FUNDAMENTELE VRAGEN OVER EXAMENS EN TOETSING

Eindrapportage

Jaap Scheerens, Arnold Brouwer, Piet Sanders,
Bernard Veldkamp, Anne Luc van der Vegt

Oberon
onderzoek | advies

Oberon B.V.
Slachtstraat 12, Postbus 1423, 3500 BK Utrecht
Tel. 030-2306090
www.oberon.eu

Dit onderzoek is gefinancierd door het Nationaal Regieorgaan Onderwijsonderzoek.
NRO-projectnummer: 40.5.18461.001

NRO NATIONAAL REGIEORGAAN ONDERWIJSONDERZOEK

Inhoudsopgave

Samenvatting

Jaap Scheerens

Het eindexamen voortgezet onderwijs heeft in Nederland een betrekkelijk lange traditie. Sinds het eind van de tachtiger jaren van de vorige eeuw is ook geleidelijk een stelsel van studietoetsen tot stand gekomen. In dit rapport ligt het accent op de verplichte eindtoetsen en examens, maar wordt af en toe eveneens gerefereerd aan toets- en evaluatieactiviteiten die op initiatief van onderwijsinstellingen plaatsvinden. Momenteel gelden als verplichte toetsen en examens: het eindexamen VO, leerlingvolgsystemen en de eindtoets in het primair onderwijs. In het eerste hoofdstuk wordt het evaluatie- en assessmentsysteem in Nederland nader beschreven. Recente externe evaluaties van de OECD oordelen in zeer positieve termen over het Nederlandse onderwijsevaluatiesysteem en er worden geen ingrijpende voorstellen tot verandering gedaan.
Toch ligt het toets- en examenstelstel in Nederland onder vuur. Over een breed front is er kritiek op 'een smalle kijk op onderwijskwaliteit' die door het huidige stelsel van examens en verplichte toetsen in stand zou worden gehouden. Tevens wordt in bloemrijke termen (toets-gekte, prestatiepijn) gesproken over negatieve gevolgen van toetsen en examens.

Een centraal begrip in de discussie over examens en toetsen is 'onderwijskwaliteit'. Om de argumentatie over 'breed' en 'smal' goed te begrijpen, is het daarom noodzakelijk stil te staan bij typeringen van onderwijskwaliteit die in het debat gebruikt worden. Door vrijwel alle betrokken organisaties, de Onderwijsraad, de VO-raad en het Platform Onderwijs2032, wordt kwaliteit getypeerd volgens een indeling van functies uit de onderwijssociologie. Als hoofdfuncties worden onderscheiden een kwalificatie-, een socialisatie- en een persoonsvormingsfunctie. De lijn van denken op basis van deze drie functies is dat er momenteel in onderwijs, examens en toetsen teveel nadruk ligt op de kwalificatie-functie en dat dit ten koste gaat van de beide andere functies. De functionalistische indeling is misschien bruikbaar om een globale richting aan te geven over een gewenste ontwikkeling in het onderwijs, maar is verder weinig operationeel. Kunnen de functies als onderscheiden onderwijsdoelstellingen worden opgevat? In hoeverre zijn ze beïnvloedbaar door het onderwijs? Zijn ze meetbaar? Behoren ze een gelijkwaardige rol te hebben, gegeven de basismissie van de school, die we voorlopig toch maar blijven zien in de sfeer van overdracht en leren van kennis en vaardigheden?

Om met dit soort vragen beter uit de voeten te kunnen wordt er een model van onderwijskwaliteit gepresenteerd dat eveneens breed is, maar de mogelijkheid biedt tot het onderscheiden van een aantal meetbare facetten. Het basisstramien is een systeemmodel, bestaande uit een contextdimensie, inputfactoren, productieproces en opbrengsten. De onderwijskundige invulling van dit model ziet leerlingresultaten als realisatie van onderwijsdoelstellingen, input, processen en randvoorwaarden als middelen en de context als een geheel van deels toeleverende en deels vragende condities. Als meetbare facetten van onderwijskwaliteit worden onderscheiden:
- Productiviteit (zijn de leerlingresultaten van voldoende niveau?)
- Effectiviteit (zijn de middelen doeltreffend om de leerlingresultaten te realiseren?)
- 'Equity' oftewel kansengelijkheid (differentiatie van leerlingresultaten naar subgroepen)
- Efficiency (keuze van de meest doelmatige middelen en werkwijzen, waarbij de verhouding tussen inzet van middelen en opbrengsten het gunstigst is)

- Responsiviteit (zijn de doelstellingen en werkwijzen zodanig gekozen dat voldaan wordt aan behoeften van belanghebbende externe partijen, zoals ouders, vervolgonderwijs en de arbeidsmarkt).

Dit conceptuele kader van onderwijskwaliteit benadrukt het belang van meetbare opbrengsten, "kwaliteit moet blijken uit de opbrengsten", en kan de inhoudelijke keuzes die samenhangen met de onderscheiden sociologische functies incorporeren.
Hoe kwaliteit wordt beoordeeld, hangt ook af van het perspectief van de beoordelaar. Ouders wegen de opbrengsten van het onderwijs anders dan de Inspectie van het Onderwijs. De school zelf is eerst verantwoordelijk voor de onderwijskwaliteit en van de manier waarop de opbrengsten worden getoetst door het schoolexamen. De school doet dat binnen de kaders van de landelijke eindtermen, maar heeft veel vrijheid bij de invulling.

Examens en eindtoetsen zijn op te vatten als bepalend voor de institutionele structuur van onderwijsstelsels. Met 'institutioneel' wordt in dit verband bedoeld dat ze de voornaamste interne spelregels en de communicatie over en weer met de ruimere maatschappelijke context vastleggen. Men zou dit ook kunnen uitdrukken door te spreken van de functie van examens voor de kwaliteitsborging van het onderwijssysteem. Daarnaast kan er een kwaliteitsbevorderende functie van examens en eindtoetsen worden onderkend, in die zin dat de werking ervan het onderwijs op een hoger peil brengt. Er wordt verwezen naar empirisch onderzoek dat deze werking ondersteunt.
Kwaliteitsborging manifesteert zich in de volgende functies:

- Eindtoetsing faciliteert doorstroom naar vervolgonderwijs;
- Eindtoetsing in het voortgezet onderwijs garandeert het civiel effect van diploma's;
- Eindtoetsing garandeert een gedeelde brede vorming met minimaal een basisniveau.

De rol van examens in het kader van kwaliteitsborging op de bovengenoemde punten accentueert tevens de allocatie- en selectiefunctie van eindtoetsen en examens.

Examens zijn met deze functies een belangrijke pijler van onderwijsstelsels. Het wijzigen van examens is daarmee een gewichtige kwestie waarvoor men een gedegen 'evidence based' onderbouwing zou wensen. Het is de vraag of bijvoorbeeld de recente voorstellen van de VO-raad tot flexibilisering en differentiatie van examens aan deze eis voldoen (zie ook de stellingname hierover van de Onderwijsraad, anno 2015).

De gedachte dat eindtoetsing en examens ook tot kwaliteitsbevordering kunnen leiden, is ingegeven door onderzoek waaruit bleek dat landen met een op standaarden gebaseerd examen beter scoren op internationale toetsen dan landen zonder examens. Hoewel deze uitkomsten niet helemaal onomstreden zijn (zo zijn er bijvoorbeeld landen die geen examen hebben maar wel hoog scoren, zoals Vlaanderen), zijn er wel redenen die dit aannemelijk maken. Als mechanismen worden genoemd:

- Het stimuleren van de extrinsieke motivatie van leerlingen, docenten en scholen, door belangen die in het spel zijn bij de uitkomsten.
- Het stimuleren van leerprocessen, op basis van instrumentele feedback, het aanduiden van sterke en zwakke punten in het functioneren.
- Het bieden van richting aan het onderwijs in het kader van 'curriculum alignment', in het bijzonder de goede aansluiting tussen nationale standaarden (eindtermen/referentieniveaus) enerzijds en de inhoud van examens en eindtoetsen anderzijds.

Het laatste punt, het gezichtspunt van 'curriculum alignment', wordt van speciaal belang geacht, omdat het de discussie over 'teaching to the test' (een van de meest gehoorde kritieken op examens en eindtoetsen) in een ander daglicht plaatst.
Maar laten we eerst de kritiek op examens en eindtoetsen nog even op een rijtje zetten.
De titel van het Advies van de Onderwijsraad uit 2013 'Een smalle kijk op onderwijskwaliteit' vat een belangrijke lijn van kritiek samen. Deze lijn van kritiek is niet alleen afkomstig van de Onderwijsraad, maar wordt ook aangetroffen in beleidsnota's van de VO-raad. In de terminologie van de hierboven besproken kernfuncties van onderwijs komt de kritiek erop neer dat er een te groot accent ligt op de kwalificatie-functie, en dat dit ten koste gaat van socialisering en persoonlijke ontwikkeling. Het vigerende stelsel van eindtoetsen en examens versterkt deze tendens, omdat kennis en vaardigheden (kwalificatie) beter meetbaar worden geacht, daardoor de examens sterk bepalen, waarmee op zijn beurt het onderwijs zich ook vooral op kennis en cognitieve vaardigheden zou richten. De Onderwijsraad vindt dat socialisering en persoonlijkheidsvorming meer aandacht moeten krijgen, maar pleit niet direct voor vermindering op cognitief vlak. Het Platform Onderwijs2032 doet dit wel, en wil het cognitieve curriculum afslanken. Verder zien alle genoemde instanties 'teaching to the test' als een groot probleem. Hierdoor krijgt de 'versmalling' van het onderwijs een extra stimulans. In Amerikaans onderzoek naar neveneffecten van strikt 'accountability' beleid, wordt een meer fundamentele analyse gegeven van de werking van 'teaching to test'. Uitgangspunt is dat toetsen en examens altijd beperkte en gebrekkige representaties zijn van wenselijke onderwijsopbrengsten, zoals aangegeven in onderwijsdoelstellingen. Vooral in situaties waarin er sprake is van grote consequenties van goede toetsuitkomsten zouden scholen een sterke prikkel ervaren om leerlingen te trainen voor het examen, ten koste van andere belangrijke onderwijsdoelen. Een tweede lijn van kritiek is dat er eigenlijk teveel en te rigide getoetst wordt. De Onderwijsinspectie gaf hier in oktober 2017 in een conferentie een bloemlezing van, waarbij veel aandacht was voor toetsstress bij leerlingen en leerkrachten die zich 'beklemd' voelen door de overheid.

Op de onderbouwing en de uitwerking van deze kritiek is veel aan te merken. Aangezien het niet alleen om vrijblijvende beschouwingen gaat, maar er ook toegewerkt wordt naar drastische veranderingen in het stelsel van examens en toetsen, is het van belang een 'fact check' te doen op de assumpties die worden gedaan en ook de basisconcepten kritisch te bekijken. Tornen aan examens moet per definitie gezien worden als een stelselherziening. Voor zover de afspraken over op evidentie gebaseerd beleid, gedaan door de Commissie Dijsselbloem nog gelden, is er alle aanleiding voor grondige empirische en analytische voorstudies. Enkele voorbeelden zijn: Hoe representatief zijn de negatieve houdingen tegenover toetsen die worden gesignaleerd werkelijk? Is er een feitelijke basis voor de aanname dat de huidige kernvakken zoveel tijd in beslag nemen dat er geen ruimte is om aan bredere kennis en vorming aandacht te geven? De meest fundamentele kritiek op het verschijnsel 'teaching to the test' heeft een technische basis, namelijk gebrekkige inhoudsvaliditeit; waarom lijkt bij voorbaat te worden uitgesloten dat er ook technische oplossingen zijn voor dit probleem? Meer voorbeelden worden verderop genoemd, waar deze studie uitmondt in een geheel van onderzoekschetsen, waarin deze ex ante analyses van examenherziening een plaats hebben.

Zoals gezegd wordt het gezichtspunt van 'curriculum alignment' als vruchtbaar gezien om een verdere constructieve wending te geven aan het debat over toetsen en examens. We gaan hier nu verder op in.

Het perspectief van 'curriculum alignment' kan als volgt puntsgewijs worden samengevat:

1) Onderwijs is doelgericht.
2) Onderwijsdoelstellingen kunnen in algemene termen worden weergegeven om hoofdrichtingen aan te geven, maar zijn nader te concretiseren en te operationaliseren.
3) Operationalisering van algemene onderwijsdoelstellingen kan in twee richtingen gaan: a) *didactische specificatie*, waarbij het aanbod aan leerinhouden en het proces van kennisoverdracht nader worden uitgewerkt en b) *evaluatieve specificatie*, waarbij onderwijsdoelstellingen worden gespecificeerd als na te streven leeropbrengsten.
4) De uitwerking op basis van didactische specificatie heeft als eindproduct een curriculum, de uitwerking van de evaluatieve specificatie heeft als eindproduct een geheel van registreerbare onderwijsopbrengsten, uitgewerkt in de vorm van eindtoetsen, examens of op andere wijze aantoonbare vooruitgang in kennis, vaardigheden en disposities van leerlingen.
5) Kerngedachte van 'alignment' of 'uitlijning' is consistentie tussen didactische en evaluatieve specificatie. Er wordt gesproken van 'horizontale alignment' als het gaat om de aansluiting van doelstellingen en opbrengstregistraties in de vorm van eindtoetsen en examens (in dit kader is de inhoudsvaliditeit van een eindtoets van cruciaal belang). De uitwerking van doelstellingen tot meer specifieke curriculumcomponenten (leergangen, leerboeken, studiemethoden, eventuele schoolwerkplannen, formatieve toetsen en gerealiseerde onderwijsleerprocessen) wordt 'verticale alignment' genoemd.

Volgens de denklijn van 'alignment' is het volstrekt logisch dat het onderwijs zoveel mogelijk wordt afgestemd op de doelstellingen, zowel algemene als geoperationaliseerde doelstellingen. Dit wordt alleen problematisch wanneer de eindtoets de doelstellingen niet goed afdekt. Bij een optimaal inhoudsvalide eindtoets kan toetsvoorbereiding als een legitieme vorm van 'teaching to the test' worden gezien.

De kritiek op examens en eindtoetsen vanuit de argumentatie van 'versmalling' van het curriculum en 'teaching to the test' wekt vaak de indruk dat de echt belangrijke doelen of functies van het onderwijs in de knel komen. Soms wordt er zelfs voor gepleit om de aansluiting tussen curriculumcomponenten juist kleiner te maken en bijvoorbeeld formatieve toetsen niet in lijn te brengen met de inhoud van eindtoetsen (vgl. Koretz, 2017). De mogelijkheid dat eindtoetsen een goede inhoudsvaliditeit hebben lijkt al bij voorbaat van de hand gewezen te worden (ibid). Een andere lijn van argumentatie tegen curriculum alignment is het benadrukken van autonomie; verticale alignment moet het vaak zonder regie of directe coördinatie doen en komt tot stand door afstemming tussen losjes gekoppelde organisaties. De vraag is of er sprake is van een patstelling, tussen enerzijds een theoretisch/technologisch ideaal (alignment) en anderzijds een ongeleid krachtenspel, waarin uiteindelijk modieuze invallen de dienst uitmaken.

In Nederland worden verschillende systemen gebruikt voor het beoordelen van de kwaliteit van toetsen en examens. In het hoofdstuk hierover wordt een overzicht van deze systemen gegeven en vervolgens gefocust op het RCEC-beoordelingssysteem voor de kwaliteit van studietoetsen en examens vanwege de geschiktheid van dit systeem voor toepassing binnen het voortgezet onderwijs. Een samenvattende beschrijving is opgenomen in Annex 2. Het systeem onderscheidt zes aandachtsvelden, die verwoord zijn in criteria: doel en gebruik, kwaliteit van toets- en examenmateriaal, representativiteit, betrouwbaarheid, standaardbepaling en normhandhaving, en tenslotte afname en beveiliging. Om elk van deze criteria te kunnen beoordelen biedt het systeem een aantal vragen die als onvoldoende,

voldoende of goed beoordeeld en gescoord kunnen worden. Per criterium resulteert het systeem tenslotte in een beoordeling. De centraal schriftelijk examens in het voortgezet onderwijs zijn de enige landelijke verplichte toetsen waarvan de kwaliteit niet door een externe instantie wordt beoordeeld. De vraag is hoe zij beoordeeld kunnen worden. Een tweede aandachtspunt is de representativiteit van toetsen en examens. Hoe speelt dit criterium een rol bij de alignment van toetsen en examens in het voortgezet onderwijs?

De toegenomen beschikbaarheid van ICT heeft tot verschillende innovaties op het gebied van toetsen en examens geleid in het voortgezet onderwijs. Allereerst biedt het digitaal afnemen van toetsen en examens de mogelijkheid om ze te personaliseren. Tijdens de toetsafname kan op basis van de vragen die al beantwoord zijn een inschatting gemaakt worden van het beheersingsniveau. Deze schatting kan vervolgens gebruikt worden om de moeilijkheid van de toets aan te passen aan het niveau van de leerling. Dit leidt tot kortere toetsen en voorkomt dat leerlingen gefrustreerd of verveeld raken door te moeilijke of te makkelijke vragen. Een tweede innovatie betreft de inhoud en vormgeving van het toetsmateriaal. Er kan gebruik gemaakt worden van digitale media of onlinebronnen, automatische feedback kan worden geïntegreerd en de psychometrische kwaliteit kan worden gemonitord. Een derde innovatie is dat authentieke toetstaken steeds vaker toegepast worden binnen het voortgezet onderwijs. De vierde innovatie is gerelateerd aan het gebruik van big data voor assessment. Al deze innovaties maken het mogelijk dat toetsen en examens efficiënter, aantrekkelijker en met een toegenomen validiteit afgenomen kunnen worden. Wel brengen ze vragen mee op het gebied van implementatie, beveiliging, technologieontwikkeling en bijvoorbeeld privacy.

In het afsluitende hoofdstuk wordt beargumenteerd dat een verandering van het programma van examens en eindtoetsen, die verder gaat dan een eenvoudig update, gezien moet worden als een stelselherziening en dus '*evidence based*' gefundeerd zou moeten zijn. In het inleidende deel van het hoofdstuk wordt uiteengezet dat aanpassing van examens en toetsen vanuit verschillende motieven overwogen wordt en dat het van groot belang is dat er vooraf tot helderheid gekomen wordt over de uiteindelijke ambities. Verder is aangegeven hoe de spanningsverhouding tussen standaardisatie en autonomie in het Nederlandse stelsel aanleiding geeft tot specifieke uitdagingen. Aan het eind van het hoofdstuk wordt een aantal gebieden genoemd, waarop nadere 'evidence based' onderbouwing van het beleid rondom examens en toetsing wenselijk wordt geacht. De rest van het hoofdstuk schetst de contouren van een voorbereidend onderzoeksprogramma, dat in staat is om de benodigde 'evidentie' voor nieuw beleid aan te leveren, of dit nu leidt tot consolidatie, een lichte bijstelling in de vorm van modernisering, of tot meer ingrijpende veranderingen.

