we leerlingen over de verschillende leerjaren willen vergelijken, maar ook dat deel van de itembank dat voor dat leerjaar relevant is[17].

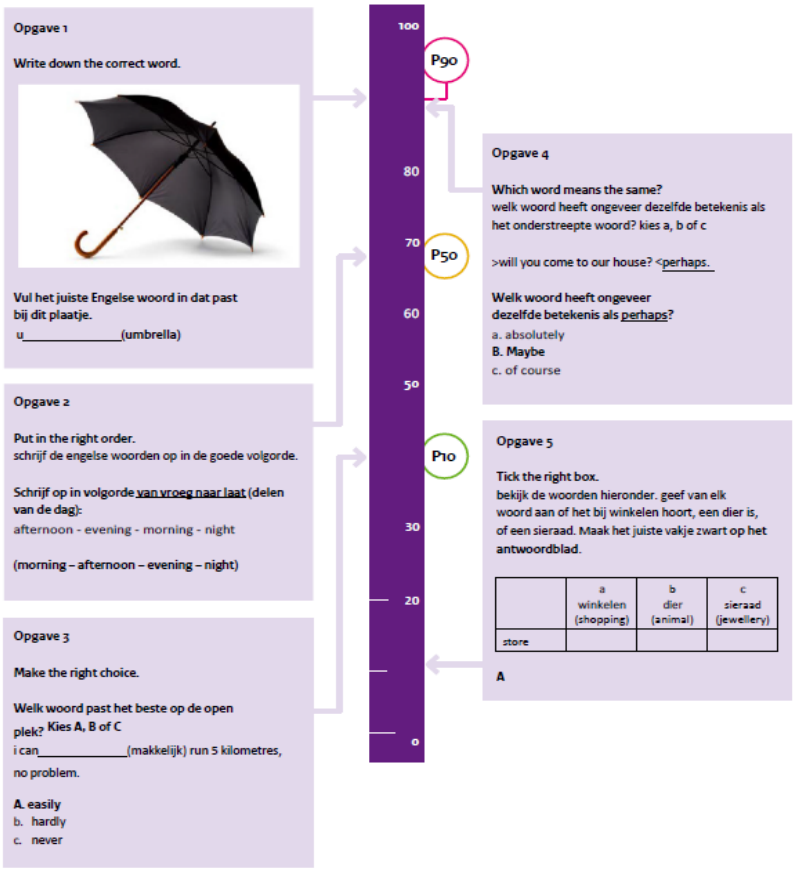
Het voordeel van deze schaal is dat deze betekenisvoller voor de gebruikers is dan de vaardigheidsscore. Een leerkracht kan zich meer voorstellen bij het aantal goede antwoorden op een verzameling vragen, dan bij een getal op intervalschaal dat zeker in het begin wanneer men nog geen vaardigheidsscores kent, geen betekenis heeft. Voor ieder gerapporteerd vaardigheidsniveau is een verzameling opgaven te geven dat representatief is voor dat niveau. Dat betekent dat het niet alleen gaat over hoeveel een leerling kan, maar ook over wat de leerling kan. Dat laatste helpt de leerkracht aan te zetten tot handelen.

Meer nog dan het aantal opgaven kan het representeren van inhoud van de opgaven een schaal betekenis geven. Een voorbeeld hiervan is gegeven in Figuur 6.3[18]. In deze figuur is de schaal Engels woordenschat[19] gegeven. De schaal krijgt inhoudelijk waarde

[17] Met name als de itembank in loop van de tijd aangevuld wordt met nieuwe opgaven, én deze score wordt gebruikt om trends over de tijd te representeren, kan het verstandig zijn om deze itembankscore alleen betrekking te laten hebben op een constante set van opgaven, om ervoor te zorgen dat de betekenis van de score niet ieder jaar verandert. Eens in de vijf jaar kan daar wellicht een update van plaatsvinden wat betreft de opgaven die in die set zitten. Er moet dan wel duidelijk aangegeven zijn hoe die scores aan elkaar relateren waarbij het met behulp van IRT mogelijk is om de scoreschalen sterk op elkaar te laten lijken, bijvoorbeeld door in ieder geval het aantal opgaven niet te wijzigingen.

[18] Figuur is afkomstig uit het publieksverslag van de peiling Engels in het (Nederlands) basisonderwijs. Zie `https://www.onderwijsinspectie.nl/onderwerpen/peil-onderwijs/documenten/rapporten/2019/11/08/peil.engels-einde-basisonderwijs-2017-2018`

[19] De schaal woordenschat is hier gegeven als het percentage goede antwoorden in de itembank gegeven de vaardigheid van de leerlingen. Meer over deze schaal in de technische rapportage van de peiling Engels 2017-2018: Ritzema. L., Hemker, B., Naayer. H., Lowie, W., Corda, A. Van Aken, M. & Rekers-Mombarg, L. (2018). Peiling Engels einde basisonderwijs, 2018, Technische Rapportage. `https://www.onderwijsinspectie.nl/onderwerpen/peil-onderwijs/documenten/rapporten/2019/11/08/peiling-engels-einde-basisonderwijs-2018-technische-rapportage`



Figuur 6.3: Voorbeeld van representatie van opgaven op een vaardigheidsschaal (Engels woordenschat, groep 8

door diverse opgaven op deze schaal af te beelden. Het niveau van de opgave geeft aan bij welke vaardigheid de helft van de leerlingen deze vraag goed hebben en de helft van de leerlingen een fout antwoord geven. Bij een lagere vaardigheid hebben meer leerlingen de opgave fout, en bij een hogere vaardigheid hebben meer leerlingen de opgave goed beantwoord. Er is gekozen voor dit punt om een opgave af te beelden, omdat dit het punt is waar de opgave het best onderscheid maakt tussen de leerlingen rond dat niveau. Inhoudelijk is dit het meest interessante punt omdat dit het beste aangeeft waar de opgaven uitdagend zijn: ze zijn niet te moeilijk en ze zijn niet te makkelijk.

Naast een figuur is het ook mogelijk een lijst van opgaven te geven met de schaalwaarden waar deze opgaven zich bevinden. Met behulp van IRT is een dergelijke lijst gemakkelijk te maken. Als gesteld wordt dat een opgave beheerst wordt op een bepaald vaardigheidsniveau als 80% van de leerlingen op dat niveau de opgave goed maakt[20], kan in een dergelijke lijst ook gegeven worden op welk vaardigheidsniveau beheersing van een dergelijke opgave verwacht wordt. Een dergelijke lijst geeft een inhoudelijke betekenis aan de schaal.

[20]De definitie van beheersing bij 80% is een definitie die bij de peilingen in Nederland gebruikt wordt. Het is ook mogelijk een lager niveau (zeg 75%) of hoger niveau (bijvoorbeeld 90%) als beheersingsniveau aan te merken, als dat gewenst is.

In Figuur 6.3 wordt ter referentie ook aangegeven op welke niveaus de populatie zich bevindt. De percentielpunten P90, P50 en P10 zijn aangegeven om de positie op de schaal aan te geven waar respectievelijk de zeer hoog vaardige, gemiddelde en zeer laag vaardige leerlingen op de schaal zitten. Ook dat kan helpen betekenis te geven aan de schaal, zij het dat dit een relatieve omschrijving van de schaal is.

Een voorbeeld van een meer inhoudelijke betekenis geven aan de schaal, aanvullend op de opgaven die afgebeeld zijn, wordt gegeven in Figuur 8.2[21]. In die figuur zijn ook inhoudelijke niveaus weergegeven in termen van drie niveaus binnen het Europees Referentiekader (ERK)[22] voor Engels. In Tabel 6.2 zijn de globale beschrijvingen van de ERK-niveaus A1 tot en met B1 voor leesvaardigheid opgenomen.

Beheersingsniveau	**Beschrijving**
A1	leerling kan vertrouwde namen, woorden en zeer eenvoudige zinnen begrijpen, bijvoorbeeld in mededelingen, op posters en in catalogi.
A2	leerling kan zeer korte, eenvoudige teksten lezen. leerling kan specifieke voorspelbare informatie vinden in eenvoudige, alledaagse teksten zoals advertenties, folders, menu's en dienstregelingen en kan korte, eenvoudige, persoonlijke brieven begrijpen.
B1	leerling kan teksten begrijpen die hoofdzakelijk bestaan uit hoogfrequente, alledaagse of aan zijn/haar werk gerelateerde taal. leerling kan de beschrijving van gebeurtenissen, gevoelens en wensen in persoonlijke brieven begrijpen.

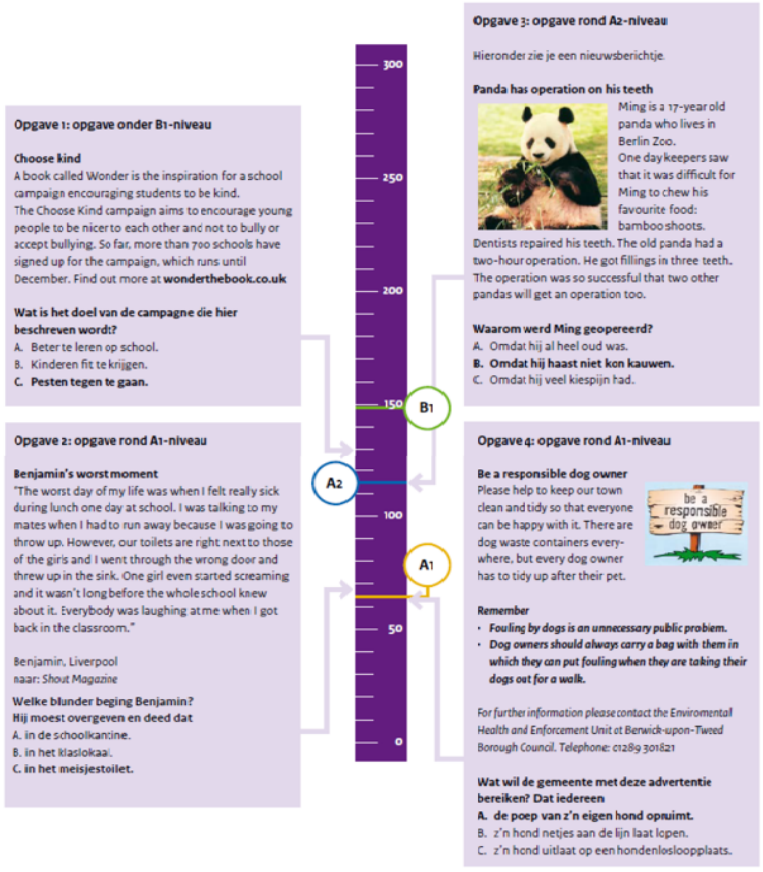
Tabel 6.2: ERK-referentieniveau lezen

Door middel van een standaardbepalingsprocedure kunnen deze omschrijvingen gerelateerd worden aan de vaardigheidsschaal[23]. Zodoende kan bepaald worden welke vaardigheid nodig is om te kunnen stellen dat een leerling het omschreven beheersingsniveau gehaald heeft. Doordat de inhoudelijke omschrijving gerelateerd wordt aan de vaardigheidsschaal krijgt deze schaal ook een inhoudelijke betekenis. Wanneer een schaal langere tijd bestaat, krijgen leerkrachten ook ervaringen met de schaal. Naar verloop van tijd kunnen de schaalwaarden zodoende ook betekenis krijgen, net zoals de waarden op een IQ-schaal.

[21] Figuur is afkomstig uit het publieksverslag van de peiling Engels in het (Nederlands) basisonderwijs. Zie `https://www.onderwijsinspectie.nl/onderwerpen/peil-onderwijs/documenten/rapporten/2019/11/08/peil.engels-einde-basisonderwijs-2017-2018`

[22] Ter referentie zie: * SLO (2011). Toetsen en beoordelen met het ERK. Geraadpleegd van: `http://www.erk.nl/docent/toetsing/toetsen-en-beoordelen-met-het-ERK.pdf/` * Europees Referentiekader Talen (z.j.) Wat is het ERK? Geraadpleegd van: `http://www.erk.nl/docent/Wat/`

[23] Voor een omschrijving van deze procedure zie hoofdstuk 7 van Ritzema. L., Hemker, B., Naayer. H., Lowie, W., Corda, A. Van Aken, M. & Rekers-Mombarg, L. (2018). *Peiling Engels einde basisonderwijs, 2018, Technische Rapportage*.`https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/rapporten/2019/11/08/peiling-engels-einde-basisonderwijs-2018-technische-rapportage/Technische+rapportage+Consortium+peiling+Engels-toegankelijke+versie.pdf`



Figuur 6.4: Voorbeeld van representatie van opgaven op een vaardigheidsschaal (Engels lezen, groep 8, Nederlands basisonderwijs met cesuurpunten voor vaardigheidsniveaus

Over het bepalen van standaarden

Op een vergelijkbare manier als met het ERK-niveau kunnen eindtermen aan de vaardigheidsschaal gerelateerd worden. In het kader "Referentieniveaus Taal en Rekenen" is een voorbeeld gegeven van een grootschalige opzet van standaardbepalingen. Daarbij was nog niet ingegaan op de kosten van dergelijke standaardbepalingen. Deze kosten zijn vooral afhankelijk van een aantal factoren. Dat is ten eerste het aantal grenspunten dat gezet moet worden (het aantal niveaus), ten tweede het aantal beoordelaars dat gebruikt wordt per gezet niveau, en ten derde de methode waarop de standaarden bepaald worden.

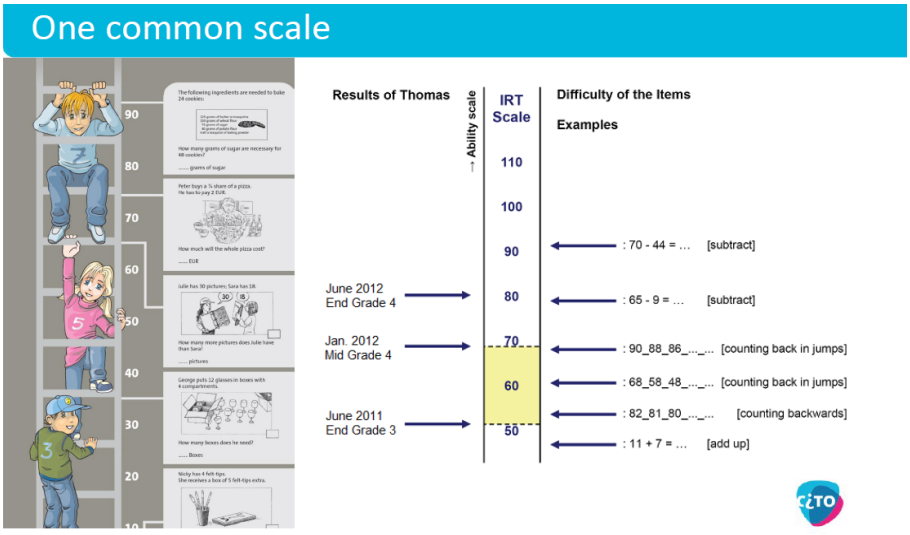
De eerste factor is afhankelijk van het aantal eindtermen per vaardigheid en in welke mate de eindtermen geclusterd worden. De tweede factor heeft te maken met de mate van vereiste nauwkeurigheid en het benodigde draagvlak van het cesuurpunt. Bij standaardbepalingen is een aantal tussen de tien en twintig experts niet ongebruikelijk om voldoende nauwkeurigheid te verkrijgen. Voor een voldoende draagvlak in het onderwijsveld kan het zeker voor zeer belangrijke beslissingen nodig zijn om een hoger aantal beoordelaars te gebruiken dan strikt noodzakelijk om iedere groep van belanghebbenden (leerkrachten, vertegenwoordigers vanuit de netten, wetenschappelijk experts) voldoende representatie te geven.

Over het bepalen van standaarden vervolg

Daar waar het de gebruikte methode betreft is dat voor een deel afhankelijk van de gemeten vaardigheid en de daarbij behorende items en taken die de leerlingen voorgelegd krijgen, en deels van de beoogde betrouwbaarheid. Wanneer een schaal gevormd wordt door meerkeuze-opgaven, of korte-antwoord-opgaven zijn de mogelijke methodes gemakkelijker uit te voeren dan wanneer een schaal gevormd wordt door niet-automatisch scoorbare opgaven. In het geval van vaardigheden zoals schrijven en spreken zijn werkwijzen nodig die meer tijd kosten. Als meer zekerheid nodig is voor het bepalen van de grenswaarde, bevat de standaardbepaling meer ruimte voor discussie en zijn er meer rondes nodig. Het is gebruikelijk om twee of drie rondes te hebben.

De kosten zijn uiteraard ook afhankelijk van hoeveel de experts betaald krijgen (waarbij ook voorbereidingstijd en reistijd meegerekend worden). Al met al kunnen de kosten variëren. Als eenmaal een standaard bepaald is, kan deze echter wel een groot aantal jaar meegaan. Door middel van IRT kan de standaard geplaatst worden op diverse versies van de toetsen en ook als de itembank verder uitgebreid wordt, kan de standaard toegepast blijven worden. Een belangrijke reden om standaarden te vervangen, is als de omschrijving van het beoogde niveau verandert. Veranderingen in de eindtermen kunnen zodoende leiden tot nieuwe standaarden.

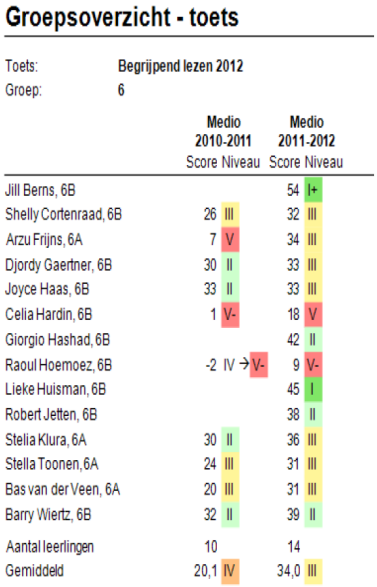
Naast de betekenis van de schaal kunnen de opgaven ook in de terugrapportage zelf een rol spelen. In Figuur 6.5 wordt van één leerling (Thomas) weergegeven wat zijn groei is in een periode van drie jaar. Deze groei is uitgedrukt in drie punten op een vaardigheidsschaal. Aan verschillende punten op deze schaal zijn inhoudelijke labels gekoppeld. Zo wordt duidelijk dat Thomas in juni 2011 eenvoudige optelsommen kon maken, maar terugtellen nog wat lastiger vond maar hier wel al mee bezig was. In juni 2012 kon hij inmiddels gemakkelijk terugtellen met sprongen.



Figuur 6.5: Voorbeeld inhoudelijke duiding vaardigheidsschaal

6.5.2 Groepsoverzichten

Het presenteren van alle resultaten van de leerlingen in één helder overzicht kan de leerkracht goed helpen de lessen voor die klas goed in te richten. Er zijn verschillende mogelijkheden om data geaggregeerd weer te geven. In onderstaand, eenvoudig voorbeeld is gekozen voor een overzicht voor leerkrachten waarin elke leerling individueel zichtbaar blijft.



Groepsoverzicht - toets

Toets: Begrijpend lezen 2012
Groep: 6

	Medio 2010-2011 Score	Niveau	Medio 2011-2012 Score	Niveau
Jill Berns, 6B			54	I+
Shelly Cortenraad, 6B	26	III	32	III
Arzu Frijns, 6A	7	V	34	III
Djordy Gaertner, 6B	30	II	33	III
Joyce Haas, 6B	33	II	33	III
Celia Hardin, 6B	1	V-	18	V
Giorgio Hashad, 6B			42	II
Raoul Hoemoez, 6B	-2	IV → V-	9	V-
Lieke Huisman, 6B			45	I
Robert Jetten, 6B			38	II
Stelia Klura, 6A	30	II	36	III
Stella Toonen, 6A	24	III	31	III
Bas van der Veen, 6A	20	III	31	III
Barry Wiertz, 6B	32	II	39	II
Aantal leerlingen	10		14	
Gemiddeld	20,1	IV	34,0	III

Figuur 6.6: Voorbeeld groepsoverzicht

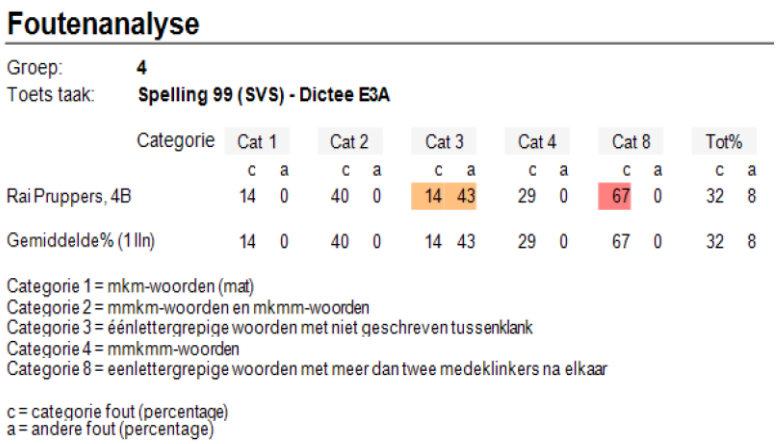
In deze rapportage (Figuur 6.6) wordt enerzijds de score van een leerling weergegeven, maar daarnaast ook een categorie die samenhangt met de percentielscore van de leerling[24]. Door middel van kleurcodering worden signalen gegeven worden over zorgleerlingen. Dit type rapportage wordt bijvoorbeeld gebruikt om niveaugroepen samen te stellen waarin leerlingen instructie krijgen met leerlingen met hetzelfde ontwikkelingsniveau.

6.5.3 Diepere inhoudelijke analyse

Naast de vaardigheidsscore en de relevante items horende bij die vaardigheid kan de terugrapportage ook de eindtermen betreffen. Een dergelijke rapportage is formatief in te zetten doordat het niet behalen van een eindterm aangeeft waar verder aan gewerkt moet worden. Voor deze rapportage worden een aantal opgaven bij elkaar genomen, maar vaak zijn dit kleinere eenheden dan een volledige toets. In het voorbeeld hieronder zijn opgaven gecombineerd op basis van de fout die gemaakt is. Het betreft een spellingstoets en de mogelijke fouten die leerlingen maken zijn in te delen in categorieën. Elke opgave is ingedeeld in een categorie, vervolgens worden de fouten geclassificeerd: is het een fout die past bij de foutencategorie of een andere fout.

In Figuur 6.7 is te zien dat de leerling (Rai) in de spellingstoets vooral in categorie

[24]Iedere 20% krijgt een categorie: 0-20: V, 20-40: IV, 40-60: III, 60-80: II, 80-100: I. Waarbij I de 20% best presterende leerlingen zijn. Overigens zijn dit vooraf bepaalde categorieën en kan het in een jaar dus zo zijn dat meer of minder leerlingen dan 20% in een categorie geplaatst worden.



Foutenanalyse

Groep: **4**
Toets taak: **Spelling 99 (SVS) - Dictee E3A**

Categorie	Cat 1		Cat 2		Cat 3		Cat 4		Cat 8		Tot%	
	c	a	c	a	c	a	c	a	c	a	c	a
Rai Pruppers, 4B	14	0	40	0	14	43	29	0	67	0	32	8
Gemiddelde% (1 lln)	14	0	40	0	14	43	29	0	67	0	32	8

Categorie 1 = mkm-woorden (mat)
Categorie 2 = mmkm-woorden en mkmm-woorden
Categorie 3 = éénlettergrepige woorden met niet geschreven tussenklank
Categorie 4 = mmkmm-woorden
Categorie 8 = eenlettergrepige woorden met meer dan twee medeklinkers na elkaar

c = categorie fout (percentage)
a = andere fout (percentage)

Figuur 6.7: Voorbeeld foutenanalyse

8 fouten maakt. Deze categorie zijn woorden als ‘schep’, ‘spruit’, ‘arts’, etc. Daarnaast schrijft deze leerling ook relatief veel woorden die als categorie 3 aangemerkt zijn fout (‘tulp’, ‘wolf’, ‘warm’), maar maakt hij daar juist weinig fouten die specifiek passen bij deze categorie.

6.5.4 Meetfout rapporteren

Het spreekt voor zich dat een rapportage de resultaten van een toets of proef weergeeft. In voorgaande paragrafen is hier verder op ingegaan. Het is naast een zo goed mogelijke weergave van de resultaten ook mogelijk om extra informatie toe te voegen die de interpretatie van de gegevens verbeteren. Dit kan bijvoorbeeld door de onzekerheid van de resultaten te presenteren, waardoor gebruikers de resultaten beter naar waarde kunnen schatten.

Gebruikelijk is om informatie toe te voegen over de betrouwbaarheid van de resultaten. Bijvoorbeeld via het presenteren van betrouwbaarheidsintervallen[25]. Afhankelijk van de betrouwbaarheid van de meting is deze groot of klein. Dit is een manier om ‘onzekerheid’ van scores te rapporteren en is bij het grote publiek bekend van het weerbericht. In het weerbericht wordt geregeld een bereik aan mogelijke waarden getoond om aan te geven dat het weerbericht onderhavig is aan onzekerheid. De interpretatie van het begrip onzekerheid bij een onderwijskundige meting is echter complexer dan bij het weerbericht.

De complexiteit van het concept meetfout maakt dat de rapportage ervan controversieel is. In een recent uitgevoerde studie rondom het leerlingvolgsysteem in Nederland is expliciet onderzoek[26] gedaan naar de wenselijkheid van het rapporteren van meetfouten (zie volgend tekstblok). In dat onderzoek bleek dat hoe deze ook wordt weergegeven, leerkrachten

[25] Hoe groot een dergelijk betrouwbaarheidsinterval moet zijn is lastig van tevoren te bepalen. Als gebruik gemaakt wordt van een 90%’s of 95%’s betrouwbaarheidsinterval geeft dat vaak een dusdanig groot bereik aan scores aan dat de meting als weinigzeggend kan worden ervaren. In leerlingvolgsystemen is een 70%’s betrouwbaarheidsinterval niet ongebruikelijk. Het bereik aan scores is de helft van dat bij een 95%’s betrouwbaarheidsinterval.

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

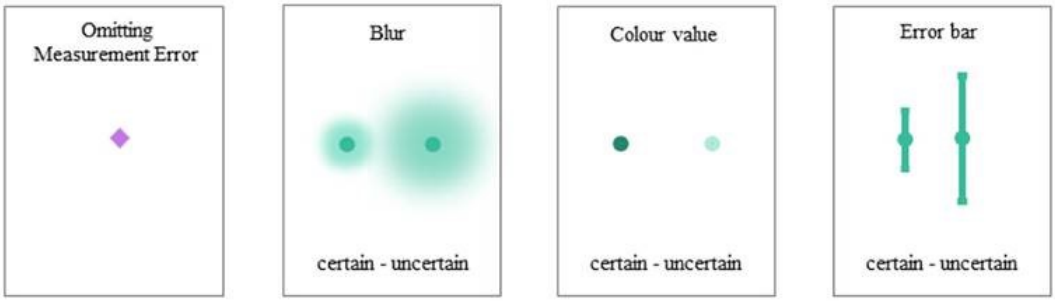
moeite hebben dit goed te interpreteren. Als gekeken wordt hoe deze onzekerheid helpt bij hun handelen, dan is dat beperkt. De belangrijkste aanbeveling uit het onderzoek is dat, indien meetfout gerapporteerd wordt, hier bij de training van leraren expliciet aandacht voor is. Dit hangt ook samen met het begrip assessment literacy, wat in een latere paragraaf aan de orde komt.

Onderzoek presentatie meetfout

In het onderzoek van Hopster-Den Otter e.a. (2018) stond centraal in hoeverre leraren betere beslissingen zouden nemen wanneer ook meetfout gepresenteerd werd. Het onderzoek werd verbonden aan het leerlingvolgsysteem dat in Nederland veelvuldig gebruikt wordt op scholen zodat leraren in het onderzoek zich een goede voorstelling konden maken van de praktijksituatie.

In het onderzoek werden drie verschillende visualisaties van meetfout gebruikt. Deze werden vergeleken met visualisaties zonder meetfout weergave. Een voorbeeld hiervan is gegeven in Figuur 6.8 hieronder.

De resultaten van het onderzoek lieten zien dat de *error bar* (uiterst rechts in Figuur 6.8) leidde tot de beste beslissing over leerlingen, maar dat dit niet altijd kwam door de beste interpretatie van de weergave. De onderzoekers vonden ook een groot aantal misconcepties in de interpretatie van deze weergave. Een voorbeeld van een misconceptie is dat gedacht wordt dat de meetfout laat zien of de leerlingen veel twijfelen over het antwoord of niet. Deze onjuiste interpretatie van de meetfout kan ook averechts werken bij de interpretatie van de toetsresultaten. Het volledige artikel is open access beschikbaar via: `https://doi.org/10.1080/0969594X.2018.1447908`.



Figuur 6.8: Visualisatie van meetfouten

Als we kijken naar uitspraken doen over het al dan niet behalen van een absolute cesuur (zoals het halen van een eindterm) is er ook sprake van onzekerheid. Als de meetfout afgebeeld wordt, dan kan die onzekerheid duidelijk worden doordat het cesuurpunt binnen het bereik van de meetfout valt. Als er echter, vanwege de genoemde uitdagingen bij het gebruik van een meetfout, gekozen wordt die niet te rapporteren, is het ook op een andere manier mogelijk om aan te geven met hoeveel zekerheid de uitspraak gedaan kan worden of de leerling het niveau wel of niet haalt[27].

[27]Heeft een leerling een (zeer) lage vaardigheid dan is zeker dat deze leerling de eindterm niet haalt, maar hoe dichter de met de toets getoonde vaardigheid richting het cesuurpunt gaat, hoe minder zeker het is dat

Bij summatieve uitspraken over individuen kan het rapporteren van meetonzekerheid een stevige uitdaging betekenen in de communicatie met de leerling, en wordt om die reden bij een dergelijk gebruik vermeden. Dit kan ook als voordeel gezien worden als summatief gebruik van resultaten tegengegaan moet worden. Door bijvoorbeeld bij de rapportage over scholen ook de meetonzekerheid mee te nemen, kan de vergelijking tussen scholen in perspectief geplaatst worden. Door het rapporteren van de meetfout per school is te zien dat verschillen tussen scholen, die al dan niet gecorrigeerd zijn voor achtergrondvariabelen, minder groot zijn dan wellicht in eerste instantie gedacht. Door het rapporteren van de significantie van de verschillen in gemiddeld vaardigheidsniveau, zijn mogelijke verschillen te relativeren als toevallig[28]. Dat is van belang voor instanties met superviserende taak over scholen, maar kan ook toegepast worden bij rapportages voor ouders[29]. Op welke wijze dit het best afgebeeld kan worden hangt ook af van de doelgroep[30].

6.6 Randvoorwaarden voor goed gebruik

In dit hoofdstuk is aan de orde gekomen dat het doel van de toets in overeenstemming moet zijn met het ontwerp van de rapportage. Tot nu toe lag daarmee de focus op de aansluiting van de rapportage op het toetsontwerp en de onderliggende analyses. Uiteindelijk geldt natuurlijk dat de rapportage ingezet dient te worden in de onderwijspraktijk. Om een goed gebruik te borgen zijn een aantal randvoorwaarden te formuleren.

deze leerling onder het beoogde niveau zit. Wanneer de gerapporteerde vaardigheid van de leerling exact op het cesuurpunt zit, dan is er ongeveer even veel zekerheid dat de leerling het beoogde niveau gehaald heeft of niet. Hoe meer de getoonde vaardigheid boven het cesuurpunt ligt, hoe zekerder het is dat de leerling het vaardigheidsniveau heeft dat nodig is om te zeggen dat de leerling de eindterm gehaald heeft. Deze zekerheid heeft te maken met de meetfout en met de vaardigheid. In de rapportage kan dan bij de vaardigheid komen te staan of de eindterm met X% zekerheid wel/niet gehaald is.

[28]Minder gebruikelijk, maar ook een mogelijkheid om de verschillen tussen scholen te relativeren, is door deze verschillen te relateren aan de aan de samengevoegde (*pooled*) standaardafwijkingen van de scholen. Hiermee kan de effectgrootte bepaald worden, die inzicht geeft of het gevonden verschil ook betekenisvol is. Als een verschil statistisch significant is, betekent dit niet noodzakelijk dat het groot, belangrijk of nuttig is bij de besluitvorming. Het betekent gewoon dat u erop kunt vertrouwen dat er een verschil is. Als er maar voldoende observaties zijn, is ieder werkelijk verschil significant. Dat houdt echter niet in dat het werkelijk gevonden verschil ook impact heeft. Als de leerlingen op school A significant 1 vaardigheidspunt hoger scoren dan de leerlingen op school B, dan is dat een werkelijk verschil. Als echter de standaardafwijking van de vaardigheid binnen ieder van de scholen 50 vaardigheidspunten is, dan is de effectgrootte $(1/50 =)$ 0,02. Dit (werkelijke) verschil tussen de scholen heeft geen impact in de praktijk. Er zijn verschillende vuistregels om een kwalitatieve interpretatie te geven aan de gevonden effectgrootte, maar effectgroottes van onder de 0,10, en vaak ook onder de 0,20 worden als niet betekenisvol gezien.

[29]Wanneer ouders voor de schoolkeuze scholen met elkaar vergelijken kan voor iedere selectie van scholen weergegeven worden of deze significant van elkaar verschillen in prestaties.

[30]Voor een doelgroep als beleidsmakers, die gewend zijn met cijfers te werken, en meer dan oppervlakkig resultaten moeten evalueren, is er meer mogelijk dan voor ouders. Wanneer ouders scholen kunnen vergelijken via een internetpagina kan er bij het weergeven van de resultaten voor gekozen worden om alleen de verschillen te tonen als het (al dan niet gecorrigeerde) verschil tussen de scholen significant en betekenisvol is. Het maken van een dergelijke tool hoeft niet heel kostbaar te zijn, en is statistisch ook niet ingewikkeld. Voor de werkelijke vormgeving is het zeker aan te raden om ook niet-statistisch geschoolde ouders te betrekken om te zien wat werkt en wat niet.

6.6.1 Doel is bekend

Allereerst is het van belang dat het doel van een toetssysteem of toets bekend is bij de gebruikers. De interpretatie van gegevens die worden weergegeven in rapportages zijn afhankelijk van de specifieke vragen waarmee naar de rapportage gekeken wordt. Wordt met een formatieve bril naar de rapportages gekeken dan zullen andere conclusies getrokken worden dan wanneer met een summatieve bril naar dezelfde rapportage wordt gekeken. Het is belangrijk dat ontwerpers van rapportages zich hiervan bewust zijn en potentiële misinformatie of verkeerde conclusies afhankelijk van het perspectief van de gebruiker proberen te voorkomen.

Veelal zien we dat vanuit efficiëntie overwegingen formatieve en summatieve doelen gecombineerd worden. Vaak levert dit bij de plaatsing van het toetssysteem in het stelsel problemen op. In Nederland werd hier de regering ook over geadviseerd door de Onderwijsraad . Dat geldt ook voor de doelen bij een systeemmeting, en de bijbehorende rapportage. Als centrale toetsen de functie van de Vlaamse onderwijspeilingen deels overnemen, heeft dat een aantal nadelen. Een ervan is dat bij de peilingen de metingen veel meer gedetailleerde informatie leveren, deels omdat de gemeten vaardigheden over het algemeen verder uitgewerkt zijn, en dat bij de centrale toetsing meer informatie ingewonnen wordt over de lessen en de werkwijze op scholen. De resultaten van de centrale toeten kunnen gebruikt worden voor een algemene indruk, maar kunnen niet de peiling als zodanig overnemen. Peilingsonderzoek kan wel aansluiten bij de centrale toetsing.

Als er informatie van leerkrachten en scholen (bijv. instructietijd, mate van professionele ontwikkeling) wordt verzameld, kan dat ook meegenomen worden in de rapportage naar de scholen. Op basis van die informatie kan de systeemevaluatie een formatieve functie hebben voor best practices op scholen. Dergelijke informatie hoeft niet rechtstreeks met alle leerkrachten besproken te worden. Dit lijkt eerder geschikt voor het niveau waar het onderwijskundig beleid op een school wordt bepaald. Het kan echter ook meegenomen worden in het communiceren over het interpreteren van rapportages, en handvatten bieden op basis waarvan men kan handelen.

Bij een summatief gebruik van de eindtermen op schoolniveau kan men per eindterm het percentage leerlingen rapporteren dat die eindterm haalt, waarbij geëvalueerd wordt of dit percentage voldoet aan de verwachting. Ongeacht of deze verwachting bepaald wordt door middel van een relatieve vergelijking met andere (vergelijkbare) scholen of door standaardbepaling heeft het rapporteren van een dergelijk percentage een risico. Wanneer scholen qua prestaties vooral hierop afgerekend worden, kan dat namelijk tot onwenselijk strategisch gedrag leiden. Bij een school die zich hierdoor sterk op de statistieken richt, zal de focus van het onderwijs liggen op leerlingen die nét de eindterm niet dreigen te halen, en om deze leerlingen over die grenswaarde te krijgen. Er zijn ook leerlingen die ver onder het niveau zitten, en weliswaar met gericht onderwijs sterk vooruit kunnen gaan, maar de eindterm waarschijnlijk niet zullen halen. Deze leerlingen kunnen mogelijk minder aandacht krijgen omdat de prestatie van de leerling in termen van eindtermen toch zeer waarschijnlijk de statistiek niet helpt. Evenzo zijn er leerlingen die redelijk makkelijk de eindterm zullen halen. Door deze leerlingen gericht onderwijs te geven, kunnen zij excelleren, maar die toegevoegde waarde komt ook niet tot uiting in “eindtermenstatistiek”. Deze nadelige invloeden zijn niet te voorkomen als de behaalde eindtermen de belangrijkste statistiek voor de school wordt, of er nu gecorrigeerd wordt voor allerlei relevante factoren

of niet: in alle gevallen is meer behaalde eindtermen altijd beter.

6.6.2 Verstaanbare rapportages

Zoals eerder vermeld, is het belangrijk toekomstige gebruikers te betrekken bij het ontwerpen van rapportages. Uit onderzoek blijkt dat de beschikbaarheid van data niet automatisch leidt tot het gebruik van deze data door leerkrachten of scholen[31,32]. Dit wordt veroorzaakt door problemen bij de interpretatie van gegevens en beperkte competenties van leerkrachten om de rapportages te lezen. Deze competentie, ook wel *assessment literacy* genoemd, is nodig om de rapportages effectief in te zetten. Voor rapportages die op grote schaal gebruikt worden is het noodzakelijk aan te sluiten bij de assessment literacy van leerkrachten in het onderwijs. Hoewel inzetten op professionalisering uiteraard aan te bevelen is, is het niet waarschijnlijk dat daarmee alle leerkrachten bereikt kunnen worden.

Eén van de mogelijkheden voor het ontwerp van rapportages die aansluiten bij de kennis en vaardigheden van leerkrachten is om aan te sluiten bij ervaringen met de eerdere terugrapportages. Het is verstandig de rapportages die nu al bestaan bij de net-eigentoetsen, zoals de IDP[33], de OVSG-toets, of de paralleltoetsen[34] mee te nemen in het ontwerp van de rapportage voor de centrale toetsen. De rapportages van de nieuwe centrale toetsen kunnen daar mogelijk nog functionaliteiten aan toevoegen door de beste elementen van deze terugrapportages te combineren.

In het onderzoek naar het gebruik van rapportages in het kader van formatief gebruik van toetsen[35] kwam ook uit literatuuronderzoek[36] naar voren dat het enige vaste recept voor succes is dat je telkens, voor iedere doelgroep opnieuw, een rapportage in co-creatie met de doelgroep moet vormgeven. Voor een succesvolle co-creatie is het van belang dat het sentiment in die groep ook een plaats krijgt.

Wanneer het doel is om met behulp van de toetsresultaten het onderwijs te verbeteren, dan is het essentieel dat de terugrapportage de leerkrachten en schoolleidingen daar handvatten voor geeft. De resultaten moeten geen eindstation zijn, maar een startpunt waarop het onderwijsveld kan handelen. Dat betekent dat deze actie-gedreven is: het moet duidelijk zijn wat de onderwijzende moet doen bij een specifiek resultaat. Zoals aangegeven betekent dat ook een relatie tot de inhoud, door opgaven mee te geven die voor een bepaald vaardigheidsniveau relevant zijn, of een duidelijk handelplan wanneer een signaalfunctie indiceert dat er iets moet gebeuren.

De keuze voor bijvoorbeeld een grafische weergave van resultaten of een tekstuele

[31] Vanhoof, J., Verhaeghe, G., Verhaeghe, J.P., Valcke, M., & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. *Educational Studies*, 37(2), 141–154.

[32] Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2011). Effecten van ondersteuning bij schoolfeedbackgebruik [Effects of support in school feedback use]. *Pedagogische Studiën*, 88(2), 90–106

[33] `https://pro.katholiekonderwijs.vlaanderen/evaluatiebox-basisonderwijs/praktische-informatie`

[34] De rapportages die geleverd zijn bij het peilingsonderzoek, zie `https://www.paralleltoetsen.be/voorbeelden`.

[35] Dorien Hopster-den Otter (2019). Formative assessment design: a balancing act (academisch proefschrift). Enschede University of Twente.

[36] Een van de werken dat met name interessant was, betrof het boek Needs Analysis and Programme Planning in Adult Education (2012) van Simona Sava (Publisher: Verlag Barbara Budrich Series: Study Guides in Adult Education) met als open access hoofdstuk Methods of Needs Analysis in Educational Context.`https://doi.org/10.2307/j.ctvbkjvs2.8`).

duiding is ook afhankelijk van de gebruiker. Uit een studie van Reiter en Dale (1997)[37] blijkt bijvoorbeeld dat het goed interpreteren van grafisch weergegeven resultaten vaak een bepaalde mate van expertise en achtergrondkennis van de gebruiker vereist.

Behalve gebruik maken van een intuïtief begrijpelijke rapportagevorm, is het van groot belang dat de wijze waarop de resultaten gepresenteerd worden, aansluiten bij de informatiebehoefte van de leerkrachten en scholen. Daartoe moet bekend zijn wat de leerkrachten en scholen willen weten. Welke informatie hebben ze nog meer nodig om richting te geven aan handelen. Daar kom je achter als de groep die moet gaan werken met de rapportage bij de ontwikkeling van de rapportage betrokken is. Wetenschappelijke kennis is dan niet voldoende: praktijkkennis is evenzeer van belang.

Het is daarmee van belang te realiseren dat het ontwikkelen van de terugrapportages een stapsgewijs proces is waarbij de PCDA-cyclus (zie ook paragraaf 5.2.9) een belangrijke rol speelt. De bestaande rapportages zijn ook het gevolg van een jarenlange ontwikkeling, en ook de terugrapportage voor de centrale toetsen zal een dergelijke verbetercyclus doormaken. Door middel van studies naar de effectiviteit van de rapportages kunnen die via deze cyclus ook bijgesteld worden zodat de rapportages het gewenste gedrag uitlokken.

Naast het ontwikkelen van de rapportages is het ook aan te raden de communicatie over de rapportages samen met het onderwijsveld te ontwikkelen. Om een juiste interpretatie van de rapportages te bevorderen moet namelijk ook aandacht besteed worden aan uitleg over die rapportages. Niemand weet beter waar leerkrachten verdere toelichting bij nodig hebben, dan leerkrachten zelf.

Het is ook van belang dat leerkrachten in de gelegenheid gesteld worden om zich te ontwikkelen in het interpreteren van de terugrapportages. Dat kan door goede instructievideo's, flyers met simpele uitleg, of handleidingen en cursussen. Wat ook helpt bij dergelijke instructies zijn voorbeelden van goed handelen.

In het leerproces bij de ontwikkeling kan ook gevonden worden dat dit context afhankelijk is. Dat betekent dat niet alleen de vaardigheid van de leerlingen een rol speelt, maar ook andere kenmerken van de school. Hier wordt het dan weer interessant de school te vergelijken met andere vergelijkbare scholen. Op basis van de verzamelde informatie over de school (algemene kenmerken, kenmerken leerkrachten, kenmerken over didactisch handelen) kunnen mogelijk richtlijnen of suggesties gegeven worden die voor die school specifiek relevant zijn. De verkregen groei, en de percentages leerlingen die de verschillende niveaus bereikt hebben, zijn goed te duiden door leerkrachten en schoolbesturen, mits de resultaten gepresenteerd worden op een eerlijke manier. Dat is ook onderdeel van de studies naar toegevoegde waarde van een school.

Een laatste punt met betrekking tot de verstaanbaarheid van rapportages betreft een reflectie van complexiteit van analyses en de weergave hiervan. In principe geldt dat de complexiteit van analyses en statistische modellen onderliggend aan de resultaten voor veel gebruikers niet helpend zijn. Sterker nog, de complexiteit kan misinterpretatie in de hand werken omdat gebruikers niet goed weten wat zij zien. Hoewel dit een eenvoudig uitgangspunt lijkt, leidt het in de praktijk soms tot dilemma's. Neem als voorbeeld het weergeven van een concept als meetfout. De weergave van meetfout kan bijdrage aan het besef dat een toetsscore beïnvloed wordt door toevallige fouten en niet als absoluut

[37] Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1), 57–87.

gegeven kan worden geïnterpreteerd. Dit is echter een complex begrip voor veel gebruikers. Het resultaat is dat, hoewel meetfout veelal weergegeven wordt als ondersteuning van de interpretatie, het juist kan leiden tot grotere verwarring[38].

6.6.3 Inbedding in schoolbeleid

In de aanvraag is aangegeven dat de toetsen zowel op leerkracht- als op schoolniveau gebruikt moeten kunnen worden voor interne kwaliteitszorg. Om dit daadwerkelijk vorm te geven dient er ook een voedingsbodem op scholen te zijn om dit in te richten. Er zijn meerdere studies bekend ten aanzien van het gebruik van data om schoolbeleid vorm te geven. Hieruit komt vooral de complexe aard van het vraagstuk naar voren. Figuur 6.9[39] laat bijvoorbeeld zien welke factoren naar verwachting invloed zullen hebben op het gebruik van data voor onderwijskundige beslissingen.

Het is dan ook noodzakelijk om een implementatieplan te maken waarin niet alleen aandacht is voor de ontwikkeling van de centrale proeven, maar expliciet ook voor de inbedding in het veld, professionalisering en de schoolcultuur waarin het systeem ingezet wordt. Dit zijn allen aspecten die de kans van slagen van de inzet van de resultaten voor interne kwaliteitszorg vergroten.

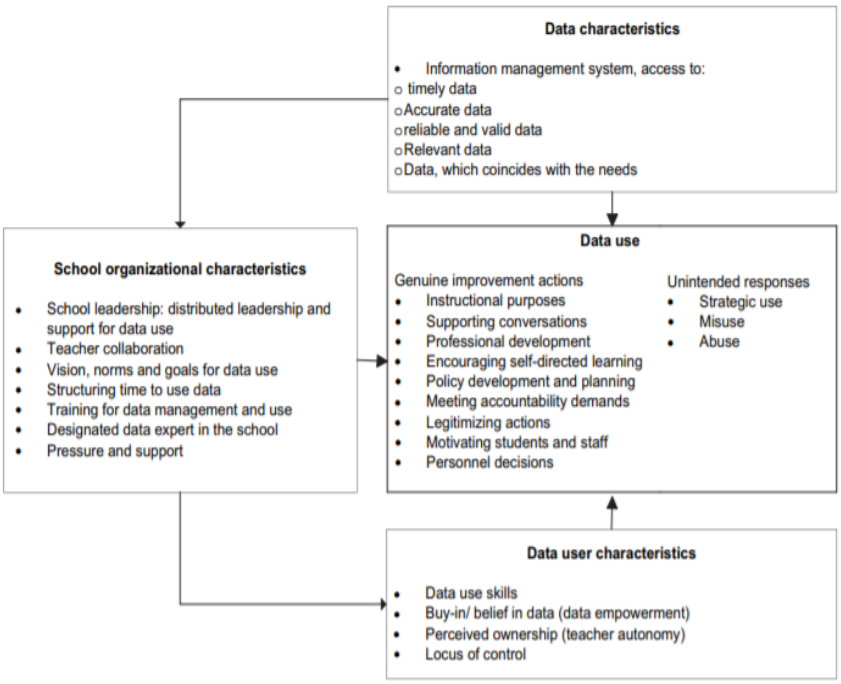


Fig. 1. Factors hypothesized to influence data use.

Figuur 6.9: Verwachte factoren van invloed op gebruik data in onderwijs

[38]Hopster-den Otter, D., Muilenburg, S.N., Wools, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2019). Comparing the influence of various measurement error presentations in test score reports on educational decision making. *Assessment in Education: Principles, Policy & Practice*, 26(2), 123-142.

[39]Figuur uit 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. `http://dx.doi.org/10.1016/j.tate.2009.06.007`.

Op basis van het schoolbeleid kan ook verder bepaald worden waar de school een grotere behoefte aan heeft, waar wil de school meer over weten ten aanzien van de rapportage: Hoe moet ik de toetsresultaten analyseren en interpreteren? Hoe toets ik 'op maat'? Hoe herken ik schoolbrede zorgsignalen? Wat kan ik nog meer uit de rapportage modules halen? Hiervoor kunnen cursussen ingezet worden voor leerkrachten, interne begeleiders en directieleden. De cursussen kunnen verschillende vormen aannemen. Het is mogelijk om aan de hand van interactieve werkvormen een team (onder begeleiding van een trainer) met elkaar in gesprek te laten komen over het waarom en hoe van toetsen op hun school. Met als doel om samen tot een goed en doordacht toetsbeleid te komen.

Dergelijke workshops kunnen twee uur duren, maar ook meer dagdelen. De kosten die hieraan verbonden zijn, betreft vooral de tijd die de trainer en het team ermee bezig zijn[40]. De effectiviteit is moeilijk sterk de kwantificeren aangezien er geen voor- en nametingen bekend zijn. Wel zijn cursussen geëvalueerd door middel van evaluatieformulieren, waarbij onder andere gevraagd is naar de kwaliteit en relevantie en naar de inhoud van cursus en de ervaren leerwinst. De resultaten zijn overwegend positief tot zeer positief.

Een meer grootschalige aanpak is via *Massive open online courses* (MOOCs). Ook aan het opzetten van een dergelijke cursus zitten kosten verbonden. Als de cursus vooral instructies betreft dan is het opzetten ervan goedkoper, maar het is bekend dat deze cursussen effectiever zijn als er sprake is van enige interactie. Deze interactie kan opgezet worden door cursisten onderling interactief aan de slag te laten gaan met het cursusmateriaal. Dit kan het simpelst georganiseerd worden binnen een school, maar uiteraard is het ook mogelijk om een netwerk op te zetten waarbij cursisten van verschillende scholen samenwerken[41]. Meer interactie kan verkregen worden door opdrachten te maken, die de clusters van leerkrachten samen kunnen maken, en die door cursusleiders nagekeken kunnen worden. Het is evident dat hoe meer van dergelijke opdrachten meegenomen worden, de cursussen ook duurder zijn.

Ook is online zelfstudie mogelijk. Materiaal (instructies, teksten, video's) kan dan online bekeken worden. Dit kan dan gecombineerd worden met online zelf-tests over verschillende onderwerpen met automatisch scoorbare items. Op die manier kunnen gebruikers zien waar zij extra ondersteuning kunnen gebruiken.

De kosten van een dergelijke opzet van ondersteuning zal mede afhangen van de eisen die gesteld worden. Ook hier geldt dat hoe meer deze zelftests als high-stakes worden gezien, hoe duurder de ontwikkeling is. In de meest goedkope vorm kunnen online allerlei gratis tools[42] ingezet worden. Als er meer eisen gesteld worden, zoals dat de ontwikkeling gevolgd moet worden, of dat het beheer ervan in handen moet liggen van de overheid, dan lopen de kosten uiteraard op. Tot slot kan opgemerkt worden dat het opzetten van ondersteunende cursussen ook niet in een dag af moet zijn. Als er een budget voor ontwikkeling is dan kan dit stap voor stap opgebouwd worden.

[40] Meer informatie over mogelijke cursussen is te vinden op `https://www.cito.nl/onderwijs/primair-onderwijs/training-advies/aanbod`. Informatie over de kosten per soort cursus is te vinden in `https://formulieren.cito.nl/common/formulieren/2018/po_so_training_en_advies_18_19`.

[41] Dit is door de huidige realiteit waarbij men steeds meer gewend is aan online samenwerking steeds minder een vreemd concept dan het ooit was en helpt ook om scholen van elkaar te laten leren.

[42] Voor een overzicht van simpele vormen hiervan: `https://www.primaonderwijs.nl/educatieve-apps-tools/online-quiz-tools-voor-in-de-klas`.

III

Scenario's

7. Scenario's voor toetsing in Vlaanderen

Het streven is om centrale toetsen af te nemen bij alle leerlingen op een viertal momenten in hun schoolloopbaan. Daarbij wordt hun vaardigheid in Nederlands (begrijpend lezen, schrijven en grammatica) en wiskunde gemeten. Het doel van de centrale toetsen in Vlaanderen is het verbeteren van de onderwijskwaliteit. Het bovenstaande geeft globaal antwoord op de drie kernvragen die we in Hoofdstuk 2 geïdentificeerd hebben: wie, wat, waarom. Van deze drie vragen is de vraag naar het waarom (ofwel het doel van de toetsen) het belangrijkst. Een proefafname leidt altijd tot een rapportage. Die rapportages leiden tot vervolghandelingen en beslissingen. Welke daarvan leiden tot het doel "onderwijskwaliteit verbeteren"? Hoe zouden de rapportages goed gebruikt kunnen worden? Antwoorden op dergelijke vragen bepalen wat er in de rapportages moet komen te staan. Het 'wie' en 'wat' er getoetst wordt, moet aansluiten bij het doel van de toetsen, om de juiste beslissingen en handelingen te verkrijgen. Het doel van de toetsen bepaalt ook het aggregatieniveau van de rapportages (leerlingen, klassen, scholen, en systeem) en het type analyses dat gedaan moet worden. De uit te voeren analyses betreffen de analyses noodzakelijk voor de bepaling van leerwinst en toegevoegde waarden (Hoofdstuk 4), maar ook of de resultaten relatief, absoluut of anderszins afgebeeld moeten worden (Hoofdstuk 6). Om deze analyses te kunnen doen, moeten ook de opgaven waaruit de toetsen bestaan aan voorwaarden voldoen. De toets moet dusdanig zijn samengesteld dat de gewenste analyses mogelijk zijn (Hoofdstuk 5). Dus om de beoogde doelen te bereiken, zou je voor het ontwerp van de centrale toetsen eigenlijk achteraan moeten beginnen: welke handelingen moeten er volgen uit de rapportages?

In dit hoofdstuk (Sectie 7.2) keren we weer terug bij de vragen uit het bestek waar we in Hoofdstukken 4, 5 en 6 uitgebreid antwoord op hebben gegeven. In die hoofdstukken zijn per vraag antwoorden gegeven vanuit een wetenschappelijk kader, met de Vlaamse praktijk als kader voor de beantwoording. Tevens zijn per vraag ook scenario's beschreven, ieder met voor- en nadelen, evenals kostenindicaties. In dit hoofdstuk komen we terug op

deze vragen, maar nu vanuit een meer praktisch oogpunt. Wat zijn nu de dilemma's, de keuzes en mogelijke stappen en oplossingen (scenario's) die passen in de Vlaamse context? Dat betekent dat de vragen die hier beantwoord worden niet altijd geheel gelijk zijn aan de vragen uit het bestek, die in de eerdere hoofdstukken beantwoord zijn. Het zijn deels ook de vervolgvragen die volgen uit de antwoorden, en de praktische dilemma's waar de gegeven antwoorden toe kunnen leiden. De kosten die gerelateerd zijn aan de scenario's zijn te verdelen in vaste en de variabele kosten, waarvan een overzicht wordt gegeven in Sectie 7.3.

Voordat we deze scenario's en dilemma's beschrijven, kijken we eerst opnieuw naar de dimensies die in Hoofdstuk 2 besproken zijn (Sectie 7.1). Dat zijn de dimensie van de formatieve toetsing en summatieve toetsing, en die van de toetsen met lage en hoge belangen voor de belanghebbende. In Hoofdstuk 2 leidde dit tot extreme scenario's, als tegenstellingen, die toegepast kunnen worden op diverse onderwijsniveaus (leerling, leerkracht, school). In dit hoofdstuk kijken we ook naar de middenweg daartussen. Hierbij wordt balans gezocht door op diverse aspecten tussen de twee uitersten in te kiezen. We beginnen met de schets 'achteraan', dus bij het resultaat van de toets (de rapportage). Daarna worden diverse aspecten die al aan de orde kwamen genoemd. Afhankelijk van de voorkeuren bij de vervolghandelingen, kan de koers voor de onderliggende aspecten dan uitgezet worden.

Het slot van dit hoofdstuk wordt gevormd door een bespreking van een grove tijdsplanning van toetsontwikkeling op basis van afhankelijkheden (Sectie 7.4), enkele praktische aanbevelingen (*best practices*) bij de implementatie en randvoorwaarden voor succesvolle implementatie die bij alle gekozen scenario's het slagen van de centrale toetsen kunnen bevorderen.

7.1 De scenario's naast elkaar

7.1.1 Handelingsscenario's op basis van rapportages

In deze paragraaf schetsen we enkele vervolghandelingen of acties die gedaan kunnen worden naar aanleiding van rapportages op diverse niveaus van aggregatie. Alle hebben het doel om onderwijskwaliteit te verbeteren. Op alle niveaus (van leerling tot en met systeem) kunnen rapportages gebruikt worden om het onderwijs te ondersteunen (formatief toetsen) of te evalueren (summatief toetsen). De handelingen die passen bij dergelijke extreme scenario's worden hieronder geschetst. Op ieder rapportageniveau is natuurlijk ook een middenweg denkbaar, waarbij handelingen zowel kenmerken van een formatief scenario als van een summatief scenario hebben. Een overzicht van deze scenario's is te vinden in Tabel 7.1.

Rapportages op leerlingniveau

Op basis van een rapportage op leerlingniveau zou je in een formatief scenario een individueel leerplan kunnen uitstippelen. De rapportage geeft aan welke taken net boven de huidige vaardigheid van een leerling liggen, op ieder van de (sub-)domeinen. In een summatief scenario wordt het leerproces achteraf geëvalueerd. Een rapportage op leerlingniveau bepaalt dan of een leerling voldoende heeft opgestoken in de afgelopen periode, om de leerstof van een volgend jaar of een vervolgopleiding aan te kunnen. Een middenweg daartussen is

Rapportage-niveau	Formatief scenario	Middenweg	Summatief scenario
Leerling	Uitstippelen individueel leerplan	Advies bij keuze uit een beperkte set van vervolgtrajecten	Bepalen van doublering of diplomering
Leerkracht*	Sterkte-zwakte-analyse (SWOT) uitvoeren	Bonus bij uitvoeren van zelfreflectie	Bepalen van salariëring of baanzekerheid
School	Ouders kiezen een type school dat bij hun kind past; Regionaal of lokaal overleg over; verklaringen voor prestaties	Schoolbegeleiding	Sluiting van zeer zwakke scholen. Beloning van goede scholen
Landelijk	Infrastructuur voor kwaliteitsverbetering opzetten en onderhouden	Transparantie bieden over landelijke prestaties	Richtlijnen voorschrijven; Financiering herverdelen

* Rapportages op leerkracht-niveau kunnen misleidend zijn.

Tabel 7.1: Rapportage per doelgroep voor de drie verschillende doel-scenario's

dat er geen individuele leerplannen zijn, maar een beperkt aantal vervolgtrajecten, en dat de leerlingrapportage adviseert welk vervolgtraject het best past bij een leerling.

Vooralsnog staat in het bestek van het Steunpunt dat de resultaten op de toets kunnen worden meegenomen in de globale beoordeling van de leerling, maar niet doorslaggevend zijn in het kader van studievoortgang en -oriëntering. Wel wordt gesteld dat de rapportage door scholen gebruikt moet kunnen worden als een van de mogelijke informatiebronnen in de beoordeling van de leerling. Wat hier verdere consequenties van zijn, welke mogelijke keuzen dit oplevert komt aan bod in Paragraaf 7.2.1.

Rapportages op leerkrachtniveau

In een formatief scenario zou de leerkracht zijn of haar rapportage kunnen gebruiken om zichzelf een spiegel voor te houden en te kijken welke sterke en zwakke punten hij of zij didactisch gezien heeft. De sterke punten kan de leerkracht dan inzetten om collega's te ondersteunen, en voor de zwakke punten zoekt hij of zij zelf ondersteuning. Een leerkrachtrapportage zou anderzijds ook door zijn of haar leidinggevende gehanteerd kunnen worden om te bepalen of de leerkracht contractverlening of salarisverhoging kan verwachten. In dat geval schetsen we een summatief scenario. Bij een middenweg tussen deze twee uitersten kan het uitvoeren van een sterkte-zwakte analyse aangemoedigd worden, door een bonus toe te kennen indien een leerkracht deze aantoonbaar heeft uitgevoerd.

Rapportages op leerkrachtniveau kunnen echter al snel misleidend zijn om meerdere redenen. Een leerkracht is ten eerste afhankelijk van het niveau van de leerlingen die hij/zij onderwijst. Een oordeel op leerkrachtniveau kan dus pas na meerdere jaren gegeven worden. Een afrekening op prestaties kan ook het nadelige neveneffect hebben

dat leerkrachten niet meer in achterstandsgebieden willen werken, ook na correctie voor achtergrondvariabelen. Dat is omdat ten eerste bij lange na niet alle variabelen verzameld en meegenomen kunnen worden om de achterstand voldoende te verklaren[1]. Ten tweede, zeker in het secundair onderwijs of als er meer dan één jaar tussen de metingen zit, is het moeilijk om één leerkracht als veroorzaker van de prestaties van de leerlingen aan te wijzen, omdat een leerling door meerdere leerkrachten onderwezen wordt. Heel strikt genomen zou dan het niveau van de leerling als deze start met lessen bij de leerkracht bekend moeten zijn, en niet van een jaar daarvoor. Ten derde zal bij een relatieve beoordeling er altijd één de zwakste zijn, maar dat betekent niet dat deze leerkracht slecht functioneert: ervaring kan ook een rol spelen. Een ander probleem voor een oordeel op basis van een leerkrachtrapportage is dat de taken van de leerkracht verder gaan dan alleen de vaardigheden voor Nederlands en wiskunde. Ten slotte hangt het gedrag van leerkrachten sterk af van het systeem waarbinnen hij/zij werkt, en de keuzevrijheid die hij/zij daarbinnen heeft.

Een summatieve rapportage op basis van resultaten van leerlingen, met directe consequenties voor salariëring of baanzekerheid, is zodoende af te raden. Zelfs als het niet de bedoeling is dat de beoordeling van leerkrachten enkel op basis van de toetsresultaten zou gebeuren. De prestaties zoals gemeten binnen de centrale toetsen kunnen beter in het geheel geen onderdeel van de individuele beoordeling van leerkrachten zijn. Een beoordeling van leerkrachten kan wel gaan over de manier waarop een leerkracht omgaat met de leerkrachtrapportage, waarmee het een procesbeoordeling wordt. Gebruikt de leerkracht de resultaten op een nuttige manier? Voert de leerkracht een SWOT-analyse uit? Helpt dit de leerkracht de lessen beter maken? De beoordeling van de leerkracht betreft dan niet de resultaten in de rapportage, maar wat de leerkracht doet naar aanleiding van de rapportage.

In een formatief scenario kan een leerkracht, ook als de resultaten goed zijn, met behulp van een leerkrachtrapportage kijken hoe de lessen verbeterd kunnen worden. Als de rapportage een klassenoverzicht is, met de resultaten van individuele leerlingen naast elkaar, kan de leerkracht op basis daarvan aandacht geven aan leerlingen op het niveau dat de leerling behoeft. Als de rapportage informatie omvat op subschaalniveau – wat voor een meting van een individuele leerling niet betrouwbaar genoeg is, maar geaggregeerd over een klas wel bruikbaar kan zijn – kan ook specifiek binnen een vak gekeken worden. Als bijvoorbeeld blijkt dat bij wiskunde in het lager onderwijs[2] de prestaties binnen de klas voor getallen, meten en meetkunde op een vergelijkbaar niveau zijn, maar dat de strategieën en probleemoplossende vaardigheden[3] daarbij achter blijven, dan zou de leerkracht daar in de lessen (voor komend jaar) rekening mee kunnen houden. Door een dergelijke interventie kan ook die vaardigheid op niveau komen. Hoe de toetsen en pedagogisch handelen beter met elkaar te integreren, wordt besproken in Paragraaf 7.2.4.

[1] Bij een regressiemodel is meestal nog zoveel variantie onverklaard dat het evident is dat de correctie niet dusdanig is dat de correctie voldoende is om het overgebleven geheel een en al toe te schrijven aan het handelen van de leerkracht.

[2] We gebruiken hiervoor nu de huidige omschrijving binnen lager onderwijs zoals die nu op `https://www.onderwijsdoelen.be/` gegeven worden als eindtermen bij wiskunde. Deze zullen in de toekomt aangepast worden, met ook consequenties voor de te gebruiken schalen.

[3] We gaan er nu van uit dat "attitude" geen onderdeel van het meten van de vaardigheden zoals beoogd met de centrale toetsen zal zijn.

Rapportages op schoolniveau

Schoolhoofden of –besturen kunnen in een formatief scenario ook een sterkte-zwakteanalyse uitvoeren. Hierbij kan ook intercollegiaal overleg met andere scholen plaatsvinden om verklaringen voor afwijkende prestaties te vinden, en mogelijk betere werkwijzen te ontwikkelen. De focus is hierbij vooral op het onderwijs in relatie tot de inhoud van de toets. Het betekent dat de informatie op klassenniveau meer in samenhang onderzocht wordt ("betreft het een enkele klas of jaarlaag, of is dit iets wat over klassen heen gevonden wordt?"), met als doel om schoolbreed de kwaliteit van het onderwijs te verbeteren ("wat kan de school doen om de kwaliteit te verbeteren of te behouden?").

In een sterk summatief scenario wordt een school ter verantwoording geroepen aan de hand van de schoolrapportage, met mogelijk financiële consequenties of sluiting als gevolg. Zeer goede scholen kunnen aan de andere kant in een dergelijk summatief scenario beloond worden. Toch zal er altijd meer kwalitatieve informatie ingewonnen moeten worden, en scholen moet de kans geboden worden zich te verbeteren. Resultaten van een enkel jaar kunnen namelijk uitschieters bevatten. Een school kan een jaar bijvoorbeeld een "moeilijke klas" hebben wat kan leiden tot een negatieve uitschieter in de prestaties. Een correctie op de achtergrondvariabelen is niet altijd voldoende om te vatten waarom die klas moeilijk is. Pas na de aggregatie over een aantal jaar kan pas echt beoordeeld worden of er sprake is van stabiele tegenvallende resultaten. Dergelijke summatieve scenario's zijn niet aan te raden, en zijn ook niet beoogd.

In de beoogde toepassing van centrale toetsen moeten scholen met significant tegenvallende prestaties in een vrij te kiezen begeleidingstraject stappen om de kwaliteit van hun onderwijs te verhogen. Of dit als summatief of als formatief beschouwd wordt, zal ervan afhangen of een dergelijk begeleidingstraject ervaren wordt als hinderlijk, of ondersteunend.

De vraag die open blijft staan, is ook of voor het aanbieden van een dergelijk traject gecorrigeerd moet worden voor achtergrondvariabelen of niet. Dus, als er budget is om 10% van de scholen ondersteuning te geven, moeten dat dan de 10% van de scholen zijn met de slechtste prestaties in Vlaanderen, of de 10% van de scholen zijn met de slechtste prestaties, gecorrigeerd voor belangrijke achtergrondvariabelen? Er zal overlap tussen deze groepen zijn, maar ze zijn niet identiek. Dit dilemma is eerder ook al beschreven in Hoofdstuk 6. De vraag welke scholen extra begeleiding krijgen, wordt niet opgelost door de proeven anders in te richten. Het antwoord is afhankelijk van de visie op wat de meeste vooruitgang voor het Vlaamse onderwijs als geheel biedt. Het is een afweging tussen hoeveelheid kosten, verwachte effectiviteit en maatschappelijke en onderwijsvisie. Uitgebreid onderzoek dat precies op maat is voor deze vraag is niet beschikbaar.

Het weergeven van de schoolresultaten heeft ook een functie die ligt tussen summatief en formatief. Dit betreft het gebruik door ouders van de weergave van resultaten. Door scholen kan dit als sterk summatief beschouwd worden, wanneer zij bij lagere prestaties meer of minder leerlingen kunnen verwachten. In een formatief scenario kunnen rapportages op schoolniveau gebruikt worden door ouders om te kijken op welke aspecten een school anders is dan andere scholen, en of dat bij hun kind past. Denk daarbij bijvoorbeeld aan de relatieve aandacht die een school aan de diverse domeinen of subdomeinen geeft, welke kan blijken uit de prestaties op de diverse (sub-)domeinen. In dit geval wordt een schoolrapportage ingebed in een bredere rapportage over de school, die ook andere

kenmerken beschrijft, naast de behaalde prestaties.

Een schoolrapportage wordt meer summatief wanneer de resultaten in de rapportage relatief gepresenteerd worden. Dat betekent dat de prestaties van de school vergeleken worden met andere scholen, al dan niet rekening houdend met de kenmerken van de school. Hoe deze relatieve vergelijking kan plaatsvinden en wat de voor- en nadelen hiervan zijn, is eerder beschreven. Als een ranking van scholen maximaal vermeden dient te worden, dan is een relatieve weergave van de schoolresultaten sterk af te raden. Een degelijke relatieve rapportage werkt op basis van een ordening van (gemiddelde leerling-)prestaties, en is daarmee per definitie een ranking. Dit kan genuanceerd worden door scholen te stratificeren op basis van achtergrondvariabelen, of door de rangordening te categoriseren ("rood, oranje, of groen"), maar zal tot een vorm van ranking leiden. In Hoofdstuk 6 zijn al enkele alternatieven geboden (meetfout weergeven). In Paragraaf 7.2.4 zullen enkele scenario's en mogelijkheden besproken worden.

Rapportage op systeemniveau

Op landelijk (deelstaat) niveau kan een rapportage summatief gebruikt worden om richtlijnen voor te schrijven of wijzigingen in onderwijsaanpak door te voeren door financiering anders te verdelen. Zeker in het geval er door de tijd dalingen in niveau worden ervaren, of wanneer de rapportages in een internationaal kader worden geplaatst. Dat betekent dat meer of minder aandacht en geld gegeven wordt aan diverse fasen in het onderwijs, of voor diverse (sub-)domeinen.

In een formatief scenario wordt een landelijke rapportage gebruikt om gericht een infrastructuur in te richten om scholen en leerkrachten te ondersteunen. Op basis van de resultaten kunnen praktische aanwijzingen en richtlijnen ontwikkeld worden voor goed pedagogisch handelen, mits het pedagogisch handelen meegenomen wordt in de dataverzameling. Dit kan tevens maatwerk opleveren, doordat er inzichten verkregen worden onder welke omstandigheden (gezien de achtergrondvariabelen) bepaald pedagogisch handelen beter werkt dan ander gedrag.

Als een soort middenweg kunnen de rapportages gebruikt worden om publiekelijk verantwoording af te leggen over de prestaties in het onderwijs. Het beoogde gebruik van opvolging van de onderwijskwaliteit op systeemniveau is dat (trends in) resultaten op de toetsen aanknopingspunten moeten bieden voor de evaluatie en eventuele bijsturingen van het beleid. Waarschijnlijk zullen alle genoemde vervolghandelingen naar aanleiding van de landelijke rapportage wel terugkomen, waarmee de landelijke rapportage veelal niet doorslaggevend wordt voor een formatieve of summatieve benadering.

In zekere zin nemen de centrale toetsen de functie van peilingen over. Die hebben ook tot doel trends in kaart te brengen, en samenhang tussen resultaten, achtergrondvariabelen en pedagogisch handelen in kaart te brengen. In de meest rudimentaire vorm zullen de resultaten van de centrale toetsen niet eenzelfde inhoudelijke fijnmazigheid hebben als de peilingen. In Paragraaf 7.2.1 worden scenario's beschreven hoe de centrale toetsen meer ook meer de functie van peilingen kunnen hebben, daar waar het gaat om het verzamelen en rapporteren van relevante systeeminformatie.

7.1.2 Vormen van rapportages op basis van doelen per niveau

Naast de opdeling formatief – summatief is er ook de opdeling high-stakes – low-stakes. Deze maakt onderscheid tussen de belangen die gehecht worden aan de rapportages. Wanneer de toetsen grote gevolgen hebben, wordt er over het algemeen een groter belang aan gehecht (high-stakes toetsen). Over het algemeen worden summatieve toetsen als high-stakes gezien, en formatieve toetsen als low-stakes. Bij formatieve toetsen worden fouten gezien als zaken om van te leren. Uiteraard is dat juist belangrijk, echter in de het oog van de belanghebbende kan dat als minder "zwaar" gezien worden dan een summatieve toets. Voor nu beschouwen we summatieve toetsen als high-stakes en formatief gebruik als low-stakes. Het tussenliggende gebruik zit er wat betreft belang tussenin. In Paragraaf 7.2.1 gaan we erop in dat de belangen voor verschillende belanghebbenden ook verschillend ervaren kunnen worden.

In deze paragraaf schetsen we hoe de rapportages op leerling- en schoolniveau eruit komen te zien in een summatief en formatief scenario. Ook het gebruik door de leerkracht komt hierbij aan bod. Deze opzet van de rapportages staat dan optimaal ten dienste van de vervolghandelingen die hierboven geschetst zijn.

- **"Formatief - leerling"**
  De rapportage biedt de richtlijnen om het onderwijs aan te passen aan de onderwijsbehoeften (en denkfouten) van de leerling. Er staan illustraties in met voorbeeldopgaven die een leerling wel of niet beheerst. Dit soort rapportages is vooral toepasbaar als de school het vervolgonderwijs aanbiedt, of de rapportage deel uitmaakt van een 'warme' overdracht naar een andere school, c.q. bij doorlopende leerlijnen.
- **"Formatief - school"**
  De rapportage biedt richtlijnen voor de school om het onderwijs doelmatiger in te richten. Bij welke domeinen zien we betere resultaten dan bij andere? Welke denkfouten komen veel voor? Veelal zal naast de afname van de proeven aanvullend onderzoek gedaan worden bij scholen, zodat succesvolle onderwijspraktijken geïdentificeerd kunnen worden.
- **"Formatief - leerkracht"**
  De rapportage is de tussenweg tussen dat van de leerling en dat van de school. Door middel van een klassenoverzicht met gegevens per leerling kan de leerkracht iedere leerling de benodigde ondersteuning bieden, terwijl geaggregeerde groepsgegevens ook in de schoolrapportage kunnen terugkeren.
- **"Summatief - leerling"**
  De rapportage wordt een soort toegangsbewijs voor vervolgonderwijs, met herkenbare cijfers. Dit kan motiverend werken, en kansengelijkheid bieden op individueel niveau. De (andere) school moet er wel belang aan hechten.
- **"Summatief - school"**
  De rapportage wordt een cijfermatige kwaliteitsbeoordeling van scholen. Dit kent een risico op ranglijsten en potentieel frauduleus gedrag van scholen[4]. Deze rapportages bieden de school zelf geen handvaten voor verbetering.
- **"Summatief - leerkracht"**

[4] In de eerdere hoofdstukken is vermeld hoe de beide nadelige gevolgen deels in te perken kunnen worden. Bij een summatief gebruik is het echter niet geheel uit te sluiten.

De rapportage wordt een cijfermatige kwaliteitsbeoordeling van leerkrachten op basis van de leerlingresultaten. De risico's en nadelen zijn in aard vergelijkbaar, maar in impact nog groter dan bij pure summatieve rapportage op schoolniveau. Deze vorm van rapportage is dan ook sterk af te raden.

Er zijn vier meetmomenten in de schoolloopbaan van leerlingen. Op individueel niveau kan per meetmoment een formatieve of summatieve rapportage beter geschikt zijn. Voor de school zelf lijkt een pure summatieve rapportage op schoolniveau op geen enkel moment van toegevoegde waarde, en is een formatieve rapportage geschikter om de kwaliteit van het onderwijs te verbeteren. Externe belanghebbenden kunnen wellicht wel behoefte hebben aan een dergelijke summatieve rapportage. Dit kan vanuit het perspectief van toezicht een rol spelen, of vanuit de rol van de ouders die een school voor hun kinderen uitzoeken. De nadelen van een dergelijk gebruik zijn genoemd, evenals mogelijke oplossingen om over-interpretatie van (tijdelijke) negatieve resultaten tegen te gaan.

7.1.3 Samenvatting aspecten van toetsontwerp binnen scenario's

Diverse aspecten van toetsontwerp, die in Hoofdstukken 4, 5 en 6 aan de orde kwamen, kunnen anders ingevuld worden, al naar gelang de voorkeur bij formatieve of summatieve toetsing ligt. De gewenste vervolghandelingen impliceren een bepaald type rapportage. Om een formatieve of summatieve rapportage te kunnen verstrekken, moeten bepaalde aspecten anders ingericht worden binnen de toetsing. Niet alle aspecten die besproken zijn, komen voor zo'n tweedeling in aanmerking. De onderstaande aspecten kennen wel zo'n andere invulling. Een overzicht van deze aspecten en de scenario's is te vinden in Tabel 7.2.