In Annex 3 wordt gerapporteerd over een beperkte internationaal vergelijkende case-studie naar de toepassing van examens en toetsen in Italië, Zweden en Vlaanderen. In Zweden en Vlaanderen gebeuren examinering en toetsing onder regie van autonome onderwijsinstellingen. In Italië is een uitvoerig toetsprogramma opgezet, dat voornamelijk een formatieve functie heeft. Uit de case-studies komt naar voren dat het contrast tussen landen met en zonder een op standaarden gebaseerd examen, vooral betrekking heeft op het externe karakter van de examens. Zo heeft Vlaanderen geen extern centraal examen, maar wel een geheel van certificaten, die op basis van schoolinterne toetsing berusten. Verder is gedeeltelijk aannemelijk te maken dat curriculumstandaardisatie kan compenseren voor de outputstandaardisatie, die geboden wordt met externe centrale examens. Alles overziend lijkt de verklaring van het, door de bank genomen, beter functioneren van onderwijsstelsels met een

centraal op standaarden gebaseerd examen de accentuering van prestatiemotivatie en het bieden van curriculaire focus in het onderwijs.

Tot slot resumeren wij de hoofdlijnen van de inhoud van dit rapport en staan kort stil bij overeenkomsten en verschillen met het recente advies van de Onderwijsraad, 'Toets wijzer' (Onderwijs Raad, 2018)

1. Uitgangspunt is een geïntegreerd model van evaluatieve en didactische specificatie dat aanleiding geeft tot het met elkaar in verband zien van examen- en curriculumontwikkeling. Dit sluit goed aan bij de huidige context in Nederland, waar beide facetten ter discussie staan. Verder stimuleert dit om 'curriculum alignment' als een vruchtbaar perspectief te zien.
2. In het rapport staat het begrip onderwijskwaliteit centraal en wordt ingegaan op de kwaliteitborgende en kwaliteitverbeterende functies van examens en toetsen.
3. Een belangrijk aandachtspunt is de kritiek op de 'smalle kijk' op onderwijskwaliteit, zoals die duidelijk is verwoord in enkele recente adviezen van de Onderwijsraad. Het nader analyseren van de meetbaarheid en onderwijsbaarheid van sociaal-emotionele vaardigheden is een gemeenschappelijk thema voor zowel curriculum ontwikkeling als eventuele examenherziening.
4. In het rapport wordt aandacht besteed aan technische ontwikkelingen en innovaties rondom examens en toetsen en ook in de beoordeling van toetsen en examens; waar dit een aspect is dat in de vigerende discussies misschien te weinig is belicht.
5. Getracht wordt een aanzet te geven om te komen tot een beter inzicht in de motieven die ten grondslag liggen aan een ambivalente houding die in het onderwijsveld tegenover toetsen en examens bestaat.
6. Het rapport staat ook stil bij de bestuurlijk-organisatorische context van het Nederlandse onderwijs, waarbinnen een zekere spanning bestaat tussen autonomie en standaardisatie. Er wordt aangesloten bij onderzoeksresultaten die suggereren dat gegeven een hoge mate van 'proces autonomie' gecentraliseerde standaardisatie van output de effectiviteit verhoogt. In het rapport wordt aangesloten bij de visie van de OECD in recente beoordelingen van het Nederlandse systeem dat verscherping van nationale doelstellingskaders wenselijk is.
7. De centrale boodschap van dit rapport is dat een meer geprononceerde evidence-based aanpak bij de vigerende discussie over examens en eindtoetsen (maar ook bij de curriculumvernieuwing) aan te bevelen is. Het rapport mondt dan ook uit in enkele aanzetten tot een onderzoeksprogramma rondom examens en toetsing, in de context van curriculumvernieuwing.

In het Advies van de Onderwijsraad 'Toets wijzer; Naar een eigen(tijdse) wijze van toetsen en examineren' (december, 2018) wordt op een aantal punten op een overeenkomstige wijze gedacht over de examen- en toetsproblematiek als in dit rapport. Dit betreft:

Het benadrukken van het in lijn brengen van onderwijsdoelstellingen, leerlijnen, examens en toetsen
Letterlijk wordt gesteld**: "**Goed toetsbeleid begint in de ogen van de raad met de integratie van toetsing, onderwijsdoelen, onderwijsinhouden en onderwijsmiddelen" (p.36). De raad brengt dit specifiek naar voren in het kader van het toetsbeleid van onderwijsinstellingen, maar het is duidelijk dat deze gedachtegang ook wordt gevolgd in de analyse van het toetsbeleid op systeemniveau. Deze denklijn sluit geheel aan bij de centrale plaats die in dit rapport gegeven is aan curriculum alignment.

Het benadrukken van heldere kaders voor examinering en toetsing door de overheid
Op dit punt volgt de raad de OECD beoordelingen van het Nederlandse onderwijssysteem, die ook in dit rapport met instemming zijn aangehaald. In een zogenoemd briefadvies over curriculumvernieuwing

(Onderwijsraad, december 2018) stelt de raad **"**De overheid stelt als verantwoordelijke voor de kwaliteit van het onderwijsstelsel op landelijk niveau inhoudelijke kaders op". De raad benadrukt dat de overheid verantwoordelijk is voor de herijking van kerndoelen en eindtermen.

Het belang van standaardisatie in examens en eindtoetsen
De raad wijst op de voordelen van gestandaardiseerde eindtoetsing bij belangrijke overgangen in het stelsel (p. 29).

Ambiguïteit in het veld over de functie van toetsen, wat betreft formatief en summatief (de raad spreekt van beslissingsgericht) gebruik van toetsen
Op dit gebied worden overeenkomstige analyses weergegeven als in dit rapport.

Een standpunt van de raad dat centraal staat in het advies is dat enkele kernfacetten van toetsing en examens meer in evenwicht zouden moeten worden gebracht.
De raad onderkent spanning in drie dimensies:
- De mate waarin toetsing een beslissende of formatieve functie heeft;
- De mate waarin toetsing op decentraal niveau of (meer) centraal niveau wordt vormgegeven en
- De mate waarin wordt gestreefd naar kwantitatieve meting of naar het meer op kwalitatieve wijze zichtbaar maken van onderwijsopbrengsten.

Volgens de raad komt formatieve toetsing nog te weinig aan bod. Op de dimensie centraal-decentraal signaleert de raad zowel onder- als overbenutting. Tenslotte breekt de raad een lans voor het gebruik van kwalitatieve methoden.
In dit rapport ligt het accent op examens en eindtoetsen en komt formatieve evaluatie zijdelings ter sprake. Daarbij bestaat de indruk dat formatief *gebruik* van toetsen versterkt kan worden, maar dat er in Nederland bepaald geen gebrek is aan formatief te gebruiken toetsen (vgl. Oomen, Veldkamp en Scheerens, 2018). Wat de dimensie centraal-decentraal betreft, constateerden we dat in het voortgezet onderwijs er in feite maar één verplichte centrale toets is, nl. het eindexamen. Daarbij geldt dan nog voor het onderdeel schoolexamen dat scholen veel vrijheid hebben bij de invulling ervan. Over de verhouding tussen het gebruik van kwantitatieve en kwalitatieve methoden zijn wij van mening dat 'evenwicht' in de toepassing niet het belangrijkste criterium is, maar dat het meer gaat om passende toepassing, gegeven het doel en inhoud van de evaluatie en toetsing.

1 Examens en toetsen als onderdeel van evaluatie- en assessmentvoorzieningen in het Nederlandse onderwijs

Jaap Scheerens en Anne Luc van der Vegt

Toetsen en examens in het onderwijs staan ter discussie. Vanaf de toetsen die op kleuterleeftijd worden afgenomen tot en met het eindexamen in het voortgezet onderwijs. Wat moeten we toetsen, hoe vaak, op welke leeftijd? Aan de andere kant: we hebben in Nederland een lange toetstraditie, waarmee we een internationale reputatie hebben opgebouwd. Over de wenselijkheid van het toetsen zijn allerlei meningen en standpunten, voorstanders van standaardisering of flexibilisering, kwantitatief of kwalitatief toetsen, breed of smal, veel of weinig. In deze studie streven we er niet naar deze discussie uitvoerig weer te geven, maar om een beeld te geven van de actuele kennis over toetsen in het onderwijs. Daarbij gaat de aandacht vooral uit naar het voortgezet onderwijs en in het bijzonder naar het examen.

1.1. Historische ontwikkeling van het examen in Nederland

Met het eindexamen sluiten leerlingen hun opleiding in het voortgezet onderwijs (vmbo, havo of vwo) af. Het examen bestaat uit twee onderdelen: het schoolexamen (SE) en het centraal Examen (CE). Het schoolexamen heeft overigens niet altijd zo geheten. Tot 2007 spraken we het van het 'school-onderzoek'. Met de naam 'examen' wordt duidelijker weergegeven dat het als onderdeel van het eindexamen gelijkwaardig is aan het centraal examen. Voor de meeste vakken tellen beide onderdelen even zwaar mee bij het examencijfer en bepalen dus in gelijke mate of een leerling geslaagd is en een diploma ontvangt. Voor de vakken waarvoor zowel een SE als een CE wordt afgenomen, zijn beide onderdelen complementair.

Schoolexamen en Centraal Examen

Schoolexamens bestaan al veel langer dan centrale examens. Het schoolexamen vormde tot in de 19e eeuw de afsluiting van de Latijnse school. Het diploma van deze school gaf toegang tot de universiteit. Pas in 1845 werd voor het eerst het eindexamen op centraal niveau vastgesteld, als onafhankelijk door de staat afgenomen examen. Het staatsexamen overleefde aanvankelijk slechts vijf jaar. De liberaal Thorbecke vond dat niemand verhinderd mocht worden hoger onderwijs te volgen. Na de Wet op het Middelbaar Onderwijs van 1863, geïnitieerd door dezelfde Thorbecke, kwamen gecentraliseerde eindexamens toch weer terug. Eerst werden de examens per provincie georganiseerd, na 1920 landelijk. Met de Wet op het Voortgezet Onderwijs (de 'Mammoetwet' in 1968) werd het beroepsonderwijs gekoppeld aan het algemeen vormend onderwijs. De positie van de Inspectie was inmiddels prominent geworden en het was duidelijk dat de Inspectie het schriftelijk examen zou organiseren. Na 1970 schakelde de inspectie ook steeds meer het Cito (Centraal Instituut voor Toetsontwikkeling) in bij de eindexamens. Het Cito ondersteunde – aanvankelijk voor de talen en later voor alle vakken – de productie en de afname van de examens. Tegenwoordig worden de examens samengesteld door het Cito, onder verantwoordelijkheid van het College voor Toetsen en Examens (CvTE).
Deze ontwikkeling van het CE verliep niet zonder oppositie. Veel docenten geschiedenis waren bijvoorbeeld tegen een schriftelijk in plaats van een mondeling examen, omdat dit de vrijheid van de

school te zeer zou beperken. Uiteindelijk werd ook voor geschiedenis het CE verplicht, maar dat zou nog duren tot 1982 (Boom, 2003).
Het eindexamen moet in de eerste plaats gezien worden als een vorm van institutionele standaardisatie binnen het onderwijs, waarmee het civiel effect van onderwijs wordt gereguleerd en leerlingen een formele kwalificatie, in de vorm van een diploma, kunnen verwerven. Meer inhoudelijk gezien is een examen een vorm van evaluatie of *assessment,* waaraan vooral voor de leerlingen die het examen doen grote belangen verbonden zijn. Sinds de jaren tachtig van de vorige eeuw werden examenresultaten geleidelijk aan tevens gebruikt in het kader van de evaluatie van scholen door de Inspectie van het Onderwijs en informatievoorziening door de overheid. Vanaf die periode deden ook andere vormen van leerlingevaluatie hun intrede en kan het geheel aan voorzieningen van examens en toetsen worden gezien als een systeem dat meerdere functies vervult in het kader van beoordeling, sturing en kwaliteitsbevordering. Als beschrijvende basis voor deze studie wordt daarom dit hele systeem van voorzieningen als uitgangspunt genomen. De contouren worden in de volgende paragraaf weergegeven.

De infrastructuur voor evaluatie en toetsing in Nederland
Het onderwijsbeleid van Minister Van Kemenade was een belangrijke fase voor de ontwikkeling van empirisch evaluatieonderzoek en de toetsing van leerresultaten in de jaren zeventig. Het ‘constructieve onderwijsbeleid’ voorzag in centraal ontwikkelde landelijke innovatieprojecten die ‘experimenten’ werden genoemd. Het was de bedoeling dat de vernieuwingsideeën over de ‘middenschool’, verandering van het basisonderwijs, het zogenoemde ‘participatie onderwijs’ en het Open School proefproject als ‘pilots’ ontwikkeld werden en dat er op basis van evaluatieonderzoek zou worden besloten om al dan niet tot grootschalige implementatie over te gaan. Om meerdere redenen kwamen de projecten moeizaam van de grond (vgl. Scheerens, 1983); het empirische evaluatieonderzoek mislukte grotendeels, omdat het een speelbal werd van een machtsstrijd tussen allerlei partijen en met name te lijden had van gebrek aan medewerking van de betrokken onderwijsinstellingen. Toch bleef beleidsgericht evaluatieonderzoek sindsdien een rol spelen en ontwikkelde het zich tot het uitgebreid geheel van activiteiten dat nu bestaat (Scheerens, 2013, OECD, 2014).
Gedurende de jaren tachtig kreeg het onderwijsbeleid een meer incrementeel karakter, waarbij zich geleidelijk aan een evaluatiepraktijk vormde die, in plaats van het model van programma-evaluatie, gebaseerd was op constante voorziening van gegevens en ‘monitoring’. In deze periode kwamen belangrijke instrumenten, zoals het Periodiek Peilingsonderzoek (PPON), de nationale cohort studies in het primair en secundair onderwijs, beleidsrelevante onderwijsstatistieken en indicatoren tot stand. Deze ontwikkelingen werden gestimuleerd door actieve deelname van Nederland in het OECD indicatoren project (INES) en initiatieven van de EU, met name EURYDICE.
Hoewel de eerste toepassingen vooral op systeemniveau plaatsvonden, werd dit gevolgd door ontwikkelingen in schoolevaluatie, vooral in de vorm van een vernieuwd inspectietoezicht. De ruggengraat van deze toepassingen op het niveau van het nationale systeem en het schoolniveau waren prestatiegegevens die op het niveau van individuele leerlingen werden verzameld; namelijk eindexamenresultaten en resultaten van de Cito-toets aan het eind van het basisonderwijs. Ook in meer recente toepassingen van evaluatie, gekoppeld aan nieuwere inzichten over ‘school governance’, zoals risicogericht schooltoezicht en ‘Scholen op de kaart’ behouden toetsgegevens een belangrijke plaats. Dit werd mede gestimuleerd door beleidsinitiatieven op het gebied van de stimulering van onderwijskwaliteit (de Kwaliteitsagenda’s) en van ‘opbrengstgericht werken’. Leerlingvolgsystemen werden verplicht gesteld en er is groeiende belangstelling voor formatief toetsen (Vgl. Oomen, Veldkamp en Scheerens, 2015). In tabel 1.1 (Scheerens, 2013) worden de belangrijkste verplichte vormen van leerlingevaluatie nog eens systematisch samengevat.

Tabel 1.1: Overzicht van verplichte examens en toetsen (naar Scheerens, 2013)

Type evaluatie	*Korte beschrijving*	*Formele verantwoordelijkheid*	*Implementatie en gebruik*
Examens	Formele beoordeling aan het einde van het voortgezet onderwijs. Doel is individuele certificering.	Het ministerie van OCW, verantwoordelijk gedelegeerd aan CvTE. Scholen zijn, onder toezicht/monitoring door de Inspectie van het Onderwijs, verantwoordelijk voor interne school examens.	Verplichte implementatie. Gebruik en toepassing zijn eenvoudig.
Eindtoets basisonderwijs	Wordt afgenomen aan het einde van het primair onderwijs. Meet de kennis van taal en rekenen en geeft aan welk type vervolgonderwijs bij een leerling past.	Scholen zijn verantwoordelijk voor deelname. Toetsontwikkelaars verzorgen de technische aspecten.	De test wordt gebruikt als extra gegeven naast het schooladvies. Op basis van de eindtoets kan schooladvies worden bijgesteld. Op geaggregeerd niveau geven de testen input aan evaluaties op school- en systeemniveau.
Cito LVS	Een leerlingvolgsysteem voor primair onderwijs, voor alle niveaus en vakken.	Scholen zijn verantwoordelijk voor deelname. Zij kopen het systeem. Cito is verantwoordelijk voor de technische aspecten.	De testen worden gebruikt voor didactische diagnose en om de voortgang van leerlingen te volgen. Daarnaast bieden de geaggregeerde data input voor zelf-evaluaties op schoolniveau.

Het is van belang te onderstrepen dat examens en eindtoetsen (met name de eindtoets basisonderwijs) een sleutelpositie hebben in het totale systeem van evaluatie en toetsing. In de eerste plaats gaat het om de evaluatie van de prestaties van individuele leerlingen, maar daarnaast worden toets- en examengegevens geaggregeerd tot prestatie-indicatoren van scholen en zelfs nationale onderwijsprestaties. Een specifieke toepassing is het openbaar maken van prestatiegegevens van scholen in ranglijsten. In andere landen worden toetsgegevens ook gebruikt om leerkrachten te beoordelen (vgl. bijvoorbeeld de bekende MET studie van de Bill and Melinda Gates Foundations (Kahn et al. 2014).

In de reviewstudie van de OECD (Nusche et al., 2014) wordt het geheel van voorzieningen voor evaluatie en toetsing in Nederland als bijzonder sterk gewaardeerd: “The Dutch evaluation and assessment approach stands out internationally as striking a good balance between school-based and central elements, quantitative and qualitative approaches, improvement and accountability functions and vertical and horizontal responsibilities of schools” (p.13). Het belangrijkste punt van kritiek is dat de structuur van onderwijsdoelstellingen versterkt zou moeten worden om een goede aansluiting voor de verschillende toepassingen van evaluatie en toetsing mogelijk te maken.