Doel van toetsing

Ten eerste is er het doel van de toetsing. Bij formatieve toetsing is het doel om het onderwijs zo aansluitend mogelijk te maken op waar de leerling behoefte aan heeft, zodat deze zich zo optimaal mogelijk verder kan ontwikkelen. Het niveau van het aangeboden les- en oefenmateriaal past dan bij de vaardigheid van de leerling: het is waar de leerling dan aan toe is. Bij summatieve toetsing is het doel van toetsen om eerlijke kansen te bieden aan leerlingen, bij kwalificatie en selectie. Toetsen kunnen functioneren als poortwachters, maar als er geen toetsen zijn, betekent het niet dat er geen poortwachters meer zijn voor doorstroom. Die beslissingen worden dan alleen niet genomen op basis van gestandaardiseerde, gevalideerde betrouwbare metingen, maar op basis van oordelen, bijvoorbeeld van de docent. Het is overigens verre van af te raden om het docentoordeel te negeren omdat toetsen ook niet alles meten wat relevant is, maar ook moet onderkend worden dat het menselijk oordeel ook niet perfect is en ondersteuning kan gebruiken, zoals ook in Hoofdstuk 2 is betoogd. Toetsen kunnen bijvoorbeeld leerlingen helpen die door docenten onderschat worden. Het biedt hen een gelegenheid om te laten zien wat ze kunnen. Een ander doel van toetsen is om leerlingen te motiveren voor onderwerpen waarvoor ze uit zichzelf mogelijk niet gemotiveerd zijn. Veelal vrezen leerkrachten dat leerlingen de relevantie van een formatieve toetsing niet zien, en daardoor niet gemotiveerd zijn. Bij een summatieve toets wordt de relevantie eerder ervaren door de leerling zelf. De consequenties van een toetsing zullen dus goed moeten worden uitgelegd aan leerlingen om hen te motiveren. Het combineren van een formatieve functie van een toets voor een

Aspect niveau	Formatief scenario	Middenweg	Summatief scenario
Doel toetsing	Passend onderwijs	Motivatie verhogen met consequenties	Eerlijke kansen bij kwalificatie en selectie
Validiteit	Construct validiteit	Construct validiteit; predictieve validiteit	Predictieve validiteit; construct validiteit
Vereiste betrouwbaarheid	Matig tot hoog	Hoog	Hoog tot zeer hoog
Toetsverversing	Lage frequentie, lage beveiliging	Matige frequentie, matige beveiliging	Alle leerlingen maken dezelfde doelen voor vergelijkbaarheid
Selectie onderwijsdoelen	Rotatie van doelen over leerlingen en jaren	Deels rotatie, deels vaste onderdelen	Alle leerlingen maken dezelfde doelen voor vergelijkbaarheid
Adaptieve toetsing	Zinvol om aan te sluiten bij niveau van leerling	Beperkte adaptiviteit	Mogelijk, mits er goede communicatie is over de betekenis van de rapportage
Net- of koepel-specifiek materiaal	Vervangend gebruik	Deels vervangend, deels vaste onderdelen	Aanvullend gebruik
Rol papieren toetsing	Altijd bij jonge of niet-digitaal-ervaren kinderen	Beschikbaar, naast digitale toets	Calamiteitenvariant
Illustratie toetsresultaten	Illustratie met vaardigheidsschaal en daadwerkelijk gedrag	Landelijke prestaties illustreren met daadwerkelijk gedrag	Illustratie door vergelijking met landelijke prestaties
Rapportage van subdomeinen	Veel aandacht voor diversiteit van onderwerpen	Beperkte aandacht voor diverse onderwerpen	Grote nadruk op één onderliggende vaardigheid

Tabel 7.2: Invulling van de kwaliteitsaspecten voor de drie verschillende doel-scenario's

leerling, met een summatieve functie voor een school kan daarbij voor uitdagingen zorgen. Het dilemma wanneer de toets voor de leerling als low-stakes gezien wordt, maar voor de school min of meer high-stakes is, komt verder aan bod in Paragraaf 7.2.1.

Validiteit

Bij een formatieve toetsing is het gewenst dat de toets vooral constructvaliditeit heeft. Het is belangrijk dat de toets de vaardigheid meet die onderwezen wordt. Dit betekent dat de toetsinhoud en items vrij nauw zullen moeten aansluiten bij het onderwijsmateriaal. De constructvaliditeit is ook van belang als een summatieve toetsing controleert of bepaalde eindtermen in voldoende mate beheerst worden: de eindtermen moeten goed geoperationaliseerd worden. Er zijn echter ook summatieve toepassingen waarbij de aansluiting van toetsinhoud op de gegeven onderwijsinhoud minder van belang is. Dat is vooral wanneer een hoge predictieve validiteit (hoge voorspellingswaarde) van een toetsing van belang

is, en de toetsing goed onderscheid maakt tussen leerlingen die in de volgende fase van het onderwijs meer of minder succesvol zullen zijn. Vooralsnog is dat niet het doel van centrale toetsen in Vlaanderen.

Betrouwbaarheid

De consequenties bij formatieve toetsing zijn iets minder groot, en kunnen door de leerkracht in de volgende fase van het onderwijs bijgesteld worden: er is de mogelijkheid de deficiëntie weg te werken. Mocht de geschatte vaardigheid iets groter of kleiner zijn dan waargenomen, dan is dat niet heel erg. Zodoende is de betrouwbaarheid van een toets binnen een formatieve toetsing iets minder van belang dan bij een summatieve toetsing[5]. Lagere eisen aan de betrouwbaarheid van toetsen, betekent dat de toetsen over het algemeen korter kunnen zijn en wat minder kritisch ontwikkeld kunnen worden. Dit scheelt in de kosten. Wanneer de betrouwbaarheid van een gerapporteerde score op individueel niveau voldoende groot is om een uitspraak te doen, dan is dat zeker het geval op geaggregeerd niveau. Als het op individueel niveau niet mogelijk is om betrouwbare metingen te doen, kan het mogelijk zijn dat dit wel mogelijk is op geaggregeerd niveau[6]. Bijvoorbeeld bij de rapportage van subschalen is het relevant om hier onderscheid in te maken.

Toetsverversing en beveiliging

Toetsverversing en beveiliging van materialen en afnamen zullen bij kleine belangen minder intensief hoeven te zijn dan bij grote belangen. Zoals hiervoor al is aangegeven, zijn de belangen bij formatieve toetsing over het algemeen lager dan bij summatieve toetsing. De kosten kunnen daarom ook op dit vlak bij formatieve toetsing lager uitvallen. Als het om vrees voor fraude gaat, is het dan ook minder noodzakelijk dat er een heel erg grote itembank is, omdat opgaven minder snel gedeeld worden.

Selectie onderwijsdoelen

Om in een formatieve context meer betekenisvolle feedback aan leerlingen te kunnen geven, is een grote en diverse itembank handig. Er kan dan in meer detail ingegaan worden op de behoeften van de leerling. Dan kan zowel de leerling met een lage vaardigheid die extra ondersteuning nodig heeft, geholpen worden, als de leerling met een hoge vaardigheid die een extra uitdaging wil hebben. Uiteraard helpt dit ook alle leerlingen daar tussenin, die wellicht op deelvaardigheden een gedifferentieerde aanpak nodig hebben. Een grote itembank maakt een dergelijke brede aanpak mogelijk. Het kan de leerlingen goed helpen en daarmee de kwaliteit van het onderwijs versterken. Ook als de toetsen niet lang genoeg zijn voor dergelijke gedetailleerde informatie op individueel niveau geeft een grote itembank wel de mogelijkheid op schoolniveau dergelijke informatie te verzamelen en te rapporteren. Het geaggregeerde karakter van de data maakt het mogelijk op gedetailleerd niveau voldoende betrouwbare ondersteuning te geven. Subschalen krijgen dan een bredere inhoudelijke dekking, in plaats van één die gebaseerd is op enkele items. Een ruimere selectie aan opgaven in een grote itembank maakt het mogelijk om een vaardigheid goed

[5] In de COTAN-richtlijnen worden toetsen met als doel het formatief bijhouden van de vaardigheid, gezien als toetsen met een minder groot belang, en de eisen voor de betrouwbaarheid zijn daarbij 0,10 punt lager.

[6] Zie de COTAN-richtlijnen Hoofdstuk 5; `https://www.cotandocumentatie.nl/cotan/beoordelingssysteem/`.

te kunnen operationaliseren.

Bij summatieve toetsen is de vraag naar gestandaardiseerde toetsen groter. De vergelijking tussen leerlingen wordt eerlijker ervaren als de toetsen identiek zijn. Als toetsvarianten alle op eenzelfde IRT-schaal passen, is dat technisch gezien niet noodzakelijk. Er zijn diverse factoren om voor enkele of meer varianten te kiezen. Een ervan is gemak waarmee verschillende varianten gemaakt kunnen worden: dat is digitaal simpeler dan op papier. Een ander is de acceptatie door het publiek, waarbij zeker wanneer de toetsen als zeer high-stakes worden gezien de neiging eerder zal zijn een enkele versie te gebruiken[7]. Dat is echter niet noodzakelijk aangezien er ook veel high-stakes toetsen met verschillende versies bestaan[8]. Vanuit de vergelijkbaarheid is het in ieder geval aan te raden om de verdeling van de aantallen en typen opgaven per onderwijsdoel vergelijkbaar te houden over de toetsen, zeker als de benodigde vaardigheden om deze doelen te bereiken onderling niet perfect met elkaar correleren. Ook bij de tussenvorm, waarbij de toetsen enigszins een summatieve functie hebben, is het aan te raden de dekking van de onderwijsdoelen gelijk te houden, met evenveel opgaven per onderwijsdoel per toets. Om de formatieve functie te dienen, kunnen specifieke opgaven wel over de toetsen verschillen. Dit kan ook zeer functioneel zijn voor rapportage op systeemniveau (zie ook Paragraaf 7.2.1).

Adaptieve toetsen

Bij een adaptieve toetsing krijgen leerlingen niet allen dezelfde variant voorgelegd, sterker nog, die kan zelfs individueel verschillen. Zoals reeds hierboven opgemerkt, kunnen leerlingen (en ouders) het bij high-stakes summatieve toetsing oneerlijk vinden als zij andere opgaven hebben gemaakt dan hun medeleerlingen. Zij denken dan geen gelijke kansen op een goed resultaat te hebben gehad. Als leerlingen van diverse cohorten ook met elkaar moeten concurreren voor kwalificaties of posities, dan zou de toetsinhoud ook tussen jaren zo min mogelijk mogen variëren. Dit betekent dat het lastig is om te roteren in toetsdoelen of subdomeinen van de toetsinhoud. Als echter de opgaven van de verschillende toetsen allen op een unidimensionele schaal passen, is het wel mogelijk om verschillende versies goed met elkaar te vergelijken, ook in een high-stakes summatieve context[9]. Het is dan wel noodzakelijk dat de uitleg en interpretatie van de gerapporteerde (vaardigheids)scores goed gecommuniceerd zijn naar alle betrokkenen (leerlingen, ouders en leerkrachten). Het is hierbij van belang dat voor leerlingen (en ouders) duidelijk is dat zij dezelfde uitdagingen hebben gekregen als de anderen. Weliswaar zijn toetsresultaten prima op één onderliggende vaardigheid af te beelden, maar het begrijpen en accepteren van dergelijke rapportages vergt een goede communicatiestrategie.

Bij formatieve toetsing biedt adaptiviteit vooral voordelen. De leerling krijgt immers opgaven op zijn of haar eigen niveau, wat de kans op toetsfrustratie sterk vermindert. Bij

[7] In Nederland zijn de centrale examens (laatste jaar voortgezet onderwijs; high-stakes toetsen) voor alle leerlingen die op één van de hogere niveaus examen doen per vak identiek. In het verleden waren de eindtoetsen (laatste jaar lager onderwijs) ook voor alle leerlingen identiek. Bij een veranderende functie van de toetsen (minder high-stakes) worden verschillende (versies van) toetsen gebruikt. Wel zijn de huidige eindtoetsen duidelijk voorschriften over de dekking van de vaardigheden (hoeveel opgaven per eindterm).

[8] Vooral bij digitale examens en toelatingstoeten wordt veel met verschillende versies gewerkt. Dit betreft examens in diverse vormen van onderwijs.

[9] Een voorbeeld hiervan zijn centrale examens in het middelbaar beroepsonderwijs (mbo) in Nederland. Ook bij de door de Nederlandse overheid uitgegeven Centrale Eindtoets wordt gewerkt met een adaptieve vorm van toetsen.

formatieve toetsen worden vaardigheidsscores dan ook makkelijker geaccepteerd als er geëxtrapoleerd wordt van geobserveerde antwoorden naar verwachte prestaties op (een relevante selectie van) de rest van de items in de itembank. Dit biedt de mogelijkheid om leerlingen niet alle subdomeinen voor te leggen, en toetsen dus relatief kort te houden, terwijl op geaggregeerde niveaus wel een breed scala aan subdomeinen gerapporteerd kan worden. Bij de tussenvorm tussen summatief en formatief geldt ook dat hoe meer de toetsen een formatief karakter hebben, hoe meer de toetsen adaptief kunnen zijn. Bij deze tussenvorm is het zoals eerder aangegeven wel van belang dat de inhoudelijke dekking vergelijkbaar is. Wanneer het doel is de vaardigheid van de leerling zo goed mogelijk te bepalen, kan door middel van adaptiviteit die opgaven geselecteerd worden die het meest geschikt zijn gegeven de (geschatte) vaardigheid van de leerling.

Tot slot kan opgemerkt worden dat als alleen het doel is om te controleren of een leerling een bepaald vastgesteld niveau al dan niet gehaald heeft, wat zowel in een formatieve als een summatieve vorm het geval kan zijn, dan is adaptief toetsen niet echt zinvol. De meest informatieve opgaven zijn dan toch de opgaven van het niveau rond die vaardigheidsgrens. De enige adaptiviteit ligt er dan in, in hoeveel opgaven men nodig heeft om vast te kunnen stellen of dat niveau al dan niet gehaald is (zie ook Paragraaf 7.2.3).

Net- of koepelspecifiek materiaal

Het gebruik van net- of koepelspecifiek materiaal wordt bij een summatieve toetsing vaak uitsluitend als aanvullende toetsing gebruikt. Dit omdat, zoals eerder gesteld, bij een summatieve toetsing het vaak als oneerlijk wordt ervaren als leerlingen andere opgaven maken dan hun medeleerlingen. Een gezamenlijk deel, dat landelijk vergelijkbaar is, zorgt dan voor de acceptatie van de summatieve toetsing. Bij een formatieve toets kunnen delen van het gezamenlijke programma vervangen worden door specifiek materiaal. Door analyse van de resultaten in combinatie met gerelateerde landelijke materialen kunnen de prestaties op het specifieke materiaal vertaald worden in prestaties op het landelijke materiaal.

Rol papieren toetsing

De rol van papieren toetsing is ook verschillend binnen een formatief scenario en een summatief scenario. Binnen een formatieve toetsing wordt de toetsing aangepast aan de ontwikkeling van een leerling. Als een leerling niet-digitaal geletterd is, bijvoorbeeld omdat hij of zij nog te jong is om met het afnameapparaat om te gaan, dan kan er de voorkeur aan gegeven worden om de toets op papier af te nemen. Een papieren en een digitale variant kunnen eventueel naast elkaar bestaan. Bij een summatieve toetsing worden leerlingen voorbereid om met een digitale afnameomgeving om te gaan. Het moet dan wel een deel van de voorbereiding op de toets zijn om ervoor te zorgen dat het probleem dat leerlingen onvoldoende digitaal geletterd zijn, minder speelt. Het is een risico dat de digitale omgeving niet functioneert op het moment van de high-stakes toetsing. Bij een dergelijke calamiteit biedt een papieren noodvariant dan uitkomst. Deze kan vrij snel op tafel gelegd worden, en de leerling ondervindt dan weinig hinder van het uitvallen van het systeem. Equivalentie van de papieren en de digitale variant moet dan wel aangetoond worden. In het eerste jaar van afname van centrale toetsen in Vlaanderen, waarbij er nog geen praktische ervaring is met de grootschalige uitrol van dergelijke proeven, is er een kans dat er ergens iets is mis gaat dat niet simpel op te lossen is. De kans dat er in het eerste

jaar van afname er scholen zijn die moeten terugvallen op papieren varianten van de toets is relatief groot[10]. Om ervoor te zorgen dat de papieren en de digitale variant equivalent zijn, moeten de opgaven wat betreft vorm ook niet te veel van elkaar verschillen. Het gebruik van de mogelijkheden die digitale afnamen bieden, zoals het per opgave toepassen van beeld en geluid, of opgaven waarbij de digitale handelingen gevolgd worden voor de scoring van de opgaven, zijn bij de papieren varianten problematisch. Om die reden is het aan te raden om zeker in het eerste jaar de digitale opgaven niet te veel te laten afwijken van wat er op papier mogelijk is. Als met de tijd bekend is dat de digitale afnamen bijna altijd een succes zijn, en papieren toetsen (vrijwel) niet meer nodig zijn om op terug te vallen, dan biedt dit ook ruimte om meer gebruik te maken van de digitale mogelijkheden bij de proeven.

Als bij een formatieve toets het systeem niet werkt op het geplande moment van toetsing, is het vaak niet erg als de toets op een later moment wordt afgenomen. De redenen om een papieren variant te ontwikkelen, verschillen dus tussen de twee scenario's. Hoe minder summatief de functie van de toets is, hoe minder cruciaal het is dat de toetsen strikt equivalent zijn. Dat betekent ook dat het dan mogelijk is opgaven op te nemen die gebruik maken van de digitale mogelijkheden, die niet op papier te repliceren zijn. Met een beperkte summatieve rol van de toetsen, bijvoorbeeld bij het vergelijken van scholen, blijft een gebrekkige equivalentie mogelijk wel een probleem.

Illustratie toetsresultaten

Rapportages zullen er anders uitzien in beide scenario's. De prestaties worden bij een formatieve toetsing vaak gerelateerd aan denkstappen, of denkfouten die met opgaven geïllustreerd kunnen worden. Bij summatieve toetsing worden prestaties vaker relatief of ten opzichte van één standaard geïllustreerd. Voorbeelden hiervan zijn gegeven in Hoofdstuk 6.

Rapportage van subdomeinen

Een rapportage bij een formatieve toetsing omvat vaak veel (sub)domein informatie. Op deze al dan niet onderling gerelateerde subdomeinen kunnen profielanalyses uitgevoerd worden. Daarentegen is een rapportage bij een summatieve toetsing veelal gericht op één eindoordeel, waarbij de onderliggende vaardigheden voldoende samen moeten hangen om een betrouwbaar eindoordeel op te kunnen baseren.

7.1.4 Praktische aanbevelingen bij de implementatie

In deze paragraaf geven we enkele tips om diverse scenario's succesvol te kunnen implementeren. Op individueel niveau is zowel het formatieve als summatieve scenario toepasbaar en is hier per meetmoment een keuze in te maken. Op schoolniveau lijkt het formatieve scenario geschikter en passender bij het doel om de kwaliteit van het onderwijs

[10]Ervaring in Nederland was dat bij de geplande digitale afname van de eindtoets in 2018, waarbij er ook een papieren toets was waar de scholen op terug konden vallen, de scholen in meerderheid de papieren variant gebruikten. Dit had ook te maken met de digitale infrastructuur in het lager onderwijs voor afname van toetsen met een summatieve functie. De ontwikkelingen op dat gebied hebben natuurlijk niet stil gestaan, en wanneer nu digitale toetsen op deze manier ingevoerd zouden worden, zal het aantal scholen dat de papieren toets gebruikt kleiner zijn (en dat is zeker zo nu scholen er meer ervaring mee hebben), maar te verwachten dat er een 100% digitale afname zou kunnen plaatsvinden op alle scholen is niet geheel realistisch.

te verbeteren, maar zal ook ingegaan worden op summatief gebruik van de toets.

Formatief – leerling

Bij het scenario 'formatief - leerling' lijkt het ons verstandig om een ruime afnameperiode te gebruiken. Dit maakt het mogelijk dat iedere school de toetsen op zinvolle momenten in hun onderwijs kunnen afnemen en de resultaten benutten. Een adaptieve toetsafname sluit beter aan bij het ontwikkelingsniveau van een leerling dan een vaste toetsvariant. Dit hoeft geen volledig adaptieve toets te zijn, een multi-stage toets biedt vaak al afdoende aanpassingsvermogen om geschikt te zijn voor alle leerlingen. Voor leerlingen met beperkingen is binnen dit scenario het advies om de afname individueel en door de leerkracht te laten plaatsvinden. De leerkracht kent de beperkingen van een leerling en de beste manier om daarmee om te gaan namelijk vaak het best. De resultaten kunnen dan ook door hem of haar binnen de context van de afname geïnterpreteerd worden. Voor de constructie van opgaven binnen een formatief-leerling scenario, is het belangrijk om de toetsing aan te laten sluiten bij de leerdoelen die op het betreffende moment in het curriculum onderwezen worden. Ook is gerichte constructie van foutieve antwoorden (afleiders) bij meerkeuzevragen aan te raden, omdat de foutieve antwoorden kunnen wijzen op systematische denkfouten bij leerlingen. Bij de rapportage kan op individueel niveau gebruik gemaakt worden van profielanalyses om de sterke en zwakke punten van een kandidaat goed in beeld te krijgen. In aanvulling daarop kan gekeken worden welke denkfouten systematisch voorkomen bij een leerling.

Summatief – leerling

Voor een meer summatieve toetsing op leerlingniveau is het cruciaal om de proeven breed gedragen te krijgen in de maatschappij en dat brede draagvlak te behouden. Het vermindert de kans op frauduleus gedrag als iedereen van mening is dat de proeven recht doen aan de vaardigheid van leerlingen en proportioneel benut worden. Een geleidelijke implementatie kan daarbij helpen. Bijvoorbeeld door eerst een jaar bij enkele scholen een pilot-afname te organiseren. Een jaar later kan bij alle scholen een pilot-afname plaatsvinden. In een derde jaar kan een integrale afname plaatsvinden, maar nog met enige vrijheid per school, voordat over gegaan wordt op een strikte, vergelijkbare afname-wijze bij alle scholen. Na iedere (pilot-) afname kan dan nog bijgesteld worden naar aanleiding van de opmerkingen vanuit het veld. Het is hierbij ook mogelijk de summatieve functie, oftewel de mate waarin het resultaat mee mag tellen bij de evaluatie van de leerlingen, te variëren over meetmomenten. Een moment waarbij een summatieve functie het meest toepasbaar lijkt, is aan het einde van de derde graad in het secundair onderwijs. De formatieve functie voor de leerling lijkt voor die groep ook beperkt. In hoeverre een score op een centrale toets dan formeel mee zou tellen, is er uiteraard een voor het (publieke) debat. Indien summatieve toetsing het gewenste scenario is, moeten leerlingen zich goed kunnen voorbereiden op de proeven, en kans hebben op een eerlijke afname en verwerking ervan. Er moeten dus voorbeeld-proeven beschikbaar zijn. Net zoals inzage-mogelijkheden om te controleren of de scoring op juiste wijze heeft plaatsgevonden. En herkansingsmogelijkheden als de leerling bij de eerste afname niet optimaal heeft kunnen presteren vanwege tijdelijke, persoonlijke omstandigheden. Indien de proeven digitaal en high-stakes zijn, is het aan te raden om een fall-back optie op papier achter de hand te hebben. Indien het digitale systeem uitvalt, kunnen de leerlingen zonder ernstige belemmeringen toch doorgaan met

hun proeven. Bij een summatieve toetsing zullen er voor leerlingen met beperkingen ook eerlijke afname-condities moeten zijn, die vergelijkbaar zijn met die voor reguliere leerlingen. Dit betekent dat er een set van aangepaste varianten moet worden ontwikkeld, die gestandaardiseerd kan worden afgenomen. Ook een leerling met beperkingen mag niet afhankelijk zijn van zijn of haar leerkracht voor de interpretatie van de geleverde prestatie.

Formatief – school

Ook op schoolniveau biedt een rapportage met een vergelijking tussen de prestaties op de diverse subdomeinen vaak een goede formatieve basis. Ook is het aan te raden te rapporteren of bepaalde denkfouten of -strategieën relatief veel voorkomen bij de leerlingen. Een identificatie van succesvolle strategieën is dan eenvoudiger uit te voeren. Een koppeling van de resultaten aan een onderzoek naar onderwijspraktijken en achtergrondvariabelen, biedt eveneens aanknooppunten om het onderwijs te verbeteren. Tot slot is het in een formatief scenario aan te raden om de schoolrapportages in een beveiligde omgeving te delen of bespreken met andere scholen. Dit verhoogt de kans dat succesvolle onderwijspraktijken geïdentificeerd worden, en overgedragen kunnen worden.

Summatief – school

De mate waarin een toets een summatief karakter heeft voor een school heeft grotendeels te maken met het gebruik. Zodra de gegevens van een school publiek gemaakt worden, en derden deze kunnen inzien, is de toets in ieder geval in zekere mate summatief. In het rapport is op diverse plekken beschreven hoe deze impact te verkleinen omdat een formele summatieve functie van de meting vooral nadelen en hogere kosten met zich mee zal brengen. Een strikt summatieve functie helpt de kwaliteit van het onderwijs ook niet verbeteren. Daarvoor is toch echt feedback nodig die kan leiden tot verbetering van de resultaten, waarmee de toets voor de school toch een meer formatieve functie heeft. Uiteraard, als een school jaren achter elkaar slechte resultaten haalt en niet tot verbetering komt, ook na voldoende ondersteuning, kan de informatie van de centrale proeven meewegen in een beslissing over een school. Het toetsresultaat zal echter slechts een zeer klein deel kunnen zijn van de beslissing, waarbij vele andere factoren binnen het gegeven onderwijs op de school een rol spelen. Net zoals bij het beoordelen van de leerlingen, kan een beslissing nooit alleen vallen op basis van de toetsresultaten. Een aanvullend punt zijn de kosten die volgen uit een stevige summatieve component waardoor de toetsen voor de school high-stakes testen worden. Daar waarbij een summatieve functie enkel op leerlingniveau, de leerkracht nog een bewakende rol kan hebben, kan bij een summatieve functie op schoolniveau een school dan ook baat bij fraude hebben. Dat noopt tot diverse vormen van controle die centraal gecoördineerd moeten worden, waarbij controle op locatie (bezoek van toetsleiders) dan wel op afstand (*proctored* testen op afstand, via camera's en controle van handelingsdata) van belang kan zijn. Het grote nadeel daarvan is dat de kosten die daarmee gepaard gaan zeer hoog zijn, het afnamemogelijkheden beperkt[11], of beiden. Dat zijn belangrijke redenen om aan de toetsresultaten op schoolniveau voor een school niet te veel consequenties te hangen die door een school als zeer negatief worden ervaren.

[11] Een mogelijke beperking is dat alle afnamen op een school op een en hetzelfde moment gehouden moeten worden om het mogelijk te maken dat een controleur langskomt. Het kan dan zelfs uit kostenoverweging bepaald worden dat alle toetsen op een school op dezelfde dag gemeten worden. De toetsduur is dan ook beperkt doordat ieder toetsuur ook een uur controle kost.

Ook voor de haalbaarheid, in termen van draagvlak vinden om de centrale proeven in het onderwijsveld geaccepteerd te krijgen, is het verstandig om met dit punt rekening te houden.

7.2 Dilemma's en scenario's

In deze paragraaf worden dilemma's besproken die naar voren kunnen komen rondom de thema's die aangesneden zijn in de Hoofdstukken 4, 5 en 6. Deze dilemma's worden nu verder uitgewerkt vanuit de Vlaamse context (zie tekstblok).

Beoogde planning van centrale toetsen

Vanuit het bestek voor de haalbaarheidsstudie en het bestek voor het steunpunt voor het thema "Ontwikkeling van gestandaardiseerde, genormeerde en gevalideerde net- en koepeloverschrijdende toetsen in Vlaanderen" zijn een aantal richtlijnen gegeven aangaande deze centrale toetsen. Het betreft in principe digitale toetsen die ieder jaar worden afgenomen bij duidelijk omschreven groepen. De toetsen moeten jaarlijks een vaste set van eindtermen meten, met een selectie van eindtermen die jaarlijks wijzigt. De verhouding in aantallen items is voorzien op 3 staat tot 1 (75% vaste, 25% wisselende eindtermen). In ieder geval moeten in alle (toetsbare) eindtermen basisgeletterdheid opgenomen worden in de vaste kern. Naast nieuwe opgaven voor de wijzigende eindtermen, moet er ook in de vaste set sprake zijn van toetsverversing om toetsfraude tegen te gaan.

Bij de ontwikkeling van toetsen moet er rekening gehouden worden met een brede afname waarbij leerlingen die binnen de doelgroep vallen, niet buitengesloten worden. De toetsen moet dusdanig betrouwbaar zijn dat het mogelijk is om vaardigheidsniveau van individuele leerlingen te kunnen bepalen, waarmee duidelijk wordt of de leerlingen de eindtermen bereiken, en ook over tijd de leerwinst bepaald kan worden. Deze informatie moet ook op geaggregeerd niveau (klas, school, systeem) helpen het onderwijs in Vlaanderen te verbeteren. De optie van (beperkte) adaptieve toetsing moet hierbij open gehouden worden.

We gaan uit van de planning zoals die nu bekend is wat betreft de fasering van de afnamen. De eerste afname zal in 2023 plaats vinden aan het einde van de eerste graad. In onderstaande schema (Figuur 7.1) is dat het achtste leerjaar. Deze meting betreft een leerling-cohort A waarvan het merendeel van de leerlingen in 2009 geboren is. In 2024 zal het volgende cohort (B) aan het einde van de eerste graad centrale toetsen maken. Het jaar 2025 is het eerste jaar dat op twee niveaus een centrale toets wordt afgenomen. Naast een meting aan het einde van de eerste graad (bij cohort C) zal dan ook een meting in het vierde jaar van het lager onderwijs plaats vinden. Dat betreft cohort G waarvan de meeste leerlingen in 2015 geboren zijn. In het jaar 2026 zullen er dan drie afnamen zijn: een aan het einde van de eerste graad (cohort C), een in het vierde jaar van het lager onderwijs (cohort H), en de eerste afname in het laatste jaar van het lager onderwijs (cohort F).

Beoogde planning van centrale toetsen vervolg

Het jaar 2027 is in een aantal opzichten een cruciaal jaar. Ten eerste is dan de eerste afname aan het einde van de derde graad (cohort A), maar het is ook het eerste jaar waarbij er leerlingen zijn die aan twee centrale metingen hebben meegedaan. Dat betref cohort A met twee metingen in het secundair onderwijs, en cohort G met twee metingen in het lager onderwijs. Dientengevolge is dat het eerste jaar dat op cohortniveau leerwinst te bepalen is, namelijk van leerjaar 4 naar 6 in het lager onderwijs, en van het einde van de eerste naar het einde van de derde graad in het secundair onderwijs. In 2028 zal het voor het eerst mogelijk zijn op cohortniveau leerwinst te bepalen tussen het zesde leerjaar in het lager onderwijs en het einde van de eerste graad (cohort F). Het jaar 2029 zal het eerste jaar zijn met een cohort (G) met drie metingen (twee keer lager onderwijs, een keer secundair onderwijs). Cohort F zal in 2032 het eerste cohort zijn met een meting in het lager onderwijs en twee in het secundair onderwijs. Cohort F is dan in 2033 weer het eerste cohort zijn dat alle vier de centrale metingen krijgt. Cohorten A tot en met E zullen allen twee centrale metingen krijgen (beiden in het secundair onderwijs), cohort F krijgt er drie en bij cohort G en verder zullen er vier metingen zijn.

Aantal jaren onderwijs *			school	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032
Cohort			jaar	2021	2022	**2023**	2024	**2025**	**2026**	**2027**	**2028**	**2029**	2030	2031	**2032**	**2033**
geboortejaren**		label	N mt.***	0	0	1	1	2	3	4	4	4	4	4	4	4
2008	**2009**	**A**	2	6	7	**8**	9	10	11	**12**						
2009	**2010**	B	2	5	6	7	**8**	9	10	11	**12**					
2010	**2011**	C	2	4	5	6	7	**8**	9	10	11	**12**				
2011	**2012**	D	2	3	4	5	6	7	**8**	9	10	11	**12**			
2012	**2013**	E	2	2	3	4	5	6	7	**8**	9	10	11	**12**		
2013	**2014**	**F**	3	1	2	3	4	5	**6**	7	**8**	9	10	11	**12**	
2014	**2015**	**G**	4	0	1	2	3	**4**	5	**6**	7	**8**	9	10	11	**12**
2015	**2016**	H	4	-1	0	1	2	3	**4**	5	**6**	7	**8**	9	10	11
2016	**2017**	I	4	-2	-1	0	1	2	3	**4**	5	**6**	7	**8**	9	10

* Vanaf lager onderwijs (groen); secundair onderwijs start bij 7 (geel/oranje); jaren voor lagere onderwijs: -2, -1 en 0 (blauw)
** merendeel van de leerling uit het cohort komt uit het vergedrukte geboortejaar
*** N mt: aantal metingen; kolom: per cohort; rij: per schooljaar

Figuur 7.1: Schematische weergave van de geplande centrale toetsen in Vlaanderen

Al met al ligt veel al vast bij de centrale toetsen. Er zijn echter ook nog verschillende mogelijkheden hoe dit voor elkaar te krijgen. In de hier gepresenteerde scenario's gaan we uit van een zich ontwikkelende, lerende organisatie van toetsontwikkeling, waarbij we uitgaan van wat nu haalbaar is voor de eerste afname in 2023, en ons richten op wat er in een later stadium mogelijk is. Tijdens de ontwikkeling en invoering van de centrale toetsen zal altijd gebruik gemaakt moeten worden van professionals op het gebied van psychometrie, pedagogisch handelen, onderwijsmethoden, schoolklimaat of andere aspecten van het onderwijs.

7.2.1 Dilemma's rond de doelen van de toets

Hoe de verschillende doelen met elkaar in balans te brengen?

Het is in voorgaande stukken evident geworden dat de doelen van de toetsen onbetwist duidelijk moeten zijn. Een overkoepelend doel als het bevorderen van de kwaliteit van het onderwijs moet praktisch geoperationaliseerd worden door de doelen van de belanghebbenden in balans te brengen. Volgens de plannen[12] is het doel voor de leerling enigszins vrijblijvend. Het is niet expliciet gesteld dat het formatief moet zijn, zoals dat uit het doel van de toets voor de leerkracht wel op te maken is[13]. Het summatieve karakter is voor de leerlingen vrijblijvend. Het kan meegenomen worden –dat is geen verplichting–, maar als het meegenomen wordt dan mag het niet doorslaggevend zijn[14]. Voor scholen is het doel het meest uitgewerkt, en lijkt het doel het meest summatief: "Zwakkere toetsresultaten vormen een stoplicht. Scholen waarvan de leerlingen significant minder leerwinst genereren op die proeven, moeten in een vrij te kiezen begeleidingstraject stappen om de kwaliteit van hun onderwijs te verhogen." Tot slot zijn er de systeemdoelen, waarover meer in Paragraaf 7.2.2. In voorgaande stukken is duidelijk geworden dat het bij elkaar brengen lastig is. Het is duidelijk dat als de proeven door de leerlingen als low-stakes ervaren worden, maar door de scholen als high-stakes ervaren worden, dit een uitdaging oplevert. Scholen willen dat leerlingen gemotiveerd zijn en het best mogelijke laten zien, zodat de school kan laten zien dat ze kwaliteit leveren. Leerlingen hebben daar minder belang bij. De oplossing ligt voor een belangrijk deel in het verhogen van het belang van de toets voor de leerling, maar hoe dat in te voeren?

- **Scenario 1: Scholen kiezen zelf**
  In dit scenario kiezen scholen zelf hoe zij het toetsresultaat meenemen in de globale beoordeling van een leerling. Dit scenario heeft als voordeel dat dit vrijheid aan de school biedt. De school kiest hoe de centrale toetsen te gebruiken, zolang het maar niet doorslaggevend is voor studievoortgang en –oriëntering. Een belangrijk element is dat scholen in dit scenario van elkaar kunnen verschillen in het gewicht van het toetsresultaat. Scholen met een gelijke kwaliteit maar met verschillende wijzen waarop de proeven meetellen, kunnen verschillende resultaten krijgen door hoe de leerlingen de proeven benaderen. Een school met tegenvallende toetsresultaten kan deze bijvoorbeeld verhogen door de proeven meer mee te laten tellen, terwijl er in het pedagogisch handelen niets wijzigt. Ook kunnen scholen in het lager onderwijs de proeven in het zesde leerjaar een hogere summatieve waarde voor de leerlingen geven dan die in het vierde leerjaar. Zo kunnen zij mogelijk een grotere leerwinst aantonen dan scholen die de summatieve functie gelijk houden over de leerjaren. Op een zeer belangrijk punt zijn de afnamen in dit scenario dus niet gestandaardiseerd. Een mogelijke oplossing is dat scholen in de overdracht van de gegevens aangeven hoe zij de toetsen laten meetellen in de beoordeling, en dat men achteraf probeert het gebrek aan standaardisatie via modellen te corrigeren. Aan een dergelijke hersteloperatie

[12] Zie het bestek voor de aanvraag van het Steunpunt: `https://data-onderwijs.vlaanderen.be/documenten/bestand.ashx?id=12646`.

[13] "In de eerste plaats is het een instrument voor de betrokken leerkrachten om te reflecteren over de resultaten en zo de kwaliteit te verhogen."

[14] "De beoordeling van de leerling: de resultaten op de toets kunnen worden meegenomen in de globale beoordeling van de leerling, maar zijn niet doorslaggevend in het kader van studievoortgang en -oriëntering."

zitten dusdanig veel haken en ogen dat dit niet aan te raden is.