In de volgende hoofdstukken zullen we kritische vragen over het stelsel van examinering en toetsing vanuit verschillende gezichtspunten belichten.

2 Examens en toetsing tegen de achtergrond van maatschappelijke functies van onderwijs

Jaap Scheerens

Er bestaat nogal wat variëteit in de manier waarop basisfuncties van het onderwijs worden gedefinieerd. We zien dit in de publicaties van de Onderwijsraad, waar het thema onderwijskwaliteit doorgaans wordt ingeleid door uit te gaan van maatschappelijke functies waaraan het onderwijs moet voldoen. In de diverse adviezen wordt de indeling in kernfuncties niet helemaal consistent gebruikt, maar wellicht moet dit gezien worden als een voorbeeld van voortschrijdend inzicht door de jaren heen.

Ook in de voornamelijk sociologische literatuur waarin de maatschappelijke functies van onderwijs behandeld worden, bestaat nogal wat variëteit in de manier waarop basisfuncties worden gedefinieerd, en hoe de onderlinge relaties worden gezien. Peschar en Wesselingh, (1985) onderscheiden als kernfuncties de kwalificatie-, de selectie- en de allocatiefunctie.

De *kwalificatiefunctie* wijst op de betekenis van het onderwijs voor de toerusting van studenten voor vervolgonderwijs en de arbeidsmarkt. Het onderwijs dient op de juiste vaardigheden en competenties gericht te zijn, en leerlingen moeten deze vaardigheden en competenties ook daadwerkelijk beheersen.

De *selectiefunctie* heeft betrekking op het toewijzen van leerlingen aan de leerwegen en schoolsoorten die bij hen passen, zodat leerlingen op hun eigen niveau een diploma kunnen verwerven. De selectiefunctie impliceert dat leerlingen verschillende vaardigheidsniveaus hebben en dat het onderwijs een zodanig gelede structuur heeft dat hierop 'gesorteerd' kan worden. Een vraag hierbij is of gelede structuren, oftewel categorale onderwijsstelsels, in het voortgezet onderwijs wel optimaal zijn om zo hoog mogelijke gemiddelde prestatieniveaus te bereiken, en of deze stelsels niet juist afkomst-gerelateerde prestaties en 'ongelijkheid' stimuleren.

De *allocatiefunctie* is te zien als een combinatie van kwalificatie en selectie, dat wil zeggen een zodanige differentiatie in vaardigheidsniveaus dat maatschappelijke taken en rollen zo goed mogelijk bediend worden. Little (2014) presenteert het onderstaande overzicht (Tabel 2.1).

Tabel 2.1 Manifeste en latente functies van onderwijs. Volgens de functionalistische theorie heeft onderwijs duidelijke, zichtbare en ook latente functies.[1]

Manifeste functies: openbaar aangehaalde functies met geplande doelen	Latente, minder zichtbare functies: verborgen, niet benoemde functies met soms ongeplande doelen
Socialisatie	Ontstaan van relaties tussen leerlingen/studenten
Overdragen van cultuur	Sociale netwerken
Sociale controle	Werken in groepen
Sociale plaatsing	Ontwikkeling van generatiekloof
Culturele innovatie	Politieke en sociale integratie

1 Vertaling door de auteurs.

In dit overzicht komt 'sociale plaatsing' overeen met allocatie, en de selectie van leerlingen op grond van academische verdienste en potentieel. Tegelijkertijd valt dit samen met kwalificatie, waarbij op basis van toetsing van leerprestaties de meest capabele leerlingen worden geïdentificeerd.

Van de Werfhorst formuleert vier centrale doelstellingen van het onderwijs (Van de Werfhorst e.a., 2011; Van de Werfhorst & Mijs 2010):[2]

1. Bevorderen van gelijke kansen voor kinderen van verschillende achtergronden (gelijke-kansenfunctie).
2. Het plaatsen van leerlingen in een schooltype gebaseerd op talenten en interesses (selectiefunctie). De selectiefunctie gaat ervan uit dat efficiënt leren wordt bereikt bij een optimaal selectieproces. De 'totale' vaardigheden- en kennisproductie is dan geoptimaliseerd (bij een gegeven budget voor onderwijs).
3. Voorbereiden op de arbeidsmarkt (toewijzingsfunctie). Deze functie gaat ervan uit dat onderwijs vaardigheden leert die van belang zijn voor werk en daarmee schoolverlaters helpt in het proces om de juiste plek te vinden (toewijzing) op de arbeidsmarkt. Tegelijkertijd helpt het werkgevers in het optimaliseren van hun productie.
4. Trainen van leerlingen en studenten in een actief burgerschap (socialisatie functie). Onderwijs kan een actieve rol spelen in het ontwikkelen van actieve en betrokken burgers. Daarnaast kan onderwijs gelijkwaardigheid in burgerschapsvaardigheden bevorderen (iets wat van andere partijen, zoals ouders, niet verwacht kan worden).

In tabel 2.2 wordt aangesloten bij de indeling die gebruikt is in het advies van de Onderwijsraad uit 2007, getiteld 'Sturen van vernieuwende onderwijspraktijken'. De selectie- en allocatiefuncties zijn hier samengenomen, als twee kanten van dezelfde medaille: het onderwijst selecteert en sorteert op een wijze die relevant is voor maatschappelijke rollen en posities. Persoonsvorming is toegevoegd op basis van de indeling die de Onderwijsraad in 2016 gebruikt in het rapport 'Een ander perspectief op professionele ruimte in het onderwijs'.

De toevoeging van 'persoonsvorming' in het rapport van de Onderwijsraad, getiteld 'De volle breedte van onderwijskwaliteit' (gepubliceerd in 2016) is opmerkelijk. Waar de functie-indeling een van de kernstukken is uit de onderwijssociologie en ook op onderdelen met de onderwijseconomie is te associëren (efficiëntie, sociale cohesie), is er opeens een pedagogische functie bijgekomen. In het rapport wordt dit als volgt verwoord:
"Het moderne pedagogische denken betoogt (…) dat socialisatie in onderwijs en opvoeding nooit voldoende is: de mens is niet alleen maar een product van een bepaalde cultuur, traditie of praktijk, maar verhoudt zich ook altijd kritisch tot die cultuur, traditie of praktijk. Dit betekent dat mensen verantwoordelijkheid moeten kunnen en willen nemen voor de mate waarin ze hun leven vorm willen geven binnen bestaande tradities en praktijken. Onderwijs dient daaraan bij te dragen en dient leerlingen daarvoor handvatten aan te dragen. Dat is wat de raad onder persoonsvorming verstaat" (Onderwijsraad, 2016, p.22).

[2] Vertaling door de auteurs

Tabel 2.2 Maatschappelijke functies van onderwijs

	ONDERWIJSRAAD, 2007	ONDERWIJSRAAD, 2016	WIKIPEDIA
KWALIFICATIE	Bij de kwalificatiefunctie gaat het om de vraag of het onderwijs de leerlingen en studenten uitrust met kennis, vaardigheden en houdingen die relevant zijn voor de arbeidsmarkt. Een breed spectrum van kennis, vaardigheden en houdingen.	Accent op meetbare cognitieve opbrengsten moet verbreed worden met bijvoorbeeld techniekonderwijs, cultuureducatie en burgerschapskunde.	Efficiëntiefunctie.
SELECTIE/ ALLOCATIE	Onderwijs bereidt mensen voor op verschillende posities in de samenleving. De school functioneert als een 'sorteer-machine' die verschillende categorieën leerlingen toewijst aan posities in de samenleving (selectie en kwalificatie zijn nauw verbonden thema's).		Arbeidsallocatiefunctie.
SOCIALISATIE	Onderwijs zorgt ervoor dat naast kennis en vaardigheden ook algemeen geldende normen en waarden worden overgedragen die mensen voor het functioneren in de samenleving nodig hebben.	Socialisatie – als iets dat bewust door onderwijs wordt nagestreefd – heeft te maken met de manier waarop leerlingen deel worden van bestaande tradities en praktijken. Socialisatie verschaft leerlingen daarmee een identiteit (p 22).	Gelijke kansen functie. Actieve participatie functie, sociale cohesie, burgerzin.
PERSOONS-VORMING		Persoonlijkheidsontwikkeling. 'character building', kritische houding[3]	

Waarschijnlijk zijn deze passages uit dit advies van de Onderwijsraad gebaseerd op het werk van Biesta, die in het debat over onderwijskwaliteit een pedagogische optiek toevoegt. Dit komt het meest tot uitdrukking in de door hem gegeven invulling van persoonsvorming, waarbij het niet eenvoudig is om te begrijpen wat daar precies mee bedoeld wordt. Uit het citaat van de Onderwijsraad moet worden opgemaakt dat er sprake is van het opvoeden tot een kritische houding. Maar er zit een optiek achter die ook andere facetten heeft: een soort fenomenologische visie op ontwikkeling, zinnen als: "Kwaliteit in brede zin ontstaat in de onderwijsprocessen en 'ontvouwt' zich in de onderwijspraktijk tussen leerkracht en leerling". Verder is er sprake van 'subjectivering' en 'in de wereld komen' (Biesta, 2012). Sandahl (2015) omschrijft Biesta's 'subjectivering' als volgt: "Biesta uses the term subjectification (derived from the German word Subjektivität) but stressed that it is a 'bit of a struggle to find the right concept' in English (Biesta 2012:13). The meaning of Biesta's concept is about the emancipation of students as humans and about providing them with agency as citizens". Andere aspecten van de

[3] Zie toelichting in de tekst

pedagogische inbreng zijn een kritische opstelling tegen toetsen en meten van onderwijsopbrengsten en een voorkeur voor het beschrijven en begrijpen van onderwijsprocessen.

De functionalistische invalshoek wordt in de rapporten van de Onderwijsraad gebruikt als een van de invalshoeken om te pleiten voor een brede visie op kwaliteit. In termen van de vier kernfuncties die gebruikt zijn in tabel 2.2, betekent dit dat men zich uitdrukkelijk niet wil beperken tot de kwalificatie-functie. In meer operationele termen wil dit zeggen dat bij kwaliteit niet alleen gekeken moet worden naar cognitieve opbrengsten, maar ook naar sociale vaardigheden en, nog nader te bepalen, persoonlijke ontwikkeling. De pedagogische inbreng zet verder vraagtekens achter een empirisch analytische benadering van onderwijsontwerp en onderwijsevaluatie en brengt de plaats van normatieve aspecten en subjectiviteit bij het beoordelen van kwaliteit naar voren.
De functionalistische categorisering van kernfuncties is geschikt voor een brede oriëntatie op de vraag waar het onderwijs toe moet dienen. Door auteurs wordt vaak gewezen op een zekere overlap tussen de onderscheiden functies en uit de gegeven citaten blijkt dan dat het soms arbitrair lijkt om bepaalde functies neer te zetten als bijvoorbeeld kwalificatie of socialisatie. Selectie en allocatie lijken twee kanten van dezelfde medaille en het zou denkbaar zijn om persoonlijke ontwikkeling als een bijzondere vorm van kwalificatie te zien. De kern van het verhaal is dat onderwijs doelgericht is en dat doelbereik, in de zin van het behalen van onderwijsopbrengsten, op verschillende manieren kan worden uitgedrukt: als cognitieve, sociale en affectieve kwalificaties van alle leerlingen en eventueel gestratificeerd naar typen leerlingen, zodat ook conclusies over gelijke kansen mogelijk zijn. Vanuit dit gezichtspunt is een andere conceptuele basis te geven voor het kwaliteitsdebat in het onderwijs en de implicaties hiervan voor het uitwerken van de evaluatiefunctie, die examinering en toetsing omvat. Bovendien maakt deze conceptuele basis het mogelijk om facetten van kwaliteit in meer operationele termen te bespreken.

3 Meetbare facetten van kwaliteit[4]

Jaap Scheerens

3.1. Inleiding

Over kwaliteit kan men interessante filosofische beschouwingen houden (zie bijvoorbeeld het indrukwekkende boek van Robert Pirsig (1999), 'Zen of de Kunst van het Motoronderhoud').Ook komt het voor dat kwaliteit van onderwijs als vrijwel ondefinieerbaar wordt voorgesteld of in ieder geval als iets dat zodanig complex en esoterisch is dat men er een *Kenner* voor nodig heeft om er een oordeel over te kunnen geven (vergelijk het idee van Elliott Eisner, van *Educational Connoisseurship*). In dit hoofdstuk wordt een analytisch kader gepresenteerd dat laat zien dat het goed mogelijk is om onderwijskwaliteit helder te definiëren, en dat de belangrijkste facetten ervan tevens meetbaar zijn. Verschillende perspectieven op onderwijskwaliteit kunnen op basis van dit kader nader geplaatst en onderling vergeleken worden.

3.2. Anatomie van de onderwijskwaliteit

Kwaliteit zou men, in de meest algemene zin, en vrij naar Van Dale, kunnen omschrijven als een hoedanigheid waaraan een positieve waardering wordt gegeven. Drie vragen moeten vervolgens gesteld worden om dit wat verder te concretiseren en toe te passen op het onderwijs:

a) Kwaliteit van wat? (wat is het object waarvan de hoedanigheid gewaardeerd wordt);
b) Kwaliteit zoals beoordeeld door wie? (welke actor is gerechtigd het kwaliteitsoordeel te vellen);
c) Hoe is 'een positieve waardering' nader te typeren? (op welke criteria beoordeelt men kwaliteit).

Kwaliteit van wat?

Hier kan een onderscheid naar organisatorisch niveau worden gemaakt. Het ligt voor de hand om in het Nederlandse bestel, gekenmerkt door grote autonomie van scholen, de school of onderwijsinstelling als geheel, dat wil zeggen de schoolorganisatie qua opbouw en functioneren, als een belangrijke focus van kwaliteitsbeoordeling te zien. Tegelijkertijd blijven vanzelfsprekend vragen over die gesteld kunnen worden over de kwaliteit van het nationale onderwijsstelsel als geheel of belangrijke deelprogramma's daarbinnen. Tenslotte kan stil worden gestaan bij de kwaliteitsbeoordeling binnen de school, bijvoorbeeld de leerkrachten.

Beoordeeld door wie?

De vraag 'kwaliteit voor wie?' wordt doorgaans beantwoord door de directe afnemers van het product of de dienst in kwestie als de beoordelende actor te beschouwen. In het geval van scholen hebben we het dan over leerlingen en ouders, maar zeker ook over verder verwijderde afnemers zoals het

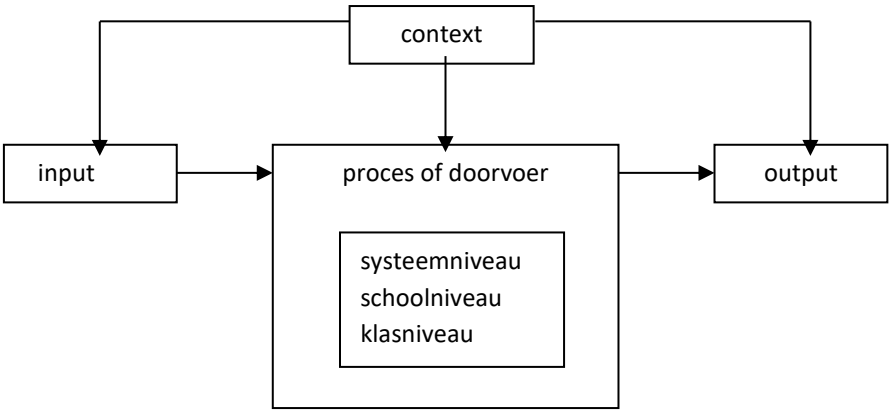
[4] Dit hoofdstuk is een bewerking en update van Hoofdstuk 1 uit: Scheerens, J. Luyten, H., and Van Ravens, J. (2011) *Perspectives on educational quality. Illustrative outcomes on primary and secondary schooling in the Netherlands.* Dordrecht, Heidelberg, New-York, London: Springer.

vervolgonderwijs, werkgevers en eventueel de maatschappij in z'n volle breedte. Toch denken we bij de beoordeling van de kwaliteit van het onderwijs doorgaans in de eerste plaats aan de Inspectie van het Onderwijs, die namens de overheid toezicht houdt op de kwaliteit.

Hoe wordt kwaliteit beoordeeld?
In de volgende paragraaf wordt een analytisch kader gepresenteerd dat vooral een nadere uitwerking is van verschillende criteria of waarderingsdimensies (aspect c). Door het model op meerdere niveaus en geledingen van het onderwijs toe te passen wordt tevens aspect a verdisconteerd. Verderop zal afzonderlijk worden ingegaan op actorperspectieven bij het centraal stellen van waarderingsdimensies (aspect b).

3.3. Analytisch kader

Het analytisch kader is gebaseerd op een systeemmodel van het onderwijs, zie figuur 3.1:



Figuur 3.1: Een eenvoudig systeemmodel van het functioneren van onderwijs

Volgens dit model kan het onderwijs worden weergegeven als een transformatie of 'productieproces', waarin invoer (*input*), door middel van een centraal productieproces (proces, doorvoer of *throughput*) resulteert in bepaalde uitkomsten of opbrengsten (*output*). Dit alles vindt plaats in een omgeving of context, die leverancier is van invoer en die eisen stelt aan de uitkomsten of opbrengsten (*context)*. De centrale 'zwarte doos' in dit model kan worden gedefinieerd op het niveau van het nationale onderwijsbestel, de school of de klas, c.q. leergroep, en zelfs op het niveau van de individuele leerling.

De beoordeling van de kwaliteit van het onderwijs kan betrekking hebben op verschillende onderdelen van dit analytisch kader. De Inspectie van het Onderwijs kijkt naar zowel het proces als de output van het onderwijs, en baseert daarop zijn oordeel over de kwaliteit. Stakeholders zoals de ouders kijken ook naar de input: leerkrachten, schoolgebouw en leerlingpopulatie.