- **Scenario 2: Voorgeschreven gebruik**
  In dit scenario wordt van tevoren gesteld op welke wijze de proeven mee moeten tellen in de globale beoordeling. Hoewel dit ten koste gaat van de vrijheid van de scholen, worden zo de toetsafnamen gestandaardiseerd en worden resultaten vergelijkbaar, hetgeen de validiteit van de meting vergroot. Het is duidelijker wat gemeten wordt, doordat de variatie door mogelijke verschillen in motivatie bij leerlingen weggenomen worden. De volgende vragen zijn dan hoe het voorschrift moet luiden, en hoe dit voorschrift tot stand komt. Het is belangrijk om scholen te betrekken bij de totstandkoming van het voorschrift, om naleving ervan in een later stadium te verhogen. Bij het voorschrift zelf kan gekozen worden voor een rekenvoorschrift of een procedureel voorschrift. Het rekenvoorschrift bepaalt hoe de uitslagen van centrale toetsen en schoolbeoordelingen gewogen moeten worden. Het procedurele voorschrift bepaalt in welke volgorde centrale toetsen en schoolbeoordelingen afgenomen moeten worden.
  - **Optie 1: Rekenvoorschrift** Bij een rekenvoorschrift is van tevoren bepaald hoe het resultaat meegewogen moet worden in het globale eindoordeel voor de leerling. Dit kan bijvoorbeeld een vastgestelde regel zijn, zoals dat het resultaat op de centrale toets voor 1/8e meetelt in het eindoordeel. Dit rekenvoorschrift kan ook verschillen per niveau. Zo kan bijvoorbeeld gesteld worden dat het 5% meetelt in het vierde leerjaar en 10% in het zesde leerjaar van het lager onderwijs, en dat het 15% en 20% meetelt voor de globale beoordeling aan het einde van respectievelijk de eerste en de derde graad. In Nederland wordt een rekenvoorschrift gebruikt bij de eindexamens waarbij het eindcijfer voor een vak voor 50% afhankelijk is van het resultaat op het centrale examen. Deze optie heeft als mogelijk nadeel dat, zeker in het lager onderwijs, misschien niet alle scholen met een cijferschaal werken, waardoor het meenemen van het toetsresultaat op een cijfermatige wijze mogelijk lastig kan worden. Wellicht problematischer is dat het gevolg van een dergelijk cijfervoorschrift eventueel toch doorslaggevend kan zijn in het kader van studievoortgang: als het resultaat, hoe laag de weging ook moge zijn, ervoor zorgt dat de leerling onvoldoende komt te staan, heeft het impact. Gezien de grote aantallen leerlingen is de kans dat dit bij minstens één leerling gebeurt, redelijk aanwezig. Een oplossing kan zijn dat het toegestaan wordt dat het alsnog negatieve impact kan hebben. In de eerste jaren zal een dergelijke aanpak voor het geaccepteerd krijgen van de centrale toetsen mogelijk problematisch zijn. Een alternatief kan het gebruik zijn van het maximum van schoolbeoordeling en resultaat op de centrale proef in plaats van een gewogen gemiddelde.
  - **Optie 2: Procedureel voorschrift** Bij een procedureel voorschrift kan gesteld worden dat het resultaat alleen positief effect mag hebben op het globale oordeel. Dat betekent dat wanneer een resultaat op een centrale toets hoger is dan het globale oordeel dat van tevoren over de leerling bekend was, het resultaat van de centrale toets een positieve impact mag hebben. Als het resultaat lager is dan heeft het geen gevolgen. In Nederland wordt een dergelijke procedure gehanteerd bij de eindtoets die helpt bij het oriënteren op het niveau van het

secundair onderwijs. Wanneer de eindtoets een hoger niveau adviseert dan de leerkracht, moet de leerkracht het resultaat van de eindtoets meenemen in een heroverweging[15].

- **Optie 3: Gecombineerd voorschrift** Een procedurele regel kan ook gecombineerd worden met een cijferregel. Een voorbeeld van een dergelijke regel is: als het resultaat op de centrale toets hoger is dan het globale oordeel dat er al van de school ligt, kan het resultaat van de centrale toets 25% meetellen. Een reden om dan de weging groter te maken dan in een standaard rekenvoorschrift, is dat dit dan een positieve bijdrage kan leveren aan de motivatie van leerlingen om goed te presteren op de toets. Een leerling die toch al een goed cijfer heeft waarbij de bijkomende toets amper invloed heeft, kan mogelijk minder gemotiveerd zijn. Merk op dat iedere vorm van summatief toetsen, ook als de school vrijgelaten wordt in hoe het toe te passen, uiteindelijk doorslaggevend kan zijn. Het is juist de mogelijkheid dat de toets impact heeft, wat de toets van groter belang maakt. Als enige impact op welke wijze dan ook niet toegestaan kan worden, dan blijven de toetsen low-stakes voor de leerlingen, met tot gevolg dat waarschijnlijk veel leerlingen niet laten zien wat ze eigenlijk wel kunnen.

- **Randvoorwaarde voor de scenario's: snelle verwerking van de resultaten**
  Wanneer de toetsen op welke wijze dan ook een summatieve invloed hebben in het schooljaar van de afname, dan moeten de belanghebbenden tijdig worden ingelicht over het resultaat. Niet alleen het aantal behaalde punten op de toets moet dan bekend zijn, maar ook de normering: hoe dat aantal punten te interpreteren. Een automatische scoreprocedure en een normering die van te voren bekend is, kunnen daar sterk bij helpen. Als er toch tijd voor normering nodig is, dan is het aan te raden de uiteindelijke leerlingrapportage binnen drie weken na de afname te laten plaatsvinden, en met voldoende tijd tot het einde van het schooljaar. Merk op dat voor de schoolrapportages de urgentie minder hoog is, en deze ook later aangeleverd kunnen worden. Ook aan het begin van een nieuw schooljaar kan een school reflecteren op zijn onderwijsresultaten. Daar waar de leerkrachtrapportages simpele samenvattingen van de leerlingresultaten kunnen zijn, kunnen deze mee met de leerlingrapportages worden geleverd. Daar waar ze meer op de schoolrapportages lijken, zouden deze ook later kunnen worden opgeleverd.

Hoe peilingen en de centrale toetsen te integreren?

Een vierde niveau van rapportage betreft een systeemevaluatie. In welke mate de centrale toetsen peilingen over kunnen nemen, is voor een groot deel afhankelijk van de inrichting van de centrale toetsen.

- **Scenario 1: Beperkte systeemevaluatie door centrale toetsen**
  De beperking zit in dit scenario niet in het achterwege laten van diverse achtergrondvariabelen of het nalaten om diverse (pedagogische) kenmerken van de scholen op te vragen middels vragenlijsten. Deze aspecten lijken ook mogelijk bij centrale toetsen.

[15] Merk op dat dit niet betekent dat de leerkracht verplicht is om het oordeel van de eindtoets over te nemen. Het kan uiteindelijk zo zijn dat het oordeel van de leerkracht blijft staan.

Leerling-vragenlijsten afnemen op grootschalig niveau lijkt een grotere uitdaging, maar zal niet de meest beperkende factor zijn. De meest beperkende factor zal liggen in de omschrijving en bevraging van het inhoudsdomein. Deze zal namelijk beperkt zijn als het aantal toetsversies beperkt is. Hoewel de toetslengte niet geheel vastligt (daarover meer in Paragraaf 7.2.3) zal een goede domeinomschrijving meer opgaven bevatten dan een leerling in redelijke toetstijd kan maken. Dat is zeker het geval als we op een meer gedetailleerd niveau van subvaardigheden kijken. Centrale toetsen met een beperkt aantal versies per leerjaar, zullen alleen een globale vinger aan de pols kunnen geven op systeemniveau. Aangaande inhoudelijke analyse zullen ze aanzienlijk minder diepgang hebben dan de huidige peilingen. Dit kan als zodanig geaccepteerd worden, of er kan gekeken worden hoe de functionaliteit van de huidige peilingen toegevoegd kan worden aan de centrale toetsen.

- **Scenario 2: Meer omvattende metingen bij centrale toetsen**
  Wanneer meerdere versies van de centrale toetsen worden aangeboden, is het mogelijk een veel breder beeld te krijgen per onderwijsdoel of eindterm. Gezien het grote aantal afnamen dat er mogelijk is, kan het aantal opgaven dat gebruikt wordt in principe ook zeer groot zijn. Mits een "gekoppeld onvolledig toetsontwerp[16]" gehanteerd wordt, kan iedere leerling op dezelfde schaal worden gerapporteerd. Rekening houdend met de dimensionaliteit en de dekking van de toetsmatrijs (zie ook de opmerkingen bij de selectie onderwijsdoelen in Paragraaf 7.2.3) is het mogelijk zowel de leerlingen als de scholen te evalueren, als ook een systeemevaluatie te doen van eenzelfde inhoudelijke diepgang als een peiling. Een voordeel van een dergelijke brede inhoudelijke dekking waarbij verschillende leerlingen verschillende opgaven maken, is dat binnen een school wel alle mogelijke opgaven aangeboden worden. Hiermee kan op schoolniveau veel gedetailleerder gerapporteerd worden, waardoor de formatieve functie op schoolniveau eveneens optimaal benut is. Een nadeel is dat een dergelijke opzet wel complex is. Een toetsontwerp waarbij binnen scholen leerlingen allemaal verschillende toetsen krijgen, is met een digitale afname makkelijker te bewerkstelligen dan met papieren toetsen. Maar ook digitaal is het een administratieve uitdaging waar de systemen op berekend moeten zijn. De analyses zijn wat complexer, al zijn de analyses met dergelijke grote aantallen leerlingen (hele cohorten) meestal wel robuust genoeg om dit aan te kunnen. Dat geldt ook voor de rapportage van de resultaten: het is goed mogelijk, maar iets complexer dan wanneer alle leerlingen dezelfde toets krijgen. De complexiteit is niet anders dan bij peilingen gebruikelijk is, zij het dat het grotere aantallen leerlingen betreft. Zeker bij aanvang van de centrale toetsen is een belangrijk nadeel dat voor een dergelijke opzet veel opgaven nodig zijn. Zeker als voor alle onderwerpen en vaardigheden de integratie van peiling en centrale toetsen beoogd wordt, zal de opgaven-productie zo veel werk zijn dat dit niet binnen de beschikbare tijd te halen is. Opties om het mogelijk te maken, zijn dan om oude opgaven te hergebruiken, of de peiling te beperken tot een van de vaardigheden, bijvoorbeeld leesvaardigheid, of wiskunde. Wanneer in de loop van de tijd er een itembank ontstaat waaruit geput kan worden, dan is de integratie van peiling en centrale toetsen wellicht ook beter haalbaar. Samenvattend

[16] Niet alle leerlingen maken alle mogelijke opgaven, maar de verschillende toetsversies zijn door overlap in opgaven wel aan elkaar gerelateerd.

kan gesteld worden dat in de toekomst, wanneer er meer ervaring is met de centrale toetsen en er meer opgaven zijn, dit scenario het overdenken waard is. In het eerste jaar, of de eerste paar jaar, afhankelijk van de ervaringen, is dit waarschijnlijk nog een stap te ver.

- **Scenario 3: Peilingen naast de centrale toetsen**
  In dit scenario krijgt een deel van de scholen, bepaald door middel van een al dan niet gestratificeerde steekproeftrekking, ook additionele toetsen. Het voordeel is dat de reguliere afname relatief simpel blijft. Ten opzichte van de reguliere peilingen kan het afnamedesign ook versimpeld worden omdat alle aanvullende toetsversies geankerd kunnen worden door middel van de centrale toetsen, en veel van de logistiek aangaande het verzamelen van de achtergrondvariabelen al geregeld is. Mogelijk hoeven er bij de peilingen naast de centrale toetsen ook minder vaak toetsleiders naar de scholen gestuurd worden dan bij reguliere peilingen, vanwege de infrastructuur die dan al klaar staat voor grootschalige afnamen[17]. Een nadeel is dat er scholen zullen zijn die meer toetsen moeten maken dan andere scholen. Ook is er is nog steeds een groot aantal diverse opgaven noodzakelijk. Hiervoor kan makkelijker uit de oude peilingen geput worden dan in scenario 2. In scenario 2 is het problematischer als de opgaven bekend zijn: Hergebruik kan in een summatieve context wellicht als problematischer gezien worden dan bij een peilingsonderzoek. In ieder geval zal deze wijze van peilingen naast centrale toetsen nog steeds een behoorlijke kosteninvestering per peiling impliceren.

Samenvattend, lijkt voor nu scenario 1 het meest realistisch. Afhankelijk van de ervaringen met de centrale toetsen, en hoe die ervaren worden, zal in de toekomst gekozen kunnen worden om scenario 2 of 3 te overwegen. Bij deze keuze kunnen ook de ervaringen met de internationale peilingen in relatie tot de centrale toetsen meegenomen worden.

7.2.2 Dilemma's rond leerwinst

Veel punten rond leerwinst zijn al behandeld in Hoofdstuk 4. Er zijn echter nog een tweetal zaken waarbij praktische overdenkingen in een Vlaamse context van belang zijn die hier nader worden beschouwd. Het eerste dilemma betreft de gemeten vaardigheden: zijn deze over de jaren heen dezelfde, of veranderen die? En als deze veranderen, hoe daar dan mee om te gaan? Het tweede dilemma betreft de gevolgde cohorten: het merendeel van de leerlingen zal binnen het cohort blijven, maar er zijn zeker leerlingen die op- en afstromen. Binnen het cohort vinden ook verschuivingen plaats: die hebben impact op het evalueren van leerwinst binnen scholen. Deze verschuivingen zijn niet willekeurig, maar hebben voor een (groot) deel een relatie met de gemeten vaardigheden. Hoe moet daarmee omgegaan worden? Dit laatste dilemma wordt nu benaderd vanuit de stap waar de kans op mutaties het grootst is, namelijk in de stap van het einde van de eerste graad naar het einde van de derde graad. Dat is met vier jaar de langste periode tussen twee metingen. Ook bij de

[17] Bij de eerste afname op die wijze zal er wel een onderzoek nodig zijn om het effect van de toetsleiders op de resultaten bij de peilingen te kunnen evalueren. Anders kunnen trends door de tijd ook te maken hebben met de veranderde afnamewijze. Bij het meten van de complexere vaardigheden zoals spreken zal een toetsleider ook nog steeds nodig zijn. Dergelijke vaardigheden zijn overigens ook al lastiger mee te nemen in de centrale toetsen (zie Hoofdstuk 5).

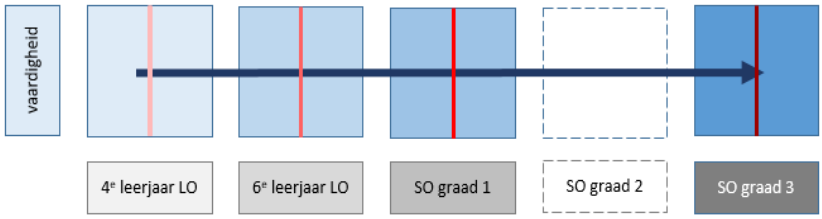
evaluatie van het bepalen van de vaardigheid over de tijd gaat speciale aandacht uit naar die grote periode tussen de twee meetmomenten.

Hoe om te gaan met leerwinst in de praktijk wanneer de gemeten vaardigheid verandert?

- **Scenario 1: Groei wordt gezien als unidimensioneel**
  Het meten van leerwinst is relatief eenvoudig als de metingen op een en dezelfde vaardigheidsschaal te brengen zijn. Als een vaardigheid gevolgd kan worden, en de vorderingen van de leerlingen kwantitatief geduid kunnen worden, is leerwinst simpelweg het verschil tussen de twee waarden op de vaardigheidsschaal (zie ook Hoofdstuk 4). Als we kijken naar zaken als sleutelcompetenties en de bijbehorende bouwstenen is een dergelijke lijn te volgen over de meetmomenten van de centrale toetsen. Neem bijvoorbeeld de bouwstenen voor Nederlands. Als we de sleutelcompetentie vertalen naar een onderwijscontext, zien we vier bouwstenen afgebakend[18]:
  - Het Nederlands receptief, productief en interactief, zowel mondeling als schriftelijk gebruiken als communicatiemiddel in relevante situaties;
  - Kenmerken en principes van het Nederlands begrijpen om ze in te zetten bij het communiceren;
  - Inzicht hebben in taal, in het bijzonder het Nederlands, als exponent en deel van een cultuur en een maatschappij;
  - Literatuur in het Nederlands beleven.

  De bouwstenen vallen verder op te delen in verschillende deelvaardigheden, zoals de eerste bouwsteen op te delen is in lees-, luister-, spreek- en schrijfvaardigheid. Deze zijn terug te vinden in de eindtermen, waaraan ook elementen van de andere bouwstenen te koppelen zijn. Dat is het gevolg van de verwevenheid van de bouwstenen binnen de verschillende eindtermen. De bouwstenen vormen zo de richting van de vaardigheid, en kunnen als richtlijn gezien worden waarop leerwinst waar te nemen is. Op die lijn, te zien als een algemene definitie van vaardigheid, zijn de eindtermen voor verschillende niveaus gedefinieerd als de minimumdoelen die leerlingen dienen te bereiken. Als al deze doelen op een schaal te ordenen zijn, en vooral kwantitatief van elkaar verschillen (een leerling heeft meer of minder van een bepaalde vaardigheid), dan is het bepalen van de leerwinst met deze schaalwaarden uit te voeren. De operationalisatie van de eindtermen vindt plaats in de vorm van opgenomen opgaven in de centrale toetsen. Al deze opgaven zijn op een unidimensionele IRT-schaal te plaatsen. Voor een vak als Nederlands kunnen we mogelijk de vaardigheidsmetingen voor bijvoorbeeld lezen afbeelden als in Figuur 7.2.
  De vaardigheid valt te volgen vanaf het vierde leerjaar in het lager onderwijs tot en met het einde van de derde graad van het secundair onderwijs. Wanneer de eindtermen per niveau gedefinieerd zijn, kunnen de grenzen op de vaardigheidsschaal bepaald worden door middel van standaardbepalingen (zie Hoofdstuk 6) waarmee bepaald is wat het te verwachten niveau op de vaardigheidsschaal is in die fase van het onderwijs. Merk op dat niet iedere eindterm een eigen standaardbepaling zal krijgen en dat deze geclusterd worden. In Figuur 7.2 zijn de eindtermen geclusterd

[18] `https://www.onderwijsdoelen.be/uitgangspunten/4806`



Figuur 7.2: Afbeelding van een unidimensionele vaardigheidsschaal voor leerwinst

tot een niveau per meetmoment aangegeven met de rode verticale strepen. Het is ook mogelijk meer niveaus per meetmoment aan te geven[19]. Of een dergelijke doorgaande leerlijn unidimensioneel te operationaliseren is, is voor een belangrijk deel een empirische vraag. Om die te beantwoorden zal het onderzoeksdesign zo opgezet moeten worden dat deze groei waar te nemen is. In eerdere hoofdstukken is daar ook al op ingegaan. Een dergelijk design betekent dat er verzamelingen opgaven geïdentificeerd moeten worden die de link leggen tussen twee afnamemomenten. Hier worden twee opties onderscheiden. De eerste optie is de minimale optie, waarbij alleen gebruik gemaakt wordt van de metingen van de centrale toetsen. Bij de tweede optie zijn er ook metingen in tussenliggende momenten. In Hoofdstuk 5 zijn al voorbeelden gegeven van dergelijke ontwerpen, maar deze worden nu toegepast, specifiek in de Vlaamse context.

– **Optie 1: Eén meetschaal, alleen gebruik makend van centrale metingen**
Het ontwerp van de afnamen is gegeven in Figuur 7.3.

		Lager onderwijs		Secundair onderwijs	
moeilijkheid	itemset	LO 4	LO 6	SO gr1.2	SO gr3.2
laag	Itemset 1	■			
	Itemset 2	■	■		
	Itemset 3		■		
	Itemset 4		■	■	
	Itemset 5			■	
	Itemset 6			■	■
hoog	Itemset 7				■

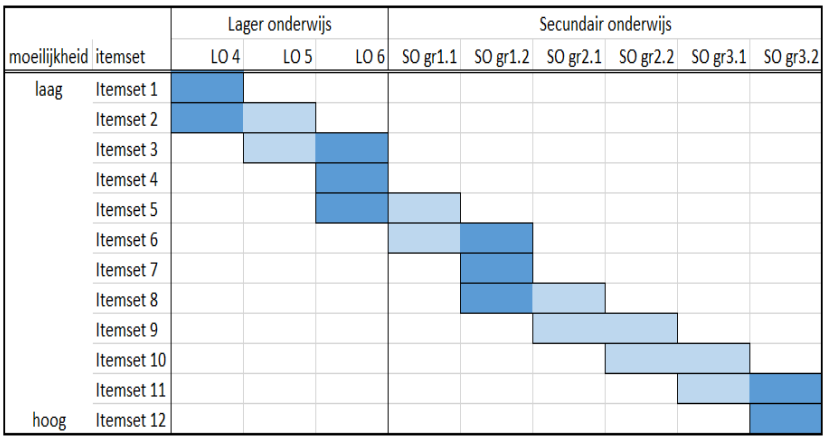
Figuur 7.3: Gekoppeld ontwerp met alleen metingen op de centrale toetsmomenten

Hierin zijn 7 itemsets te definiëren, geordend van makkelijk naar moeilijk op de vaardigheidsschaal. Merk op dat itemsets van elkaar kunnen verschillen in grootte: het is niet gezegd dat iedere itemset evenveel items bevat. Itemset 1 is de itemset met de gemakkelijkste opgaven op het niveau van het vierde leerjaar van het lager onderwijs. In de tweede itemset zitten opgaven die zowel in het vierde als het zesde leerjaar van het lager onderwijs afgenomen kunnen worden.

[19]Eindtermen kunnen representatief zijn voor verschillende niveaus, ook binnen een leerjaar. Zo zijn er makkelijker te behalen eindtermen, en wat moeilijker te behalen eindtermen. Door deze verschillend te clusteren kunnen verschillende niveaus binnen een leerjaar onderscheiden worden. Ook zouden op de gehele set van eindtermen verschillende niveaus van beheersing aangegeven kunnen worden zoals een fundamenteel (minimum) niveau en een streefniveau. Meer in detail kunnen we niet treden aangezien op dit moment de ontwikkeling van de eindtermen in volle gang is. Alleen de eindtermen van graad 1 zijn gemoderniseerd.

Dat zijn opgaven die moeilijk zijn in het vierde leerjaar en makkelijk in het zesde leerjaar. De definitie van de andere itemsets volgt dezelfde logica. Het voordeel van dit ontwerp is dat alleen met de metingen tijdens de centrale toetsen gewerkt wordt. Dat scheelt in de kosten. Een belangrijk nadeel is echter dat het moeilijk is betekenisvolle gecombineerde itemsets[20] te maken. De opgaven in die sets zijn óf (veel) te moeilijk op het lagere niveau zijn, óf (veel) te makkelijk op het hogere niveau. Het gevolg is dat op beide niveaus met deze items geen goede meting plaats kan vinden. Dit probleem zal zeker spelen bij de stap van het einde van graad 1 (SO gr1.2) naar het einde van graad 3 (SO gr3.2), maar naar alle waarschijnlijkheid ook voor de andere stappen. Overigens kan de mate waarin dit probleem speelt per onderdeel verschillen. Bij leesvaardigheid kunnen teksten en bijbehorende opgaven bijvoorbeeld kinderachtig overkomen qua onderwerp voor een hoger niveau, maar nog steeds redelijk goed onderscheid maken tussen zwakke en goede lezers. Bij rekenen/wiskunde echter kunnen er eindtermen zijn, die twee of drie jaar na het aanleren ervan, geroutineerd worden uitgevoerd door (bijna) alle leerlingen, en dus geen onderscheid meer maken tussen zwakke en goede rekenaars op het hoger niveau.

– **Optie 2: Eén meetschaal, gebruik makend van metingen in alle leerjaren**
Om nu rekening te houden met de grootte van de vaardigheidsstappen tussen de momenten van de centrale toetsen, worden ook metingen uitgevoerd op tussenliggende niveaus. We breiden het ontwerp van Figuur 7.3 zodoende uit met deze niveaus die niet meegenomen worden met de centrale toetsen. Dit levert het ontwerp op dat gegeven is in Figuur 7.4.



Figuur 7.4: Gekoppeld ontwerp met metingen in alle leerjaren

Bij een dergelijk ontwerp zal in het lager onderwijs ook een meting plaats vinden in het vijfde leerjaar. Hiervoor hoeven geen nieuwe items ontwikkeld te worden, omdat de moeilijke items uit het vierde leerjaar en de makkelijke uit het zesde leerjaar, wel goed toepasbaar kunnen zijn in leerjaar 5. Alleen om de afstand te overbruggen tussen het einde van de eerste en de derde graad zullen nieuwe opgaven ontwikkeld moeten worden, ongeveer op het niveau van

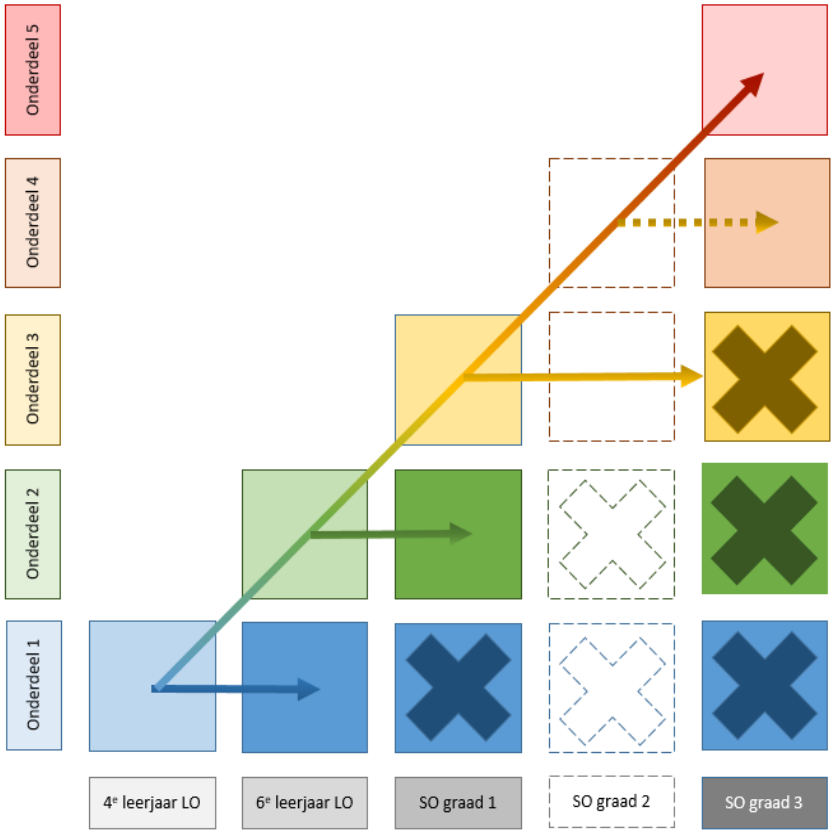
[20] Dat zijn in Figuur 7.3 de items met de even nummers.

het einde van de tweede graad. Het is aannemelijk dat als we werkelijk een echte meetschaal willen ontwikkelen, deze tweede optie gekozen moet worden. Hierbij kan opgemerkt worden dat de aantallen observaties in de tussenliggende jaren niet extreem groot hoeven te zijn. In de eerdere hoofdstukken zijn de benodigde aantallen genoemd. Daar zijn uiteraard kosten aan verbonden, maar bij werkende systemen op de scholen om centrale toetsen af te nemen, zullen die meevallen. Deze metingen hoeven ook niet in het eerste of tweede jaar van afnamen van de centrale toetsen gedaan te worden. Ze kunnen - op z'n laatst - ook plaats vinden in het jaar dat twee niveaus aan elkaar gekoppeld moeten worden. Volgens de huidige planning zal de eerste leerwinst meting pas in 2027 zijn[21], wat het vierde jaar is dat er centrale toetsen moeten worden afgenomen. Er kan tegen die tijd een goede itembank ontwikkeld zijn en voor de meeste stappen hoeven dan ook geen nieuwe items ontwikkeld te worden. Een uitzondering zal echter het tussenliggende niveau zijn tussen het einde van de eerste en de derde graad, wat betekent dat er waarschijnlijk kosten gemaakt moeten worden om nieuwe opgaven te ontwikkelen ongeveer op het niveau van het einde van de tweede graad. Een alternatief is om de meting aan het einde van de derde graad, en de bijbehorende leerwinst, anders te benaderen, waarover later in deze paragraaf wordt ingegaan.

- **Scenario 2: Groei wordt (ook) kwalitatief bezien**
  Niet bij alle vaardigheden zal een unidimensioneel IRT-model passend zijn. Dit kan mogelijk blijken uit empirisch onderzoek, als resultaat van de analyses volgend uit scenario 1. In sommige gevallen is het al te voorzien. Het is te verwachten dat bijvoorbeeld bij wiskunde het lastig kan zijn een puur unidimensionele schaal te maken. Dat wat de leerlingen kunnen, is niet alleen kwantitatief anders, maar ook kwalitatief anders. Er zit zeker een opbouw in, en de wiskundevaardigheden hebben ook enige relatie, maar als we kijken naar de eindtermen per niveau zoals die nu bekend zijn, dan zien we dat een doorlopende lijn lastiger te definiëren is. Vaardigheden die hun aanvang hebben in het lager onderwijs, hebben een ontwikkeling die niet verder gaat: een vaardigheid als optellen en aftrekken tot honderd (eindterm 1.13 lager onderwijs) ontwikkelt zich niet meer. Ook als we kijken naar een vaardigheid als vermenigvuldigen dan ontwikkelt deze zich niet verder. De sommen kunnen dan wel iets moeilijker worden, maar op een gegeven moment worden de leerlingen niet meer duidelijk vaardiger: de leerwinst stokt op die vaardigheid. Ondertussen leren de leerlingen wel steeds nieuwe zaken, die gebruik maken van de eerder verworven vaardigheden, maar wel anders zijn. Neem zaken als het oplossen van vergelijkingen (analyse) en statistiek. Dat zijn vaardigheden die niet zozeer "meer van hetzelfde" zijn, maar echt kwalitatief nieuwe vaardigheden zijn. De leerwinst zit dan niet in de groei van eerder geleerde vaardigheden, maar in de verwerving van nieuwe vaardigheden. Voor het verwerven van die vaardigheden

[21] De planning zoals die nu bekend is, zou een eerste afname hebben aan het einde van de eerste graad in 2023 (cohort A), en in 2025 een eerste afname het vierde leerjaar (cohort G). Dat betekent dat in 2027 met een afname aan het einde van de derde graad, en een in het zesde leerjaar, deze cohorten de eerste cohorten zijn met twee metingen, en dus waarbij leerwinst op cohortniveau bepaald kan worden.

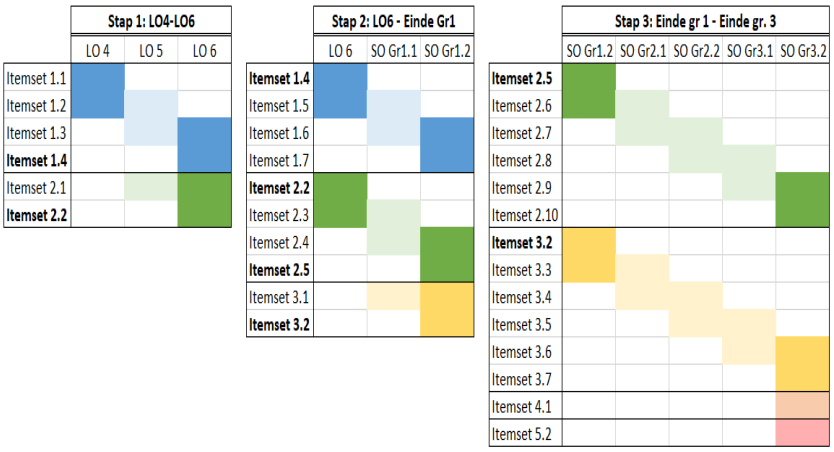
moet er meestal wel een basis zijn bij de eerder verworven vaardigheden, maar deze zijn moeilijk echt op dezelfde dimensie te plaatsen. Deze ontwikkeling is weergegeven in Figuur 7.5.



Figuur 7.5: Kwantitatieve en kwalitatieve groei

In Figuur 7.5 is aangegeven dat voor onderdeel 1 de groei in het lager onderwijs te volgen is. Het is de verwachting dat leerlingen over het algemeen dit onderdeel aan het einde van het lager onderwijs beheersen, waar bij veruit de meeste leerlingen geen groei verwacht wordt in het secundair onderwijs (aangegeven met het kruis). Leerwinst wordt niet meer op dat onderdeel verwacht. Onderdeel 2 wordt ingezet aan het einde van het lager onderwijs, en voert verder naar het secundair onderwijs, enzovoorts. In deze opzet kan de leerwinst gegeven worden per onderdeel, of gezien worden als de groei die door middel van de diagonale lijn gegeven wordt. Bij de huidige graadspecifieke eindtermen is –van lager onderwijs naar de derde graad–een dergelijke structuur ook te herkennen bij wiskunde.
De designs die bij een dergelijke aanname horen, lijken op die van optie 2 zoals hierboven gegeven, alleen worden deze opgedeeld naar de drie stappen die er te nemen zijn. Een voorbeeld hiervan wordt gegeven in Figuur 7.6. Het is hierin duidelijk dat de vaardigheid door de loop van de tijd verschillend gedefinieerd wordt. Het is ook duidelijk dat in dit scenario een deel van de groei lastig waar te nemen is. Ten eerste is dat doordat de groei van een eerder onderwezen onderdeel (onderdeel 1 hierboven) minder groot is van de eerste naar de tweede stap. Ten tweede doordat, wanneer de onderdelen als onafhankelijke unidimensionele schalen gezien worden,

van een nieuw onderdeel geen meting in een eerder jaar beschikbaar is en zo de groei onzichtbaar blijft.



Figuur 7.6: Ontwerpen voor afnamen bij kwalitatieve groei

Een mogelijke oplossing is om binnen de korte tijdsspanne van twee jaar (stap 1 en stap 2; zie Figuur 7.6) de verschillende onderdelen te behandelen alsof ze unidimensioneel zijn. Over een dergelijke tijdspanne is dat vaak nog wel mogelijk, omdat de onderdelen dan nog samenhangen. De eindtermen lijken dan net nog genoeg op elkaar. Dan wordt zo ook de groei op het nieuwe onderdeel meegenomen. De schaal voor de groei is voor de eerste stap echter wel anders dan voor die van de tweede stap: het betreft een net iets andere dimensie. De grotere uitdaging zit in de derde stap. Voor de nieuwste onderdelen is geen link meer te maken met het einde van de eerste graad, en ook de unidimensionaliteit van die schaal kan onder druk komen te staan. Hiervoor zijn drie opties.

- **Optie 1: Stap drie op een gelijke manier oppakken als de eerste stappen** Het is mogelijk dat de schending van unidimensionaliteit meevalt, of weinig impact heeft. Van tevoren is dat niet geheel te voorzien, en zal ook deels uit empirisch onderzoek moeten blijken. Het nadeel is dat er een risico is dat de betekenis van de leerwinst onder druk kom te staan doordat deze niet geheel correct opgepikt wordt. Het voordeel is dat er niet veel meer additionele kosten aan verbonden zijn, anders dan het meenemen van tussenliggende jaren.
- **Optie 2: Einde van Graad 2 meenemen als officieel meetmoment bij de centrale toetsen** Bij deze optie zijn alle stappen tussen de metingen even groot, namelijk twee schooljaren. Dit zorgt ervoor dat groei zo beter gevolgd kan worden in het kader van leerwinst. Ook de verandering in de populatie in kader van de mutaties zijn zo beter mee te nemen in de analyses. In een tijdspanne van vier jaar zullen er ook meer mutaties plaatsvinden in de schoolpopulatie dan in een periode van twee jaar: meer leerlingen doubleren, veranderen van school, niveau of richting. Zeker in de periode tussen de eerste en de tweede graad zal dat spelen. De groep die van einde tweede graad naar einde derde graad doorgroeit binnen een school, zal stabieler zijn, en de leerwinst zal dan ook beter te interpreteren zijn. Het is evident dat deze optie zeer kostbaar is. Met

vijf in plaats van vier metingen zullen de kosten aanzienlijk hoger liggen. Het is ook nog maar de vraag of een invoeging van nog een centraal meetmoment als acceptabel gezien wordt door het veld. Het is daardoor aannemelijk dat deze optie niet als eerste gekozen zal worden. Het is mogelijk in een latere fase van de invoer van centraal toetsen in Vlaanderen, wanneer het veld ermee bekend is en de toetsen als meerwaarde gezien worden, dat deze optie wellicht serieuzer overwogen kan worden.

– **Optie 3: Een andere aanpak van leerwinst tussen einde eerste en derde graad** Deze optie wordt verder uitgewerkt in de volgende paragraaf. Die gaat dieper in op de mogelijkheden met leerwinst om te gaan wanneer er sprake is van een groot gat in de tijd en in de vaardigheid. Deze zal nu vooral uitgewerkt worden voor de stap van de eerste naar de derde graad, maar zou deels ook betrekking kunnen hebben op de eerste twee stappen. De mogelijkheden die in de volgende paragraaf genoemd worden, kunnen met name ook relevant zijn bij de tweede stap (van lager naar secundair onderwijs) wanneer leerlingen naar een andere school gaan.

Hoe om te gaan met de leerwinst in de praktijk wanneer de tijdsspanne groot is?

De afname aan het einde van de derde graad wijkt op een aantal punten af van de eerdere drie metingen. Doordat deze toetsen aan het einde van de leerplicht zitten, is het formatieve karakter op leerlingniveau beperkt. Daarnaast zijn de klassen homogener dan in de eerdere metingen. In het reguliere lager onderwijs zijn leerlingen niet over klassen verdeeld op basis van niveau of inhoudelijke keuzes. In de eerste graad begint een dergelijke opdeling, maar die is minder uitgesproken dan die in de derde graad. De klas als eenheid is daarmee ook anders. Tot slot, is in het kader van leerwinst een groot verschil dat de meting ervoor vier jaar eerder plaatsvond, in plaats van twee jaar.

In de eerdere paragrafen is al aangegeven dat een dergelijke lange tijdspanne tussen twee metingen in het kader van leerwinst uitdagingen met zich mee draagt. Ten eerste is dat hoe meer tijd er zit tussen meetmomenten, hoe meer de gemeten vaardigheid net iets anders gedefinieerd wordt, en de eindtermen tussen de twee metingen minder strak op elkaar aansluiten. Dit probleem is zowel voor het bepalen van leerwinst op leerlingniveau en op schoolniveau een probleem. Een mogelijke oplossing is gegeven in de eerdere paragrafen, en in deze paragraaf komt een alternatief aan bod.

De tweede uitdaging betreft vooral het bepalen van leerwinst op schoolniveau. Hoe langer er tussen twee meetmomenten zit, hoe meer de populatie op de scholen verandert. Dat is zeker het geval wanneer we kijken naar de verschillen tussen de eerste graad en de derde graad, waarbij leerlingen ook nog kunnen divergeren in studierichting. In de traditionele leerwinstbenadering zijn er statistische procedures die enige correcties kunnen doorvoeren, maar deze modellen zullen de wijzigingen nooit geheel goed vangen in de cijfers. Bij wisselende aannames zullen wisselende resultaten het gevolg zijn. Het kan er ook toe leiden dat scholen, om beter uit de statistieken te komen, er geen baat bij hebben om leerlingen waar de leerwinst onder het gemiddelde van de school zit, aan boord te houden. De school kan er voordeel bij hebben dat deze leerlingen de school verlaten.

Als deze leerling naar een ander school gaat waarop deze leerling juist een iets grotere leerwinst heeft dan gemiddeld is de verandering van school voor beiden scholen goed[22]. Zonder een werkelijke verbetering in kwaliteit van het Vlaamse onderwijs zou het wel zo kunnen lijken.

Het bijkomende nadeel van deze zeer grote tijdspanne is dat de causaliteit – de school als oorzaak van mogelijke tegenvallende leerwinst, of juist leerwinst die boven verwachting is – moeilijk te bepalen is. Er is in de tussentijd te veel veranderd.

- **Alternatief scenario: brede modelering van de resultaten einde derde graad**
  In plaats van de traditionele leerwinst meting is een brede modelering van de resultaten aan het einde van de derde graad een nuttig alternatief scenario. Dat betekent dat de toetsresultaten gelegd worden naast de cijfers van de doorstroom waar de school mogelijk invloed op kan hebben (leerlingen die de school verlaten, nieuwe leerlingen die binnenkomen, veranderde onderwijsvormen en studierichtingen en dergelijke), en meer algemene schoolkenmerken. In dergelijke modellen worden de beginmetingen als benadering gebruikt van de vaardigheid van de leerling. Dat deze vaardigheid niet geheel overeenkomt met exact dezelfde vaardigheid die aan het einde van de derde graad gemeten wordt, is dan minder een probleem aangezien de schaalwaarde als zodanig niet direct een rol speelt. De kwaliteit van de school wordt zo ook in een breder perspectief gezien: die wordt ook niet alleen bepaald door de leerwinst in de vakken Nederlands en wiskunde, maar kent veel meer aspecten die in een dergelijk model meegenomen worden.

  Een ander voordeel van deze werkwijze is dat de moeite die gedaan zou moeten worden om op een vaardigheidsschaal de afstand tussen de twee meetmomenten te overbruggen niet nodig zijn. Tussenmetingen zijn niet meer nodig, en de zorg van de verandering in de vaardigheidsschaal is minder relevant, zolang de meting aan het einde van de eerste graad voldoende kan dienen als prior. De prior is een indicator van de vaardigheid Nederlands of wiskunde aan het einde van de eerste graad, maar hoeft niet noodzakelijk op dezelfde schaal te liggen als de meting aan het einde van de derde graad. In principe is deze werkwijze ook toepasbaar voor de overige stappen, maar bij de stap van het einde van de eerste graad naar die van de derde graad, lijkt deze werkwijze in kosten-baten-overweging en in de psychometrische context vooral aan te raden.

[22]Rekenkundig valt aan te tonen dat de verplaatsing van de beschreven leerling op beiden scholen tot een additionele leerwinst leidt. Dit leidt tot een overschatting van de werkelijke leerwinst, die ten koste van de leerling op een lager niveau het onderwijs volgt. Dit lijkt statistisch een verbetering van de kwaliteit, maar is dat niet in werkelijkheid.

7.2.3 Dilemma's rond toetsontwikkeling

Over het maken van opgaven

Voordat de dilemma's verder uitgewerkt worden kan nog een algemene opmerking gemaakt worden over het maken van opgaven. Bij het maken van opgaven voor de centrale toetsen zullen de eindtermen van groot belang zijn. Deze geven richting aan de inhoud en zijn ook direct gerelateerd aan het beoogde curriculum. Een andere belangrijke bron voor het maken van de opgaven zullen de (meest gebruikte) methoden zijn. Deze zijn een goede weergave van hoe het curriculum daadwerkelijk in de scholen aan de leerlingen onderwezen wordt. Methoden bevatten ook oefenopgaven, die een belangrijke bron zijn om te bepalen welke type opgaven de leerling mee vertrouwd is. Om net- en koepeloverschrijdende toetsen te verkrijgen is het ook van groot belang om huidige net- en koepeleigen toetsen in de ontwikkeling mee te nemen. Deze laten zien hoe scholen gewend zijn de vaardigheid van de leerlingen te bepalen. Het is evident dat de centrale toetsen niet te sterk moeten leunen op een specifieke methode, of een specifieke net- en koepeleigen operationalisatie. In die zin lijkt het maken van (opgaven voor) de centrale toetsen sterk op het maken van goede toetsen voor peilingen.

Als veel versies van een toets uitgegeven moeten worden van een toets, kan het handig zijn om opgaven te 'klonen'. Nieuwe opgaven daarbij worden gebaseerd op bestaande opgaven, bijvoorbeeld bij wiskunde door slechts enkele getallen aan te passen. Op deze wijze kunnen relatief snel nieuwe opgaven gemaakt worden. Echter, een kloon is niet equivalent aan het originele item. De opgave kan moeilijker of makkelijker geworden zijn, of beter of slechter onderscheidend. Van te voren is lastig te voorspellen wat de psychometrische kenmerken zijn, ook als men de kenmerken van het originele item kent. Gekloonde opgaven moeten daarom als nieuwe opgaven gezien worden en de psychometrische kenmerken moeten altijd opnieuw geschat worden.

Hoe om te gaan met toetsverversing?

In deze paragraaf bespreken we hoe en hoe vaak de toetsen ieder jaar vernieuwd kunnen worden om het uitlekken van opgaven tegen te gaan. Hoe de resultaten over de verschillende afnamejaren toch met elkaar te vergelijken zijn, wordt besproken in de volgende paragraaf. We beschouwen eerst de kern van de eindtermen (vaste set) die binnen een afnamemoment[23] te zien is als een vaardigheid die op een unidimensionele IRT-schaal te brengen is[24]. Daar waar het gaat om een verandering in de inhoud, bijvoorbeeld door rotatie, wordt het verder besproken in Paragraaf 7.2.3. In deze paragraaf zijn ook twee intermezzo's ingevoegd met praktische overdenkingen aangaande de ontwikkeling van een itembank, en trekkingen van toetsen uit een itembank.

- **Scenario 1: Vaste toetsen die periodiek geheel vervangen worden**
  In het eerste scenario worden de proeven periodiek geheel vernieuwd. Bij jaarlijkse verversing krijgt ieder cohort een nieuwe toets. Binnen dit scenario is er nog de

[23] Hierbij worden de vier leerjaren (afnamemomenten) onderscheiden: vierde en zesde leerjaar lager onderwijs, en einde eerste en derde graad secundair onderwijs.

[24] Merk op dat over afnamemomenten dit anders kan zijn. Dit is besproken in Paragraaf 7.2.2

mogelijkheid om één enkele variant te hebben of verschillende versies op één afnamemoment. Indien de proeven low-stakes zijn, en de kans op het uitlekken van opgaven zeer gering is, kan besloten worden om toetsversies meerdere jaren te gebruiken, en pas te verversen als deze inhoudelijk of technisch verouderd zijn. Als de belangen hoog zijn, is het belangrijk dat het afnamevenster juist klein gehouden wordt. Door binnen een afnameperiode met meerdere, onderling overlappende varianten te werken, kan de kans op uitlekken verkleind worden.

- **Optie 1 - Enkele toetsversie** Alle leerlingen binnen een leerjaar krijgen dezelfde verzameling opgaven. Dit heeft als voordeel dat het aantal te maken opgaven beperkt is: slechts voor 1 versie hoeven opgaven gemaakt te worden. Een ander voordeel is dat de vergelijkbaarheid van de resultaten binnen een afnamemoment relatief simpel is: de vaardigheid wordt weergegeven door het aantal gescoorde punten dat een leerling heeft. De leerlingen worden op exact dezelfde opgaven vergeleken. Dit voelt eerlijk voor de betrokkenen vanwege de *face validity*.

  Een nadeel van deze optie betreft de kans op fraude. Als de toets voor een specifieke vaardigheid niet tegelijkertijd op een vast aangegeven moment wordt afgenomen, dan kunnen de opgaven bekend worden. Wanneer opgaven bekend zijn, kunnen leerlingen en scholen die de proef later afnamen daar onevenredig baat bij hebben. De bekendheid van de opgaven kan mogelijk enigszins beperkt worden door de opgaven in willekeurige volgorde aan te bieden. Bij digitale toetsen is dat technisch veelal geen enkel probleem (zie Perceel 3). Het bemoeilijkt het "lekken van opgaven", maar zal het waarschijnlijk niet geheel voorkomen. Een digitale afname op een enkel, vast aangegeven moment in heel Vlaanderen heeft ook praktische uitdagingen die onder meer besproken zijn bij Perceel 2 en Perceel 3.

  Een inhoudelijk nadeel van een enkele versie is dat de gehele vaardigheid door slechts een beperkt aantal opgaven omschreven is. Daarmee wordt slechts een beperkt beeld verkregen van het domein. Een versie wordt onacceptabel lang qua afnameduur als een veel bredere verzameling opgaven voorgelegd wordt. Op schoolniveau is dat goed op te lossen door meer versies parallel te hanteren.
- **Optie 2 -Verschillende toetsversies voor een vaardigheid** Een beperkt aantal toetsversies, die onderling deels overlappen, worden op één afnamemoment verdeeld onder de leerlingen van iedere school. Op deze wijze kan op schoolniveau een vaardigheid breder gemeten worden. Door gebruik te maken van IRT kan de vergelijking tussen leerlingen eerlijk zijn, en de terugkoppeling naar de scholen veel gedetailleerder. Met een digitale afname is het ook niet moeilijk bij te houden wie welke versie gemaakt heeft.

  Een nadeel van verschillende versies is dat er ook meer opgaven geschreven moeten worden. De itemproductie moet dus omhoog, en niet alleen in aantallen. Het klonen van items (gebruik maken van items die sterk op elkaar lijken) voegt namelijk niets toe aan de verbreding van de gemeten vaardigheid. Een ander nadeel is dat leerlingen doordat ze verschillende versies krijgen het gevoel kunnen hebben dat de onderlinge vergelijkbaarheid niet eerlijk is. Door IRT is

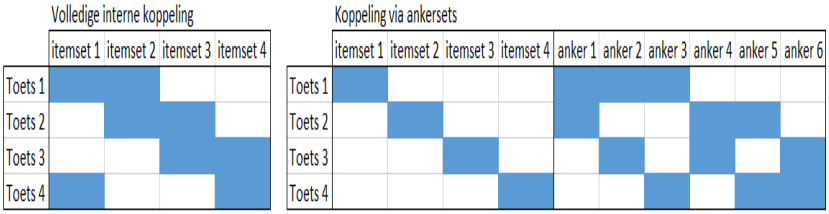
dit wel te waarborgen, maar de uitleg ervan naar leerlingen en ouders moet niet onderschat worden.

De verschillende versies kunnen, zeker in combinatie met willekeurige volgordes, de kans op fraude verkleinen: er zijn meer opgaven om te onthouden, en het kennen van een opgave levert geen voordeel op als de opgave niet in de voorgelegde versie zit. Hoe groot de impact op fraude is, hangt ook af van de opzet van de verschillende versies. Er zijn verschillende varianten mogelijk om de versies onderling te laten overlappen. De versies kunnen onderling volledig gekoppeld zijn, bijvoorbeeld doordat een versie altijd bestaat uit twee halve andere versies (zie linker ontwerp in Figuur 7.7).
Ook kan gewerkt worden met toetsversies waarbij iedere versie een groot uniek eigen deel heeft, in combinatie met kleinere onderling overlappende ankersets (zie rechter ontwerp in Figuur 7.7). Hoe groot die overlap is, zal afhangen van het aantal items dat per ankerset is opgenomen. Als dat relatief klein is, maar groot genoeg om een goede koppeling te verkrijgen, dan is de kans op fraude geringer. Bij een volledige interne koppeling is het waarschijnlijk verstandig het toetsvenster beperkt te houden, terwijl bij de koppeling via ankersets, de versies gefaseerd uitgebracht kunnen worden (toets 1 op dag 1, toets 2 op dag 2, etc.).

Over het ontwikkelen van een itembank

Als een aantal jaren achter elkaar met vaste toetsen gewerkt wordt, groeit de verzameling beschikbare opgaven. Het is evident dat als binnen een afnamejaar met meer verschillende versies per leerjaar gewerkt wordt, de verzameling beschikbare opgaven sneller groeit. Bij een snellere groei van een itembank is de kans dat opgaven verouderen ook kleiner. Dit betekent dat de verzameling van beschikbare opgaven voor hergebruik toeneemt. Als het administratief systeem van de itembank (zie Hoofdstuk 5) goed op orde is, biedt dit goede mogelijkheden voor de afname met behulp van een itembank. In het kader van itemverversing is het ook aan te raden een itembank goed up-to-date te houden. Dat houdt in dat ieder jaar gekeken moet worden of de opgaven niet verouderd zijn qua inhoud of voor wat betreft de vorm. Ook betekent dat dat ieder jaar nieuwe opgaven ontwikkeld moeten worden om in de itembank op te nemen. Voor dergelijke opgaven kan het verstandig zijn een klein pilotonderzoek te organiseren bij experts en leerlingen. Na een beoordeling door vakinhoudelijke experts, en mogelijke aanpassingen van het item op basis van het commentaar, kunnen de toetsvragen afgenomen worden bij een (zeer) beperkte steekproef van leerlingen. Als het alleen gaat om opgaven uit de toetsen (en bank) te houden die echt disfunctioneren, dan is een proefafname bij 10 tot 15 scholen voldoende. Het kalibratieonderzoek kan uitgevoerd worden op basis van de werkelijke afname. Dit wordt verder besproken onder Scenario 2.

- **Scenario 2: Toetsen op basis van een itembank verversen**
  Er zijn meerdere wijzen mogelijk waarop toetsversies uit een itembank getrokken kunnen worden (zie volgend intermezzo). In alle gevallen draait verversing van de



Figuur 7.7: Twee ontwerpen van de verschillende versies

toetsversies dan om de keuze van opgaven in de bank. Als een itembank gebruikt wordt om toetsversies uit te trekken (zie Hoofdstuk 5), dan moet deze groot genoeg zijn om bekendheid met de opgaven tegen te gaan. Het opnemen van nieuwe opgaven, het controleren of opgaven gedeeld worden via het internet, en het bijhouden van de statistische kenmerken van de opgaven (parameterdrift) zijn ook eerder genoemde manieren om bekendheid met de bank te controleren. Het verversen van de bank kan overigens niet alleen vanuit het oogpunt van fraude wenselijk zijn, maar ook vanwege inhoudelijke of technische veroudering van opgaven.

Om de statistische eigenschappen van nieuwe items te relateren aan die van de items in een itembank, is kalibratieonderzoek nodig. Dit kalibratieonderzoek kan in de werkelijke afname plaatsvinden, of in een afzonderlijk kalibratieonderzoek met kleinere groepen leerlingen. Er is een aantal redenen om het kalibratieonderzoek en de werkelijke afname tegelijkertijd te doen. De eerste betreft de kosten die bespaard kunnen worden. Het tweede is dat de kans op het lekken van opgaven voorafgaand aan de werkelijke afname toeneemt als met afzonderlijk kalibratieonderzoek wordt gewerkt. Een derde is dat de motivatie tijdens een kalibratie-onderzoek en tijdens de werkelijke afname van elkaar kunnen verschillen[25].

Bij het trekken van toetsversies uit een itembank worden er vaak opgaven hergebruikt. Voor hergebruik kunnen regels opgesteld worden. Bijvoorbeeld dat opgaven die in een afnamejaar ingezet zijn, daarna twee jaar uitgesloten worden van hergebruik. Dit verkleint de kans dat de opgaven bekend worden, met name omdat de motivatie voor doorvertellen kleiner wordt. Ook kunnen regels voor hergebruik gekoppeld worden aan het aantal leerlingen dat een opgave gemaakt heeft, of de tijdspanne waarin een opgave hergebruikt kan worden. Alles om de kans op uitlekken te verkleinen. Bijvoorbeeld zodra meer dan 1000 leerlingen de opgave gemaakt hebben, wordt het hergebruik voor een bepaalde tijd bevroren. Of de periode waarin een opgave mag voorkomen in toetsversies wordt ieder jaar tot één week beperkt. Een langere afnameperiode kan dan gedekt worden met voldoende andere opgaven uit de itembank.

[25]Hiernaar is uitgebreid onderzoek gedaan dat ook leidde tot een academische promotie: Keizer-Mittelhaëuser, M-A. (2014). Modeling the effect of differential motivation on linking educational tests. [s.n.]; `https://pure.uvt.nl/ws/portalfiles/portal/5017508/Mittelhaeuser_modeling_12_12_2014.pdf`.

Over het selecteren van itemversies uit een itembank

Optie 1: vaste toetsversies uit de itembank Deze toepassing lijkt sterk op de verschillende versies zoals bij Scenario 1 omschreven, alleen wordt er (deels) gebruik gemaakt van eerder gebruikt materiaal. Het voordeel van deze werkwijze is dat versies van te voren goed uitgebalanceerd kunnen zijn wat betreft inhoud en moeilijkheid. Het nadeel kan zijn dat als de versies vast zijn, het afnamevenster kleiner moet zijn omdat anders bekend is welke opgaven het jaar worden afgenomen.

Optie 2: willekeurige toetsversies Uit de gehele itembank krijgt de leerling een willekeurige verzameling opgaven. De trekkingen moeten wel gestratificeerd gebeuren om er zeker van te zijn dat de toetsmatrijs wel gevolgd blijft worden. Het voordeel hiervan is dat het aantal mogelijke verschillende versies zeer groot is en zo fraude ingeperkt kan worden. Een nadeel is dat de versies –als daar niet ook voor gestratificeerd wordt– sterk kunnen verschillen in moeilijkheid. De impact van dat nadeel wordt geminimaliseerd door de toepassing van IRT om de resultaten op de verschillende versies vergelijkbaar met elkaar te krijgen.

Optie 3: multi-stage testing (MST) Deze variant is eerder beschreven in Hoofdstuk 5. De toets bestaat uit minstens twee fases. In de eerste fase wordt een toets(versie) van gemiddelde moeilijkheid aangeboden, waarna in de tweede fase een toets op maat wordt aangeboden. Als de leerling in de eerste fase een lagere vaardigheid laat zien, krijgt deze leerling een gemakkelijkere toets, die mogelijk ook eindtermen omvat die de meeste leerlingen al behaald hebben, maar deze leerlingen niet. Zo kan gecontroleerd worden of zij deze eindtermen toch beheersen. Dat is met name interessant als in een eerdere meting was gebleken dat zij die eindtermen nog niet beheersten. Leerlingen met een gemiddelde of hogere vaardigheid krijgen een toets die bij hun niveau past. Voor het toepassen van een dergelijke afname moeten de kenmerken van de opgaven in de bank gekend zijn. Dat geldt ook voor de nieuwe opgaven die niet eerder afgenomen zijn binnen de centrale proeven. Dit zou een reden kunnen zijn om wel over te gaan op een uitgebreid kalibratieonderzoek van tevoren. Een alternatief zou de volgende opzet kunnen zijn waarbij de opgaven opgenomen zijn in de eerste fase. Voor deze fase is een beperkte tijd beschikbaar, variërend van zeer beperkt (vaste dag, vaste tijd) tot een week. Na afloop van die fase worden alle opgaven gekalibreerd, worden de vaardigheden van de leerlingen bepaald en de versies voor fase twee overeenkomstig klaargezet. Deze alternatieve opzet is alleen mogelijk bij een goede automatisering van de kalibraties en de rest van het proces. Dit zal eerst nog een paar jaar ervaring met centraal toetsen kosten. Aan de andere kant moet de itembank toch eerst opgebouwd worden. Het voordeel van MST is dat de leerlingen zeer gericht gemeten worden. Met name in de tweede fase zijn de toetsen goed toegerust om exact op niveau te meten. Dat maakt de toetsing betrouwbaarder, wat ook de meting op geaggregeerd (school)niveau verbetert. Daarnaast zijn de toetsen inhoudelijk relevanter. Een voorbeeld daarvan was hierboven gegeven voor leerlingen met een wat lagere vaardigheid: zo kan bekeken worden wat ze nu wel kunnen, in plaats van dat focus ligt op wat ze niet kunnen.

Over het selecteren van itemversies uit een itembank vervolg

Optie 4: computer adaptieve toetsen (CAT) Bij deze variant wordt ieder item dat de leerling voorgelegd wordt, gebaseerd op de antwoorden van (alle) voorgaande items. Voor deze vorm van toetsen is een goede gekalibreerde itembank nodig. Ook nieuwe opgaven moeten van tevoren grondig gekend zijn, wat extra kosten met zich meebrengt. Het stelt ook grote eisen aan het afnamesysteem waar ook kosten mee gemoeid zijn. Ondertussen is de toename in meetnauwkeurigheid ten opzichte van MST beperkt. Vooralsnog is dit af te raden. Wellicht dat ICT-ontwikkelingen deze vorm van toetsen in grootschalige afnamen in de toekomst aantrekkelijker kunnen maken.

Hoe afnamejaren met elkaar te vergelijken?

Voor beide scenario's is het relevant dat de afnamen van de verschillende afnamejaren met elkaar vergeleken moeten kunnen worden. Indien gewerkt wordt met jaarlijks een nieuwe versie (of set van versies) verloopt dit anders dan indien gewerkt wordt met een itembank. Het kunnen koppelen van de versies van de verschillende afnamejaren op dezelfde vaardigheidsschaal is een noodzakelijke voorwaarde voor het werken met een itembank. Er zijn verschillende manieren om ervoor te zorgen dat IRT toe te passen is, zodat de versies van verschillende jaren aan elkaar te relateren zijn, ieder met eigen voor- en nadelen.

- **Optie 1: koppeling buiten de toets om (pretest/posttest)**
  Deze werkwijze is een vorm van kalibratieonderzoek naast de daadwerkelijke afname. Nieuwe opgaven worden gelijktijdig met ankeropgaven of oude opgaven afgenomen, bij een beperkte groep leerlingen. Een dergelijk onderzoek is groter dan een pilotonderzoek, waarbij alleen het functioneren van items gescreend wordt, en zullen voor goede schattingen bij 40 tot 50 scholen uitgevoerd moeten worden. Het is een werkwijze die enigszins robuust is. Het heeft ook enige nadelen (kosten, kwaliteit) die eerder zijn beschreven.
- **Optie 2: koppeling via een (geheim) intern anker**
  Bij deze werkwijze worden bij een aantal leerlingen opgaven aangeboden die andere leerlingen niet krijgen. Stel dat er een aparte set van 20 wiskundeopgaven is. Deze wordt bij de allereerste afname aangeboden bij duizend à tweeduizend leerlingen, bij voorkeur willekeurig verdeeld over scholen in Vlaanderen in plaats van 20 vergelijkbare wiskundeopgaven in de reguliere versies[26]. Als met verschillende versies gewerkt wordt, zijn de leerlingen gewend dat de opgaven kunnen verschillen met hun klasgenoten. Deze 20 beperkt afgenomen opgaven zullen door de beperkte afname niet snel bekend worden en kunnen het jaar erop op een vergelijkbare manier ingezet worden. De vergelijking over de jaren vindt dan plaats via deze opgaven. Deze werkwijze van het geheime anker is in een papieren format vele jaren succesvol toegepast bij de (high-stakes) eindtoets in Nederland. De opgaven werden niet bekend, de kosten zijn relatief laag en de kwaliteit van de gegevens is hoog. Bij het trekken

[26]Alternatief is dat in plaats van een blok van 20 opgaven bij 1500 leerlingen de 20 opgaven op een vergelijkbare manier in blokjes van 4 opgaven bij 7500 leerlingen afgenomen worden. Zo zijn er nog meer varianten te bedenken die van dit principe uitgaan.

van toetsversies uit een itembank is deze werkwijze zeer gebruikelijk. Een deel van de bank wordt aangewezen om na een bepaalde periode hergebruikt te worden. Een nieuwe versie bestaat deels uit nieuwe opgaven en deels uit ankeropgaven. Dit hergebruik zorgt voor een jaarlijkse ankering van nieuwe versies met de bank.

- **Optie 3: aanname van een gelijke populatie**
  Om de itemgegevens met elkaar te kunnen vergelijken is het ook mogelijk te stellen dat de vaardigheid over de verschillende jaren niet fundamenteel verandert: de leerlingen aan het einde van graad 1 in 2023 zijn verondersteld even vaardig te zijn als in 2024. Op zich is dit geen enorm vreemde aanname aangezien, zeker jaar op jaar, de verschuivingen in vaardigheid niet spectaculair zullen zijn. Deze werkwijze is ook gemakkelijk toe te passen en gekoppelde afnamen zijn niet nodig, wat betekent dat er echt telkens met 100% unieke toetsen gewerkt zou kunnen worden. Binnen IRT zijn er schattingsmethoden beschikbaar (*Marginal Maximum Likelihood*) die dit ook mogelijk maken. Het heeft echter wel belangrijke nadelen. Ten eerste is de aanname niet te controleren, dus als er wel verschuivingen zijn dan is dat niet te zien. Ten tweede is het meten van trends door de tijd niet mogelijk: de veronderstelling is juist dat er geen trends zijn. Ten derde kan de verschuiving tussen twee opvolgende jaren beperkt zijn, de trends over meerdere jaren hoeven dat niet te zijn, en die neem je op deze manier ook niet waar.
- **Optie 4: combinatie van de bovenstaande opties**
  Het is mogelijk om deze drie opties naast elkaar te gebruiken. Optie 2 kan dan leidend zijn, waarbij optie 1 kleinschaliger wordt uitgevoerd (zeg bij 10 tot 15 scholen), en optie 3 wordt gebruikt om te controleren of de met de eerste twee opties toegepaste kalibratie niet een te grote verschuiving oplevert.

Over het meten van eindtermen en het toepassen van rotatie van inhoud

Bij het meten van eindtermen speelt rotatie een rol. Er zijn scenario's met een vaste set van eindtermen binnen een afnamejaar. Er zijn ook scenario's waarin jaarlijks een wijzigende selectie van eindtermen aan de vaste set van items wordt toegevoegd (zeg, 25% op het totaal van de items). In het geval van roterende eindtermen, is het bovendien van belang of die wisselende eindtermen (zeer) hoog samenhangen met de vaste eindtermen of niet. Dit belang lichten we toe in onderstaande scenario's.

- **Scenario 1: hoge samenhang tussen de vaste en rotende eindtermen**
  Als de samenhang tussen vaste en roterende eindtermen hoog is, dan kunnen deze vaardigheden binnen een unidimensioneel model opgenomen worden. Dat is zeker mogelijk als de latente correlatie[27] tussen de vaardigheid die gemeten wordt met de vaste set van eindtermen, en de roterende eindterm groter is dan 0,90. Als de latente correlatie onder de 0,80 valt, is dat niet mogelijk. Voor de waarden daartussen moet robuustheidsonderzoek gedaan worden als de opgaven toch als één vaardigheid beschouwd worden. Er is dan sprake van een schending van unidimensionaliteit en de impact van die schending moet onderzocht worden. Ook kan er direct overgestapt worden op Scenario 2. Het voordeel van de werkwijze bij Scenario 1 is dat deze

[27] Dit is de correlatie tussen de latente trek na correctie voor meetfout. Dit is de IRT equivalent van correlatie na attenuatiecorrectie in Klassieke Test Theorie.

gemakkelijker is.

- **Scenario 2: lage samenhang tussen de vaste en roterende eindtermen**
  Wanneer de roterende eindtermen niet hoog genoeg samenhangen met de vaste eindtermen dan meten deze kwalitatief wat anders. Dat is vergelijkbaar met de situatie die besproken is rond leerwinst wanneer eindtermen over de verschillende leerjaren vooral kwalitatief verschillen: er wordt niet meer van hetzelfde, maar meer vaardigheden gemeten. Sterker nog: bij de beschrijving van die mogelijkheid gaat het om de situatie waarin de vaste set van eindtermen uiteindelijk ook kwalitatief verschilt. Mocht het zo zijn dat de vaste set van eindtermen echt niet kwalitatief verandert over de leerjaren, dan kan de leerwinst over de vaste set van eindtermen bepaald worden. Bij een lage samenhang tussen de vaste en de wisselende eindtermen kunnen deze beter uit de (traditionele) leerwinst analyses gehaald worden. In plaats van een deel van de onderwijsdoelen ieder jaar te laten roteren, is een alternatief om naast de vaste kern van eindtermen, de overige eindtermen per leerling met een zeer beperkt aantal opgaven te meten. Doordat deze roterende set anders over de jaren heen toch zou verschillen, is leerwinst op individueel niveau niet goed haalbaar. Zeker niet als deze roterende set niet sterk samenhangt met de vaste set. Ook een aparte score op leerlingniveau voor een specifieke set roterende eindtermen is niet aannemelijk, omdat dat zou vereisen dat er voldoende van dergelijke opgaven opgenomen zouden zijn om deze betrouwbaar te meten. Hoewel niet vastligt hoeveel opgaven dat zouden zijn, zal dat in de richting van 25 opgaven liggen. Als dat 25% van de opgaven is, zou iedere leerling per vaardigheid 100 opgaven moeten maken. Het is de vraag of dat wenselijk en haalbaar is. Om bovengenoemde redenen, zou het zinvol kunnen zijn om niet het leerlingniveau als belangrijkste rapportageniveau te zien, maar het school- en systeemniveau. De rapportage op de roterende eindtermen vindt dan alleen plaats op de geaggregeerde niveaus. Het is evident dat het aantal opgaven per eindterm per toets zeer beperkt is, echter als er met diversie versies gewerkt wordt binnen een leerjaar dan kunnen deze eindtermen op schoolniveau, en in ieder geval op systeemniveau goed bepaald worden. Die maakt het mogelijk voor de te roteren eindtermen op geaggregeerd niveau alsnog leerwinst en trends door de tijd te bepalen. Een alternatief om deze gegevens alsnog mee te nemen is de aanpak die beschreven is in Paragraaf 7.2.2 waarbij de focus vooral gericht is op de meest recente meting. De eerdere metingen – zowel de vaste als de losse schattingen van de roterende vaardigheden – kunnen dan afzonderlijk als voorspellende priors meegenomen worden in de analyses voor leerwinst.

Hoe om te gaan met brede afnamen en afwijkingen van de gestandaardiseerde afname?

Om leerlingen en scholen goed met elkaar te kunnen vergelijken, verdient het de voorkeur om de afnamen te standaardiseren. Dat betekent dat de condities van de afnamen gelijk zijn en de toetsen grotendeels gelijk zijn. We hebben al gezien dat door middel van IRT van de gelijke toetsen afgeweken kan worden, mits aangetoond wordt dat de opgaven op een schaal liggen. Aan welke eisen toetsen dan moeten voldoen, is beschreven in Hoofdstuk 5. Standaardisatie kan betrekking hebben op meerdere aspecten van de afname. Hieronder

gaan we in op omstandigheden, instructies en het gebruik van oefentoetsen.

- **Omstandigheden**
  Als we kijken naar standaardisatie, dan hebben we het over de omstandigheden waarin de leerlingen de toetsen maken. Dat heeft te maken met praktische zaken zoals de vastgestelde toetstijd, en de kenmerken van de omgeving (klassikaal/individueel; soort ruimte). Bij digitale toetsen heeft dat ook te maken met de technische omstandigheden. Als de technische voorzieningen op de ene school beter zijn, of de toetsen beter draaien op specifieke devices dan op andere en de scholen daarin verschillen dan gaat dat zeker invloed hebben op de resultaten.
- **Instructies**
  Standaardisatie heeft ook te maken met de instructies. Daar vallen de instructies naar de leerlingen onder, en de toegestane hulp die de leerkracht mag geven, maar ook hoe de school met de resultaten omgaat. De impact van dat laatste punt is ook uitvoerig aan bod gekomen in Paragraaf 7.2.1.
- **Oefentoetsen**
  Standaardisatie heeft ook betrekking op de mate van noodzakelijke bekendheid met de afnamevorm. Het is aan te raden oefentoetsen te maken om leerlingen te laten oefenen met de afnamevorm, zodat bekendheid met computertoetsen niet sterk onderling verschilt. Hoe de toetsen exact gestandaardiseerd moeten worden is moeilijk van tevoren geheel te voorzien. Dat zal ook moeten blijken in het ontwikkelproces van toetsen. Wat in ieder geval van tevoren duidelijk gecommuniceerd moet worden, zijn de voorwaarden en de regels die gelden rond de toetsafname. Bij een brede afname zal al snel afgeweken moeten worden van de standaard-afnamewijze. Het is verstandig om in de itemconstructie al rekening te houden met de diverse kenmerken van leerlingen die beperkingen kunnen ervaren bij de afname. Dit kan afwijkingen van de standaard-afnamewijze voorkomen. Voorbeelden zijn al gegeven van opgaven die even geschikt zijn voor kleurenziende als kleurenblinden, of digitale afnamen waar rekening gehouden wordt met linkshandigen. Andere voorbeelden zijn simpel taalgebruik bij wiskundeopgaven, om zo zuiver mogelijk de wiskundevaardigheid te meten en zo min mogelijk de taalvaardigheid. Er blijven nog wel wat scenario's over voor verschillende leerlingen, waarbij in alle gevallen equivalentieonderzoek noodzakelijk is.
- **Laagvaardige leerlingen**
  In sommige gevallen zijn de verschillen niet zozeer kwalitatief van aard, maar kwantitatief. Dat betreft leerlingen die sterk achterlopen in hun vaardigheid. In die gevallen kunnen beter eenvoudigere varianten van de toetsen worden aangeboden voor betrouwbare metingen. Als MST beschikbaar is, dan is dat gemakkelijk. In de eerste jaren kunnen ook eenvoudigere versies aangemaakt worden, die op voorhand aan die leerlingen worden aangeboden. Zolang er overlap met andere versies is, zijn deze leerlingen door middel van IRT te vergelijken met andere leerlingen.
- **Dyslectische leerlingen**
  Twee opties die hier gebruikt kunnen worden, zijn voorgelezen (verklankte) toetsen en verlengde toetstijd. Voor voorgelezen toetsen moet de afname-omstandigheid daar wel geschikt voor zijn (andere locatie, koptelefoon). De verlengde toetstijd is minder invasief. Het zou niet zo moeten zijn dat leerlingen die geen dyslexie hebben

ook baat zouden hebben bij verlengde toetstijd. De toetstijd zou zo ingericht moeten zijn dat voor reguliere leerlingen dat in ieder geval genoeg is.

- **Leerlingen met dyscalculie**
  Een mogelijke optie bij deze leerlingen is hen toe te staan om een (al dan niet digitaal ingebouwde) rekenmachine te gebruiken. De vraag die in die gevallen openstaat, is of de toets dan nog wel hetzelfde meet als bij anderen leerlingen. Kortom, equivalentie kan nog wel een probleem zijn. De oplossing kan dan zijn om alle leerlingen in de gelegenheid te stellen een rekenmachine te gebruiken. Alleen als de validiteit van de meting teniet wordt gedaan door het opnemen van een rekenmachine, hetgeen afhankelijk is van de omschrijving van de eindterm, zou die rekenmachine afwezig moeten zijn. Vanuit het oogpunt van validiteit zou die dan ook moeten ontbreken voor dyscalculische leerlingen. Vanuit het oogpunt van 'fairness' zou er een pleidooi gehouden kunnen worden dat zij daar ook beschikking hebben over een rekenmachine[28].
- **Blinde leerlingen**
  Voor blinde leerlingen zijn voorzieningen beschikbaar die hen in staat stellen om ook digitaal toetsen te maken. Het zijn voorzieningen waar de niet-blinde leerlingen geen baat bij hebben. Bij deze aanpassingen kan equivalentie lastiger zijn omdat de vorm van de opgaven soms ook aangepast moet worden. Gezien de beperkte aantallen is de impact van de (lichte) afwijkingen van equivalentie minder problematisch. Wanneer er leerlingen zijn met (zeer) zeldzame kenmerken is het vaak te kostbaar om een aparte toetsversie te maken die daarvoor accommodeert. In veel gevallen is equivalentie dan ook moeilijk te bereiken of moeilijk aan te tonen, onder andere door de lage aantallen.
- **Beperkingen in digitale afname**
  Het kan zijn dat leerlingen niet in staat zijn de toets digitaal af te nemen. Dit element van de brede afname hoeft niet zozeer aan de leerling te liggen, maar kan ook aan de omstandigheden liggen. Als de toets afgenomen moet worden, maar de digitale infrastructuur maakt het onmogelijk, dan is het mogelijk dat de toets ook op papier uitgegeven wordt. Zeker in het eerste jaar van de afname –dat is voor het secundair onderwijs 2023 en voor het lager onderwijs 2025– is het verstandig om van tevoren bij de scholen te inventariseren of de school er klaar voor is. Scholen die er niet klaar voor zijn, kunnen wellicht ondersteund worden, maar als zij zelfs dan op het moment dat de centrale proeven moeten plaats vinden niet klaar zijn voor digitale afnamen moeten papieren varianten beschikbaar zijn. Bij papieren toetsen kunnen een heleboel zaken niet, die digitaal wel kunnen. Dat betreft zaken als het aanbieden van opgaven in willekeurige volgorde en het gemak waarin verschillende versies aangeboden worden. Ook bepaalde opgaven zijn digitaal wel aan te bieden, maar op papier niet, wat ook al aangehaald is in Paragraaf 7.2.3. Het is zodoende aan te raden in het eerste jaar, of in de eerste jaren, opgaven te maken in een vorm waarin gemakkelijk een papieren equivalent te maken is. Ook in dat geval is nog steeds equivalentieonderzoek nodig[29]. Hoe meer de scholen ervaring hebben met de

[28] In dit rapport zal niet de keuze gemaakt worden tussen deze twee. Dit zal in overleg met het veld beslecht moeten worden.

[29] Onderzoek naar papier en digitale equivalentie is gebruikelijk. Idealiter wordt daar in oefenonderzoek een

digitale centrale proeven, en als bekend is dat de infrastructuur op alle scholen erop berekend is, is het toepassen van de digitale mogelijkheden bij de opgaven zeker aan te raden. Wanneer dat zal zijn, is niet bekend, maar voor het eerste jaar van afnamen is het echter aan te raden geen opgaven te maken die digitaal ambitieus zijn. Als alle scholen de digitale toetsen moeten kunnen maken, dan moeten deze ook op de scholen met een minder sterke IT-structuur te doen zijn. Naast de scholen waarvan van tevoren bekend is dat zij niet klaar zijn voor digitale toetsen, zullen er mogelijk ook scholen zijn met onverwachte technische problemen. Een noodscenario waarbij een speciale toetsvariant op papier toegestuurd kan worden, is dan noodzakelijk. Het is de verwachting dat met de bekendheid met de digitale centrale proeven, de noodzaak voor papieren varianten afneemt[30].

Hoe om te gaan met toetslengte en toetsduur?

Om een betrouwbare meting te krijgen, moeten de leerlingen voldoende opgaven maken per vaardigheid (zie Sectie 5.4.1). Hoe meer opgaven de leerling moet maken, hoe langer de toetsduur. In de loop van de tijd kan een deel van de nauwkeurigheid ook verkregen worden uit een vorm van adaptief toetsen, maar daarvoor moet eerst een goede itembank ontwikkeld worden. Het is lastig in te schatten wanneer dat is, maar dat zal waarschijnlijk niet voor 2025 zijn. Wat het benodigde aantal opgaven is voor een betrouwbare meting is van tevoren lastig in zeer specifieke scenario's te vatten. Wel is van tevoren aan te geven welke factoren in ieder geval een rol spelen.

- **De kwaliteit van de opgaven**
  Hoe minder de kwaliteit, hoe meer opgaven nodig zijn om toch tot een goede meting te komen. Om efficiënt te meten is het dus van belang goede opgaven te maken, die de vaardigheid goed meten. Dit heeft ook met de inhoudsvaliditeit te maken.
- **Mate van unidimensionaliteit van de te meten vaardigheid**
  Naarmate de vaardigheid meer unidimensioneel te meten valt, hoe meer de opgaven zullen samenhangen, en hoe hoger de betrouwbaarheid. De keuze van de te meten eindtermen, als ook de operationalisering van die eindtermen zijn dus van invloed op de betrouwbaarheid van de meting. Het kan dus schelen om een aantal unidimensionele schalen te maken, in plaats van alles bij elkaar als één meting te geven.
- **Spreiding van de vaardigheid in de populatie**
  Hoe meer de leerlingen op een te meten eigenschap van elkaar verschillen, hoe hoger de betrouwbaarheid. Dit is een aspect waar een toetsproducent minder invloed op heeft. Het is dus wel van belang dat zaken gemeten worden waarvan verwacht wordt dat leerlingen van elkaar te onderscheiden zijn.
- **Individuele meting versus geaggregeerde meting**
  Een individuele meting heeft meer meetfout dan een meting waarbij een aantal

deel van de toets digitaal en een deel op papier gedaan, wat vier varianten oplevert: deel A&B papier, deel AB digitaal (reguliere afname), deel A papier, deel B digitaal, en deel A digitaal, deel B papier. De correlaties tussen A en B zouden in alle vier de gevallen gelijk moeten zijn.