Het is bekend dat tegen de gebruikte terminologie in dit model in onderwijskundig/pedagogische kring bezwaren bestaan. Dat 'de leerling' in het model wordt geplaatst als onderdeel van de invoer of als ruw onbewerkt materiaal stuit op weerstand. Toch kan men zich dit voorstellen als leerlingen die de school binnenkomen met bepaalde karakteristieken en er na enige tijd weer uitstromen met op bepaalde

gebieden veranderde karakteristieken. Eenvoudiger gezegd: aan het begin van het proces weten ze wat minder, aan het eind wat meer.
Gevoelige kwesties in deze zijn gedachten over maakbaarheid, een technocratische of mechanistische inslag en reductionisme in de betekenis van een overdreven focus op meetbare opbrengsten. Het algemene en abstracte karakter van het kader biedt echter ruimte voor brede interpretatie. Ook binnen dit model is het geen probleem te erkennen dat leerlingen eerder als primaire producent dan als consument of passieve grondstof van kennis worden gezien, te erkennen dat er veel onzekerheid bestaat over de effectiviteit van processen op het terrein van onderwijsbestuur en organisatie en lesgeven en onder ogen te zien dat moeilijk meetbare gewenste opbrengsten niet principieel worden uitgesloten.
Op basis van dit analytische kader is het mogelijk om belangrijke facetten van onderwijskwaliteit nader te preciseren. Dat gebeurt door het accentueren van bepaalde onderdelen van het model – input, proces, output en context – en door het bezien van specifieke relaties tussen de onderdelen.

a) Productiviteit (accent op output)
Volgens dit gezichtspunt wordt het succes van het systeem als geheel bepaald door het opleveren van de gewenste opbrengsten of uitkomsten. Bijvoorbeeld in de zin van een bevredigend percentage schoolverlaters dat, zonder studievertraging, het diploma behaalt, of in de zin van een voldoende gering niveau van werkloosheid onder degenen die een bepaald diploma behaald hebben. Een derde voorbeeld is een voldoende geachte gemiddelde score van een land op een internationaal vergelijkende studietoets (vergelijk de resultaten van projecten als PISA en TIMSS). In het Engels wordt onderscheid gemaakt tussen *achievement outputs (*prestaties op studietoetsen)*, attainment outcomes* (indicatoren die te maken hebben met gerealiseerd onderwijsniveau en numeriek rendement van opleidingen) en *impact* indicatoren (die effecten op de verdere school- of maatschappelijk loopbaan aangeven).

Opgemerkt moet worden dat het model open laat welke opbrengstindicatoren gekozen worden. Vooral als het gaat om de leerresultaten die aan het eind van een opleiding gemeten worden, zijn er vele mogelijkheden. Doorgaans ligt het accent op cognitieve prestaties in basisvakken als rekenen, taal en lezen. Het is echter ook denkbaar dat er een ruimer geheel aan leerresultaten wordt gemeten, dat er meer nadruk ligt op algemene cognitieve vaardigheden dan op vakgebonden kennis en dat aan attituden grenzende concepten als bijvoorbeeld ‘burgerschap’ en sociale competenties worden gekozen.

b) Effectiviteit (causale relatie tussen input, processen en context enerzijds, en output anderzijds)
Effectiviteit betekent doeltreffendheid. Binnen het effectiviteitsperspectief staat de vraag centraal welke context, input en proces factoren doeltreffend zijn, dat wil zeggen een positieve samenhang vertonen met opbrengsten, in de zin van in een bepaalde mate gerealiseerde onderwijsdoelstellingen. Kennis over onderwijseffectiviteit is van grote beleidsmatige en praktische betekenis, omdat er informatie geboden wordt over in principe beheersbare factoren die tot betere onderwijsopbrengsten leiden. Men zou kunnen volstaan met het meten van input- en procesindicatoren als volledig bekend zou zijn welke processen tot maximale doelbereiking leiden. Dit is in feite het ideaal van *total quality management*, dat verderop behandeld wordt. In werkelijkheid is er slechts beperkte kennis over dit soort instrumentele relaties, maar dit neemt niet weg dat het een verstandige aanpak is om zich te richten op die beheersbare factoren die, zoals blijkend uit onderzoek, doorgaans een positieve associatie met onderwijsopbrengsten hebben.

c) Onderwijs(on)gelijkheid (accent op context: de verdeling van middelen, goede onderwijspraktijk en onderwijsopbrengsten)

Hier gaat het om een eerlijke verdeling van middelen, goed onderwijs en onderwijsopbrengsten tussen verschillende groepen, bijvoorbeeld leerlingen uit milieus met verschillende sociaal-economische, of sociaal-culturele status, jongens en meisjes, en autochtone en allochtone leerlingen. Ideaal is dat alle groepen in gelijke mate van het onderwijs kunnen profiteren, en dat verschillen in opbrengsten niet, of in zo gering mogelijke mate, zijn toe te schrijven aan dit soort achtergrondkenmerken van leerlingen. Structuurkenmerken van onderwijsstelsels, zoals de leeftijd waarop gekozen moet worden voor een bepaald soort vervolgonderwijs, en de verscheidenheid van schoolsoorten in het voortgezet onderwijs blijken een behoorlijke invloed te hebben op de selectiviteit van het onderwijs en op de clustering van minder bevoorrechte groepen in de lagere schooltypen (zie verder hoofdstuk 4).

d) Efficiency (effectiviteit tegen de laagst mogelijke kosten)

Efficiency of doelmatigheid gaat om de economische meest voordelige keuze van op zich doeltreffende middelen om relatief hoge opbrengsten te verkrijgen. Meer doelmatigheid kan worden bereikt door met dezelfde middelen meer opbrengsten te behalen of door met minder middelen dezelfde opbrengst-niveaus te realiseren. Ter illustratie: klassenverkleining heeft doorgaans een geringer effect dan verbetering van onderwijsleermateriaal, maar is ook nog eens veel duurder. Reductie van de gemiddelde klassengrootte met enkele leerlingen heeft een verwaarloosbaar effect op de leeropbrengsten, maar levert flinke bezuinigingen op.

e) Het aanpassingsperspectief (accent op context: gewenste opbrengsten, c.q. doelstellingen, aanpassen aan de vragen vanuit de omgeving)

Dit perspectief gaat minder dan de eerdere behandelde gezichtspunten uit van gegeven doelstellingen, maar gaat juist over de vraag welke prioriteiten en doelstellingen het onderwijs zou moeten kiezen, gegeven een bredere sociaal, cultureel en economische maatschappelijke context. Voor een deel kan dit inderdaad in termen van adaptatie en aanpassing gezien worden, waarbij het onderwijs geacht wordt om functioneel te zijn ten opzichte van ruimere maatschappelijke ontwikkelingen. Het economisch perspectief en de kenniseconomie zijn hierbij zeer nadrukkelijk aanwezig. Vergelijk ook recente publicaties waarbij effecten van verbetering van leerprestaties en numerieke rendementen in economische groei op de lange termijn worden uitgedrukt (vgl. Hanushek en Woessmann, 2005, 2009). Andere maatschappelijke impulsen stellen juist eisen aan sociale ontwikkeling en internationaal burgerschap. Op dit niveau van analyse zijn allerlei visies op de relatie tussen onderwijs en de ruimere maatschappelijk context aan de orde, bijvoorbeeld de vraag hoe nauwgezet men concrete eisen van de arbeidsmarkt moet zien te vertalen in ‘educational objectives’, de keuze voor een gerichtheid op meer algemene disposities en ‘competenties’, alsmede het belang van kennis en vaardigheden die minder direct op economische utiliteit zijn gericht. Vanuit bepaalde gezichtspunten wordt ook de eigen verantwoordelijkheid van het onderwijs als een bron voor maatschappijkritiek, creativiteit en innovatie onderstreept. Visies dus, waarvoor de term aanpassingsperspectief minder op zijn plaats is (zie verderop de discussie over alternatieve visies op kwaliteit).

Als het gaat om het meetbaar maken van condities die te maken hebben met het aanpassings-perspectief kan gedacht worden aan nationale voorzieningen die maatschappelijke eisen aan onderwijs continu peilen (bijvoorbeeld onderzoek naar de relatie onderwijs-arbeidsmarkt) en op schoolniveau aan activiteiten als schoolmarketing en contacten met lokale stakeholders.

f) Een gefragmentariseerde benadering van onderwijskwaliteit (afzonderlijke onderdelen van het model – input, proces en output – worden op zichzelf staand op hun kwaliteit beoordeeld)

De eerst in aanmerking komende kandidaat voor een dergelijke ontkoppelde aanpak van onderdelen uit het model is het productiviteitperspectief, waarbij gekeken wordt naar het niveau van de onderwijsopbrengsten, zonder dat er verbanden worden gelegd met input- of procescondities. Andere voorbeelden zijn: het beoordelen van de kwaliteit van scholen op basis van niet alleen opbrengstcriteria, maar tevens procescriteria. In het werk van de Inspectie van het Onderwijs zijn hier voorbeelden van, het beoordelen van onderwijsbeleid aan de hand van de gedane investeringen, het beoordelen van de kwaliteit van leerkrachten en het beoordelen van kwaliteit in de context van certificering op basis van het naleven van een aantal procedures. Als we kijken naar uiteenlopende praktijken, zoals het gebruik van onderwijsindicatoren, kwaliteitscertificering, Inspectie en de schoolkeuze van ouders, dan moet geconstateerd worden dat de gefragmentariseerde benadering sterk de overhand heeft. Het relateren van processen, input en output (het effectiviteitperspectief) is niet alleen onderzoekstechnisch gecompliceerd, maar communicatief moeilijk over te brengen. De gefragmentariseerde benadering van onderwijskwaliteit heeft het voordeel van de betrekkelijke eenvoud, maar het nadeel van de betrekkelijke willekeur, wanneer andere dan opbrengstindicatoren worden gebruikt.

3.4. Meetbare facetten van kwaliteit in relatie tot maatschappelijke functies van het onderwijs

Zoals in hoofdstuk 2 beschreven wordt in de onderwijssociologie de kwaliteit van het onderwijs getypeerd door na te gaan of het in de maatschappij bepaalde kernfuncties vervult (Peschar en Wesselingh, 1985). Deze kernfuncties zijn de kwalificatie-, de selectie- en de allocatiefunctie.

De kwalificatiefunctie wijst op de betekenis van het onderwijs voor de toerusting van studenten voor vervolgonderwijs en de arbeidsmarkt en is dus in verband te brengen met wat in de vorige paragraaf is aangeduid als de adaptiviteit van het onderwijs (de juiste doelen kiezen) en de productiviteit (het realiseren van die doelstellingen).

De selectiefunctie heeft betrekking op het toewijzen van leerlingen aan de leerwegen en schoolsoorten die bij hun passen, zodat leerlingen op hun eigen niveau een diploma kunnen verwerven. De vraag is echter of gelede structuren, oftewel categorale onderwijsstelsels, niet juist afkomstgerelateerde prestaties en 'ongelijkheid' stimuleren. De selectiefunctie heeft te maken met landelijke onderwijs-structuren, binnen het eerder gepresenteerde kader gaat het daarbij om de *throughput* van het systeem, gedefinieerd op macroniveau.

De allocatiefunctie is te zien als een combinatie van kwalificatie en selectie, dat wil zeggen een zodanige differentiatie in vaardigheidsniveaus dat maatschappelijke taken en rollen zo goed mogelijk bediend worden.

Als het gaat om de bovengenoemde functies te vervullen spelen zowel prestatie als rendements-indicatoren een centrale rol.

Samenvattend kan worden gesteld dat productiviteit (a) van centrale betekenis is voor deze visie op het typeren van onderwijskwaliteit. Verder is er een relatie gelegd met adaptiviteit (in de zin van het kiezen van de juiste doelstellingen) (b) en gelijkheid (c), waarbij het de vraag is of stelsels met veel selectie

drempels afkomstgerelateerde clustering van leerlingen in de ‘lagere’ schoolsoorten niet juist zou stimuleren.

3.5. Kwaliteit vanuit het gezichtspunt van verschillende belanghebbenden

In deze paragraaf wordt stilgestaan bij de accenten die verschillende actoren of belanghebbenden leggen bij het beoordelen van de kwaliteit van onderwijs. Achtereenvolgens wordt stilgestaan bij de ouders van leerlingen, de overheid, c.q. het Ministerie van OCW, de onderwijsinspectie en de school zelf. In de volgende paragraaf gaan we apart in op de certificering business. Ter illustratie van de optiek van de overheid wordt gebruik gemaakt van de zogenoemde kwaliteitsagenda’s die in de periode 2007 tot 2009 werden opgesteld door het ministerie.

De kwaliteitsagenda’s van het Ministerie van OCW
Voor het basis-, voortgezet en beroepsonderwijs zijn door de verschillende directies van OCW kwaliteitsagenda’s gepubliceerd (OCW, 2007, 2008 en 2009).

De kwaliteitsagenda voor het basisonderwijs was sterk gericht op het verbeteren van de leerprestaties in taal en rekenen. Hoewel Nederlandse leerlingen het best goed doen in internationale assessment-onderzoeken, werd verdere verbetering nodig geacht, zeker wat betreft de proportie studenten die in het topsegment van de verdeling van de internationale toetsen scoort. Een eerste stap om de gewenste verbetering tot stand te brengen bestond uit het formuleren van leerstandaarden voor taal en rekenen. De groep leerlingen die beneden hun kunnen presteert, en die op 10% van de totale leerlingenpopulatie wordt geschat, zou met 40% moeten afnemen. Ook het gemiddelde prestatieniveau moest omhoog. Het aantal scholen dat door de Inspectie als ‘zwak’ wordt gekwalificeerd moet met 50% gereduceerd worden. Tenslotte werd gesteld dat in 2011 80% van de scholen een goed functionerend systeem van kwaliteitszorg zou moeten hebben.

De volgende maatregelen om een en ander tot stand te brengen werden gepropageerd en financieel ondersteund:
- een effectief gebruik van de officiële schooltijd;
- het stimuleren van een opbrengstgerichte schoolcultuur;
- het gebruik van leerlingvolgsystemen;
- de toepassing van schoolverbeteringsprogramma’s die wetenschappelijk onderbouwd zijn (evidence based), speciaal in scholen met veel achterstandsleerlingen;
- het creëren van rijke leeromgevingen (ICT toepassingen);
- het stimuleren van ouderbetrokkenheid;
- disseminatie van goede praktijken op het terrein van lesgeven;
- vrijheid en autonomie voor de scholen bij het tot stand brengen van hun pogingen tot verbetering;
- professionele ontwikkeling van de leerkrachten en het organiseren van schoolnetwerken;
- hogere eisen (c.q. standaarden) bij het rekenonderwijs op de pabo’s.

De Kwaliteitsagenda voor het voortgezet onderwijs legt de nadruk op hogere testscores op internationale toetsen op het gebied van wiskunde en taal, optimale prestaties van alle leerlingen, het behoren tot de top in de Internationale Kennissamenleving, het scheppen van een aantrekkelijke werkomgeving voor docenten en ervoor zorgen dat het publiek weer vertrouwen krijgt in het voortgezet onderwijs (na alle kritiek van de Commissie Dijsselbloem

Er werden zeven beleidsprioriteiten gesteld:
- rekenen en taal;
- hoger rendement;
- stimuleren dat burgerschap een volwaardig schoolvak wordt;
- ruimte voor de leerkracht, 'eigenaarschap', maar met een duidelijk accent op leerinhouden met lesgeven als kernactiviteit van leerkrachten;
- goede en betrouwbare examens;
- een verbetercultuur op scholen, onder meer gestimuleerd door goed leiderschap, en geleid door concrete *targets*, hetgeen moet leiden tot een geringer percentage zwakke scholen;
- toename van het aantal scholen die een goed systeem van interne kwaliteitszorg gebruiken.

De nota gaat ervan uit dat de hoge mate van autonomie van de Nederlandse scholen optimaal is voor het realiseren van de kwaliteitsagenda (Ministerie van OCW, 2007, p. 9).
De 'Strategische Agenda voor het Beroepsonderwijs en de Volwassenen Educatie', voor de periode van 2008 tot 2011, zoekt de verbetering van het beroepsonderwijs binnen de volgende vijf thema's:
- verbetering van de aansluiting tussen onderwijs en arbeidsmarkt;
- verbetering van de kwaliteit van het onderwijs (door middel van vernieuwde kwalificatiedossiers, goede beheersing van taal en rekenen, 850 lesuren per jaar, standaardisering van examens in de beroepsgerichte vakken en mogelijke vereenvoudigingen in de structuur van het beroepsonderwijs);
- betere aansluiting binnen de beroepsgerichte kolom (lager, middelbaar en hoger beroepsonderwijs);
- actieve en duurzame deelname in zowel het onderwijs, het werk en de maatschappij (hierbij gaat het om maatregelen om ongediplomeerd schoolverlaten tegen te gaan);
- betere consistentie in het beleid, waarbij het onder meer gaat om een heroverweging van de budgettaire verantwoordelijkheden van gemeentes.

Een belangrijke innovatie op het terrein van het beroepsonderwijs was de introductie van 'competentiegericht onderwijs', als onderdeel van de vernieuwde kwalificatiestructuur. Inmiddels is deze term alweer door 'beroepsgericht onderwijs' vervangen. In 2011 werd de beroepsgerichte kwalificatiestructuur ingevoerd voor het beroepsonderwijs.
In vergelijking met de kwaliteitsagenda's voor basis- en voortgezet onderwijs legt de kwaliteitsagenda voor het beroepsonderwijs een aanvullend accent, namelijk de aansluiting tussen onderwijs en arbeidsmarkt. Evaluatief gezien betekent dit een focus op *impact* naast rendement- en prestatie-indicatoren.

Een rode draad door de drie kwaliteitsagenda's is de aandacht voor verbetering van de prestaties in de basisvakken, taal en rekenen. Nieuwe onderwijsopbrengsten die worden nagestreefd zijn burgerschap in het voortgezet onderwijs en beroepsgerichte competenties in het beroepsonderwijs. Numeriek rendement en het voorkomen van voortijdig schoolverlaten hebben ook hun plaats in de kwaliteitsagenda's. De procesfactoren die de verbeterde opbrengsten teweeg zouden moeten brengen liggen vooral op het terrein van professionalisering van leerkrachten en de schoolorganisatie. In de kwaliteitsagenda's werden ook meer specifieke categorieën van 'hefbomen' voor kwaliteitsverbetering genoemd: toetsen, monitoring van leerprestaties, examens, kwaliteitszorg, de opleiding van leerkrachten, continue professionele ontwikkeling, evidence-based innovatie, en een betere aansluiting tussen de diverse schooltypen. De grote autonomie van scholen binnen het Nederlandse onderwijsbestel wordt omarmd, de sterk gesegmenteerde schoolstructuur wordt geheel onbesproken

gelaten. Een laatste kenmerk van het Nederlandse onderwijs dat minder uitgewerkt wordt in de kwaliteitsagenda's is de uitgebreide ondersteuningsstructuur, waarin zo'n miljard euro per jaar omgaat (zie voor een kritische bespreking: Scheerens, 2009).