[30] Merk op dat ook wanneer leerlingen een eerdere meting op papier gedaan hebben en een latere meting digitaal equivalentieonderzoek, samen met IRT schaling noodzakelijk is om goed leerwinst te bepalen. Ook voor die situatie is goed afnamemodus-equivalentie onderzoek nodig.

individuele metingen bij elkaar genomen worden. Het kan dus zijn dat met name in het geval dat er sprake is van meer gedetailleerde meetschalen, het niet mogelijk is op individueel niveau betrouwbaar te rapporteren, maar wel op schoolniveau. Een meting wiskunde in het lager onderwijs voor een leerling blijft dan voor een leerling een enkele vaardigheidsscore, terwijl op schoolniveau de eindtermen ook nader uitgesplitst kunnen worden om zo voldoende betrouwbare schoolscores te krijgen voor getallen, meten, meetkunde en toepassingen.

- **Mate van tijd die aan toetsen besteed mag worden**
  Als men ook op leerlingniveau de vaardigheid op subschalen afzonderlijk wil kennen, is er meer toetstijd nodig. Veelal betekent dit meerdere afnamesessies, omdat de aandachtspanne van leerlingen niet onbeperkt opgerekt kan worden. Daarvoor moet ook draagkracht in het veld zijn. Wat daarbij helpt is dat de leerkrachten en de scholen het nut van de toetsen ook ervaren in hun wens om de kwaliteit van het onderwijs te verbeteren. Goede rapportages kunnen daar sterk bij helpen.

Hoe om te gaan met externe toetsinformatie?

Het is mogelijk en in een aantal gevallen zelfs wenselijk dat informatie van toetsen die onder supervisie van scholen afgenomen zijn, gecombineerd worden met de metingen van de centrale proeven. Er zijn meerdere bronnen van externe informatie die hieronder uitgewerkt worden. Ongeacht de bron zal er een goed administratief systeem moeten zijn waarin de scores per leerling ingevoerd kunnen worden, opdat een goede koppeling met de resultaten van de centrale proeven mogelijk is. Bij voorkeur worden de resultaten van de leerlingen op het meest gedetailleerd niveau ingevoerd. Dat betekent dat iedere score op het beoordelingsformulier wordt opgenomen in de data. Als dat handmatig moet, zal dat te veel werk zijn. Wat in ieder geval opgeslagen moet worden is het eindoordeel van de beoordelaar van de toets.

Er kan bij veelvoorkomende externe toetsen voor gekozen worden om de responsen door middel van optisch leesbare formulieren te verzamelen, al dan niet centraal. Maar dat zal echter tijd en geld kosten, onder andere doordat dan een eigen logistieke stroom moet worden opgezet. Het is aannemelijk dat in de beginjaren van de centrale proeven dit geen prioriteit zal krijgen in de reguliere centrale afnamen[31].

De beoordeling van externe toetsen kan binnen de school plaatsvinden, door de eigen leerkracht. Dat is een relatief goedkope optie, die uitgaat van goed vertrouwen in de eigen leerkracht. Er zijn ook externe alternatieven mogelijk, als dezelfde toets bij leerlingen van verschillende scholen wordt afgenomen.

Als het mogelijk is om de afzonderlijke responsen van leerlingen goed op te slaan, dan kunnen deze responsen ook over verschillende Vlaamse leerkrachten verdeeld worden. Daarbij kan gekozen worden om alle responsen van één leerling naar dezelfde beoordelaar te sturen, of de responsen van één leerling te segmenteren en naar verschillende beoordelaars te sturen. In het eerste geval krijgt een leerkracht van wie een bepaald aantal leerlingen de toetsen maakt, de werken van precies dat aantal leerlingen te beoordelen,

[31] Het is wel sterk aan te raden in vooronderzoek op een kleinschaliger niveau deze informatie op detailniveau te verzamelen. Dergelijk onderzoek is nodig bij het ontwikkelen van dergelijke opdrachten. Zonder die informatie is het namelijk lastiger de kwaliteit van de opdrachten te kunnen evalueren en verbeteren.

maar dan wel van verschillende Vlaamse scholen. Een dergelijke werkwijze heeft ook tot gevolg dat het niet een enkele beoordelaar is die de beoordelingen aan een school geeft, maar dat het aantal beoordelaars gelijk is aan het aantal leerlingen. Dit heeft tot gevolg dat de impact van de strengheid van de beoordelaars verdeeld wordt, wat de vergelijkbaarheid van de scholen vergroot. Een nadeel is dat de leerlingen wel last kunnen hebben van verschillen in de mate van strengheid van beoordelaars, en zo binnen een school mogelijk lastiger te vergelijken zijn[32].

Er zijn ervaringen met dergelijke systemen: in Nederland worden bijvoorbeeld de schrijfwerken voor Nederlands als tweede taal zo gesegmenteerd gescoord. Daar wordt standaard met twee beoordelaars per respons gewerkt[33]. Aan het toevoegen van meer beoordelaars, zijn natuurlijk ook meer kosten verbonden. Er zijn uiteraard ook kosten verbonden aan het inzetten (aankoop, of zelf bouwen) van een dergelijk systeem dat de toewijzing van responsen aan beoordelaars organiseert.

Als de resultaten van de toetsen geen high-stakes impact hebben op de scholen, en de oordelen van de eigen leerkrachten vertrouwd worden, kunnen al die kosten bespaard worden. Bij de eerste afname van de centrale proeven is het beperken van de belangen in ieder geval al aan te raden, waardoor de invoer van een dergelijk systeem in de eerste jaren geen prioriteit hoeft te kennen.

Indien er al externe toetsinformatie in andere (bijvoorbeeld net-eigen, of school-administratie-) systemen opgeslagen staat, is het bouwen van een koppeling tussen het externe systeem en de dataopslag van de centrale proeven een relatief makkelijke manier om gegevens aan elkaar te kunnen relateren.

De externe toetsen waarvan de resultaten gekoppeld zouden kunnen worden aan de resultaten op de centrale proeven kennen meerdere bronnen. Enkele daarvan worden hieronder besproken.

- **Moeilijk meetbare vaardigheden**
  Een goed voorbeeld hiervan zijn toetsen van moeilijk meetbare vaardigheden. De vergelijking daarmee is met name functioneel als deze toetsen en beoordelings-voorschriften centraal worden ontwikkeld, maar de afname en beoordeling onder supervisie van de school gebeurt. Productieve vaardigheden zoals spreken zijn lastiger om valide in een centrale toets af te nemen, maar zijn wel zeer belangrijk. Als de spreekvaardigheid van de leerling niet meegenomen wordt in de evaluatie van de school, bestaat de kans dat sommige scholen daar dan ook minder aandacht aan gaan besteden. Op die manier zou dat leiden tot een onwenselijke verschraling van het onderwijs. In Paragraaf 5.2.4 is beschreven hoe moeilijker toetsbare vaardigheden ook centraal ontwikkeld kunnen worden. Wanneer vanuit het steunpunt richtlijnen voor moeilijk te toetsen vaardigheden aan de scholen aangeboden worden, dan kan de afname en scoring door scholen worden uitgevoerd. Dat kan ook toetsen betreffen

[32] Een simpele oplossing is toch één (externe) beoordelaar per school toe te wijzen, maar dat gaat dan weer ten koste van de vergelijkbaarheid van scholen. De keuze tussen deze twee opties zal afhangen waar meer focus op komt te liggen: de school of de leerling. Vooralsnog wordt er vanuit gegaan dat de vergelijkbaarheid van de scholen belangrijker is. Een optie leerlingen binnen de school vergelijkbaar te krijgen is dat de eigen docent voor de beoordeling van de eigen leerlingen het laatste woord heeft, daar waar het gaat om de impact voor de leerling. Hiervan zijn verschillende varianten, met verschillende voorwaarden uit te werken.

[33] In het geval dat de twee beoordelaars onderling sterk verschillen in oordeel, wordt een derde beoordelaar toegevoegd.

waarbij de afname mogelijk wel centraal is, maar het automatisch scoren van de toetsen moeilijk is, en aan scholen zelf wordt toevertrouwd. Een voorbeeld daarvan kunnen de schrijftaken zijn in het secundair onderwijs[34]. Het werk dat vanuit het steunpunt moet worden gedaan is dat dergelijke taken (het pakket met opdrachten, instructies, en scoringsvoorschriften) ontwikkeld moet worden, en dat er een systeem ontwikkeld moet worden om de leerkrachtoordelen goed in te voeren. De centrale administratie van moeilijk meetbare vaardigheden heeft als voordeel dat de resultaten van alle leerlingen in Vlaanderen op deze vaardigheden geëvalueerd kunnen worden, en tot betere metingen van deze vaardigheden kunnen komen. Door de resultaten van deze vaardigheden te relateren aan de (digitale) centrale proeven kunnen ook scholen met opvallende scores op deze vaardigheden gevonden worden. Als deze relatief hoog zijn, kan er bij de school langsgegaan worden om te zien wat de school zo succesvol maakt in dergelijke vaardigheden. Het kan namelijk ook zo zijn dat de school het beoordelingsmodel niet geheel zoals bedoeld heeft toegepast en zo overschattingen van bijvoorbeeld spreekvaardigheid geeft. Een school kan dan ondersteund worden in het wel goed toepassen van de beoordelingsmodellen.

- **Net-eigen toetsen**
  Naast de metingen van moeilijk meetbare vaardigheden kan ook nog steeds veel waarde gehecht worden aan de net-eigen toetsen. Deze gegevens kunnen ook centraal geïmporteerd worden, zodat de resultaten makkelijk vergeleken kunnen worden met de resultaten op de centrale proeven. Door middel van regressiemodellen kunnen toetsen van verschillende netten aan elkaar gerelateerd worden, waarbij de centrale proeven kunnen functioneren als anker. Daar waar de net- of koepelspecifieke toetsen moeilijk meetbare vaardigheden meten kunnen deze ook zeker naast de centrale toetsen afgenomen worden. Zolang er geen andere extern ontwikkelde meetinstrumenten zijn voor deze vaardigheden is dat ook aan te raden. Ook hier helpt het curriculumvernauwing tegen te gaan. Het doet ook recht aan de eigenheid van de school. Ook andere toetsen met niet automatisch gescoorde opgaven kunnen een meerwaarde hebben als aanvulling naast de centrale toetsen. Het vervangen van delen van de centrale toetsen door net- of koepelspecifieke toetsen is af te raden. Het integreren van delen van de net- of koepelspecifieke toetsen in de ontwikkeling van de centrale toetsen is voor te stellen, maar levert voor de vergelijkbaarheid van de toetsen en resultaten wel de nodige uitdagingen. Het is zeker voorlopig af te raden. In hoeverre de koppeling van net-eigen toetsen aan de centrale proeven in een vroeg stadium van het centrale proeven uitgevoerd moet, of kan, worden hangt af van de beleidswensen, en de overleggen met de netten. Technisch is een dergelijke koppeling weliswaar een stevige klus, maar niet onmogelijk.
- **School-eigen gegevens**
  De voorbeelden hierboven betreffen de verwerking van gegevens waarvan de toetsen centraal of op net-niveau gekend zijn. Als de scholen zelf items en toetsen ontwikkelen, en die gegevens ingevoerd moeten worden, dan moeten er ook gegevens

[34] Als we de nieuwe eindtermen bekijken voor de eerste graad dan kan het lastig worden om hier automatisch scoorbare opgaven voor te maken. Bij de huidige eindtermen voor derde graad lijkt het het helemaal moeilijk (zeker zonder moderne machine learning technieken) om zinvolle, valide taken te maken die automatisch te scoren zijn.

over deze toetsen mee opgeslagen worden. Een dergelijk systeem is lastiger op te zetten. Op het niveau van de toets moet in ieder geval opgeslagen worden wat de meetpretentie is (wat wordt er gemeten), het aantal opgaven, de mogelijke scorerange, en de grens tussen de voldoende en de onvoldoende op die scoreschaal. Het is het gemakkelijkst om dit sterk voor te structureren. Op deze manier kunnen de schooltoetsen direct gerelateerd worden aan de centrale proeven. Een (aanzienlijk) ambitieuzere stap is het invoeren van itemgegevens. Hiermee zou een Vlaamse itembank ontwikkeld kunnen worden. Scholen voeren zelf hun opgaven in, evenals de scores op die items. Een dergelijke optie, maar ook de optie om de toetsgegevens op schoolniveau in te voeren zullen waarschijnlijk niet hoog geprioriteerd worden, en is meer iets voor de (verre) toekomst.

7.2.4 Dilemma's rond rapportage

Zoals eerder is aangegeven is voor goed toetsgebruik de rapportage van zeer groot belang. Pas als de rapportage goed begrepen wordt, kan het beoogde doel bereikt worden. Per doelgroep kunnen de eisen aan de rapportagevorm verschillen, om deze begrijpelijk te krijgen. Dat heeft te maken met het feit dat doelgroepen verschillen in handelingen op basis van de rapportages. Zo zal een leerkracht andere acties moeten ondernemen, dan dat een schoolbestuur dat moet doen. De groepen kunnen onderling, maar ook intern, verschillen in expertise, achtergrondkennis en datageletterdheid. Als de keuzes gemaakt worden zal er dus niet één vorm van terugrapportage zijn, maar verschillende vormen die voor ieder van de doelgroepen begrijpelijk genoeg moeten zijn. Een belangrijke vraag is hoe nu de rapportagevorm en de kennis, expertise en datageletterdheid van de belanghebbende bij elkaar gebracht kunnen worden. Dat kan door de expertise van gebruikers aan te passen aan de rapportages (handleidingen, cursussen), maar ook door hen meer bij het ontwerp van de rapportages te betrekken. Dat laatste kan ook helpen om ervoor te zorgen dat de belanghebbenden, bijvoorbeeld de leerkrachten, precies weten wat ze moeten doen op basis van de rapportages. Voordat we dit bespreken gaan we eerst in op de keuze tussen relatief en absoluut rapporteren. Deze paragraaf besluit met een manier om rankings te voorkomen. Deels is dit laatste punt ook al in Hoofdstuk 6 besproken, maar hier wordt een kort overzicht gegeven van de mogelijkheden.

Uit welke vormen van terugrapportages te kiezen?

Op deze vraag wordt uitgebreid antwoord gegeven in Hoofdstuk 6. Daarin is een overzicht gegeven, van de mogelijkheden die bij verschillende doelen passen. Deze rapportagevormen hangen ook sterk samen met de toegepaste vorm van normering. Zoals aangegeven past een absolute normering voor alle aggregatieniveaus goed bij een formatieve functie van toetsen, maar kan het ook toegepast worden bij de summatieve toetsen. Een relatieve normering, en bijbehorende rapportagevormen, lokt vooral summatief gebruik van de toetsen uit, waarbij de formatieve functie van een toets vaak wegvalt. Dit levert drie hoofdscenario's op voor representatie van de resultaten, die hieronder besproken worden. Het is een keuze of op alle rapportageniveaus dezelfde keuze gemaakt wordt. Het is mogelijk om voor de rapportage op leerlingniveau andere keuzes te maken dan op het niveau van de school. Het zou ook kunnen dat de rapportagevorm aangepast wordt aan de doelgroep

(bijvoorbeeld een afzonderlijke rapportage voor ouders en voor leerkracht over dezelfde leerling). Dat is voor de begrijpelijkheid van de rapportages aan te raden, maar daarbij moet wel gezocht worden naar afstemming tussen die rapportages. Als de informatie namelijk divergeert, delen de verschillende belanghebbenden niet langer hetzelfde beeld, wat bijzonder verstorend kan werken bij het verbeteren van de onderwijskwaliteit.

- **Scenario 1: Relatieve normering – rapportage gebaseerd op ordening van resultaten**
  Zoals beschreven in Hoofdstuk 6 is een voordeel dat deze wijze van normeren en de daaruit volgende rapportage relatief makkelijk te verkrijgen en te interpreteren is. Veel betrokkenen vinden deze vorm vanwege kosten en bekendheid vaak prettig. Een nadeel is dat deze vorm tot rangordeningen leidt, en afleidt van de inhoud, en daarmee de formatieve functie van de toetsen niet gebruikt wordt. Binnen deze vorm van normeren is er nog een keuze te maken over de vergelijkingsgroep. Die keuze gaat in de praktijk over het wel of niet corrigeren voor achtergrondvariabelen (leerling, leerkracht, school), en zo ja, welke achtergrondvariabelen. Ook is het mogelijk zowel een gecorrigeerde als een niet-gecorrigeerde rapportage te geven, of een aantal verschillende gecorrigeerde rapportages. Meer rapportages geven een verfijnder beeld, maar het is dan ook ingewikkelder om te interpreteren.
- **Scenario 2: Absolute normering – rapportage gebaseerd op inhoud van de toetsen**
  Het voordeel van deze normering is dat deze sterk op de inhoud van de toetsen is gericht en daarmee ook werkelijk op de inhoud van het onderwijs. De absolute normering kan aangeven wat het te verwachten niveau is, ongeacht of nu veel of weinig leerlingen dit niveau halen. Het is daarmee een constant niveau. Het is ook niet gebaseerd op vergelijkingen van leerlingen en scholen en lokt daardoor ook minder rangordeningen uit. Het nadeel is dat voor het bepalen van het grenspunt of de grenspunten die aangegeven of het beoogde niveau bereikt is, additioneel onderzoek nodig is. Een ander nadeel is dat de belanghebbenden vaak iets minder hiermee bekend zijn, en de ervaren houvast van de relatieve normering soms missen. Er moet bij deze wijze van rapporteren gekozen worden welke (en hoeveel) grenswaarde(n) bepaald worden. Ook zijn er meerdere manieren, ook wel standaardsettingsmethoden genoemd, om de grenswaarden te bepalen (meer hierover in Hoofdstuk 6). Een grenswaarde kan geïllustreerd worden met opgaven. Er zijn vele mogelijkheden om dit af te beelden, zeker digitaal. Bij het vormgeven van een rapportage moet hier een keuze tussen gemaakt worden.
- **Scenario 3: Zowel relatief als absolute normering**
  Het is ook mogelijk om zowel een relatieve als een absolute rapportage te geven, om zo alle voordelen van beide vormen te kunnen gebruiken. Het nadeel is dat in deze gevallen de ervaring is dat na verloop van tijd belanghebbenden vooral kijken naar de relatieve vergelijking, en de inhoudelijke interpretatie op de achtergrond komt[35]. Het gebruik van de formatieve functie, die meer moeite kost, komt daarmee ook meer

[35]Een opmerkelijke paradox is dat veel leerkrachten een voorkeur hebben voor een formatieve functie van toetsen, maar ondertussen ook vaak vragen om een relatieve normering. De oorzaak is hier niet geheel bekend, maar de wens om te weten waar de leerling, klas of school staat ten opzichte van een vergelijkbare populatie is zeer sterk, ook als de inhoudelijke interpretatie daardoor ondersneeuwt.

op de achtergrond.

Hoe te komen tot de juiste rapportage?

Of een instrument een positief effect heeft op de vaardigheid van de leerlingen en dus ook op de kwaliteit van het onderwijs, hangt voor een groot deel af van het gebruik van het toetsen en de opvolging van de toetsresultaten. Die hangt weer af van de rapportage. Vaak is voor het nut van de toets de (vorm van de) rapportage minstens zo belangrijk als het instrument zelf, zo niet belangrijker. Desalniettemin is de tijd die in de ontwikkeling van het instrument gestoken wordt vaak vele malen groter dan de tijd die in de ontwikkeling van de rapportage gestoken wordt. In het traditionele scenario wordt pas een tijd na de ontwikkeling van de toets gestart met de ontwikkeling van de rapportages. Met een beperkte groep personen, met beperkt onderzoek worden (vaak traditionele) keuzes gemaakt, waarbij de invloed van degenen die de toetsen moeten gebruiken beperkt is. Veel rapportageontwikkeling vindt zo plaats.

In een scenario waarbij de rapportage centraal staat[36], worden dezelfde stappen gezet bij de ontwikkeling van de rapportages als bij de ontwikkeling van toetsen. Denk aan stappen als de selectie van de te toetsen domeinen en eindtermen, het opstellen van een toetsmatrijs, het ontwikkelen van toetsvragen, pilotonderzoek en kalibratieonderzoek. Vergelijkbare stappen kunnen dus ook worden genomen rond de ontwikkeling van de rapportages:

1. Selectie van de te rapporteren toetsdoelen en rapportage eenheden: dit gebeurt in samenspraak met de opdrachtgever, rekening houdend met de andere belanghebbenden in het veld;
2. Opstellen van een rapportagevorm: dit betreft in ieder geval de keuzes die eerder in deze paragraaf genoemd zijn. Deze rapportage-matrijs bepaalt voor iedere doelgroep een blauwdruk voor de rapportage, en zorgt ervoor dat de rapportage representatief is voor de handelingen die moeten volgen op basis van de toetsen. In deze fase moeten de doelgroepen die moeten handelen op basis van de rapportage betrokken worden, om ervoor te zorgen dat de rapportagevorm gericht is op het juiste handelen;
3. Ontwikkelen van rapportagevormen: de eerste ontwerpen en noodzakelijke onderzoeken worden uitgevoerd, rekening houdend met de rapportage in een digitale omgeving;
4. Pilotonderzoek bij belanghebbenden die met de rapportages moeten werken. Kleinschalig onderzoek met een beperkt aantal experts. Op basis van de analyse van deze resultaten kunnen de rapportages verder aangepast of weggelaten worden indien nodig;
5. Grootschalig onderzoek bij een representatieve steekproef van belanghebbenden. Hierbij wordt naar de begrijpelijkheid van de rapportages gekeken, en of duidelijk is wat de vervolgacties zouden moeten zijn. Op deze manier wordt optimaal rekening gehouden met hoe de diverse doelgroepen de toetsen gaan gebruiken, maar kan ook rekening gehouden worden met de expertise, achtergrondkennis en datageletterdheid

[36]Deze aanpak sluit aan bij de aanpak van het Evidence Centred Design, waarbij de rapportage ook een integraal onderdeel is van de toetsontwikkeling. Zie onder ander Mislevy et al. (2003) A Brief Introduction to Evidence-centered Design. ETS, Princeton; `https://www.ets.org/Media/Research/pdf/RR-03-16.pdf`.

van ieder van de doelgroepen. Ook kan er rekening gehouden worden met het feit dat deze niet gelijk verdeeld hoeft te zijn, en dat sommigen gebruikers meer zouden kunnen met de toetsen dan anderen. Hier kan bij het ontwerp van de rapportages rekening mee gehouden worden, en zeker een digitale uitgave van de rapportages biedt hier mogelijkheden voor.

Het onderzoek bij (4) en zeker ook (5) moet per doelgroep worden uitgevoerd. Hoe grootschalig dat is, zal per doelgroep verschillen. De groep belanghebbenden bij (formeel) controlerende instanties, zoals de inspectie, zal aanzienlijk kleiner zijn dan de belanghebbenden die op scholen werken. Binnen die groep zal de groep die op basis van de rapportages beleid moet bepalen, zoals de schoolleiding, weer kleiner zijn dan de groep leerkrachten die op uitvoeringsniveau in de klas met de resultaten aan de slag moeten.

De groep van leerkrachten is waarschijnlijk de grootste groep van belanghebbenden die intensief met de rapportages moet werken. Het is verstandig bij de ontwikkeling te kijken hoe de rapportages hen kunnen helpen. De ontwikkeling kan het best gedaan worden samen met hen. Dan kan beter ingespeeld worden op hoe zij de rapportages willen en kunnen gaan gebruiken, en kan de "ergonomie" van het ontwerp zo ingericht worden dat deze meer intuïtief, direct begrepen wordt. Het is daarbij verstandig vooral veel aandacht te hebben voor de leerkrachten met een relatief lage mate van datageleterdheid. Als zij het uitgangspunt voor het ontwerp van de rapportage zijn, dan kunnen alle leerkrachten het gebruiken. Voor leerkrachten die dieper op de stof in willen gaan, kunnen eventueel verdiepende rapportages ontwikkeld worden. Verdiepende rapportages kunnen leerkrachten ook aanvullende handvatten geven voor juist handelen op basis van rapportages.

Er zijn verschillende opties om met verdiepende, moeilijker te interpreteren, rapportages om te gaan. De eerste is deze informatie voor niemand beschikbaar te maken. Het scheelt kosten, omdat de rapportages niet ontworpen hoeven te worden, hetgeen ook analyses scheelt. Het nadeel is dat leerkrachten voor wie het een welkome aanvulling is, van deze informatie verstoken blijven. De tweede optie is het pas beschikbaar te stellen als leerkrachten hebben aangetoond dat ze deze informatie kunnen interpreteren. Of dat via een cursus en een officieel examen gaat, of via zelfstudie en een online toets zal afhangen van hoeveel kosten er aan besteed kunnen worden, en hoeveel vertrouwen er is dat de leerkrachten hier goed mee omgaan. Het nadeel is dat hier enige additionele kosten aan verbonden zijn, maar het voordeel is dat gekend is dat de leerkrachten die deze informatie krijgen ook weten hoe ze hiernaar moeten handelen. De derde optie is om deze verdiepende rapportages in principe voor iedereen beschikbaar te stellen. Als een leerkracht er behoefte aan heeft, dan kan deze die gebruiken. Net zoals bij optie 2 zal er in ieder geval online informatie beschikbaar zijn. Het voordeel is dat er geen kosten gemaakt hoeven te worden om te bepalen of een leerkracht deze informatie kan gebruiken of niet. Het nadeel is dat de kans op verkeerde interpretatie van informatie vergroot wordt. Een alternatief is dat er een extra helpdesk nodig is, wat ook kosten met zich meebrengt.

Goed voorwerk scheelt veel geld in de nazorg. Het is beter 200 leerkrachten te betrekken in de ontwikkeling van de rapportage dan 10.000 achteraf te leren wat je ermee kan doen, waarbij niet eens zeker is of wat ze ermee kunnen perfect past bij hun behoeften. Zelfs als er nog een deel van de leerkrachten extra ondersteuning of aanvullende informatie nodig zal hebben, dan zal deze groep kleiner zijn, en minder tijd nodig hebben om een systeem te

begrijpen. Een ontwerp van de rapportage dat gericht is op een verhoogde bruikbaarheid, maakt ook dat de toetsen meer geaccepteerd zullen worden.

Dit pleidooi geldt uiteraard ook voor het ontwerp van schoolrapportages die door de schoolleiding gebruikt kunnen worden om de kwaliteit van het onderwijs te verbeteren. Het is verstandig ook dergelijke groepen mee te nemen bij de ontwikkeling van de voor hen bestemde rapportages. Tijdens het ontwerp van die rapportages wordt zo ook duidelijk welk ondersteund materiaal wellicht nuttig is. Als veel ondersteuning nodig is, zijn de rapportages nog niet geoptimaliseerd. Andere groepen die te maken krijgen met de rapportages zijn de leerlingen en de ouders. Voor de leerlingen moet de rapportage uiteraard niet te ingewikkeld zijn. Zij kunnen daarbij ook door de leerkrachten ondersteund worden. De ouders zijn een lastige groep om aan te rapporteren, met name omdat deze zeer heterogeen is wat betreft voorkennis. Ook hier is het verstandig bij het ontwerp van de rapportages vooral te richten op de groep die laag-datageletterd is. Zij moeten zowel de rapportages van hun kinderen begrijpen, maar zodra er ook schoolgegevens openbaar gemaakt worden, dan moet deze ook op de juiste wijze door hen begrepen worden. Wat betreft dit laatste punt, is er ook het algemeen publiek, dat via de journalistiek geïnformeerd wordt. Aangezien het de bedoeling is dat de rankings niet gerapporteerd worden, moeten de rapportages daar ook geen aanleiding toe geven, waarover later meer.

De belangrijkste boodschap van deze paragraaf is dat de rapportageontwikkeling minstens zo serieus te nemen is als de toetsontwikkeling, zo niet serieuzer. Betrek daarbij in grote mate de belanghebbenden. Het is verstandig in de kostenverdeling van toetsontwikkeling en rapportageontwikkeling daar rekening mee te houden.

Hoe juist handelen op basis van rapportages te bevorderen?

De groepen voor wie de rapportages bedoeld zijn, zijn ook de groepen die moeten handelen op basis van die rapportages. Uit voorgaande paragraaf is ook duidelijk dat de type handelingen en dus ook de type rapportages van elkaar kunnen verschillen. Ook is duidelijk dat de juiste vorm van rapportage ook de juiste vorm van handelen moet uitlokken: het is een rapportage die begrepen wordt, en samen met de doelgroep ontwikkeld is om een gedeelde visie van juist handelen te bevorderen.

Het doel van het juist handelen is duidelijk: de kwaliteit van het Vlaamse onderwijs vergroten. Echter, op deze plek voor iedere doelgroep aan te geven wat die juiste vorm van handelen is, is niet functioneel. Dat is juist iets wat samen met die doelgroep geoperationaliseerd moet worden. Deels kan wetenschappelijke literatuur hints geven in welke richting we het moeten zoeken, maar wat werkelijk werkt, zal toch context afhankelijk blijven. Om tot "evidence informed" onderwijs[37] te komen, zal de wetenschap samen met de praktijk, dus het steunpunt samen met de doelgroepen uit het Vlaamse onderwijsveld, aan tafel moeten gaan. Dat is ook onderdeel van de ontwikkeling van de juiste rapportagevorm (stap 2).

Er zullen altijd leerkrachten zijn die ondersteuning nodig hebben om de voor hen ontworpen rapportages goed te doorgronden. In Hoofdstuk 6 is al aangegeven dat workshops, instructiefilmpjes, en flyers kunnen helpen. Ook een helpdesk voor vragen is een optie. Welke optie gekozen wordt, zal deels ook afhangen van het budget. Het is duidelijk dat

[37] Zie bijvoorbeeld `https://www.ru.nl/docenten/onderwijsonderzoek/evidence-informed-onderwijsinnovatie/`.

workshops en een helpdesk duurder zijn dan online filmpjes en instructies, maar het is ook bekend dat deze interactieve instructie effectiever is. Het is echter lastig op voorhand vast te stellen of deze effectiviteit opweegt tegen de extra kosten. Cijfers over hoeveel ondersteuning een school precies gaat vragen, zijn moeilijk te geven. Deze kosten zijn vaak ook versnipperd, omdat deze hulp nu vaak bij diverse bronnen wordt gehaald, waardoor ook van de huidige praktijk geen eenduidige cijfers te geven zijn. Het is wel zo dat, als bij het ontwerpen van de rapportages in grote mate de belanghebbenden betrokken zijn geweest, er minder additionele maatregelen, en dus kosten, nodig zijn om het voor iedereen bruikbaar te maken.

De benodigde ondersteuning zal voor een deel ook afhangen van de mate van datageletterdheid. Deze zal per type leerkracht verschillen: een leerkracht die lesgeeft aan groep 4 in het lager onderwijs zal minder ervaren zijn met het lezen van grafieken en tabellen dan een wiskundeleraar die lesgeeft in de derde graad aso. Daar kan in het ontwerp ook rekening mee gehouden worden. De datageletterdheid zal echter ook binnen deze groepen variëren. Een van de ontwerpcriteria was dat de rapportage ook duidelijk moest zijn voor de minder datageletterde leerkracht. Het zou echter zonde zijn dat leerkrachten die "moeilijkere" rapportages aankunnen, deze informatie niet zouden krijgen.

Zoals aangegeven kan onderscheid gemaakt worden tussen basisrapportages en verdiepende rapportages. Deze kunnen tijdens hetzelfde proces ontwikkeld worden. Dit scenario wordt hier genoemd omdat voor dit scenario ook bepaald moet worden of en hoe dergelijke verdiepende rapportages beschikbaar gesteld worden.

Hoe rankings van scholen te voorkomen?

Het oneigenlijke rangordenen van scholen, kan op verschillende manieren worden voorkomen. In de eerdere hoofdstukken zijn al maatregelen genoemd die hierbij kunnen helpen. Wat zeker is, is dat een interpretatie in termen van rangordeningen zal ontstaan als er (digitaal) een lijst gepubliceerd wordt waarbij van iedere school een enkele gemiddelde score gegeven wordt. Er zijn verschillende opties om dit te voorkomen. Het is verstandig om een combinatie van deze opties toe te passen.

- **Betrouwbaarheidsintervallen rapporteren**
  Alle metingen hebben enige onzekerheid. Deze kunnen weergegeven worden door betrouwbaarheidsintervallen rondom de observaties. Die geven een bereik van hoe zeker we zijn van een bepaalde observatie. Hoe meer observaties we hebben, hoe kleiner de meetfout. Dat betekent dat voor grote scholen de meetfout kleiner is dan bij kleine scholen. De breedte van het betrouwbaarheidsinterval hangt ook af van hoeveel onzekerheid toegestaan wordt. Een 95%-betrouwbaarheidsinterval is ongeveer twee keer zo groot als een 70%-betrouwbaarheidsinterval, en zal dus meer overlap tussen scholen opleveren. Als het doel is om rankings tegen te gaan, helpt een groter betrouwbaarheidsinterval. Door ook geen gemiddelden weer te geven, is het duidelijker dat er ook echt overlap is.
- **Score-ranges rapporteren**
  Naast de meetonzekerheid is ook de spreiding van de scores in de school relevant. Als de best en de slechtst presterende leerling als grenspunten aangegeven worden, levert dit bij scholen een zeer grote range op aan scores, waarbij scholen een grote overlap zullen hebben. Het kan tot gevolg hebben dat de scholen de zwakker presterende

leerlingen mogelijk uit zullen sluiten. Het kan daardoor verstandiger zijn een andere scorerange te gebruiken, die weliswaar kleiner is (en dus minder overlap oplevert), maar waarbij uitsluiting van een leerling minder impact heeft. Dat kan de scorerange zijn van het 20e tot en met het 80e percentiel binnen de school, of de iets kleinere range van het 25e tot en met het 75e percentiel, maar ook andere ranges zijn mogelijk. Deze waarden zijn ook stabieler dan de minimum en de maximum score. Ook hier geldt dat, als er ook geen gemiddelden worden weergegeven, het duidelijker is dat er ook echt overlap is.

- **Diverse correcties rapporteren**
  Er kan gekozen worden om zowel de resultaten te presenteren waar wel gecorrigeerd wordt voor diverse achtergrondvariabelen als de resultaten te presenteren waar niet gecorrigeerd wordt voor deze achtergrondvariabelen. Bij gecorrigeerde waarden zijn ook diverse mogelijkheden daar waar het de keuze van de achtergrondvariabelen betreft, maar ook de keuze van de correctiemethode. Als een aantal van de mogelijke representaties van de school weergegeven wordt, dan is het evident dat er geen ordening bestaat, maar dat deze deels afhangt van keuzen en aannames.
- **Alleen gegevens over meerdere jaren rapporteren**
  De onzekerheid van de meting en de impact van toevalligheid zijn een onderdeel van de aanpak bij de eerste optie. Deze onzekerheid geldt ook over de verschillende jaren. Als de school een keer een minder vaardige klas heeft, dan wordt dat niet altijd door de eerste drie aanpakken opgevangen. Het kan dus verstandig zijn om pas te rapporteren als er gegevens verzameld zijn over meerdere jaren. Met de toepassing van IRT is het goed mogelijk om resultaten van verschillende jaren op één schaal te brengen. Er ligt niet vast hoe lang er gewacht moet worden voordat er gepubliceerd kan worden. Eerder dan na drie jaar is niet aan te raden, en het is in dergelijke situaties niet ongebruikelijk een periode van vijf jaar te gebruiken. Het nadeel is dan wel dat ouders pas na verloop van tijd de resultaten van centrale toetsen kunnen gebruiken als ondersteuning bij de schoolkeuze. Het is echter wel zo dat de gerapporteerde gegevens dan aanzienlijk beter gefundeerd zijn. Daarom is deze aanpak ook aan te raden.
- **Gegevens beperkt beschikbaar stellen**
  In plaats van een overzichtelijke lijst van alle scholen tegelijkertijd aan bieden, kan de rangordening bemoeilijkt worden door ouders slechts uit een beperkte selectie per keer te laten kiezen. In het meest extreme geval kan ervoor gekozen worden om per keer alleen een paarsgewijze vergelijking toe te staan die afgebeeld kan worden. Met name bij een digitale vrijgave van de informatie is dit goed mogelijk. Het beperken van de selectie bemoeilijkt het samenstellen van een algemene rangordening die publiek gemaakt wordt. Merk op dat dit alleen gaat over het bemoeilijken, en niet het voorkomen, aangezien het technisch altijd mogelijk is als deze gegevens openbaar zijn.

7.3 Kosten

De kosten voor het introduceren van gecentraliseerde toetsen en examens kunnen onderverdeeld worden in kosten die min of meer onafhankelijk zijn van het aantal afnamen, de zogenaamde vaste kosten, kosten die schaalbaar zijn, dat wil zeggen afhankelijk van het aantal deelnemende leerlingen en scholen, en variabele kosten, afhankelijk van de keuzes voor een bepaald toetsontwerp. Bij de vaste kosten zijn nu zaken opgenoemd die in ieder geval onderdeel uit moeten maken van centrale proeven waaraan kosten verbonden zijn. Binnen deze categorieën is nog steeds variabiliteit mogelijk, waardoor het moeilijk is exacte kosten op te geven. Dat is helemaal het geval bij de variabele kosten.

7.3.1 Vaste kosten

Vrijwel onafhankelijk van het aantal leerlingen en scholen dat een toets zal afnemen en de keuzes die gemaakt worden voor een bepaald toetsontwerp, zijn er een aantal kostenposten die bij elke toetsafname meegenomen moet worden.

Item constructie

Het construeren van items is misschien wel de belangrijkste kostenpost bij de introductie van gecentraliseerde toetsen. Het aantal items dat geconstrueerd moet worden voor een toetsafname hangt af van de hoeveelheid items die hergebruikt kunnen worden en de gekozen betrouwbaarheid van de toets (zie Hoofdstuk 5).

Afname-systeem

De items die ontwikkeld zijn, zullen aan de leerlingen voorgelegd worden. Bij een papieren afname zullen kosten ontstaan voor het drukwerk, opslag en verspreiding van de toetsen. Bij een digitale afname zullen kosten ontstaan voor het kunnen afspelen van de items. In beide gevallen zijn er kosten die ontstaan rondom de logistiek omtrent de afname.

Data-opslag-systeem

De responses die de leerlingen hebben gegeven op de items in de test, moeten worden opgeslagen, zodat deze later gescoord kunnen worden. De kosten voor een opslagsysteem zijn uiteraard afhankelijk van de afnamemodus.