Inspectie van het Onderwijs

De Inspectie van het Onderwijs heeft als taak toezicht te houden op de onderwijskwaliteit. De Inspectie stelt in het kader van zijn *waarborgfunctie* vast of scholen voldoen aan de wettelijke eisen aan de basiskwaliteit. Op basis van de beoordeling stelt de Inspectie vast of een school 'voldoende', 'onvoldoende' of 'zeer zwak' wordt beoordeeld. De Inspectie beoordeelt scholen op de volgende kwaliteitsstandaarden:

- Onderwijsproces;
- Schoolklimaat;
- Onderwijsresultaten;
- Kwaliteitszorg en Ambitie;
- Financieel beheer.

In termen van figuur 3.1 (zie paragraaf 3.3) hebben de meeste standaarden betrekking op het proces of doorvoer; de standaard Onderwijsresultaten betreft de output. De Inspectie kijkt hierbij naar de volgende indicatoren:

1. Positie in leerjaar 3 ten opzichte van advies van de basisschool (Onderwijspositie t.o.v. advies po);
2. Percentage onvertraagde studievoortgang in leerjaar 1 en 2 (Onderbouwsnelheid);
3. Percentage onvertraagde studievoortgang vanaf leerjaar 3 per afdeling (Bovenbouwsucces);
4. Gemiddeld cijfer Centraal Examen van alle vakken per afdeling (Examencijfers).

Het oordeel over de onderwijsresultaten is dus mede gebaseerd op de cijfers voor het CE. De Inspectie kijkt naar het gemiddelde cijfer voor alle vakken van alle leerlingen die geslaagd of gezakt zijn voor de betreffende onderwijssoort. Cijfers voor vakken die niet meetellen voor het slagen, worden niet meegewogen. Om te voorkomen dat het oordeel wordt gebaseerd op kleine aantallen, wordt het gemiddelde alleen berekend voor onderwijssoorten waarvan tenminste 30 CE-cijfers beschikbaar zijn en wordt een gewogen driejaargemiddelde bepaald.
De cijfers voor het SE worden *niet* meegewogen bij de beoordeling. Tot 2016 speelden de SE-cijfers nog wel een rol: de Inspectie nam het verschil tussen resultaten van SE en CE mee in de beoordeling van de onderwijsresultaten. Als de SE-cijfers drie opeenvolgende jaren meer dan 0,5 punt hoger lagen dan het gemiddelde CE-cijfer, werd dit als onvoldoende beoordeeld.
Als aanleiding om het SE-CE niet meer mee te nemen in de beoordeling wordt de aanpassing van de zak-slaagregeling in 2012 genoemd. Bovendien waren er signalen dat de beoordeling leidde tot meer ongewenst 'teaching-to-the-test'. De Inspectie hecht er overigens nog steeds aan dat het verschil tussen scores op SE en CE klein blijft. Als het gemiddelde SE-cijfer jaar in jaar uit hoger ligt dan het gemiddelde CE-cijfer, kan dit leiden tot het tijdelijk intrekken van de examenlicentie van de school.

De Inspectie publiceert de onderwijsresultaten op haar website (www.zoekscholen.onderwijsinspectie.nl). Van de bovengenoemde indicatoren worden de gegevens gepresenteerd van drie achtereenvolgende jaren en van het driejaarsgemiddelde. Van de driejaarsgemiddelden wordt ook weergegeven of deze 'onder de norm' of 'boven de norm' liggen (Inspectie van het Onderwijs, 2018). In het verleden waren deze normen relatief, dat wil zeggen dat ze werden bepaald op basis van het

gemiddelde in een referentiegroep. In de nabije toekomst wil de Inspectie *absolute* normen gaan hanteren. De huidige normering bevindt zich in een overgangssituatie.
De website van de Inspectie verwijst ook door naar www.scholenopdekaart.nl, waar dezelfde gegevens worden gepresenteerd. Op deze site kunnen scholen zelf een toelichting geven bij de gepresenteerde resultaten.

De school
De verantwoordelijkheid voor een goede onderwijskwaliteit ligt in de eerste plaats bij het schoolbestuur. Bij het bevorderen en bewaken van die kwaliteit gelden de kaders van wet- en regelgeving, waar de Inspectie op toeziet. De kaders bepalen wat de school *moet* doen, maar bieden de school tevens veel ruimte om keuzes te maken in wat men *wil* doen. Scholen zijn in Nederland in hoge mate autonoom, ook bij het kiezen van doelstellingen en het bepalen wanneer deze zijn behaald. Zo bepaalt de school zelf welke leerstof wordt aangeboden in welk leerjaar, in hoeverre gewerkt wordt aan de hand van een lesmethode en hoeveel onderwijstijd wordt besteed per vak.
Ook hoe de onderwijsopbrengsten worden getoetst, is grotendeels de verantwoordelijkheid van de school. Het schoolexamen (SE) wordt opgesteld door de school, aan de hand van landelijke eindtermen die per vak worden opgesteld. Elke school heeft dus zijn eigen schoolexamen. Bij het maken van toetsen hebben scholen de ruimte om eigen inhoudelijke keuzes te maken. De school stelt een programma van toetsing en afsluiting (PTA) op. Daarin wordt de inhoud van de examentoetsen vastgelegd, de toetsvorm en de normering.

Ouders
Onderzoeksgegevens die iets laten zien over de wijze waarop ouders tegen de kwaliteit van scholen aankijken hebben betrekking op de schoolkeuze. Volgens Dulmers (1988) hangen schoolkeuze-motieven af van het profiel van de school (het onderwijsconcept en de denominatie), pragmatisme (nabijheid, veiligheid) en onderwerpen die te maken hebben met kwaliteit. Daarbij wordt dan, als het gaat om het basisonderwijs, vooral gedacht aan een kindvriendelijk pedagogisch klimaat en aan onderwijsresultaten.

Andere auteurs, bijvoorbeeld Pannecoucke (2005) hanteren een input-, proces-, outputkader om kwaliteitsoverwegingen bij schoolkeuze door ouders te definiëren. Een voor ons relevante indeling, die overeenkomt met figuur 3.1. *Input*criteria hebben betrekking op het schoolgebouw, de kwaliteit van de leerkrachten en het overige personeel en de kwaliteit van het onderwijsleermateriaal. *Proces*criteria zijn onder meer de breedte van het pedagogische aanbod, de didactische aanpak en het pedagogische klimaat. *Output*criteria zijn indicatoren die iets zeggen over de leerprestaties. Een inputcriterium dat door deze auteur wordt gezien als zowel pragmatisch als betrekking hebbend op onderwijskundige overwegingen is de samenstelling van de schoolbevolking, qua leerlingen en leerkrachten. In Vlaanderen bleek de etnische samenstelling van de leerlingenpopulatie een belangrijke overweging voor ongeveer 50% van de ouders. Samenstelling bleek voor ouders van etnische minderheden nog belangrijker, waarbij voorkeur bestond voor een monoculturele in vergelijking met een multiculturele samenstelling. Door de geciteerde onderzoekers wordt gewezen op het feit dat de schoolkeuzemotieven variëren met de sociale status en het opleidingsniveau van de ouders. Wat betreft de procescriteria: ouders met een lager opleidingsniveau geven de voorkeur aan de meer ‘traditionele’ aspecten van kwaliteit (orde, kleine klassen en prestatiegerichtheid), terwijl ouders met een hoger opleidingsniveau autonomie, creativiteit en sociaal-emotionele ontwikkeling van belang vinden.

Het feit dat pragmatische overwegingen de overhand hebben bij de schoolkeuze kan voor een deel worden verklaard uit het gegeven dat ouders het meestal moeilijker zullen vinden om schoolkwaliteit te

beoordelen dan pragmatische zaken, zoals afstand, reputatie van de school en denominatie. Dit wordt onderstreept in een onderzoek van Janssens (2010) dat betrekking heeft op het gebruik van de zogenoemde kwaliteitskaarten in het Nederlandse onderwijs. Geconcludeerd wordt dat prestatie-indicatoren van scholen weinig invloed hebben op de schoolkeuze, zowel als het gaat om de initiële schoolkeuze als in situaties waarin men van school wil veranderen. Uit de onderzoeksliteratuur over motieven bij schoolkeuze ontstaat het beeld dat profiel en pedagogisch klimaat voor ouders nog belangrijker zijn dan het niveau van de leerprestaties.

3.6. Schoolkwaliteit vanuit het gezichtspunt van de certificeringindustrie

De missie van systemen voor kwaliteitsborging, zoals ISO 9001, is ervoor te zorgen dat organisaties kwaliteitsmanagementsystemen hebben die aan zeer precieze standaarden voldoen. Organisaties moeten bewijzen dat hun voorzieningen voor kwaliteitsmanagement volgens een geheel aan standaard-procedures verlopen. Dit wordt getoetst door externe auditoren, die zelf weer gecertificeerd zijn om organisaties te certificeren (de vraag blijft wie de certificeerders van de certificering certificeert).

Het basis conceptuele kader voor dit kwaliteitsmanagement heeft twee hoofdelementen:

- het beschrijven van de kernprocessen van de organisatie;
- het toepassen van een bepaalde methodologie voor kwaliteitsborging op elk van deze processen (bijvoorbeeld resource management en productrealisatie), volgens de PDCA (Plan-Do-Check-Act) methode.

Zo gaat het ISO 9001 model uit van een primair productieproces dat uiteindelijk bepaald wordt door de wensen van de klanten. De productie is op te vatten als het omzetten van hulpbronnen, of inputs, in producten of output. De primaire processen worden onderzocht op hun effectiviteit en het bieden van toegevoegde waarde. Kwaliteitsmanagement heeft drie ondersteunende processen: (1) het nemen van verantwoordelijkheid voor kwaliteitsborging door het management, (2) resources management en (3) meting, analyse en verbetering.

Het leren op basis van feedback is een centrale gedachte achter deze aanpak. In het ISO-model is de feedback enerzijds gebaseerd op het monitoren van het productieproces, maar ook op het meten van klantbehoeften en klanttevredenheid.

Kwaliteitsmanagementsystemen berusten op de premisse dat primaire productieprocessen volledig in hun werking bekend zijn, waardoor bij een nauwgezette monitoring van inputs en processen de verwachte opbrengsten gegarandeerd zijn. Deze premisse is niet vervuld voor de primaire processen van scholen, in het bijzonder onderwijs- en leerprocessen. Volgens de terminologie van onderwijseconomen is de onderwijs productiefunctie niet bekend, hoewel er wel duidelijke gedachten en empirische gegevens zijn over wat meestal werkt; het geheel aan resultaten van onderzoek en meta-analyses op het terrein van onderwijseffectiviteit is hierbij de relevante kennisbasis (Scheerens en anderen, 2007; Hattie, 2009). Een onkritisch gebruik van dit soort kwaliteitsmanagementbenaderingen in het onderwijs kan leiden tot een bureaucratisch ritueel van uitgebreide beschrijvingen van minder centrale processen of kenmerken die niet, of alleen maar losjes, gekoppeld zijn aan opbrengsten.

3.7. Alternatieve visies op onderwijskwaliteit

In het 'Education for All Global Monitoring Report' van de UNESCO voor het jaar 2005 (UNESCO, 2004), getiteld 'The Quality Imperative', worden opvattingen over kwaliteit in verband gebracht met onderwijstradities. Bij nader inzien blijken deze 'tradities' te bestaan uit filosofische, psychologische en sociologische benaderingen:

1. humanisme,
2. behaviorisme,
3. kritische theorie.

Voor een ander deel bestaan ze uit meer pragmatische keuzes die bepaald worden door specifieke contextuele condities of de specifieke onderwijssoort waar ze op worden toegepast:

4. kwaliteit in de inheemse traditie,
5. kwaliteit in het volwassenenonderwijs.

De belangrijkste kenmerken van deze vijf alternatieve benaderingen worden hieronder aangehaald uit het rapport (*ibid,* 32-35).

1. Kwaliteit in de *humanistische traditie* wordt als volgt gekenmerkt:
   - Gestandaardiseerde, voorgeschreven en extern bepaalde of gecontroleerde curricula worden verworpen. Deze worden gezien als ondermijnend voor de mogelijkheden van leerlingen om hun eigen betekenissen te construeren en als schadelijk voor programma's die willen aansluiten bij de omstandigheden en behoeften van individuele studenten of deelnemers.
   - *Assessment* is bedoeld om leerlingen en deelnemers feedback te geven over de kwaliteit van hun individuele leren. Het wordt als een integraal onderdeel van het leerproces gezien. Zelfbeoordeling en beoordeling door medeleerlingen worden toegejuicht als manieren om een diepere bewustwording van het leren tot stand te brengen.
   - De rol van de leerkracht is meer die van begeleider dan van instructeur.
   - Sociaal constructivisme wordt omarmd, omdat het de bovenstaande uitgangspunten deelt en leren opvat als een sociaal proces, in plaats van het resultaat van individuele inbreng.
2. Volgens de behavioristische benadering wordt aangenomen dat onderwijs objectief getoetst kan worden en gestimuleerd wordt door duidelijk gestructureerde externe inputs:
   - Gestandaardiseerde, voorgeschreven en extern bepaalde of gecontroleerde curricula, gebaseerd op extern voorgeschreven doelstellingen, en vastgesteld zonder inspraak van de lerende, worden onderschreven.
   - Assessment, gebruik makend van vooraf vastgestelde criteria en standaarden, wordt als een objectieve meting van aangeleerd gedrag gezien.
   - Toetsen en examens worden als centrale kenmerken van leren gezien en tevens beschouwd als middel voor planning en het toekennen van beloning en straf.
   - De leerkracht dirigeert het leerproces, als een expert die stimuli en response controleert.
   - Aan incrementele leertaken die gewenste associaties in de geest zouden versterken worden, wordt de voorkeur gegeven.
3. Kwaliteit volgens de *kritische traditie* biedt plaats aan een breed scala van benaderingen, variërend van de Marxistische ideologie tot de visie van 'ontscholers' als Illich en Freire. Volgens het EFA Global Monitoring Report ligt de focus bij de kritische theorie op ongelijkheid in deelname en onderwijsopbrengsten en de rol die het onderwijs speelt bij het legitimeren en in stand houden van

sociale structuren en kennis die bepaalde groepen bevoorrecht. Aanhangers stellen kwaliteit van het onderwijs gelijk aan:
- onderwijs dat tot sociale verandering leidt;
- een curriculum en een wijze van lesgeven die kritische analyse van sociale machtsrelaties hoog houden;
- actieve deelname van de lerende aan het ontwerp van de eigen leerervaringen.

4. Omdat het EFA Global Monitoring Report bedoeld is voor ontwikkelingslanden, wordt er een vierde visie op kwaliteit besproken die kritisch staat tegenover de adoptie van Westerse concepten, en die zich richt op zelfbeschikking, gelijkheid en werkgelegenheid op het platteland. Deze traditie wordt aangeduid als kwaliteit volgens de *inheemse traditie* en zou de volgende implicaties hebben:
   - Mainstream benaderingen, geïmporteerd uit Europa, zijn niet noodzakelijkerwijs relevant in totaal andere economische omstandigheden.
   - Het verzekeren van relevantie van het onderwijs vraagt om lokaal ontwerp van curriculummateriaal, pedagogische aanpak en van leerlingevaluatie (assessment).
   - Alle leerlingen bezitten rijke bronnen van eerder verworven kennis, die gevormd is door tal van ervaringen; opvoeders moeten daarbij aansluiten en deze kennis verder voeden.
   - Leren zou verder moeten gaan dan het klaslokaal, door middel van niet-formeel en levenslang leren.
5. Een vijfde perspectief is gekoppeld aan *benaderingen in het volwassenenonderwijs.* "Volgens de traditie van het volwassenenonderwijs zijn allerlei ervaringen en kritische reflectie bij het leren belangrijke kwaliteitsaspecten. Radicale denkers zien de lerende als sociaal gesitueerd met de potentie om hun ervaring en leren te gebruiken als uitgangspunt voor sociale actie en sociale verandering" (p. 34).

Sleuteldimensies bij deze alternatieve visies op kwaliteit zijn: de soort onderwijsdoelstellingen die centraal gesteld worden, overwegingen betreffende wenselijk geachte proceskenmerken van onderwijs en leren en premissen die betrekking hebben op de leertheorie. Om met de laatste te beginnen, constructivisten benadrukken bijvoorbeeld het belang van aanvangskennis en actief leren. Ook wanneer dit in verband wordt gebracht met het gebruik van levensechte leersituaties zou men dit kunnen zien als een didactisch principe dat in elk van de tradities een plaats zou kunnen hebben. Het is namelijk niet dwingend noodzakelijk om dit soort onderwijs altijd als 'open' en niet voorgestructureerd op te vatten, hoewel dit in de humanistische traditie wel zo wordt opgevat. In de kritische, inheemse en volwassenen-educatie traditie is het aansluiten bij de lokale situatie en het dagelijks leven van de leerlingen meer dan een didactisch principe, maar eerder een doel op zich. Het is maar de vraag of de vijf tradities werkelijk radicaal verschillende onderwijsopbrengsten nastreven. Wel is er sprake van graduele verschillen in accent. Sociale vaardigheden passen vooral bij de humanistische traditie, terwijl het aanbrengen van een maatschappijkritische benadrukt worden in de kritische, inheemse en volwassenenonderwijs traditie. Tegelijkertijd is niet aannemelijk, met uitzondering misschien van de inheemse traditie, dat veel van de alternatieve tradities zouden willen afzien van het belang van cognitieve vaardigheden en basisvakken. Met de opkomst van internationaal vergelijkende assessment studies, zoals TIMSS en PISA, krijgt de globalisering verder voet aan de grond in het onderwijs. Omdat veel ontwikkelingslanden hebben besloten om deel te nemen aan deze projecten bestaat kennelijk de wens om zich volgens dezelfde criteria en standaarden te meten met de geïndustrialiseerde wereld.

3.8. Integratie

Als we kijken naar verschillende benaderingen van onderwijskwaliteit, zoals productiviteit, effectiviteit, (on)gelijkheid, efficiency en het aanpassingsperspectief, overheerst het beeld dat kwaliteit moet blijken uit de onderwijsopbrengsten, in de zin van leerprestaties, numeriek rendement en langere termijn 'impact'. Deze visie onderstreept het belang van eindtoetsen en examens. Daarbij zijn er onderlinge verschillen tussen de behandelde kwaliteitsperspectieven naar gelang het accent ligt op een van deze drie categorieën van opbrengstindicatoren – prestaties, rendement en impact – en, als het gaat om de leerprestaties, of daarbij een wat bredere of smallere range aan kennis en vaardigheden wordt gepropageerd. In de praktijk ligt er vaak een accent op cognitieve vakken, maar nieuwe gebieden, zoals burgerschap en allerlei beroepsgerichte competenties worden daar soms aan toegevoegd.

Vanuit verschillende actorperspectieven (ouders, certificering, onderwijsinspectie) wordt ook belang gehecht aan input- en procesfacetten van kwaliteit. Bij het gebruik van input- en procesindicatoren, naast opbrengstindicatoren, worden deze 'op zichzelf' beoordeeld, volgens het in paragraaf 3.3 behandelde gefragmenteerde gebruik van kwaliteitsindicatoren. Kwaliteitsbeoordeling op basis van effectiviteit en efficiency komt eigenlijk alleen voor in de vorm van programma-evaluaties en achtergrondstudies van internationale assessmentprogramma's. In dergelijke analyses neemt de gerichtheid op vraagstellingen betreffende gelijkheid (equity) trouwens steeds meer toe (vgl. Causa en Chapuis, 2009; OESO, 2007; Woessmann en anderen, 2009).

Effectiviteit, efficiency en responsiviteit zijn analytisch ingewikkelde verschijnselen, die voor veel actoren lastig te beoordelen zijn. Dit onderstreept de eerdere conclusie dat kwaliteitsoordelen meestal het karakter hebben van het los naast elkaar bezien van afzonderlijke kwaliteitsfacetten, meestal 'ruwe', ongecorrigeerde opbrengstindicatoren. 'Netto-effecten' van scholing, in de zin van indicatoren van de toegevoegde waarde (*value-added)* zijn eveneens moeilijk te begrijpen voor de meeste actoren.

Dit hoofdstuk lijkt uit te monden in een nadrukkelijke bevestiging van de stelling dat onderwijskwaliteit moet blijken uit de onderwijsresultaten. Deze stelling werd decennia geleden al met verve naar voren gebracht door A. D. de Groot (De Groot, 1978).

In zijn behandeling van de vraag 'is de kwaliteit van het onderwijs te beoordelen', werd door hem in de eerste plaats gewezen op de risico's van het begrip kwaliteit; "Het begrip 'kwaliteit' kan werken als een soort vluchthaven voor wie de discussie in open water over meer concrete zaken te moeilijk en te onbevredigend vindt."(ibid, p. 124). Echter, met het credo 'kwaliteit moet blijken' gaf hij al twee pagina's verder zijn oplossing: "Bij onderwijs gaat het uiteindelijk niet om de vraag hoe mooi we het geven maar om wat het uiteindelijk uithaalt, om wat leerlingen ervan meenemen."

4 Kernfuncties van examens en eindtoetsen: kwaliteitsborging en kwaliteitsbevordering

Jaap Scheerens en Anne Luc van der Vegt

4.1 Inleiding

In dit hoofdstuk wordt nader ingaan op de kernfuncties van examens en eindtoetsen[5]. Eerder werden in dit verband genoemd het formaliseren van het civiele effect van scholing en het verzekeren van evaluatie en verantwoording (accountability). Deze functies bespreken we in paragraaf 4.2, 4.3 en 4.4. Vervolgens wordt empirisch onderzoek beschreven, waarin is nagegaan of er sprake is van effecten van examens en eindtoetsen op onderwijsprestaties. Bij dit onderzoek worden landen of regio's waarin dit soort voorzieningen wel aanwezig zijn vergeleken met landen of regio's waar dit niet het geval is. Daarna wordt stilgestaan bij mogelijke verklaringen bij de vraag waarom evaluatie en toetsing bijdragen aan prestatieverbetering. We onderscheiden daarbij a) het bieden van prikkels en stimuleren van prestatiemotivatie, b) het stimuleren van leerprocessen op basis van feedback van uitkomsten en c) het mogelijk maken van een goede onderlinge aansluiting van curriculumcomponenten ('curriculum alignment') (4.5). Alignment duidt op een derde functie van examens en toetsen, namelijk de richtingbepalende rol in het kader van curriculumontwikkeling en curriculumimplementatie. Omdat deze laatste functie de laatste decennia wat minder aandacht heeft gekregen, wordt er wat uitvoeriger op ingegaan (4.6).
In dit hoofdstuk wordt de nadruk gelegd op de positieve verworvenheden en gunstige effecten van examens en eindtoetsing. In het volgende hoofdstuk (hoofdstuk 5) wordt stilgestaan bij veronderstelde dysfunctionele aspecten en schadelijke (neven)effecten.

4.2 Examens en eindtoetsen als onderdeel van de institutionele infrastructuur van het onderwijsstelsel: accent op kwaliteitsborging

Examens en eindtoetsen zijn onderdeel van de institutionele infrastructuur van een onderwijsstelsel. Helderheid over de institutionele kant van onderwijs is van essentieel belang voor de maatschappelijke erkenning en voor de regeling en besturing van het onderwijs. Het examen reguleert het civiel effect van het onderwijs, geeft heldere ijkpunten voor persoonlijke leerroutes en doorstroming door het onderwijs en onderlinge aansluiting van schoolcategorieën. Omdat examens en eindtoetsen fungeren als evaluatiemechanisme kunnen ze tevens worden gezien als bijdragend aan de voorwaarden voor effectieve besturing van het stelsel. Zo stelt De Leeuw (2008) dat de aanwezigheid van een evaluatiemechanisme de basisvoorwaarde voor effectieve besturing is. Deze hoedanigheden maken dat bijgedragen wordt aan de voorspelbaarheid van het onderwijs en de betrouwbaarheid ervan voor belangrijke belanghebbenden, ouders, leerlingen, maatschappelijke organisaties en de overheid. Deze voorspelbaarheid betekent niet dat examens en eindtoetsen niet veranderd en bijgesteld kunnen worden, bijvoorbeeld uit het oogpunt van modernisering, en veranderende omgevingscondities, maar

[5] De term eindtoetsen wordt gebruikt als generieke term, om te refereren aan toetsen die in het Engels "high stakes worden genoemd.

wel dat de basisfunctie ervan overeind moet worden gehouden. Echter, aanpassing en modernisering van een dergelijke fundamentele institutionele voorziening vragen om grote zorgvuldigheid en op feiten gebaseerde overwegingen. Anders gezegd, wijzigingsvoorstellen zouden niet 'lichtvaardig' geponeerd moeten worden (vergelijk voorstellen van de VO-raad om 'meer maatwerk' te leveren bij examens, onder meer door de mogelijkheid te bieden om op verschillend niveau en per vak afzonderlijk examen te doen). Om dit nader te illustreren, maar vooral vanwege de uitstekende weergave van de functie van examens en eindtoetsen, wordt hieronder uitvoerig geciteerd uit het Advies van de Onderwijsraad getiteld; *Maatwerk binnen wettelijke kaders: eindtoetsing als ijkpunt voor het funderend onderwijs (Onderwijsraad, 2015)*

Het desbetreffende advies verwoordt de kernfuncties van examens en eindtoetsing als volgt:

> "Het onderwijsproces dat aan eindtoetsing voorafgaat, is van belang voor het funderend karakter van het primair en voortgezet onderwijs. Hierbij gaat het vooral om het niveau en de breedte van de gezamenlijke kennisbasis en om het gemeenschappelijke onderwijsproces, dat naast kwalificatie ook socialisatie en persoonsvorming tot doel heeft. Eindtoetsing en, in het voortgezet onderwijs, een daarop gebaseerd diploma, vormen hiervoor een dragend element. Eindtoetsing en diploma's vervullen ook een belangrijke rol bij de doorstroom binnen het onderwijsstelsel. Daarnaast hebben diploma's buiten het onderwijs een civiel effect dat van grote waarde is. Verruiming van de wettelijke kaders van eindtoetsing doet afbreuk aan deze aspecten." (p. 16)

Vervolgens worden deze functies nader toegelicht:

> *"Eindtoetsing faciliteert doorstroom naar vervolgonderwijs*
> De eindtoetsing waarborgt een goede doorstroom naar en aansluiting op het vervolgonderwijs. In het primair onderwijs is de eindtoets verbonden met de doorverwijzing naar een schooltype in het voortgezet onderwijs. Weliswaar wordt het schooladvies voorafgaand aan de eindtoets vastgesteld, maar op basis van de eindtoets kan dit advies worden bijgesteld. Daarnaast heeft de eindtoets een objectiverende functie en biedt het leraren een maatstaf waartegen ze hun advies kunnen afzetten. De eindtoets draagt bij aan een betrouwbare doorverwijzing en biedt tegenwicht tegen factoren die sociale ongelijkheid kunnen vergroten, zoals de druk die ouders op de school uitoefenen om een hoger schooladvies te geven. Ten slotte vormen de referentieniveaus van de eindtoets een duidelijke norm voor het prestatieniveau dat alle leerlingen dienen te behalen. De eindtoets draagt op die manier bij aan het behalen van een basisniveau waarmee ook de leerlingen aan de onderkant van de prestatieladder voldoende zijn toegerust voor het vervolgonderwijs.
>
> In het voortgezet onderwijs is de eindtoetsing verbonden met de doorstroom binnen het voortgezet onderwijs (leerlingen met een havo-diploma kunnen bijvoorbeeld instromen in het vwo) en de doorstroom naar het hoger onderwijs of het middelbaar beroepsonderwijs. In alle gevallen geldt dat de centrale examens objectieve normen bieden voor het prestatieniveau dat nodig is om succesvol door te stromen.
>
> Eindtoetsen hebben in het Nederlandse systeem een selecterende functie. Leerlingen worden verwezen naar of krijgen toegang tot het vervolgonderwijs op basis van hun prestaties op deze toetsen. In dat verband is herhaaldelijk gewezen op de nadelen van een vroege toewijzing in niveaustromen die plaatsvindt aan het eind van het primair onderwijs, met name voor leerlingen die minder snel op gang komen omdat hun thuissituatie een minder rijke leeromgeving

biedt. Uit onderzoek blijkt echter dat het gebruik van objectieve toetsen een dempend effect heeft op de negatieve effecten van vroege selectie naar niveaustromen. Centrale toetsing is één van de redenen voor de relatief goede prestaties van Nederlandse leerlingen aan de onderkant van de prestatieladder. Onderzoek wijst uit dat ongelijke kansen niet zozeer te maken hebben met het moment van selectie, als wel met de wijze van selectie. Objectieve eindtoetsing helpt om ongelijke kansen tegen te gaan. (…)
In onderwijsstelsels waar geen of minder sprake is van standaard eindtoetsing aan het eind van het voortgezet onderwijs, wordt door het vervolgonderwijs een ander instrument (toelatingsexamen of ander selectiemechanisme) gecreëerd en toegepast om te toetsen of de aangemelde kandidaat zich daadwerkelijk kwalificeert voor de desbetreffende opleiding. Het zogenoemde civiel effect van het vo-diploma is in Nederland dermate groot, dat vervolgopleidingen niet overgaan tot het stellen van aanvullende eisen naast een diploma. Wel wordt de afgelopen jaren vaker extra aandacht besteed aan de motivatie van de aankomende student." (p. 17-18)

"Eindtoetsing in het voortgezet onderwijs garandeert het civiel effect van diploma's
Het civiel effect van diploma's wordt bepaald door zowel subjectieve als objectieve aspecten. Bij de subjectieve aspecten gaat het om hoe betrokkenen (zoals leerlingen, ouders, vervolgopleidingen) oordelen over de aard en inhoud van de diploma's. Hieronder vallen ook de informatiekosten, dat wil zeggen de moeite die belanghebbenden moeten doen om kwaliteiten en leervermogen van een leerling adequaat in te schatten. Met de huidige vo-diploma's zijn de informatiekosten vanwege de standaardisatie relatief laag. Door de waarde die het diploma vertegenwoordigt hebben betrokkenen (zoals vervolgopleidingen en arbeidsmarkt) een belangrijke indicatie van de kennis en kunde die de gediplomeerde in generieke en/of specifieke zin heeft opgedaan. Daarnaast zegt het verkrijgen van een diploma ook iets over andere eigenschappen en vaardigheden waarover iemand beschikt, bijvoorbeeld doorzettingsvermogen, motivatie en snelheid. Een diploma geeft het vervolgonderwijs en de arbeidsmarkt een mogelijkheid om kandidaten sneller te werven en te selecteren op geschiktheid voor de betreffende opleiding of baan. Extra assessments of aanvullende testen zijn meestal niet noodzakelijk als de waarde van diploma's valide is geborgd. (…)"
De raad heeft erop gewezen dat een diploma aan een aantal algemene uitgangspunten moet voldoen, wil het civiel effect behouden blijven: 1) het moet 'vaste' informatieonderdelen bevatten (gegevens over de prestaties per vak/leeronderdeel op basis van betrouwbare examinering); 2) de informatie moet op een standaardwijze worden aangeboden (cijfers die vergelijking mogelijk maken); en 3) het moet duidelijk zijn hoe in de volgende stap (vervolgopleiding of arbeidsmarkt) met het diploma wordt omgegaan: wanneer geeft een diploma recht op doorstroom, dat wil zeggen op instroom in het vervolgonderwijs?" (p. 18-19)"

"Eindtoetsing garandeert een gedeelde brede vorming met minimaal een basisniveau
Het primair onderwijs en het voortgezet onderwijs worden funderend genoemd, omdat ze voor alle leerlingen toegankelijk zijn en hen een algemene vorming bieden. Die vorming dient als grondslag voor hun verdere ontwikkeling in het onderwijs, het beroepenveld en de samenleving. Het funderende karakter van het primair en het voortgezet onderwijs is gebaseerd op de drie domeinen van onderwijs, namelijk kwalificatie, socialisatie en persoonsvorming. Kwalificatie betekent dat leerlingen kennis, vaardigheden en houdingen verwerven die hen in staat stellen doelen na te streven zoals het volgen van een vervolgopleiding, het uitoefenen van een beroep, of functioneren in een complexe samenleving."(p. 20)

Tenslotte stelt de raad dat aanpassingen in een eindexamen alleen bij wijze van uitzondering mogelijk zijn, en wijst uiteindelijk de door de VO-raad voorgestelde flexibilisering af:

> "Aanpassingen in een eindexamen zijn alleen bij specifiek bepaalde uitzondering mogelijk. Het College van Toetsing en Examens is onder meer verantwoordelijk voor de centrale examens en bepaalt hoe deze op verantwoorde wijze worden afgenomen en welke uitzonderingen toegestaan kunnen worden."(p. 19)

> *"Mogelijke effecten van verruiming eindtoetsing onduidelijk*
> De raad wil benadrukken dat er geen empirische kennis voorhanden is over mogelijke gevolgen van verdere verruiming van de wettelijke kaders van eindtoetsing. Naast positieve kunnen ook nadelige effecten van een verruiming worden verondersteld. De raad wijst in dit verband op zijn advies Onderwijspolitiek na de commissie-Dijsselbloem. Na het uitkomen van het rapport van de commissie-Dijsselbloem bestond er brede consensus over het idee dat het overhaast doorvoeren van stelselwijzigingen risicovol was, en dat dat niet meer diende te gebeuren. In het voornoemde advies schetst de raad hoe het de onderwijspolitiek is vergaan na het verschijnen van het rapport van de commissie-Dijsselbloem. De raad komt daarin tot de conclusie dat opnieuw allerlei kleinere en grotere veranderingen in het stelsel zijn aangebracht zonder onderbouwing vooraf met effectonderzoek en zonder de effecten op het stelsel als geheel te doordenken. Ook waarschuwde de raad voor het verschil tussen politiek draagvlak en draagvlak in het onderwijsveld. De raad wil deze conclusies nog eens benadrukken in het licht van de huidige discussie en besluitvorming aangaande het maatwerkdiploma en de verruiming van de wettelijke kaders rondom eindtoetsing."(p. 21)

De raad acht verruiming van de wettelijke kaders niet nodig om maatwerk te kunnen leveren en wijst daarbij onder meer op de vrijheid om interne differentiatie in het onderwijsproces na te streven en op bestaande mogelijkheden tot flexibiliteit, vooral in het voortgezet onderwijs.

4.3 Examinering, eindtoetsing en 'accountability' als strategie tot kwaliteitsverbetering

In het kader van onderwijseffectiviteit onderzoek is er de laatste twee decennia een nieuw aandachtsgebied bijgekomen dat, ter onderscheiding van school- en instructie-effectiviteit, wordt aangeduid als 'systeemeffectiviteit'. Hierbij wordt het effect van door het overheidsbeleid beïnvloedbare factoren op leerprestaties onderzocht. Dit onderzoek heeft begrijpelijkerwijs een sterke impuls gekregen van internationaal vergelijkend assessmentonderzoek, zoals TIMSS, PISA en PIRLS. Het gaat namelijk om onafhankelijke variabelen die gedefinieerd zijn op landniveau. Scheerens (2016) geeft een overzicht van de voornaamste systeemfactoren die in het effectiviteitsonderzoek zijn gebruikt (zie tabel 4.1) Hij maakt daarbij onderscheid tussen meer algemeen maatschappelijke factoren die van belang zijn voor de prestaties van een onderwijssysteem, maar niet direct door onderwijsinstanties te beïnvloeden zijn, infrastructurele condities die wel beïnvloedbaar zijn door het onderwijs, maar niet gemakkelijk veranderbaar zijn en directe, meer korte-termijnbeleidsmaatregelen.