Scoren van items

De responses zullen vervolgens gescoord moeten worden. Dit kan geautomatiseerd plaatsvinden of door menselijke beoordelaars. De beoordeling kan zowel direct op school plaatsvinden, als op een andere locatie, door geheel onafhankelijke externe beoordelaars. Het automatisch scoren van items is veelal goedkoper dan het werken met beoordelaars die de items scoren. Of dit mogelijk is, hangt mede af van het type opgaven en de afnamemodus, hetgeen wel enige variabiliteit mogelijk maakt (zie Sectie 7.2.2). Of het wenselijk is om alles automatisch te laten beoordelen, hangt af van de wijze waarop draagvlak voor de proeven gecreëerd gaat worden (zie Randvoorwaarden-sectie hieronder).

Analyse

Analyses kunnen zowel plaatsvinden op de responses als op de gescoorde data. Voor een rapportage aan de gebruikers van de toets zal er in ieder geval een analyse moeten plaatsvinden op de gescoorde data. Op het moment dat alleen zeer eenvoudige rapportages

worden opgeleverd - zoals het aantal goed gemaakte items op een toets - kan dit min of meer geautomatiseerd worden. Op het moment dat complexere rapportages opgeleverd worden, zoals over leerwinst of vaardigheidsschattingen, zullen er ook kosten ontstaan voor analyses uitgevoerd door analisten. Merk op dat dit onafhankelijk is van het aantal leerlingen dat de toets afneemt.

Rapportagesysteem

De resultaten uit de analyse moeten vertaald worden naar een rapportage en deze rapportages moeten bij de eindgebruikers terecht komen. Rapportages kunnen in geprinte vorm bij de gebruikers terecht komen. Veel rijkere rapportages zijn mogelijk op het moment dat deze digitaal via een rapportagesysteem aan de gebruikers gerapporteerd worden.

Communicatiekanaal

Een goed communicatiekanaal verhoogt de kansen op een goed functionerend toetssysteem aanzienlijk. Een zorgvuldige toetsafname en een correcte interpretatie van de rapportages vraagt om een communicatiekanaal met leerkrachten en scholen voor instructie, monitoring en feedback. Een basale helpdesk kan aangevuld worden met bijvoorbeeld regio-bijeenkomsten. Hierbij is het praktisch en daardoor ook goedkoper om van bestaande structuren gebruik te maken, zoals overkoepelende schooloverlegorganen. Ook kan gebruik gemaakt worden van speciale teams die naar scholen zelf gaan om in te gaan op de behaalde toetsresultaten en eventuele bijbehorende didactische handelingsopties. Tenslotte kan daarnaast ook een digitaal communicatiekanaal opgezet worden met daarin een collectie best practices en de mogelijkheid voor collegiaal overleg.

Systeemtest

Bij digitale (adaptieve) toetsen is het aan te raden ook een ‘systeemtest’ op te nemen. Dit met als doel om zowel de werking van het afnamesysteem te testen als gebruikers (leerlingen/ leerkrachten/ IT-ondersteuners) bekend te laten raken met het afnameplatform, de manier van toetsen, het type items (als deze onbekend zijn voor een groep gebruikers) en de mogelijke (digitale) ondersteuningen en hulpmiddelen die aangeboden worden in de toets (e.g., verklanking van teksten, gebruik van een digitale rekenmachine, digitale liniaal e.d.). Een systeemtest heeft als voordeel dat bij de werkelijke afname zich minder onverwachte problemen voordoen op het gebied van IT en inrichting van de afnamesetting en dat de gebruikers beter weten wat van hen verwacht wordt. Er zijn wel kosten gemoeid met de invoer van een systeemtest.

Aantal leerlingen

Gegeven dat alle leerlingen in Vlaanderen getoetst worden is het aantal leerlingen een vast gegeven. Echter, additionele kosten kunnen ontstaan als een bestaande toets, binnen een bestaand toetssysteem, door extra leerlingen of scholen wordt afgenomen. Deze kosten kunnen omschreven worden als kosten per eenheid (leerling of school). Een belangrijke schaalbare kostenpost is bijvoorbeeld het aantal rapportages dat uitgeleverd moet worden. Ook de keuze voor een papieren of digitale afname heeft veel impact op schaalbaarheid, aangezien ingeschat moet worden hoeveel leerlingen een papieren of digitale afname gaan doen. Bij een aantal variabele kosten die hieronder beschreven worden (bijvoorbeeld niet automatisch gescoorde opgaven), kan het ook uitmaken hoeveel leerlingen het betreft.

7.3.2 Variabele kosten

Verschillende kostenposten die afhankelijk zijn van keuzes in het toetsontwerp staan hieronder beschreven.

Mate van betrokkenheid van het veld bij het tot stand komen van de rapportage

De ontwikkeling van een rapportage kan vergelijkbaar zijn met de ontwikkeling van toetsen (zie Paragraaf 7.2.4). Hoewel het niet gebruikelijk is daar even veel tijd en geld in te steken, zou dat veel meerwaarde hebben om de budgetten voor beiden meer in evenwicht te brengen. De kosten zijn variabel in de zin dat de mate van betrokkenheid kan variëren. Het is echter wel aan te raden veel belanghebbenden die in de praktijk met de rapportages moeten gaan werken vanaf het begin bij de ontwikkeling ervan te betrekken. Dit heeft namelijk zeer grote voordelen. De rapportages sluiten dan beter aan bij de expertise en datageletterdheid van de meeste leerkrachten, wat later kosten scheelt omdat de expertise niet aangepast hoeft te worden aan de rapportage. Een ander voordeel is dat de rapportage dan ook beter aansluit bij de wensen van de belanghebbenden, zodat zij de rapportages ook werkelijk in de praktijk kunnen gebruiken. Deze sluiten dan ook aan op de beoogde vervolghandelingen naar aanleiding van de rapportage. Al met al verhoogt dit het gebruik en de acceptatie van de toetsen, en daarmee ook de kans op een succesvolle introductie van de centrale toetsen.

Mate van fraudepreventie

Er zijn ook kosten verbonden aan veiligheid, ofwel aan de mate van zekerheid die men heeft dat de responses op een toets daadwerkelijk van de relevante leerling komen, zonder dat gebruik is gemaakt van oneigenlijke hulpmiddelen. Hoe meer zekerheid men wil hebben rondom de veiligheid van een toetsafname, hoe meer kosten ontstaan. Kosten waaraan gedacht kan worden, zijn kosten voor (online) toezicht op leerlingen, opslag van toetsmaterialen en de veiligheidseisen die men stelt aan het itembanksysteem en afnamesysteem. Denk bij het laatste bijvoorbeeld aan het afsluiten van toegang tot het internet bij een toetsafname, een zogenaamde lock-down browser.

Selectie van centraal getoetste vaardigheden

Bij sommige vaardigheden, zeker op de hogere vaardigheidsniveaus, zullen taken met complexe responspatronen gebruikt moeten worden, zoals essays voor schrijfvaardigheid of uitgebreide spreektaken voor spreekvaardigheid. Er wordt de laatste jaren meer en meer geëxperimenteerd met het automatisch scoren van responsen op dergelijke taken, maar deze zijn vaak nog niet zodanig ontwikkeld dat deze ook gemakkelijk ingezet kunnen worden. In dat geval moeten deze opdrachten door beoordelaars gescoord worden. Wanneer alleen de eigen leerkracht de taken scoort en de resultaten zelf invoert, dan zijn de kosten nog beperkt. Het nakijken door leerkrachten kan als regulier onderdeel van hun werk gezien worden, omdat ze anders een andere essay taak zouden beoordelen. De leerkrachten zijn op z'n hoogst extra tijd kwijt voor familiarisatie met het beoordelingssysteem (inhoudelijk en mogelijk ook technisch). Als de responsen van de leerlingen willekeurig door leerkrachten beoordeeld worden, liggen de kosten ook niet in het beoordelen (daar geldt hetzelfde voor als bij de eigen leerkracht), maar dan moet er wel een systeem ontwikkeld worden dat de taken verdeelt over de beoordelaars, en de beoordelingen per taak goed administreert. Als alle werken door twee beoordelaars beoordeeld moeten worden, en zeker als er externe

beoordelaars gebruikt worden, zullen er ook beoordelaarskosten zijn. Deze zullen hoog dan wel zeer hoog zijn (zie ook Hoofdstuk 5).

Type opgaven

Een belangrijk verschil in kosten ontstaat door de keuze in het type opgaven. Meerkeuze-opgaven, en korte-antwoord opgaven zijn vaak relatief goedkoop in de productie en in het verwerken van de antwoorden. Dat laatste is ook het geval bij andere vormen van automatisch scoorbare items. Als deze een wat ingewikkeldere vorm hebben, dan kunnen deze wel in de productie duurder zijn. Hotspot items, en drag-and-drop items vallen nog wel mee, in het inhoudelijk ontwerpen en digitaal uitvoeren, maar opgaven waarbij een aantal handelingen gevolgd en gescoord moeten worden, kunnen wel duurder in productie zijn. Ook kan de productie van opgaven duurder zijn als er gebruik gemaakt wordt van beeldmateriaal. Dat moet dan gefilmd en bewerkt worden, of er zijn licentiekosten aan verbonden. Ook kan beeldmateriaal bij de afname zwaardere eisen stellen aan de computer voor een goede afname. Het kan ook meer ontwikkeltijd kosten om de opgaven gestandaardiseerd te krijgen op alle afnameplatforms[38]. Wanneer er gebruik gemaakt wordt van niet-automatisch gescoorde opgaven, kan dat in de ontwikkeling kostbaar zijn doordat er ook een goed beoordelingsmodel gemaakt moet worden voor dergelijke opgaven, en dat kan tijdrovend zijn. Daarnaast is er dan ook onderzoek nodig naar beoordelaarsovereenstemming. De grootste additionele kosten zullen mogelijk liggen in de beoordelingen van de leerlingwerken, al dan niet in combinatie met de invoer van de beoordelingen. Hierover stond reeds meer bij het variabele kostenpunt *Selectie van centraal getoetste vaardigheden*.

Proeftoetsing

Proeftoetsen kunnen worden opgezet om de eigenschappen van nieuw ontwikkelde items te leren kennen. Op basis van proeftoetsen kan bijvoorbeeld worden onderzocht wat de kwaliteit van nieuwe items is, en of de items niet te moeilijk of te makkelijk voor de beoogde doelgroep zijn. Juist bij adaptieve toetsen spelen proeftoetsen een belangrijke rol, omdat bij adaptieve toetsen informatie over de moeilijkheid van items nodig is voordat de adaptieve toets afgenomen kan worden. Afhankelijk van het doel van een proeftoets hebben deze vaak een omvang van 100 tot 400 leerlingen, afhankelijk van het gehanteerde psychometrisch model[39]. De organisatie van proeftoetsing is dus een ander element dat meegenomen moet worden bij de totale kosten voor een bepaald toetsontwerp. Een alternatief voor proeftoetsing is gebruik te maken van “zaaien”. Bij zaaien of embedded field testing worden nieuw geconstrueerde items in de reguliere toetsafname opgenomen en op deze manier wordt informatie over de nieuwe items verzameld. De nieuwe items worden dan niet meegenomen in de scoring van de leerlingen. Een kostenvoordeel van zaaien is dat er geen aparte proeftoetsing hoeft te worden georganiseerd. Desondanks vraagt ook het gebruik van zaaien een (technische) investering.

[38] Dat hangt ook af van de wijze waarop de ICT ingericht is: welke standaarden er gebruikt worden voor de afnamemodule, of de afname lokaal gebeurt of via een netwerkverbinding enzovoorts. Bij standaard meerkeuze opgaven zonder te veel visuele hulpmiddelen zijn er over het algemeen minder afhankelijkheden.

[39] Zie de COTAN-richtlijnen (Hoofdstuk 5; `https://www.cotandocumentatie.nl/cotan/beoordelingssysteem/`).

Wijze van koppeling van gegevens over leerjaren

Als met behulp van een IRT model een vaardigheidsschaal gemaakt wordt om de vier afnamemomenten met elkaar te verbinden dan moet er een toetsdesign toegepast worden waarbij er sprake van overlap tussen de toetsen is. De gemakkelijkere opgaven van het hogere leerjaar komen dan samen met de moeilijkere opgaven van een lager leerjaar in de toets. Wanneer de afstand tussen twee afnamemomenten groot is, zullen de moeilijke opgaven uit het lagere leerjaar te makkelijk zijn voor het hogere leerjaar, en de makkelijke opgave uit het hogere leerjaar te moeilijk voor het lagere leerjaar. Het zal dan nodig zijn om een tussenliggend leerjaar te introduceren waarbij beide typen opgaven wel functioneel zijn. Tussen de eerste drie afnamemomenten betreft dat leerjaar 5 in het lager onderwijs en het eerste jaar van het secundair onderwijs. Bij die jaren zou dan een proeftoets afgenomen moeten worden. Hiervoor hoeven geen nieuwe opgaven ontwikkeld te worden, en het reguliere afnamesysteem kan hiervoor gebruikt worden. Met duizend afnamen over iets meer dan 40 scholen zijn de kosten beperkt[40].

Itembank-systeem

De kosten van een itembanksysteem zijn vooral te vinden in de kosten die ontstaan door het "onderhoud" van het systeem. Een belangrijke uitdaging is om consistentie in de itembank te behouden (zie Hoofdstuk 5). Om consistentie te bereiken, wordt vaak gebruik gemaakt van relationele databasemanagement systemen. Hier zijn zowel commerciële varianten, als ook vrij beschikbare software voor beschikbaar.

Accommoderen voor alle leerlingen

Er kunnen speciale toetsversies ontwikkeld worden voor leerlingen die de reguliere toetsversie niet kunnen afnemen door bijvoorbeeld slechtziendheid of een andere beperking. De kosten kunnen liggen in toegevoegde hulpmiddelen, maar mogelijk ook in het ontwikkelen van een geheel nieuwe toetsversie voor een specifieke groep. Hoe meer verschillende groepen er onderscheiden worden die een eigen versie nodig hebben, hoe duurder. Als die speciale groepen klein zijn (minder dan 400 leerlingen per leerjaar) kan onderzoek naar equivalentie van de toetsen ook niet goed worden onderzocht. Als de groepen leerlingen voor wie een eigen versie nodig is zeer klein is, kan het kosten-efficiënt zijn te kiezen voor een individuele, wellicht minder gestandaardiseerde, afname onder begeleiding. Ook in de rapportage van Perceel 2 wordt nader ingegaan op de specifieke kosten van aparte toetsversies voor specifieke doelgroepen.

Afnamemodus

De keuze voor een papieren naast een digitaal afnamesysteem heeft invloed op de totale kosten voor een toetsontwerp. De vaste kosten van het implementeren van een digitaal systeem zijn vaak hoger dan van een papieren systeem. Met name als er nieuwe systemen ontwikkeld moeten worden. Daarentegen zijn bij een digitale afname de schaalbare kosten

[40] Bij deze werkwijze is de afstand tussen het laatste jaar van graad 1 en het laatste jaar van graad drie met vier schooljaren te groot om deze werkwijze voor die groepen toe te passen. Het volgen van leerlingen tot en met het laatste jaar lijkt voor wat betreft de formatieve functie ook beperkt, ook omdat dit het eindpunt is van het formele verplichte onderwijs. Een (afsluitende) summatieve functie voor de leerlingen lijkt gezien het moment van de toetsing meer voor de hand te liggen. Een dergelijke functie geaccepteerd krijgen in Vlaanderen lijkt op dit moment lastig haalbaar, waardoor de toets in het laatste jaar vooral voor de school een functie heeft.

per leerling of school vaak lager dan bij papieren systemen. Bij een papieren afname betreft het per leerling of school extra drukwerk en transport. Ook de logistiek van extra varianten vergt binnen een papieren systeem meer coördinatie dan bij een digitaal systeem. Bij een digitaal afnamesysteem speelt ook de mate van hoe toegankelijk het systeem moet zijn een rol in het kostenplaatje. Als alle scholen, met verschillende IT faciliteiten en verschillende devices (zowel vaste computer, laptop als Chromebooks; zowel Windows, Apple als Linux e.d.), gebruik moeten kunnen maken van het platform, dan moet het platform flexibeler zijn dan wanneer er bijvoorbeeld eisen gesteld worden aan de digitale omgeving van de school. Dergelijke keuzes hebben invloed op de kosten van het digitale afnamesysteem. Dit is ook een onderwerp waarnaar gekeken is bij Perceel 3.

Mate van adaptiviteit

Kosten worden hoger naarmate er meer toetsversies zijn, of de proeven een hogere mate van adaptiviteit kennen, omdat er dan meer opgaven benodigd zijn. Voor adaptieve toetsing moet bovendien zoals hierboven al benoemd van alle items de moeilijkheid bekend zijn, dus de items moeten al eens eerder zijn afgenomen.

Ontwikkeling en scoren van centrale proeven voor moeilijk meetbare vaardigheden

Als ook moeilijk meetbare vaardigheden worden meegenomen, zijn er kosten verbonden aan de ontwikkeling van deze toetsen. Het proces van ontwikkelen van dergelijke taken is anders dan het ontwikkelen van de losse items binnen een systeem met automatisch scoorbare items. Er zal bij dergelijke toetsen veelal ook beoordelaarsonderzoek noodzakelijk zijn. Bij de inzet van machine-learning procedures bijvoorbeeld, zullen er beoordelingen door experts gedaan moeten worden om het algoritme afdoende te trainen. Ook zullen 'gewone', niet-automatisch scoorbare opgaven relatief vaak gebruikt worden bij moeilijk meetbare vaardigheden, wat voor aanvullende kosten kan zorgen, zoals hierboven reeds beschreven onder 'Selectie van centraal getoetste vaardigheden'.

Mogelijkheid voor toevoeging van externe gegeven

Als additionele gegevens toegevoegd moeten worden, zoals bijvoorbeeld leerkrachtoordelen van de moeilijk meetbare vaardigheden, dan moet er een systeem zijn om dat ook in te kunnen voeren. Daar waar het de leerkrachtoordelen zijn hoeft dat niet zeer ingewikkeld te zijn. Als externe gegeven uit andere bronnen ingelezen moeten worden, scheelt dat invoerwerk, maar voor het maken van een dergelijk data-transfer-tool moeten ook kosten begroot worden.

Type normering

De verschillende normeringen die gebruikt kunnen worden in een rapportage verschillen in de kosten. Het gebruik van relatieve normeringen, waarbij leerlingen op basis van hun prestaties geordend worden en ingedeeld worden in een bepaalde groep (bijvoorbeeld een percentiel- of kwantielgroep), brengt nauwelijks andere dan analysekosten met zich mee. Dit is anders bij het gebruik van absolute normeringen. Absolute normen worden vaak bepaald door een standaardbepaling (zie Hoofdstuk 6). Voor een standaardbepaling worden vaak vertegenwoordigers van verschillende belanghebbenden uit het onderwijs (digitaal) uitgenodigd om in een sessie een bepaald prestatieniveau op de eindtermen als norm vast te stellen. De omvang van deze groepen experts variëren - afhankelijk van het

belang van de norm - vaak van 10 tot 20 personen.

Mate van gedetailleerde rapportage

Wanneer op hoofdvaardigheden gerapporteerd wordt, zoals leesvaardigheid of rekenvaardigheid zijn de kosten lager dan als er binnen die vaardigheden ook gedetailleerde rapportages plaatsvinden op schoolniveau en mogelijk zelfs op leerlingniveau. Additionele kosten moeten gemaakt worden omdat er meer verschillende opgaven afgenomen moeten worden om ook op detailniveau te rapporteren. Dat heeft te maken met de operationalisatie van de subvaardigheid: voor de meeste vaardigheden moeten daar minstens 5 tot 20 verschillende opgaven voor beschikbaar zijn[41].

Vorm van terugrapportage

Aan de verschillende vormen van terugrapportage hangen verschillende kosten. Relatieve normeringen zijn goedkoop te verkrijgen. De gegevens zijn verzameld, en alleen bij de keuze van het formuleren van welke subgroepen er gebruikt worden (voor welke variabelen wordt gecorrigeerd) kan wat extra werk zitten. De wijze van terugrapportage is ook bij veel gebruikers bekend. Inhoudelijk is er echter wel veel op aan te merken (zie Hoofdstuk 6). Een inhoudelijke terugrapportage door middel van omschrijving van een taak in termen van beheersing van items, is ook relatief goedkoop, en kan met behulp van IRT ook gemakkelijk verkregen worden. Een duurdere vorm van terugrapportage is via de absolute normen. Die zal echter wel noodzakelijk zijn als er uitspraken gedaan moeten worden over of bepaalde (clusters van) eindtermen gehaald worden. Meer over de kosten in Hoofdstuk 6. Bij het ontwikkelen van de vorm van terugrapportage is het verstandig om ook klankbordgroepen samen te stellen van (potentiële) gebruikers van de toetsen. Daar zijn ook kosten aan verbonden.

Wijze van kennisbevordering rapportages

Bij de ontwikkeling van de rapportages moet rekening gehouden worden met de huidige kennis van rapportages die leerkrachten hebben. Het teruggrijpen naar rapportages die nu al gebruikt worden en de ontwikkeling met behulp van een klankbordgroep zullen daarbij helpen. Desalniettemin zal het nodig zijn om ook meer kennis te delen over rapportages en het gebruik ervan. Hierbij speelt een aantal variabelen een rol in de mogelijke kosten, zoals de wijze van communicatie (eenrichting, dan wel interactief), en de complexiteit van de informatie (enkele cijfers, dan wel formatieve adviezen op meerdere niveaus). Het lijkt makkelijk hierop te besparen, maar uiteindelijk is deze stap wellicht de belangrijkste als het gaat om het uiteindelijk beoogde doel van kwaliteitsbevordering te halen.

7.4 Planning en randvoorwaarden

7.4.1 Planning van toetsontwikkeling

Zoals hierboven al beschreven, zou je in de planning achteraan moeten beginnen. Dus eerst wordt beslist welke vervolghandelingen mogelijk zouden moeten zijn. Dit resulteert

[41] Het aantal benodigde opgaven heeft ook te maken met de homogeniteit van de schaal. Als de opgaven zeer inwisselbaar zijn, en allen erg op elkaar lijken, zijn weinig opgaven nodig (vaak bij interpunctie het geval). Wanneer er een grotere mate van diversiteit mogelijk is (bijvoorbeeld bij woordenschat) zijn er meer opgaven nodig.

in beslissingen omtrent de opzet van de rapportages (zie Hoofdstuk 6). Daarna volgt de keuze voor een model van leerwinst (zie Hoofdstuk 4). Als die keuzes eenmaal vast liggen kunnen de keuzes omtrent de toetsopzet en -ontwikkeling gemaakt worden (Hoofdstuk 5). Pas nadat de gemaakte keuzes helder zijn, kan gericht begonnen worden met het implementeren van het afname-, scoring-, dataopslag-, data-analyse- en rapportagesysteem. De opdracht tot constructie van items en toetsvarianten kan veelal pas als duidelijk is wat de mogelijkheden binnen het afname- en scoringsysteem zijn. Het opzetten van een communicatiekanaal met het veld kan al zodra duidelijk is welke keuzes gemaakt zijn, en kan dus gelijktijdig in gang gezet worden met de implementatie van het systeem en de constructie van items.

Bij het maken van een planning is het ook goed om alvast enkele jaren vooruit te denken en daarbij aan vragen te denken zoals: welke toetsen worden op welke momenten afgenomen? Hoe lang gaan de geïmplementeerde systemen mee? Hoe lang zijn geconstrueerde items nog relevant? Welke onderwijshervormingen zijn te voorzien? Op dit moment zijn er nog zoveel keuzen en beslissingen die genomen kunnen worden dat een uitgebreide opzet van de stappen die te nemen zijn moeilijk te maken is. De doorlooptijd is daarmee ook moeilijk te bepalen. Over het algemeen geldt echter dat de keuzen die het duurst zijn ook het meeste tijd kosten. Zeker als het gaat om de variabele kosten bij de ontwikkeling van de toetsen. Dat is deels logisch omdat de belangrijkste kosten daarbij vaak personeelskosten (manuren) zijn.

Bij het uitbrengen van dit rapport is er relatief weinig tijd te gaan voordat de eerste centrale proeven moet worden afgenomen. Dat betekent dat er snel gewerkt moet worden, waardoor in veel gevallen uit praktische overweging de keuze op de simpelste oplossing zal vallen die de minste problemen op zal leveren. Zelfs het uitrollen van die meest simpele afnamevorm zal nog een uitdaging zijn. Dat houdt in dat het wellicht verstandig is om te starten met een low-stakes afname van de toetsen. Dit scheelt veel tijd en geld die anders gestoken moet worden in fraudepreventie, en er hoeven minder opgaven gemaakt worden. In eerste instantie kan de keuze het best worden genomen voor vaardigheden die gemeten kunnen worden met automatisch scoorbare opgaven, en bij voorkeur voor opgaven die gemakkelijk te ontwikkelen zijn.

De tijd en moeite zal vooral ook gestoken moeten worden in het opzetten van de eerste systemen om items in op te slaan, in het afspelen van de opgaven, het verzamelen van de data (leerling-antwoorden), het opzetten van een geautomatiseerd scoringssysteem, en een platform om rapportages op een veilige manier te delen met leerlingen en scholen. Er moet voorgesorteerd worden op het kunnen volgen van leerlingen om daarna leerwinst en toegevoegde waarde te kunnen rapporteren. Nadat het basissysteem werkt, kan er uitgebreid worden naar meer verschillende vormen van rapportages, meer verschillende vormen van opgaven, een grotere opgavenbank, ontwikkelingen die meer high-stakes afname toestaan, enzovoorts. Bij het opzetten van het basissysteem moet wel van te voren rekening gehouden worden met uitbreidingen, maar om die allen vanaf de eerste afname werkend te hebben, zal te ambitieus zijn. Welke uitbreiding het eerst zal zijn, zal de tijd leren. Hierbij is het toepassen van een PCDA-cyclus voor het stap voor stap verbeteren van de centrale proeven van groot belang.

7.4.2 Randvoorwaarden voor succesvolle implementatie

Naast al het voorgaande zijn er tot slot enige randvoorwaarden te schetsen die de implementatie van de proeven bevorderen in ieder mogelijk scenario. Deze randvoorwaarden gelden eigenlijk altijd bij de opzet van nieuwe toetsen.

Randvoorwaarde 1: Specificeer het doel van de toetsing

In detail omschrijven wat bereikt moet worden met de toetsing, helpt bij het kiezen van een geschikt scenario. In welke mate wil je formatief of summatief zijn? In welke mate leg je nadruk op leerwinst, of juist op de bereikte resultaten op dit moment? Op welke niveaus wordt er gerapporteerd? En welke gewenste vervolgacties zouden in gang gezet moeten worden aan de hand van de rapportages?

Randvoorwaarde 2: Draagvlak creëren

Zorg dat leerkrachten en schoolbesturen zich eigenaar voelen van een probleem en de rapportages als deel van de oplossing zien. Dit valt te bevorderen door uitgebreide handreikingen te schrijven bij de rapportages, en regionale informatiebijeenkomsten over de proeven en het gebruik van de rapportages te organiseren. Geef het veld de tijd om te wennen aan de nieuwe proeven, hanteer dus een geleidelijke invoering, of koppel de proeven aan bestaande proeven. Een andere manier om eigenaarschap en betrokkenheid van leerkrachten te bevorderen is om leerkrachten invloed te geven op het uiteindelijke resultaat. Bijvoorbeeld met een rol in het scoringsproces (correctie van open vragen), of door leerkrachtoordelen of –verwachtingen mee te laten wegen naast het resultaat van een toets. Een actieve rol van leerkrachten bij kennisuitwisseling tussen scholen over succesvolle praktijken, kan hier ook aan bijdragen.

Randvoorwaarde 3: Voldoende itemproductie

Met te weinig items, kan een toetsing niet goed ingericht worden. Voldoende itemproductie is helaas relatief duur, maar wel noodzakelijk. Er moeten voldoende items zijn op alle domeinen en subdomeinen. En voor alle niveaus van leerlingen, ook binnen de meetmomenten. Al snel is binnen diverse scenario's behoefte aan meerdere varianten per meetmoment: die moeten wel goed gevuld kunnen worden met items. Items gaan niet eeuwig mee na constructie. In de planning van de itemproductie moet rekening gehouden worden met itemmortaliteit. Na een eerste afname moeten items van onvoldoende kwaliteit uit de proeven verwijderd worden, en vervangen worden door andere items. Die moeten er wel zijn. Ook goed-functionerende items kunnen na verloop van tijd verouderen, bijvoorbeeld omdat contexten niet meer relevant zijn in het huidige leven van leerlingen, licentierechten verlopen zijn, de opgave openbaar is geraakt, of omdat opgaven niet meer afspelen na aanpassingen van de afnamesoftware. Plan dus tijdig actualiteitschecks in, en voldoende productie om verouderde items te vervangen.

Randvoorwaarde 4: Voldoende ondersteuning en onderhoud

Ook op andere vlakken dan itemproductie moet het toetsproces ondersteund en onderhouden worden. Denk dan niet alleen aan de technische of ICT-ondersteuning bij een toetsafname. Maar ook aan de organisatie van de proeven, van datastromen en van rapportagestromen: planning, distributie, communicatie met scholen of andere partijen. Op het gebied van de toetsontwikkeling is er administratieve ondersteuning nodig voor het

bijhouden van itemkenmerken. Een goede itemkenmerken-administratie zorgt bij een groot project als deze proeven voor een efficiënte inzet van geproduceerde items, en gerichte productie-opdrachten. Ook is psychometrische ondersteuning bij het ontwerp van toetsen, en bij de analyse en interpretatie van toetsresultaten altijd nodig om de proeven succesvol te implementeren.

Randvoorwaarde 5: Een routekaart voor de invoer van centraal toetsen

Een goed werkend centraal toetssysteem is niet in een drietal jaar gebouwd. Zoals al aangegeven, is dat ook een leerproces waarbij zeer veel belanghebbenden betrokken zijn. Er zijn nog veel paden open en een deel van de ontwikkelingen is nog onzeker. Het is echter verstandig een routekaart te maken om daar een volgorde in aan te geven wat eerst moet gebeuren en wat later kan. In de scenario's zijn al suggesties voorgegeven. Ook bij het overzicht van de kosten is duidelijk wat in ieder geval moet, en wat optioneel is voor wat betreft hoe uitvoerig iets gedaan moet worden. De kosten bepalen voor een deel ook de routekaart, en de kosten hangen ook sterk af van het belang dat aan de toetsen gehecht wordt. In dat kader raden we sterk aan om zeker bij de eerste afnamen in het veld duidelijk te maken dat dit deels ook een leerproces is, en dat de belangen –zeker in het eerste jaar– niet al te hoog kunnen zijn voor alle belanghebbenden. Deze keuze zou de kosten ook meer beheersbaar maken, omdat de kosten hoog zijn bij de stappen die nodig zijn voor het voorkomen van fraude bij de afnamen. Kostenbeheersing kan zo deels ook door beheersing van de belangen gestuurd worden. De focus zou in het begin vooral moeten liggen op dat de basis goed is op het vlak van toetsontwikkeling, rapportages, afname en techniek. Als de leercurve stijl is, en de eerste al dan niet voorziene problemen opgelost zijn, kan het belang van de toetsen opgehoogd worden. Hoewel er bij de inrichting van het proces rekening gehouden moet worden dat de belangen hoger kunnen worden dan in het eerste jaar, hoeven de werkelijke kosten die daaraan gerelateerd worden pas dan gemaakt te worden. Dit kan ook een stapsgewijs proces zijn, dat ook afhangt van het draagvlak. Hoe meer de toetsen geaccepteerd zijn, hoe groter de belangen die er vanaf hangen kunnen zijn.

7.4.3 Tot slot

In deze haalbaarheidsstudie naar de pedagogische-psychometrische aspecten van de introductie van gecentraliseerde proeven in Vlaanderen hebben we getracht een inzicht te geven in het brede palet van mogelijkheden dat te vinden is bij het ontwerp van toetsen en examens. Toetsresultaten kunnen bij een goed gebruik een rijk en informatief beeld geven voor verschillende belanghebbenden, waarbij het naar ons inziens vooral van belang is dat de doelen van de toets voor deze verschillende belanghebbenden goed afgewogen worden en in balans zijn. Dit te realiseren, is echter geen sinecure. We hopen dan ook dat we ondersteunend hebben kunnen zijn in het relateren van de vele mogelijkheden die te vinden zijn voor een toetsontwerp aan de keuzes voor een bepaald gebruik van de toetsresultaten. En dat de voorbeelden die we gegeven hebben, inspireren bij de keuze van een haalbare strategie, en het maken van de routekaart van de invoer van het centrale proeven in Vlaanderen.

8. Samenvatting

Hoofdstuk 1 Introductie

In deze haalbaarheidsstudie naar de pedagogische-psychometrische aspecten van de introductie van gecentraliseerde proeven in Vlaanderen gaan we in op de vraag wat er nodig is om centrale toetsen vorm te geven. In de Hoofdstukken 2 en 3 wordt een algemeen kader gegeven. In Hoofdstuk 2 gaan we in op het gebruik van toetsen, i.e., verschillende doelen van toetsen die met elkaar in balans gebracht moeten worden. In Hoofdstuk 3 geven we een psychometrisch kader. We gaan dieper in op itemresponstheorie (IRT), wat het meest geschikt lijkt voor de aanpak van veel van de uitdagingen die de haalbaarheid van de centrale toetsen heeft, en waar in latere hoofdstukken naar gerefereerd zal worden. In de Hoofdstukken 4, 5, en 6 bespreken we de drie thema's: leerwinst, toetsontwikkeling en omgaan met resultaten. Tot slot gaan we in Hoofdstuk 7 in op de mogelijke scenario's met betrekking tot de praktische uitvoering van centrale toetsen in het Vlaamse onderwijs.

Deel I – Achtergrond

Hoofdstuk 2: Gebruik van toetsen in balans

Voordat men een kwalitatief goede toets kan ontwikkelen moeten eerst de uitgangspunten van de toets duidelijk zijn. De uitgangspunten van een toets zijn gedefinieerd rondom een drietal vragen. Dat betreft de vragen waarom we willen meten, wie we willen meten en wat we willen meten.

Waarom we willen meten

In grote lijnen zijn er twee verschillende doelen om een toets af te nemen. Het eerste doel is om informatie te verkrijgen die steun biedt aan het leerproces. Het tweede doel is om informatie te verkrijgen over waar de leerling staat na afronding van het leerproces. Toetsen die ons ter ondersteuning van het leerproces inzicht moeten geven

in het niveau van leerlingen noemen wij formatief. Toetsen die aan het eind van een leerproces worden afgenomen om een oordeel te vellen over het leerresultaat, noemen wij summatief. De opdeling van de doelen in formatief en summatief kan ook gemaakt worden op een geaggregeerd niveau, zoals dat van de school of op systeemniveau. Ook hier is het formatieve doel het verbeteren van het leren. Dat betreft een echte kwaliteitsbevordering van het onderwijs. Anderzijds kunnen toetsen voor scholen ook summatief gebruikt worden om een oordeel te vellen over het leerresultaat van de leerlingen van de school. Dit is meer kwaliteitsbewaking, en heeft net als bij een summatieve toets bij leerlingen meer een connotatie van "afrekenen". De opdeling van de doelen in formatief en summatief speelt dus een rol op zowel leerling-, school- als systeemniveau. Summatieve toetsen worden meestal als toetsen met grote belangen gezien (high-stakes tests). In onderstaande tabel staan de verschillen tussen formatief en summatief toetsen op een rij gezet.

Type toetsdoelen	**Vaststellen van het leerresultaat**	**Ondersteuning van het leerproces**
Kernbegrip	Oordelen	Helpen
Doel	Een beslissing nemen	Het leren verbeteren
Benadering van fouten	Vermijden	Om van te leren
Wanneer (t.o.v. leeractiviteit)	Aan het einde	Tijdens
Toetsresultaat	Eindoordeel	Terugkoppeling naar lesmateriaal
Benaming	Summatief	Formatief

Tabel 8.1: Verschil tussen summatieve en formatieve toetsen op diverse criteria

Wie we willen meten

Bij de centrale toetsen is de beoogde doelgroep leerlingen in het vierde en het zesde leerjaar van het lager onderwijs, en leerlingen aan het einde van de eerste en de derde graad van het secundair onderwijs. Naast de definiëring van de doelgroep is het ook van belang te definiëren of er uitsluitcriteria zijn die aangeven welke leerlingen de toets niet hoeven te maken.

Wat we willen meten

Bij de centrale toetsen heeft de vraag wat we willen meten betrekking op de vaardigheden wiskunde en Nederlands, met in ieder geval de onderdelen begrijpend lezen, schrijven en grammatica. Door bij de omschrijving van de toetsinhoud aan te sluiten bij de bestaande eindtermen en curricula, krijgen de toetsen en de daaropvolgende resultaten betekenis voor het onderwijs.

De terugkoppeling

Rapportages worden afgestemd op het toetsdoel, de doelgroep en de toetsinhoud van de toetsing. Bij een absolute normering wordt de prestatie van een leerling gerelateerd aan een bepaalde standaard. We evalueren of een leerling bepaalde stof beheerst of niet. Bij een relatieve normering wordt de prestatie van een leerling gerelateerd aan de prestaties van andere leerlingen. Met de toetsscore worden de leerlingen gerangschikt van zwak naar sterk, van lage vaardigheid naar hoge vaardigheid. De bedoeling is de resultaten

aan de scholen terug te koppelen op leerling- en schoolniveau. Bij de eerste meting zal de informatie vooral gaan over het al dan niet behalen van de eindtermen. De toetsen zijn dan absoluut genormeerd. De aandacht gaat in zo'n geval uit naar welke (clusters van) eindtermen wel en niet gehaald zijn, om vervolgens het onderwijs daarop aan te passen. Naast aandacht voor welke eindtermen behaald zijn, kan men ook richten op de hoeveelheid behaalde eindtemen. Wanneer de leerling voor een tweede keer gemeten wordt met centrale toetsen, minstens twee jaar na de eerste meting, wordt het mogelijk individuele leerlingen ook een terugkoppeling te geven over de individuele leerwinst.