Tabel 4.1: Beleidscontext op systeem niveau

Achtergrondfactor (sociale context)	Systeemontwikkeling en structurele hervormingen	Onderwijsbeleid, met beïnvloedbare input en processen
<u>Welvaartsniveau in het land/regio</u> BNP. BNP per hoofd. Werkloosheidspercentage (aantal werklozen als deel van de totale arbeidspotentieel). Percentage WW-uitkeringen (aantal personen met WW of andere werkloosheidsuitkeringen als deel van de totale werkzame populatie). <u>Samenstelling van de populatie</u> Proportie eerste, tweede, derde generatie immigranten. Proportie immigranten uit ontwikkelingslanden. <u>Lerarenstatus</u> Status van leraarschap ten opzicht van andere beroepen, volgens algemeen publiek en leraren. <u>Betrokkenheid van lokale gemeenschap bij onderwijs</u> Proportie van beslissingen over onderwijs die door de lokale overheid gemaakt worden (plaats van besluitvorming). <u>Culturele aspecten in relatie tot onderwijs</u> <u>Ongelijkheid in land of regio</u> Gini index (berekent de mate waarin de distributie van inkomen (of consumptie) onder individuen of huishoudens in een land afwijkt van een perfecte gelijke verdeling).	<u>Functionele decentralisatie</u> (De) centralisatie in het curriculum, primaire onderwijsproces en (financieel) management. <u>Evaluatie en accountability</u> Verscheidenheid aan methoden. Institutionele infrastructuur. Hoge of lage consequenties bij evaluatie en verantwoording (High/low stakes). Mate waarin een systeem een op standaarden gebaseerd examen heeft in het voortgezet onderwijs. <u>Structurele verscheidenheid voortgezet onderwijs</u> *Structurele voorwaarden* - Leeftijd van eerste selectie voor voortgezet onderwijs. - Verscheidenheid aan school-types op voortgezet onderwijs niveau (openbaar versus categoraal). - Percentage scholen die leerlingen op basis van capaciteit.	<u>Investeringen in onderwijs</u> Percentage onderwijs/BNP. Uitgaven per leerling volgens ISCED. Lerarensalarissen. <u>Selectie en training leraren</u> Verplicht opleidingsniveau Professionele ontwikkeling <u>Beleid rond kansengelijkheid</u> Maatregelen voor achterstandsleerlingen - Proportie van onderwijs budget bestemd voor speciale programma's. - Voorkeursbeleid voor specifieke groepen. - Strategieën om onderwijsongelijkheid te compenseren.

Het blok door het beleid beïnvloedbare factoren in tabel 4.1 dat tot nu toe de sterkste invloed lijkt te hebben op nationale onderwijsprestaties is 'Evaluatie en accountability arrangementen'. Hieronder worden relevante onderzoeksresultaten vermeld. In de volgende paragraaf (4.5) wordt een nadere typering gegeven van accountability.

Onderzoek naar de prestatiebevorderende werking van examens

Bishop (1997) concludeerde op basis van een secundaire analyse van TIMSS-data dat landen met een op standaarden gebaseerd eindexamen door de bank genomen beter presteerden dan landen zonder zo'n examen. Later werd deze conclusie bevestigd door onderzoek van Woessmann (2001) en van Fuchs en Woessmann (2004). In de Verenigde Staten concludeerden verschillende onderzoekers dat staten met een streng accountabilitysysteem beter presteerden dan staten zonder zo'n systeem, of met een milder systeem (Rand News Release, 2000, Carnoy, Elmore & Siskin, 2003). Woessmann en anderen, 2009, vonden positieve resultaten voor sommige evaluatieprocedures, maar niet voor andere. Zij concluderen (op basis van analyses gebaseerd op PISA 2003) dat leerlingen uit landen met een centraal examen beter scoren, met een effect dat te vergelijken is met een half jaar studievoorsprong (p. 30). Tevens vermelden zij een positief effect van evaluatie en accountability op kansengelijkheid. Ook Bol, Witsche, Van de Werfhorst en Dronkers (2014) laten zien dat de aanwezigheid van een centraal examen een matigend effect heeft op ongelijkheid in sterk gestratificeerde onderwijsstelsels.
Het positieve effect van centrale examens wordt opnieuw bevestigd in recentere overzichtsstudies van Woessmann (2016, 2018). Overigens vinden niet alle onderzoeken positieve effecten van examens en accountability (Scheerens et al., 2015). Verder zijn er onderwijsstelsels aan te wijzen die geen centraal examen hebben en toch zeer goede prestaties laten zien. Vlaanderen is een voorbeeld.

4.4 Nadere analyse van de mechanismen die de prestatiebevorderende werking van examens verklaren

Bevorderen van de extrinsieke motivatie van leerlingen.

Economen als Woessmann benadrukken de sterke belangen die examens hebben voor leerlingen. Examens stimuleren de extrinsieke motivatie van leerlingen om goed hun best te doen. Volgens Woessmann (2018) zorgen examens er ook voor dat er tegendruk wordt geboden ten opzichte van groepsdruk onder leerlingen om zich op school niet uit te sloven.

Accountability.

De resultaten van examens en eindtoetsen kunnen (op geaggregeerd niveau) ook gebruikt worden om leerkrachten en scholen te beoordelen. Hierbij hangt het van het administratieve regime en de beleidscultuur af door wie en hoe streng de beloningen en sancties zijn die aan de uitkomsten worden verbonden. In de Engelstalige literatuur wordt de term 'accountability' gebruikt voor deze externe evaluatie.
In algemene zin gaat het bij accountability om het ter verantwoording stellen van overheidsorganisaties en diensten, op basis van de kwaliteit van de opbrengsten die ze leveren. Glass (1972) stelt dat accountability bestaat uit drie losjes verbonden kernprocessen: (a) openbaar maken van de producten en diensten die geleverd worden; (b) het waarderen van die producten en prestaties en (c) het toekennen van sancties in het geval de prestaties onder de maat zijn. Dit derde aspect betekent dat accountability meer is dan het verzamelen van informatie en het waarderen daarvan, maar tevens consequenties kan hebben voor de betrokken organisaties. Dit betekent dat accountability verbonden is met op meritocratische gedachten geïnspireerd beleid, zoals prestatiebeloning van leerkrachten en financiering van scholen, die afhankelijk is van prestatieniveaus. Overigens heeft ook openbaarmaking van de prestaties soms al indirecte consequenties, namelijk in de vorm van reputatieschade.

Verschillende soorten accountability worden onderscheiden op basis van de vraag wie, liever gezegd welke eenheid of stakeholder, verondersteld wordt de informatie die geopenbaard wordt door scholen en leerkrachten te gebruiken, eventueel ook de vraag wie de sancties oplegt.
Elmore en collega's (1990) maken onderscheid tussen drie 'theorieën' over accountability, die gebaseerd zijn op de vraag *wie de informatie gebruikt:*

- technische accountability, waarbij (hogere) administratieve niveaus beslissingen nemen op basis van wetenschappelijk valide en betrouwbare prestatiemetingen;
- het klantgerichte perspectief, waarbij de klanten van het onderwijs, zoals bijvoorbeeld ouders van leerlingen schoolkeuze baseren op prestatie-indicatoren;
- het professionele perspectief, waarbij feedback op basis van prestatiemetingen gebruikt wordt voor professionele ontwikkeling. Elmore et al (ibid) omschrijven dit als volgt: "Accountability is, therefore, to be accomplished by deconstructing and reconstructing the meaning of schooling, collaborative planning, and co-operative teaching and learning."

Uitgaande van de definitie van Glass zou men moeten betwijfelen of 'professionele accountability' wel echt als accountability kan worden gezien, omdat het element van sancties lijkt te ontbreken. Professionele accountability past eerder in theorieën over organisatieleren, zoals bijvoorbeeld het concept van de 'reflective practitioner' ontwikkeld door Argyris and Schön (1974) en concepten van de 'lerende organisatie' die hier op gebaseerd zijn.

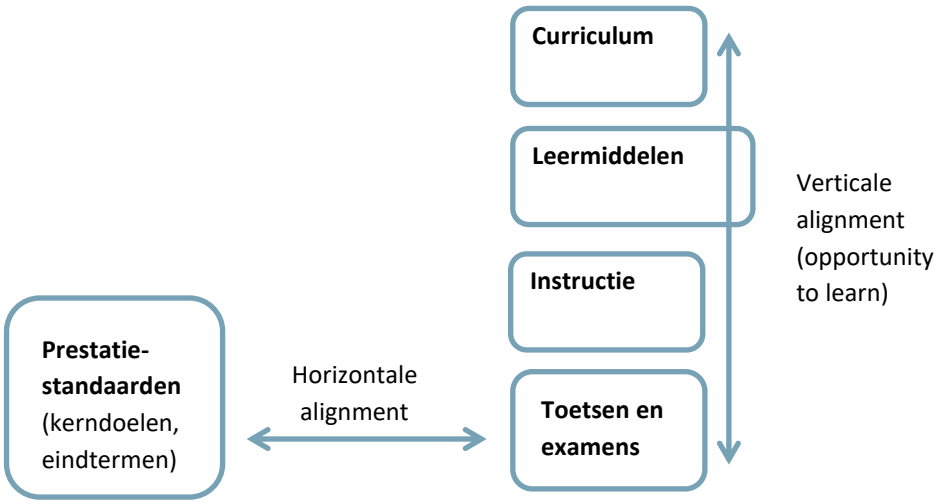
Hierboven zagen we dat er verschillende in aanmerking komende actoren zijn om de beoordeling te gebruiken en incentives toe te passen: de overheid, de 'klanten' van onderwijs en schoolinterne beoordelaars (schoolleiders en collega's). Wanneer de overheid de beoordelaar is, al dan niet gemitigeerd door een onderwijsinspectie, kunnen sancties variëren van financiële prikkels, tot openbaar making van schoolprestaties, die op hun beurt neutraal, dan wel sterk evaluatief gekleurd kunnen zijn (soms uitgedrukt als 'naming en shaming'). Tenslotte past binnen het marktdenken over onderwijs de gedachte dat 'stakeholders', met name ouders die een school kiezen voor hun kinderen, de prestaties van scholen in aanmerking nemen bij de schoolkeuze. Tenslotte kunnen toetsgegevens ook bij interne beoordeling op school gebruikt worden. Dit is een grensgeval tussen 'accountability' en leervormen als 'collegiale consultatie'; het element van sterke prikkels is hierin minder prominent. Er bestaan gerede twijfels over de stimulerende werking van zowel administratieve als 'marktgerichte' accountability. In het eerste geval, omdat 'high stakes' toetsing niet vaak wordt toegepast in de beoordeling van scholen, en in het tweede geval, omdat het zeer de vraag is hoe sterk ouders kwaliteit, in de zin van eerdere schoolprestaties, laten meewegen bij schoolkeuze. Het gebruik van prestatiegegevens van leerlingen bij de beoordeling van leerkrachten is omstreden, evenals het bieden van financiële incentives hierbij. In lijn met de argumentatie van Bosker en Scheerens, (1999) zien wij een 'milde' toepassing van accountability, in de zin van een feitelijke weergave van schoolprestaties in openbare media, als een doelmatige toepassing, die de prestatiemotivatie van scholen stimuleert.

Leren op basis van feedback.
Naast motivatie van leerlingen en scholen is 'leren op basis van feedback' een derde mechanisme dat de positieve effecten van examens en toetsen kan verklaren. Bij externe beoordeling is er naast de motivationele inwerking ook sprake van een cognitieve terugkoppeling. Toets- en examenresultaten geven een beeld van zwakke en sterke punten in het presteren van de school, die gebruikt kunnen worden bij schoolzelfevaluatie en het aangeven van targets voor verbetering.

Het bieden van curriculaire focus

Er is nog een vierde mechanisme, dat, na een overtuigende introductie in de jaren zeventig en tachtig van de vorige eeuw, decennia lang in vergetelheid is geraakt. Het gaat om *curriculum alignment*, een goede aansluiting tussen centrale onderwijsdoelen en de wijze waarop de realisering hiervan getoetst wordt, door middel van examens en eindtoetsen. Overigens wordt de koppeling tussen doelstellingen en toetsen/examens ook al min of meer expliciet gemaakt in de wijze waarop auteurs als Bishop en Woessmann examens typeren. Zij spreken uitdrukkelijk van 'standard based' examinations, waarbij de koppeling tussen standaarden en toetsen dus al verondersteld wordt. Tegelijkertijd bestaat overigens de indruk dat er in het empirisch onderzoek niet erg selectief gekeken is naar de vraag of de examens inderdaad aan deze eis voldoen. Het belangrijkste is dat examens materie- of vakgebonden zijn en niet het karakter hebben van algemene cognitieve vaardigheidstests.

De Engelse term 'alignment' betekent letterlijk 'in lijn brengen'. Het begrip wordt gebruikt om aan te duiden in hoeverre onderwijsdoelstellingen, curricula en assessment en/of examenprogramma's op elkaar zijn afgestemd. Oorspronkelijk werd het begrip vooral gebruikt om de afstemming tussen onderwijsdoelstellingen en assessment aan te duiden (Bloom, Madaus, Hastings, 1981; Tyler, 1949; Webb, 1997). We noemen dit 'horizontale' alignment. In recentere toepassingen (Case, Jurgensen & Zucker, 2004) wordt een meer 'verticale' invulling van alignment gegeven, waarbij diverse curriculaire componenten op verschillend niveau binnen onderwijsstelsels worden meegenomen: leerboeken en methoden, schoolwerkplannen, formatieve toetsen, instructie in de klas en summatieve toetsen of examens. Dit laatste segment, de overeenkomst tussen gegeven en getoetst onderwijs komt overeen met wat wordt aangeduid als 'opportunity to learn'. Een integraal model van verticale curriculum alignment, gebaseerd op De Groot's begrippen 'didactische en evaluatieve specificatie' (De Groot, 1986) is weergegeven in Scheerens en Exalto, (2017). In het onderstaande schema geven we een vereenvoudigde weergave van horizontale en verticale alignment.



Figuur 4.1 Horizontale en verticale alignment (vereenvoudigde weergave)

In publicaties van de OECD (OECD, 2010) en McKinsey (Mourshed et al., 2010) wordt alignment als een systeemkenmerk gezien dat kenmerkend is voor goed presterende onderwijsstelsels. Een interessante toevoeging is het concept 'social alignment' (Looney, 2011). Hiermee wordt enerzijds gerefereerd aan meer horizontale samenwerkingsvormen tussen en binnen scholen, is er minder sprake van een

uitdrukkelijke gerichtheid op curriculaire elementen en gaat het deels om bestuurlijke samenwerking, gedeelde perspectieven, vertrouwen en agogische relaties.
Het denken over curriculum alignment onderstreept dat examens en eindtoetsen een belangrijke inhoudelijk richtinggevende werking hebben, die ervoor kunnen zorgen dat er consequent en consistent gewerkt wordt aan de realisatie van vooraf opgestelde doelstellingen en standaarden. In de volgende paragraaf wordt de optiek van curriculum alignment nader toegelicht en in verband gebracht met de onderwijskundige context in Nederland, wat betreft het functioneren van eindtermen en referentieniveaus.

4.5 Curriculum alignment: 'teaching to and from the test'

In deze paragraaf wordt de plaatsing van eindtoetsen en examens in een ruimer kader van curriculum alignment nader toegelicht vanuit verschillende perspectieven: resultaten van onderwijseffectiviteit onderzoek, curriculumtheorie, functies van examens en toetsen en sturingsconcepten (pro-actieve en retro-actieve sturing). De inhoud van deze paragraaf is een bewerking van een onderdeel uit het rapport 'Teaching to and from the test', Scheerens en Exalto, 2017, p16-17).

Onderwijseffectiviteitsonderzoek naar Opportunity to Learn (OTL)
Een van de laatste schakels in een model van verticale alignment is de aansluiting van het gegeven onderwijs bij inhoud van toetsen en examens, alsmede het effect van die aansluiting op de daadwerkelijk gerealiseerde onderwijs leerresultaten. Dit staat in de onderzoeksliteratuur bekend als 'opportunity to learn' (OTL).
Hoewel de gemeten effecten van OTL sterk afhankelijk zijn van de methode waarmee OTL gemeten is (vgl. Luyten en Scheerens, 2018) steekt de effectgrootte niet ongunstig af in vergelijking met de effecten van andere effectiviteitbevorderende factoren (leertijd, frequente toetsing, ouderbetrokkenheid, een prestatiegerichte cultuur; Scheerens, (2017). In ieder geval is het effect van OTL veel sterker dan het effect van schoolleiderschap en de samenwerking tussen docenten.
Sterke effecten zijn betrekkelijk zeldzaam in onderwijseffectiviteitonderzoek. Veel factoren hebben ongeveer dezelfde, betrekkelijk kleine effecten. Een mogelijke verklaring hiervoor is dat veel specifieke factoren in feite representaties zijn van algemenere principes. Een voorbeeld op het terrein van het onderzoek naar effectieve instructie: meermalen is geconstateerd dat instructiepraktijken, die gebaseerd zijn op duidelijk verschillende leertheorieën (behaviorisme en cognitieve psychologie) ongeveer even grote effecten hebben. Louis (2010) heeft in dat kader het begrip 'focused teaching' voorgesteld, waarin zowel elementen van constructivistisch onderwijs en meer gestructureerd onderwijs (direct teaching) zijn opgenomen. Hattie (2009) doet in feite hetzelfde met het concept 'active teaching'. Onderliggende kenmerken van goed onderwijs zijn focus (doelgerichtheid), didactisch bewustzijn van docenten en een repertoire waarin structuur en zelfstandig werken beide aanwezig zijn.
Twee andere kernfactoren waarop de resultaten van effectiviteitsonderzoek kunnen worden samengevat zijn curriculum alignment en het cybernetisch principe van evaluatie en feedback. Als er nog verder geabstraheerd wordt zou men zelfs kunnen denken aan een combinatie van curriculaire focus, evaluatie en feedback en rationele actieplanning als belangrijkste kenmerken van effectief onderwijs.

Curriculumtheorie

Een belangrijk conceptueel kader uit de curriculumtheorie is het onderscheid tussen het beoogde, het geïmplementeerde en het gerealiseerde curriculum. Het beoogde curriculum is een plan, dat meer of minder tot in detail gespecificeerd kan zijn. Het plan bestaat uit doelen en middelen of methoden om die doelen te bereiken. Tot het plan kan ook behoren dat de doelen worden geconcretiseerd tot beoogde opbrengsten, die vervolgens kunnen worden uitgewerkt tot toetsen en/of examens. Bij implementatie gaat het erom dat het beoogde of geplande curriculum ook daadwerkelijk in praktijk wordt gebracht. In het geval van ver uitgewerkte en gespecificeerde curricula wordt implementatie beoordeeld op de getrouwheid van het in praktijk brengen van het beoogde curriculum. In het Engels heet dit de 'fidelity'-benadering van implementatie. Wanneer het beoogde curriculum minder ver is gespecificeerd, is er meer ruimte voor interpretatie door de onderwijspraktijk. Deze benadering wordt aangeduid als de 'mutual adaptation'-benadering en geniet al decennia de voorkeur, omdat de mogelijkheid geboden wordt dat de plannen op een zodanig wijze worden uitgevoerd dat rekening kan worden gehouden met de lokale situatie. Hierbij spelen overigens ook meer ideologische overwegingen over de autonomie en de 'empowerment' van leerkrachten een rol. Ball e.a. (2012) gaan nog verder en nemen aan dat er altijd sprake moet zijn van 'enactment' oftewel co-constructie van het curriculum door docenten. Wat ook meespeelt zijn opvattingen over enerzijds de rol van de leerkracht als autonome professional en anderzijds de invloed van artefacten als planningsdocumenten, methoden en evaluatie-instrumenten. Ook speelt de nationale context een rol, in de mate van centralisme of decentralisatie van het bestel.