Toetsen met grote belangen en minder grote belangen

Toetsen verschillen in de mate waarin er belang aan gehecht wordt. Wanneer een leerling een groot belang toedicht aan de uitkomst van een toets, ervaren zij deze als 'high-stakes', en bij vrijwel geen belang als 'low-stakes'. Vaak zien we in de praktijk dat er een discrepantie is tussen de ervaren belangen door de leerling en het belang van het toetsresultaat voor een leerkracht, de school of het ministerie. Het ervaren belang kan van invloed zijn op het niveau van de prestatie, op strategisch gedrag en heeft gevolgen voor toetsverversing, controle op fraude en de daarmee gemoeide kosten. In onderstaande tabel zijn deze aspecten van belangen overzichtelijk bij elkaar gezet.

	high-stakes	**low-stakes**
Niveau prestatie	Optimale prestatie	Typische prestatie
Strategisch gedrag	Grote kans	Weinig noodzaak
Toetsverversing	Vaak	Kan langer mee leren
Controle op fraude	Streng	Mild
Kosten van controle op fraude	Kostbaar in beheersing	Lager

Tabel 8.2: Belangen: hoog – laag (high-stakes vs low-stakes)

Hoofdstuk 3: Itemresponstheorie

Het grote voordeel van itemresponstheorie (IRT) ten opzichte van bijvoorbeeld klassieke testtheorie is dat het veel flexibiliteit biedt in het vergelijken van resultaten die behaald zijn op verschillende toetsen. Juist in een context waarin gezocht wordt naar mogelijkheden om te werken met verschillende varianten van toetsversies, biedt IRT dus aantrekkelijke voordelen. Om deze reden wordt in Hoofdstuk 3 een korte introductie gegeven over deze modellen die een prominente rol spelen in het moderne onderwijskundig meten. Ondanks de mogelijke risico's en uitdagingen die met de toepassing van IRT bestaan, is het aan te raden deze modellen te gebruiken binnen het centrale toetsen in Vlaanderen. IRT heeft namelijk een aantal zeer aantrekkelijke voordelen. Ten eerste, door het gebruik van IRT kan een cesuur gemakkelijk geëquivaleerd worden, wat betekent dat alle leerlingen, over scholen en door de tijd heen eerlijk met elkaar vergeleken worden. Zo zijn ook vaardigheidstrends over de verschillende afnamejaren goed te volgen. Ten tweede, leerwinst valt ook inhoudelijk te interpreteren als men IRT gebruikt en absolute voortgang op de vaardigheidsschaal weergeeft. Ten derde, itembanken gebaseerd op IRT zijn flexibel in het gebruik en bieden de mogelijkheid tot stapsgewijze aanvulling met meer opgaven zodat zowel vernieuwing als continuïteit geborgd is. Ten vierde, bij IRT kunnen gerichte verzamelingen items (toetsversies) aangeboden worden, passend bij de ingeschatte vaardigheid van de

leerling zodat de toets ook echt relevant is voor de leerling. Tot slot, de vaardigheidsschaal verkregen met IRT, is zeer goed aanschouwelijk te maken door een grote hoeveelheid opgaven zodat een bepaalde vaardigheidsscore niet alleen een abstract begrip wordt, maar werkelijk een betekenis kan krijgen in termen van wat de leerling kan.

Deel II – Onderzoeksvragen

Hoofdstuk 4: Leerwinst en toegevoegde waarde

Definities van leerwinst

Een gebruikelijke definitie van leerwinst is de toename van vaardigheden of kennis van leerlingen gedurende een bepaalde periode. Deze winst weerspiegelt dus het verschil in kennis van leerlingen op twee verschillende momenten. Een voor de hand liggende operationalisering van leerwinst is om deze te zien als het verschil in scores tussen een toets aan het einde en een toets aan het begin van de periode. Dit verschil kan ook geaggregeerd worden over groepen, bijvoorbeeld over klassen, scholen, regio's, of gewesten. Op schoolniveau wordt ook vaak gebruik gemaakt van de aan leerwinst gerelateerde term "toegevoegde waarde." De toegevoegde waarde van de school is de bijdrage van de instelling aan de leerwinst van haar leerlingen.

Leerwinst in het perspectief van leerlingen, scholen en beleidsmakers

Een goede interpretatie van leerwinst vereist het plaatsen van leerwinst in de context waarin deze behaald is. Op het niveau van leerlingen is het belangrijkste vraagstuk te bepalen of een leerling voldoende gegroeid is. Dit is te doen door te beoordelen of de leerling al dan niet bepaalde vaststaande standaarden heeft bereikt. Een manier om een indruk van de groei van een leerling te krijgen is het werken met groeicurves. Een manier om de interpretatie van groeicurves te faciliteren, is het werken met zogenoemde normeringsgegevens. In dat geval is er een combinatie van een vergelijking met de leerling met zijn of haar eerdere prestatie en wordt zijn of haar prestatie vergeleken met de prestatie van anderen. Op het niveau van de school is wellicht het belangrijkste vraagstuk hoe vast te stellen welk didactisch handelingsperspectief gebruikt kan worden om leerlingen verder te helpen. Leerwinst en toegevoegde waarden zijn instrumenten die in te zetten zijn door scholen om de kwaliteit van het gegeven onderwijs te verhogen.

Gegevens over toegevoegde waarde kunnen wel een meer genuanceerd beeld geven van het succes van de onderwijspraktijk dan alleen separate eindscores. Een school waar vooral leerlingen zitten met een bovengemiddelde intelligentie zal minder inspanningen hoeven te leveren om een hoge gemiddelde eindscore te bereiken dan een school met juist veel leerlingen met een benedengemiddelde intelligentie. Omgekeerd kan de laatste school echter wel veel "waarde" hebben toegevoegd aan de leeruitkomsten van haar leerlingen. Een hoge toegevoegde waarde en hoge eindopbrengsten kunnen dus wel samengaan, maar dat hoeft lang niet altijd het geval te zijn. Beleidsmakers, tenslotte, zullen vooral geïnteresseerd zijn in de vraag of de bijdrage van scholen aan leerwinst voldoende is.

Er spelen een aantal zaken die tot voorzichtigheid manen om leerwinst en toegevoegde waarde gegevens te gebruiken op het niveau van het onderwijssysteem. Ten eerste, om een goed beeld te krijgen welke scholen een effectief onderwijsbeleid voeren, moet je een onderscheid kunnen maken tussen de effecten op leerwinst die door de kenmerken van de

leerlingen, dan wel de omgeving van de school veroorzaakt zijn én dat deel dat wel toe te schrijven valt aan het schoolbeleid. Op het moment dat je dit onderscheid niet maakt, bestaat het risico dat er beleidsaanbevelingen worden gedaan voor alle scholen gebaseerd op scholen die in een bepaalde context opereren. Juist het goed onderscheid kunnen maken tussen omgevingseffecten en schoolpraktijk vraagt veel kennis over de context waarin de school zich bevindt. Het is complex om al deze informatie vanaf een afstand te verzamelen.

Modellen voor het in kaart brengen van leerwinst en toegevoegde waarde

In de literatuur zijn verschillende benaderingen te vinden voor het in kaart brengen van leerwinst en toegevoegde waarde. Deze kunnen grofweg ingedeeld worden in regressie-, stratificatie- en modelleringbenaderingen. Bij de verschillende methoden kan op verschillende manieren rekening gehouden worden met achtergrondvariabelen en factoren als zittenblijven, schoolveranderingen, en veranderingen van studierichting. Bijvoorbeeld door deze zaken in modellering mee te nemen, is het mogelijk de scholen te vergelijken, inzicht te krijgen in de oorzaken van de veranderingen en de toegevoegde waarde van de scholen. Variabelen die vaak als 'fairness kenmerken' worden opgenomen in de modellen zijn leerlingkenmerken zoals geslacht, migratie-achtergrond en de sociaaleconomische status. In Hoofdstuk 4 gaan we uitgebreid in op deze modellen en hoe achtergrondvariabelen meegenomen kunnen worden in deze modellen. In ieder geval is het aan te bevelen diverse methoden te gebruiken om leerwinst en toegevoegde waarde te bepalen. Juist door de verschillende indicatoren van leerwinst en toegevoegde waarde naast elkaar te beschouwen krijgt men het gehele beeld.

Hoofdstuk 5: Toetsontwikkeling

Toetsverversing

Toetsen kunnen op verschillende momenten geheel of gedeeltelijk vervangen worden. Een belangrijke reden om toetsen te vervangen, is het voorkomen dat leerlingen een oneigenlijk voordeel hebben op het moment dat items voor toetsafname bekend zijn. Daarnaast is een belangrijke reden om aan te sluiten bij een veranderende onderwijsinhoud. Wanneer toetsen een puur formatieve functie hebben, dan zijn vaste toetsen die langer meegaan goed toepasbaar. Toetsen kunnen ook bij elke afname helemaal ververst worden. Dit is vooral relevant op het moment dat er veel kans op fraude bestaat. In alle gevallen is een goed onderliggende itembank aan te bevelen. Een toetsitemdatabank of itembank is een gestructureerde verzameling van toetsvragen die een bepaald concept of inhoudsdomein meten. Als de statistische itemkenmerken uit een IRT-model opgenomen worden in de itembank, èn een rol spelen bij de samenstelling van toetsvarianten, dan spreken we van een gekalibreerde itembank. Het werken met een itembank waarvoor een IRT-model geldt, biedt veel flexibiliteit, met alle voordelen van dien. Doordat niet iedereen dezelfde toets hoeft te maken om de resultaten met elkaar vergelijkbaar te maken, is het mogelijk binnen een afname verschillende toetsen af te nemen. Ook biedt IRT de mogelijkheid om trends door de jaren heen waar te nemen zonder dezelfde toetsen te hoeven gebruiken. Hergebruik van een deel van de opgaven uit het voorgaande jaar maakt het mogelijk door middel van IRT de uitkomsten van verschillende jaren met elkaar te vergelijken. Ook het volgen van leerlingen over leerjaren heen is mogelijk met behulp van opgaven uit een IRT-gekalibreerde itembank. Om de itembank in omvang voldoende groot te houden, is

verversing van de bank nodig.

Selectie te toetsen onderwijsdoelen
Er zijn meerdere redenen om gebruik te maken van verschillende toetsversies binnen een toets. Een eerste reden kan zijn om te faciliteren dat niet iedereen op hetzelfde moment een toets kan maken. Een tweede reden om te werken met verschillende toetsversies is dat het de mogelijkheid geeft om de items te matchen aan de vaardigheid van de leerlingen. Een derde reden om met meerdere toetsversies te werken is dat het gelegenheid biedt om een compleet inhoudsdomein te meten op geaggregeerd niveau, terwijl gelijktijdig de toetstijd op individueel niveau beperkt blijft. Een meer algemeen voordeel van het gebruik van verschillende toetsversies is dat de opgaven uit een itembank minder snel bekend raken. Het verkrijgen van vergelijkbare resultaten ondanks het gebruik van meerdere toetsversies is bij het gebruik van IRT te realiseren, maar een cruciale voorwaarde is het creëren van overlap in het rotatiedesign. Dit betekent dat binnen de verschillende toetsversies dezelfde items terugkomen (zogenaamde ankeritems) die ervoor zorgen dat alle toetsversies met elkaar verbonden zijn. Voorbeelden van dit soort designs zijn het anchor item design, het block interlaced design en een balanced block design.

Kerntoets aangevuld met materiaal onderwijsverstrekkers
Een van de manieren om curriculumvernauwing tegen te gaan, is door de centrale toetsing aan te vullen met school-eigen toetsmateriaal. Er zijn drie hoofdscenario's voor de aanvulling van kerntoetsen met net- of koepelspecifiek toetsmateriaal: ten eerste door de net- of koepel-specifieke toetsen naast de kerntoetsen af te nemen; ten tweede door delen van de centrale toetsen door net- of koepel-specifieke toetsen te vervangen; en ten derde door delen van de net- of koepel-specifieke toetsen te integreren in de ontwikkeling van de centrale toetsen. Het scenario van net- of koepel-specifieke toetsen naast de centrale toetsen is aan te raden bij het meten van vaardigheden die niet makkelijk met gesloten of korte-antwoordvragen kunnen worden geoperationaliseerd. Het combineren van koepel-specifieke toetsen met de kerntoetsen heeft een aantal belangrijke voordelen. Een eerste voordeel is dat het kan helpen om curriculumvernauwing tegen te gaan. Het tweede voordeel is dat de combinatie van koepel-specifieke toetsen en kerntoetsen de scholen kan helpen bij het inschatten van het niveau van de niet-centraal gemeten vaardigheden. Een derde voordeel is dat het werken met koepel-specifieke toetsen recht doet aan de eigenheid van de school. Een school kan eigen accenten leggen daar waar zij zich binnen het curriculum extra op willen richten. Het vervangen van centrale toetsendelen door net- of koepel-specifieke toetsdelen is complexer. Voor het handhaven van de vergelijkbaarheid is het dan noodzakelijk dat een deel van de leerlingen zowel de specifieke als de volledige kerntoets maakt. Dit scenario is om praktische reden beter niet te volgen bij de invoering van het centrale toetsen in Vlaanderen. Het scenario dat er net-eigen varianten van de centrale toetsen bestaan is voor te stellen, maar levert voor de vergelijkbaarheid van de toetsen en resultaten wel de nodige uitdagingen. Op dit moment zijn net-eigen varianten van de centrale toetsen daarom niet te adviseren, maar valt dit na het opdoen van voldoende ervaring met de centrale toetsen mogelijk in de toekomst te overwegen.

De relatie tussen toetstijd, nauwkeurigheid, en betrouwbaarheid

De betrouwbaarheid van een test geeft een indicatie of er veel of weinig toevallige meetfout te verwachten is en kan dan ook gezien worden als een maat voor consistentie van een meting. Uiteraard is het wenselijk dat een meetinstrument een hoge betrouwbaarheid heeft. Verschillende factoren zijn gerelateerd aan de betrouwbaarheid van de toets. Deze factoren zijn onder te verdelen in kenmerken gerelateerd aan de toets, de afnamecondities en de groep leerlingen die de toets afnemen. Het eerste kenmerk van de toets dat van invloed is, is de lengte van de toets. De relatie met de toetslengte en betrouwbaarheid is eenduidig: hoe meer opgaven, hoe betrouwbaarder de meting. Hierbij wordt er wel vanuit gegaan dat al deze opgaven één en hetzelfde construct meten. Betrouwbaarheid is ook gerelateerd aan de kwaliteit van de opgaven. Hoe beter elk van de opgaven de te meten vaardigheid meet, hoe betrouwbaarder de toets. Naast toetskenmerken spelen de afnamecondities een rol bij de hoogte van betrouwbaarheid van een toets. Een toetsafname die gestoord wordt door omgevingslawaai, zal meer behept zijn met meetfout. Duidelijke instructies voor het afnemen van een toets dragen bij aan betere afnamecondities. De groep leerlingen tenslotte is de derde factor die invloed heeft op de hoogte van de betrouwbaarheid. Een toets kan betrouwbaarder zijn bij een grotere spreiding van de te meten vaardigheid in de populatie. De maten die binnen de context van de centrale toetsen relevant zijn, zijn betrouwbaarheid op basis van interitemrelaties die vooral uit de klassieke testtheorie komen, en de methoden op basis van IRT. Betrouwbaarheid van verschillen tussen twee metingen zijn over het algemeen lager dan die van de metingen op zich. De sterkte van die daling is voor een groot deel afhankelijk van de correlatie tussen de metingen. Om de leerwinst betrouwbaar te meten, is het dus een noodzakelijke voorwaarde dat de afzonderlijke metingen in ieder geval voldoende betrouwbaar zijn.

Veranderende adaptiviteit

Bij een adaptieve toets krijgen leerlingen items voorgelegd die aansluiten bij het niveau dat ze op een eerder onderdeel van de toets hebben laten zien. Adaptief toetsen heeft een aantal voordelen ten opzichte van het gebruik van lineaire toetsen waarin alle leerlingen dezelfde items maken. Ten eerste wordt een leerling niet of in ieder geval minder geconfronteerd met veel te moeilijke of veel te makkelijke opgaven. Dergelijke opgaven kunnen de kandidaat demotiveren en tot een minder zuivere meting leiden. Ten tweede kan een adaptieve toets tot een nauwkeurigere schatting leiden van de vaardigheid van de leerling. Een derde voordeel van adaptieve toetsing is dat het automatisch tot het gebruik van verschillende toetsversies leidt, wat bijdraagt aan een veilige toetsafname en risico's van toetsfraude reduceert. Er bestaan twee belangrijke varianten van adaptief toetsen, computer adaptief toetsen (CAT) en multistage-toetsen (MST). Bij een CAT wordt na ieder afgenomen item de vaardigheid van de leerling opnieuw geschat en op basis van die schatting het meest geschikte volgende item gekozen. Multi-stage-testing (MST) is een vorm van adaptief toetsen die niet op item-niveau plaatsvindt, maar op een verzameling items. Zo'n verzameling items wordt een module genoemd. Leerlingen beginnen allen met dezelfde module (routing test genoemd) en op basis daarvan krijgen ze een moeilijkere of juist een makkelijkere module voorgelegd in het tweede deel van de toets. In Hoofdstuk 5 worden de voor- en nadelen van de verschillende varianten van adaptief toetsen besproken.

Brede afname
Een brede afname geeft alle leerlingen de mogelijkheid om deel te kunnen nemen aan de toets, en hun vaardigheid of kennis aan te tonen. Om een brede of inclusieve afname te kunnen realiseren zijn er soms verschillende aanpassingen aan de toets nodig. Als het nodig is verschillende varianten van opgaven te gebruiken om een brede afname te realiseren, moeten de toetsen geëquivaleerd worden, wat betekent dat aangetoond wordt dat de alternatieve variant hetzelfde meet als de reguliere versie van de toets. Dat betekent dat er ook onderzoek naar vraagpartijdigheid (DIF) wordt uitgevoerd, waarbij onderzocht wordt of de aangepaste opgave nog wel hetzelfde meet als de reguliere opgave. Als de toetsvarianten voldoende equivalent zijn, kan de score van de alternatieve variant via psychometrische equivalering -meestal met IRT- vertaald worden naar de reguliere toetsscore. Ook het equivaleren van eenvoudigere varianten, waarin relatief makkelijke reguliere opgaven aangevuld zijn met eenvoudige vragen aan reguliere varianten kan met IRT. Ook door middel van standaardbepalingsprocedures zouden de toetsen en aangepaste toetsen vergelijkbaar gemaakt kunnen worden.

Leereffecten in toetsontwikkeling
Voor de continue verbetering van de toetsen is het belangrijk om het psychometrisch functioneren van items en toetsen te onderzoeken. Ook is een terugkoppeling en bespreking hiervan met de constructeurs noodzakelijk, zodat zij leren om betere items te maken. Het is aan te bevelen rapporten te produceren die itemschrijvers helpen hun opgaven goed te kunnen beschouwen. Het is aan te bevelen een plan-do-check-act-(PDCA-)cyclus te hanteren, die tot een continue verbetering van de kwaliteit van de toetsen kan zorgen. Het toepassen van de PCDA-cyclus helpt zowel bij het doorvoeren van verandering en continue verbetering van de centrale toetsen als wel het verbeteren van hoe deze toetsen gemaakt worden. Het betekent in het geval van centrale toetsen in Vlaanderen dat de haalbaarheid toeneemt als er gestart wordt met een basisvariant van deze toetsen.

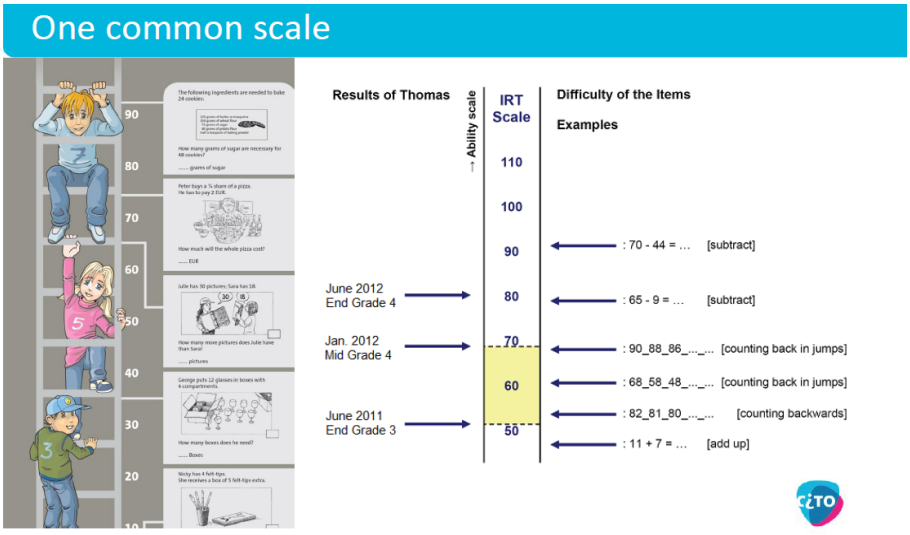
Toetsontwikkeling met papieren en digitale toetsvarianten
Het is de ambitie om alle centrale toetsen in Vlaanderen digitaal af te nemen. Dit biedt een aantal belangrijke voordelen. De grootste voordelen zijn te vinden in de schaalbaarheid in verdeling en verwerking van de toetsen. De digitale afname biedt ook de mogelijkheid om nieuwe soorten opgaven te ontwikkelen die op papier niet goed werken. Echter, de ervaring leert dat een grootschalige centrale afname waarbij tienduizenden leerlingen tegelijkertijd dezelfde toets digitaal moeten afnemen, eigen uitdagingen kent. Om die reden kan het zeer zinvol zijn ook een scenario te ontwikkelen waarbij bij de centrale toetsing er –in ieder geval deels, en in ieder geval het eerste jaar van afnamen– op papieren varianten teruggevallen kan worden.

Hoofdstuk 6: Omgaan met de resultaten
Toetsscores en resultaten worden veelal gepresenteerd in rapportages. Deze rapportages zijn het middel waarmee toetsresultaten vertaald worden naar betekenisvolle acties. Bij het beantwoorden van de vragen over hoe te rapporteren, is het van groot belang wat gerapporteerd moet worden, wie er ingelicht moet worden, en wat de ontvanger van de rapportage ermee moet doen, oftewel wat het doel van de rapportage is. Bij de vraag ‘wat’ er gemeten

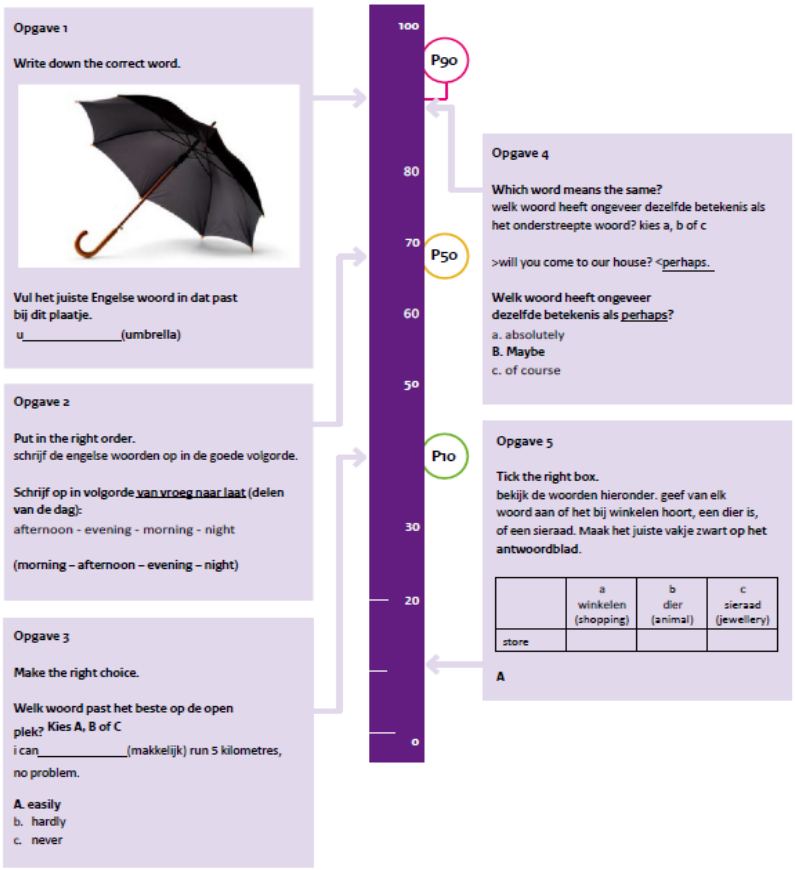
wordt, gaat het niet alleen om de omschrijving van het construct (de toetsinhoud) en of deze in lijn is met de eindterm. De mate van detail waarop gerapporteerd wordt, is ook van belang. Is het bijvoorbeeld alleen gewenst de vier hoofdvaardigheden te meten (wiskunde en Nederlands lezen, schrijven en grammatica), of zijn meer gedetailleerde uitspraken op onderliggende leerdoelen of domeinen gewenst? Daarnaast is het van belang of er gerapporteerd moet worden hóe goed een leerling in een bepaalde vaardigheid is, of dat het wellicht voldoende is te weten dát de leerling de eindterm behaald heeft. Tevens is het relevant om onderscheid te maken in rapportages waar de nadruk ligt op de vaardigheid op één moment in de tijd, versus rapportages waar de nadruk juist ligt op veranderingen in de tijd: bijvoorbeeld op de groei die een leerling heeft doorgemaakt. Naast een keuze over 'wat' er gerapporteerd wordt, is het ook belangrijk over 'wie' de rapportage gaat: over een individuele leerling, over een klas, over alle leerlingen van één docent, over alle leerlingen in een school of zelfs over het Vlaams onderwijsstelsel als geheel. Afhankelijk van het doel kan een verschillend aggregatieniveau gekozen worden om de toetsresultaten te presenteren. Het is aan te raden verschillende vormen van rapportages te geven geschikt voor verschillende doelgroepen (leerling, ouders, docent, school, bestuurder) en afgestemd op de informatiebehoefte. Daarnaast onderstrepen we het belang van relevante en begrijpelijke rapportages van toetssystemen.

Voorbeelden van rapportages

In deze haalbaarheidsstudie worden verschillende voorbeelden van rapportages besproken en geïllustreerd. We gaan in op voorbeelden van rapportages waarin een inhoudelijke betekenis van een vaardigheidsschaal gepresenteerd wordt, rapportages met groepsoverzichten en rapportages met als doel een diepere inhoudelijke (fouten)analyse te geven. Ten slotte wordt er ingegaan op het rapporteren van meetfouten. Onderstaande figuren zijn voorbeelden van rapportages waar een inhoudelijke betekenis aan de vaardigheidsschaal gegeven wordt.



Figuur 8.1: Voorbeeld inhoudelijke duiding vaardigheidsschaal



Figuur 8.2: Voorbeeld van representatie van opgaven op een vaardigheidsschaal

Randvoorwaarden voor goed gebruik

Om een goed gebruik van rapportages te borgen zijn een aantal randvoorwaarden te formuleren. Allereerst is het van belang dat het doel van een toetssysteem of toets bekend is bij de gebruikers. De interpretatie van gegevens die worden weergegeven in rapportages zijn afhankelijk van de specifieke vragen waarmee naar de rapportage gekeken wordt. Wordt met een formatieve bril naar de rapportage gekeken dan zullen andere conclusies getrokken worden dan wanneer met een summatieve bril naar dezelfde rapportage wordt gekeken. Daarnaast is het belangrijk om verstaanbare rapportages te ontwikkelen. Het is daarom belangrijk om toekomstige gebruikers te betrekken bij het ontwerpen van rapportages. Behalve gebruik te maken van een intuïtief begrijpelijke rapportagevorm, is het van groot belang dat de wijze waarop de resultaten gepresenteerd worden, aansluiten bij de informatiebehoefte van de leerkrachten en scholen. Daartoe moet bekend zijn wat de leerkrachten en scholen willen weten. Het is daarmee van belang te realiseren dat het ontwikkelen van de terugrapportages een stapsgewijs proces is waarbij de PCDA-cyclus een belangrijke rol speelt.

Deel III – Scenario's

Hoofdstuk 7: Scenario's voor toetsing in Vlaanderen

Tot slot gaan we in Hoofdstuk 7 in op de mogelijke scenario's en hoe deze scenario's

in de Vlaamse context vorm kunnen krijgen. Er zijn twee extreme scenario's te schetsen: formatieve toetsing en summatieve toetsing. Deze kunnen op diverse onderwijsniveaus (leerling, docent, school, landelijk) worden toegepast. We beginnen met de schets 'achteraan', dus bij het resultaat van de toets (de rapportage). De handelingen die passen bij de verschillende doel-scenario's en verschillende rapportageniveau 's worden in onderstaande tabel weergegeven.

Rapportage-niveau	**Formatief scenario**	**Middenweg**	**Summatief scenario**
Leerling	Uitstippelen individueel leerplan	Advies bij keuze uit een beperkte set van vervolgtrajecten	Bepalen van doublering of diplomering
Leerkracht*	Sterkte-zwakte-analyse (SWOT) uitvoeren	Bonus bij uitvoeren van zelfreflectie	Bepalen van salariëring of baanzekerheid
School	Ouders kiezen een type school dat bij hun kind past; Regionaal of lokaal overleg over; verklaringen voor prestaties	Schoolbegeleiding	Sluiting van zeer zwakke scholen. Beloning van goede scholen
Landelijk	Infrastructuur voor kwaliteitsverbetering opzetten en onderhouden	Transparantie bieden over landelijke prestaties	Richtlijnen voorschrijven; Financiering herverdelen

* Rapportages op leerkracht-niveau kunnen misleidend zijn.

Tabel 8.3: Rapportage per doelgroep voor de drie verschillende doel-scenario's

Diverse aspecten van toetsontwerp, die in Hoofdstukken 4, 5 en 6 aan de orde kwamen, kunnen anders ingevuld worden, al naar gelang de voorkeur bij formatieve of summatieve toetsing ligt. Niet alle aspecten die besproken zijn, komen voor zo'n tweedeling in aanmerking. De onderstaande aspecten kennen wel zo'n andere invulling. Een overzicht van deze aspecten en de scenario's is te vinden in onderstaande tabel.

Dilemma's en scenario's

In Hoofdstuk 7 worden de thema's uit de Hoofdstukken 4, 5 en 6 vanuit een praktisch kader beschouwd, en bespreken we een aantal dilemma's rondom deze thema's. We gaan in op de vraag hoe verschillende summatieve en formatieve doelen met elkaar in balans gebracht kunnen worden. We bespreken de mate waarin centrale toetsen peilingen over zouden kunnen nemen, of aan zouden kunnen vullen. Binnen het thema leerwinst gaan we in op de vraag in hoeverre de gemeten vaardigheden naar verwachting gelijk zijn over de jaren heen, of veranderen. We gaan in op de consequenties voor het meten van de vaardigheid over tijd. Een ander punt rond dit thema betreft de praktijk van het meten van leerwinst als de tijdspanne tussen de metingen groot is en welke oplossingsrichtingen voorhanden liggen. Bij het thema toetsontwikkeling gaan we in op de uitgave van toetsversies: hoe en hoe vaak toetsen binnen een jaar, en over jaren heen vernieuwd kunnen

Aspect niveau	Formatief scenario	Middenweg	Summatief scenario
Doel toetsing	Passend onderwijs	Motivatie verhogen met consequenties	Eerlijke kansen bij kwalificatie en selectie
Validiteit	Construct validiteit	Construct validiteit; predictieve validiteit	Predictieve validiteit; construct validiteit
Vereiste betrouwbaarheid	Matig tot hoog	Hoog	Hoog tot zeer hoog
Toetsverversing	Lage frequentie, lage beveiliging	Matige frequentie, matige beveiliging	Alle leerlingen maken dezelfde doelen voor vergelijkbaarheid
Selectie onderwijsdoelen	Rotatie van doelen over leerlingen en jaren	Deels rotatie, deels vaste onderdelen	Alle leerlingen maken dezelfde doelen voor vergelijkbaarheid
Adaptieve toetsing	Zinvol om aan te sluiten bij niveau van leerling	Beperkte adaptiviteit	Mogelijk, mits er goede communicatie is over de betekenis van de rapportage
Net- of koepel-specifiek materiaal	Vervangend gebruik	Deels vervangend, deels vaste onderdelen	Aanvullend gebruik
Rol papieren toetsing	Altijd bij jonge of niet-digitaal-ervaren kinderen	Beschikbaar, naast digitale toets	Calamiteitenvariant
Illustratie toetsresultaten	Illustratie met vaardigheidsschaal en daadwerkelijk gedrag	Landelijke prestaties illustreren met daadwerkelijk gedrag	Illustratie door vergelijking met landelijke prestaties
Rapportage van subdomeinen	Veel aandacht voor diversiteit van onderwerpen	Beperkte aandacht voor diverse onderwerpen	Grote nadruk op één onderliggende vaardigheid

Tabel 8.4: Invulling van de kwaliteitsaspecten voor de drie verschillende doel-scenario's

worden om het uitlekken van opgaven tegen te gaan, terwijl de resultaten van de versies wel goed met elkaar vergeleken kunnen worden. Ook gaan we wat dieper in op de invulling die gegeven kan worden aan het implementeren van brede afnamen. Bij het thema van rapportages gaan we in op de vorm van de rapportages, en gaan we dieper in op het belang van de ontwikkeling van de rapportages. Besproken wordt hoe om te gaan met aanwezige datageletterdheid en hoe goed toetsgebruik gestimuleerd kan worden. Tot slot geven we aandacht aan de vraag hoe rankings van scholen te voorkomen.

Kosten, praktische aanbevelingen en randvoorwaarden voor succesvolle implementatie
Voor de verschillende scenario's (formatief/summatief op leerling/schoolniveau) zijn praktische aanbevelingen te geven om deze scenario's succesvol te kunnen implementeren. In Hoofdstuk 7 bespreken we voor elk scenario best practices om het slagen van de centrale proeven te kunnen bevorderen. Daarnaast zijn er ook randvoorwaarden te schetsen die

de implementatie van de proeven in ieder mogelijk scenario bevorderen. Deze hebben betrekking op het specificeren van het doel van de toetsing, het creëren van draagvlak en ondersteuning en onderhoud. De kosten voor het introduceren van gecentraliseerde toetsen en examens kunnen onderverdeeld worden in kosten die min of meer onafhankelijk zijn van het aantal afnames (vaste kosten), kosten die wel afhankelijk zijn van het aantal deelnemende leerlingen en scholen (de variabele kosten) en kosten die afhangen van de keuzes voor een bepaald toetsontwerp. In Hoofdstuk 7 worden deze typen kosten besproken.

Tot slot

In deze haalbaarheidsstudie naar de pedagogische-psychometrische aspecten van de introductie van gecentraliseerde proeven in Vlaanderen trachten we een inzicht te geven in het brede palet van mogelijkheden dat te vinden is bij het ontwerp van toetsen en examens. We hopen met de voorbeelden die we geven te inspireren bij de keuze van een haalbare strategie.

IV

Appendix

A. Vragen uit het bestek

A.1 Leerwinst

1. Welke definities van leerwinst zijn mogelijk?
2. Hoe kunnen deze definities van leerwinst geoperationaliseerd worden en dit op niveau van de individuele leerlingen, op niveau van de school en op niveau van het onderwijssysteem. Geef hierbij voor elk van de definities aan:
   - hoe deze de interpretatie en het gebruik van de resultaten beïnvloeden op het niveau van de leerling, de school en het systeem;
   - welk effect deze hebben op de vergelijkbaarheid tussen scholen;
   - voor welke kenmerken gecontroleerd moet worden;
   - hoe omgegaan moet worden met factoren als zittenblijven, schoolveranderingen, veranderingen van studierichting, . . . ;
   - welke mogelijke tegenstrijdigheden/afwegingen kunnen opduiken en hoe hiermee omgegaan kan worden (bv scholen met zwakke instroom die hoge leerwinst hebben maar waar geen enkele leerling de eindtermen bereikt)

3. Op welke manier kan leerwinst gemeten worden aan de hand van de graadspecifieke eindtermen? Kunnen de aan de eindtermen gekoppelde bouwstenen hiervoor gebruikt worden?

A.2 Toetsontwikkeling

Alle leerlingen van een bepaald leerjaar worden jaarlijks getoetst voor één of meerdere leerdomeinen.

A.2.1 Toetsverversing

4. Moet jaarlijks een nieuwe toets worden ontwikkeld of kan er gewerkt worden met een toetsitemdatabank waaruit voor een langere periode geput kan worden? Welke andere scenario's zijn mogelijk?
5. Wat zijn voor- en nadelen en randvoorwaarden van deze scenario's? Hierbij wordt minstens gekeken naar toetsfraude en de mogelijke publieke verspreiding van de toetsitems.

A.2.2 Selectie te toetsen onderwijsdoelen

6. Om curriculumvernauwing tegen te gaan zullen niet steeds dezelfde onderwijsdoelen geselecteerd worden voor opname in de toets. Binnen het vak/leergebied zal rotatie worden toegepast. Op welke manieren kan deze rotatie uitgewerkt worden?
7. Hoe kan een rotatieprincipe verzoend worden met het meten van leerwinst op het niveau van de individuele leerling en van de school?

A.2.3 Kerntoets aangevuld met materiaal onderwijsverstrekkers

8. Op welke manieren kan een kerntoets, die voor alle leerlingen en scholen dezelfde is, worden aangevuld met toetsmateriaal vanwege de onderwijskoepels?
9. Wat zijn de gevolgen van een dergelijke aanvulling voor vb. de verwerking van de toetsresultaten?

A.2.4 De relatie tussen toetstijd, nauwkeurigheid en betrouwbaarheid

10. Hoe lang moet de toets zijn (aantal items/toetstijd) opdat deze:
    - voldoende betrouwbaar is op (a) individueel niveau en (b) schoolniveau;
    - voldoende nauwkeurig is op (a) individueel niveau en (b) schoolniveau;
    - werkbaar/haalbaar is voor zowel de individuele leerlingen als de school en dit voor de verschillende stappen in het proces?

A.2.5 Veranderende adaptiviteit

11. Indien toetsen in de toekomst meer of minder adaptief worden gemaakt, wat is dan het effect op het aantal benodigde items en de toetstijd? Zijn er zinvolle conceptuele aanpassingen mogelijk?

A.2.6 Brede afname

12. Wat zijn de gevolgen van het implementeren van redelijke aanpassingen en strategieën (bv op de kostprijs, standaardisatie, beperking in itemontwikkeling?
13. Welk effect heef het verminderen van de standaardisatie van de afname op de vergelijkbaarheid van de resultaten?

A.2.7 Leereffecten in toetsontwikkeling

14. Gezien de opeenvolgende ontwikkeling van verschillende (sets van) toetsen, welke elementen lenen zich tot continue verbetering van cyclus tot cyclus? Hoe kunnen we ervoor zorgen dat we optimaal gebruik kunnen maken van dergelijke leereffecten?

A.3 Omgaan met de resultaten

15. Wat zijn mogelijke scenario's voor het terugkoppelen van de resultaten aan leerkrachten en scholen?
    - Geef voor elk van de scenario's aan:
    - hoe de resultaten van de school op een voor onderwijsprofessionals verstaanbare en bruikbare manier worden beschreven;
    - hoe wordt omgegaan met mogelijke tegenstrijdigheden/afwegingen (bv scholen met zwakke instroom die hoge leerwinst hebben maar waar geen enkele leerling de eindtermen bereikt)
    - hoe de resultaten van de school vergelijkbaar worden gemaakt met relevante andere scholen;
    - hoe er wordt omgegaan met de balans tussen de complexiteit van de analyses waarop de resultaten gebaseerd zijn en de bruikbaarheid en verstaanbaarheid van de resultaten anderzijds.

16. Welke randvoorwaarden moeten vervuld zijn opdat scholen en leerkrachten de resultaten op een goede manier kunnen gebruiken in het kader van hun interne kwaliteitszorg? Hoe kunnen we komen tot het vervullen van deze randvoorwaarden?