Het gerealiseerde curriculum bestaat op papier uit de toetsen en/of examens die het resultaat zijn van de operationalisatie van de doelen, in de zin van beoogde onderwijsopbrengsten. In de praktijk zijn dit de leerresultaten zoals scores van leerlingen op eindtoetsen of examens.

Opgemerkt kan worden 'alignment' impliciet aanwezig is in het conceptuele kader van planning/ implementatie/realisatie van het curriculum. Het spreekt vanzelf dat het de bedoeling is dat de implementatie aansluit bij het plan, en hetzelfde geldt voor toetsen en examens. Doelstellingen vormen de centrale categorie waarom alles in feite draait. De conceptuele benadering in kwestie is dan ook niets anders dan een invulling van het rationele planningsmodel.

De centrale plaats van doelstellingen blijkt ook uit de vergelijking die A.D. de Groot (1998) maakte tussen het proces van toetsontwikkeling enerzijds, en het proces van curriculumontwikkeling anderzijds. De Groot beschrijft twee deductieve processen, waarbij onderwijsdoelstellingen vanuit tweeërlei optiek worden uitgewerkt: enerzijds met het oog op toetsontwikkeling (evaluatieve specificatie) en anderzijds met het oog op curriculumontwikkeling (didactische specificatie). Interessant zijn intermediaire producten van deze twee specificatie-processen in de vorm van subdoelstellingen, hoofdonderdelen van een schoolvak, thema's en subthema's, lessenplannen en toetsmatrijzen. De Groot verdedigt de opvatting dat de evaluatieve specificatie beter voorafgaand aan de didactische specificatie kan plaatsvinden, omdat de curriculumontwikkeling op die manier betrokken blijft op de operationele doelstellingen en zich niet verliest in de keuze van middelen en methoden, alleen omdat die mooi of aardig zouden lijken. Een nadere bespreking van dit model van De Groot wordt gegeven in Scheerens, (2016) en Scheerens en Exalto (2017).

'Alignment' op systeemniveau: proactieve en retroactieve structurering

Zoals al is opgemerkt werkt curriculumspecificatie volgens de curriculumtheorie proactief. Eerst worden doelstellingen geformuleerd en vervolgens worden er onderwijsmethoden en evaluatiemethoden bij gezocht. Hoewel deze logica van het rationele planningsmodel onontkoombaar is, kunnen er door praktische omstandigheden uitwerkingen plaatsvinden die de zaak als het ware op zijn kop zetten. In de

praktijk werkt de proactieve aanpak vaak moeizaam en is het moeilijk om vanuit brede idealistische doelen te komen tot concrete uitwerkingen. Een voorbeeld is de schoolwerkplanontwikkeling die enige tijd populair was in het Nederlandse onderwijs. Vaak bleken de papieren plannen gedoemd om een stille dood te sterven in bureauladen. Ook kunnen er principiële motieven zijn om op nationaal niveau terughoudend te zijn met het formuleren van specifieke onderwijsdoestellingen. Nederland is daar in zekere zin een voorbeeld van, met zijn aversie voor staatspedagogiek en zijn traditie van soevereiniteit in eigen kring en subsidiariteit. Uit OECD-indicatoren blijkt al jaren dat scholen in Nederland ongeveer het meest autonoom ter wereld zijn. In dergelijke gevallen is denkbaar dat examens en eindtoetsen, te beschouwen als ver geoperationaliseerde onderwijsopbrengsten, als richtinggevend gaan functioneren voor andere partners in de keten van uitwerking van plannen, construeren van onderwijsmethoden (leerboeken en ict-pakketten) en formatieve toetsen. In dat geval zou men kunnen spreken van een retroactieve structurering, waarin toets- en examenvoorbereiding een belangrijke plaats hebben. Ook dat, zou men kunnen zeggen, is een voorbeeld van optimalisering van OTL, niet zozeer door de toetsen bij het onderwijs aan te passen, maar door het onderwijs aan te passen bij de toetsen. Vanuit een dergelijke retroactieve optiek wordt alignment en OTL geoptimaliseerd door alle ballen te richten op de examenuitkomsten.

Deze weergave van de functie van examens stuit bij sommigen op kritiek, vanwege huiver voor ‘teaching to the test’. Het focussen op examens wordt gezien als reductie en tunnelvisie. Deze kritiek komt in het volgende hoofdstuk uitvoerig aan de orde.

5 Kritiek op examens en toetsen

Jaap Scheerens en Anne Luc van der Vegt

5.1 Inleiding

In dit hoofdstuk wordt in de eerste plaats stilgestaan bij twee soorten van kritiek die in Nederland op de voorgrond treden: in de eerste plaats de opvatting dat examens en toetsen leiden tot 'versmalling' van het onderwijs en in de tweede plaats het vaak gehoorde commentaar dat er sprake is van 'toetsgekte' en 'doorgeschoten rendement denken' (5.2.) Vervolgens wordt ingegaan op onderzoek naar ongunstige neveneffecten van examinering en toetsing (5.3). In de slotparagraaf wordt de kritiek nader geanalyseerd en becommentarieerd (5.4).

5.2 Kritiek die voortkomt uit de roep om een 'brede' kijk op onderwijskwaliteit

Recente adviezen van de Onderwijsraad

In het advies 'Een smalle kijk op onderwijskwaliteit' (2013) stelt de onderwijsraad dat er sprake is van eenzijdige aandacht voor meetbare doelen en prestatieverhoging. "4Ze krijgen ook bepaalde waarden en normen mee, én culturele bagage. De raad constateert dat in de afgelopen periode de aandacht eenzijdig gericht was op meetbare doelen, in het bijzonder op taal en rekenprestaties. Er was veel minder beleidsaandacht voor het bredere vakkenaanbod en voor algemene vorming, en (in het middelbaar beroepsonderwijs) voor de beroepspraktijkvorming. Meetbare doelen zijn de maatstaf geworden voor de kwaliteit van het onderwijs. De raad roept op discussie te voeren over wat het Nederlandse onderwijs leerlingen en studenten moet meegeven. Met het vaststellen van heldere doelen kunnen overheid en instellingen doelgerichter werken aan onderwijskwaliteit" (Onderwijsraad, 2013, p. 4). De Raad spreekt van een trend om in de publieke sector sterker te sturen en te controleren op gerealiseerde resultaten. Tegelijkertijd is de raad van mening dat, hoewel de basiskwaliteit van het onderwijs op orde is, verdere verbetering en vernieuwing stagneert:

"Scholen worden niet of nauwelijks uitgedaagd om 'goed' of 'uitstekend' te presteren, of om hun aanbod 'te vernieuwen'. De raad vindt dat de 'eenzijdige aandacht' op basiskwaliteit (taal en rekenen) een verschraling is. Onderbelicht zijn de kennis en vaardigheden in andere vakken (geschiedenis, economie, filosofie, cultuureducatie enzovoort), maar ook sociale competenties, burgerschapsvorming en de ontwikkeling van vakoverstijgende 'advanced skills' als problemen oplossen, samenwerken, communiceren en ict-geletterdheid" (p. 4-5). De raad is van mening dat er te weinig visie is op de doelen van het onderwijs. Tevens is de raad van mening dat er te weinig ruimte is voor instellingen om te variëren en te vernieuwen. Tenslotte vreest de raad dat de eigenwaarde van niet goed presterende leerlingen onder druk komt te staan. "De nadruk op cognitieve prestaties dreigt ten koste te gaan van leerlingen en studenten die niet goed scoren op basisvaardigheden. Deze jongeren kunnen zich ondergewaardeerd voelen. Meer aandacht voor de brede doelstelling van onderwijs kan dit voorkomen" (p. 6).

De raad ziet de volgende aangrijpingspunten om de geschetste situatie te verbeteren. In de eerst plaats moet er niet alleen gestreefd worden naar bredere kwaliteit, maar deze zou ook inzichtelijk gemaakt moeten worden. Hierbij wordt onder meer gedacht aan de beschrijving van het leerproces. Scholen zouden ook nieuwe kennis moeten ontwikkelen over 'wat er werkt' in het onderwijs. In de tweede plaats wordt de overheid aangeraden om te sturen op hoofdlijnen en een groter appel te doen op de

professionele inbreng vanuit het onderwijsveld. In de derde plaats zou er gezorgd moeten worden voor een hogere waardering van niet-cognitieve capaciteiten.
Hoewel de raad van mening is dat er in de sfeer van beoordeling en verantwoording sprake is van een 'te smal kwaliteitsoordeel', worden in dit advies geen concrete voorstellen gedaan om het bestaande stelsel van eindtoetsen en examens te veranderen. Wel wordt gepleit voor lokale gegevensuitwisseling en kaders voor scholen om de eigen kwaliteit zichtbaar te maken, zoals bij het project 'Scholen op de kaart'.

In het advies 'De volle breedte van onderwijskwaliteit' (2016) fundeert de Onderwijsraad de 'brede visie' nader, waarbij de raad zich onder meer baseert op kernfuncties van onderwijs (zie de bespreking hiervan in hoofdstuk 2 van dit rapport). In het advies uit 2016 wordt de verbreding aangevuld met moeilijk grijpbare pedagogische elementen, waardoor toetsing en verantwoording van bepaalde facetten nog kwetsbaarder wordt, zo wordt er onder meer gezocht naar een term die tussen 'merkbaar' en 'meetbaar' ligt. Ook in dit advies blijven implicaties van de brede visie op kwaliteit voor de bestaande examens en eindtoetsen achterwege, en wordt vooral gepleit voor lokale procesbeschrijvingen en procesindicatoren in het Inspectietoezicht. "Het breder opvatten van kwaliteit gaat volgens de raad samen met het breder verantwoorden daarvan. Deze verantwoording vindt plaats op het niveau van horizontale en verticale verantwoording. De eigen visie van de school op brede onderwijskwaliteit staat hierbij centraal. De raad is daarbij voorstander van procestoezicht aan de hand van een aantal ijkpunten in het toezichtkader" (p. 34). Realisatie van de brede visie op onderwijskwaliteit, maar ook de verantwoording daarvan wordt, onder erkenning van de rol van de Inspectie van het Onderwijs grotendeels op het bord van de onderwijsinstellingen geschoven: "Met inachtneming van de wettelijke voorschriften is het uiteindelijke de school die kiest wat te realiseren. Het vertrekpunt van (brede) verantwoording is de visie van de school op onderwijskwaliteit in brede zin" (p. 37). Uit het advies in kwestie zou men zelfs kunnen opmaken dat de brede visie op onderwijskwaliteit buiten het kader van meetbare opbrengsten moet blijven:
"In dat verband wijst de raad er uitdrukkelijk op dat gewaakt moet worden voor het aanvullen van verantwoordingskaders met standaarden aan de hand waarvan verantwoording over onderwijskwaliteit in brede zin kan worden gevraagd. Een set met 'af te vinken' aspecten brengt het risico met zich mee dat scholen zich hierdoor laten sturen, met als gevolg dat zij de aandacht gaan richten op zaken die daarvoor in kaart moeten worden gebracht in plaats van de eigen visie te ontwikkelen en in de praktijk te brengen. Door pogingen te doen onderwijskwaliteit in brede zin te beheersen en te vangen in kaders, wordt niet alleen de verantwoordelijkheid van de school voor het realiseren van deze kwaliteit beknot, maar wordt ook de ruimte voor het ontwikkelen en het hanteren van een brede opvatting van kwaliteit beperkt (p. 38)."

Op basis van de hierboven besproken adviezen van de onderwijsraad zou men verwachten dat er een duidelijker stellingname zou zijn tegenover de vigerende examens en eindtoetsen. Met zoveel woorden wordt hier en daar gesuggereerd dat het accent op meetbare opbrengsten de bekritiseerde smalle kijk op kwaliteit in stand houdt. Tegelijkertijd kiest men expliciet niet voor verbreding van het scala van mogelijk meetbare opbrengsten en wordt ingezet op lokale evaluatie en procestoezicht bij de onderwijsinspectie. In termen van het analytisch kader dat we presenteerden in hoofdstuk 3: de Onderwijsraad wil bij het vaststellen van de kwaliteit van het onderwijs naast *output*-indicatoren ook nadrukkelijk *proces*- of *throughput*-indicatoren een plaats geven.
Hoewel de Onderwijsraad met deze analyses de bal in het doel van afslanking van examens en toetsen voor het inkoppen legt, is dat tot nu toe niet gebeurd. Eind 2018 heeft de Onderwijsraad opnieuw een

advies uitgebracht over toetsing en examens, 'Toets Wijzer'. In dit advies wijst de raad op de voordelen van gestandaardiseerde eindtoetsing bij belangrijke overgangen in het stelsel. Een standpunt van de raad dat centraal staat in het advies is dat enkele kernfacetten van toetsing en examens meer in evenwicht zouden moeten worden gebracht.
De raad onderkent spanning in drie dimensies:

- De mate waarin toetsing een beslissende of formatieve functie heeft;
- De mate waarin toetsing op decentraal niveau of (meer) centraal niveau wordt vormgegeven en
- De mate waarin wordt gestreefd naar kwantitatieve meting of naar het meer op kwalitatieve wijze zichtbaar maken van onderwijsopbrengsten.

Volgens de raad komt formatieve toetsing nog te weinig aan bod. Op de dimensie centraal-decentraal signaleert de raad zowel onder- als overbenutting. Tenslotte breekt de raad een lans voor het gebruik van kwalitatieve methoden.

Het advies van het Platform Onderwijs2032

In het eindadvies van het Platform Onderwijs2032 (2016) wordt de gedachte van een bredere invulling van onderwijskwaliteit nader onderbouwd. In dit advies gaat men verder dan de Onderwijsraad. Het Platform trekt de consequentie dat verbreding op terreinen als socialisatie en persoonlijke vorming gepaard moet gaan met beperking en afslanking van het cognitieve onderwijsprogramma. Ook worden aan de voorstellen tot ontwikkeling van 'brede kwaliteit' duidelijke implicaties verbonden voor toetsing en examinering. Hieronder wordt een deel van de samenvatting van het rapport geciteerd:
"Het is duidelijk dat er een nieuwe koers in het onderwijs nodig is om leerlingen die nu voor het eerst naar school gaan de kennis en de vaardigheden mee te geven die ze nodig hebben wanneer ze in 2032 aan hun volwassen en werkende leven beginnen. Het Platform onderscheidt een aantal kenmerken van gewenst toekomstig onderwijs, waaronder een grotere nadruk op persoonsvorming (naast kennisontwikkeling en maatschappelijke vorming het derde hoofddoel van het onderwijs). Met een beter evenwicht tussen deze doelen kan het onderwijs leerlingen begeleiden in hun ontwikkeling tot zelfstandige volwassenen die vaardig, waardig en aardig zijn, voor zichzelf en voor hun omgeving.
Om deze visie op toekomstgericht onderwijs mogelijk te maken, pleit het Platform voor een vaste basis van kennis en vaardigheden die zich beperkt tot datgene wat alle leerlingen ten minste nodig hebben voor vervolgonderwijs en om in de maatschappij te kunnen functioneren. Door die basis te beperken en vast te leggen in een kerncurriculum, krijgen leraren meer ruimte om hun onderwijsaanbod in te richten naar de behoeften, ambities en persoon van hun leerlingen.
Het Platform beschouwt Nederlands, Engels, rekenvaardigheid (inclusief wiskunde), digitale geletterdheid en burgerschap als verplichte onderdelen van het kerncurriculum. Dat omvat daarnaast kennis die leerlingen nodig hebben om de wereld te kunnen begrijpen en eraan bij te dragen. Om het onderwijs meer betekenis voor leerlingen te geven, stelt het Platform voor die kennis in drie leerdomeinen te clusteren: Mens &Maatschappij, Natuur & Technologie, Taal & Cultuur. Leerlingen maken zich de kennis van die domeinen op een diepgaande manier eigen: niet van alles een beetje, maar meer van minder. Ze leren kennis uit verschillende vakken met elkaar in verband te brengen aan de hand van maatschappelijke vraagstukken. Scholen brengen hun leerlingen behalve kennis ook vakoverstijgende vaardigheden bij, die eveneens tot de vaste basis behoren. Het gaat om leervaardigheden, creëren, kritisch denken, probleemoplossend vermogen en samenwerken. Het kerncurriculum biedt de basis van waaruit scholen werken aan een uitdagend en relevant aanbod voor hun leerlingen. Scholen maken keuzes voor verbreding en verdieping van het aanbod die het best passen bij hun visie, de leerlingen en hun ouders en de professionaliteit van hun leraren. Verdieping en

verbreding zijn niet vrijblijvend, maar verplicht. Scholen kunnen hun onderwijsaanbod invullen door nauw samen te werken met de buitenwereld, waaronder het bedrijfsleven, maatschappelijke en culturele instellingen en sportverenigingen.
Een andere onderwijsinhoud vraagt om herijking van kerndoelen en eindtermen. De bestaande kerndoelen geven leraren te weinig richting en houvast. Het Platform wil een afgebakend, wettelijk verankerd kerncurriculum en een keuzedeel dat past bij de school en de leerling. Het kerncurriculum schept een basis voor een samenhangend onderwijsaanbod. Versterking van de doorlopende leerlijn en niveaudifferentiatie zijn aandachtspunten voor de uitwerking van het kerncurriculum.
Curriculumvernieuwing komt niet van de grond zolang de manier van toetsen en examineren niet wordt aangepast. Toetsen en examens moeten de gewenste onderwijsinhoud weerspiegelen" (p. 8-9).

De opvatting dat er teveel nadruk ligt op de kwalificerende functie van onderwijs en dat dit verankerd zit in het stelsel van eindtoetsen en examens wordt duidelijk verwoord. Waar de Onderwijsraad A (onderwijs moet breder) en B (meer nadruk op socialisatie en persoonlijke ontwikkeling) zei, zegt het Platform 2032 duidelijk ook C (minder accent op cognitieve kern, daarmee ruimte voor socialisatie en persoonlijke ontwikkeling).
"Momenteel bepalen de inhoud van de centrale eindtoets in het primair onderwijs, de determinerende toetsen in de onderbouw van het voortgezet onderwijs en het centraal examenprogramma in hoge mate de inhoud van het onderwijs. Door de grote nadruk die daardoor op het kwalificerende doel van het onderwijs komt te liggen, is er weinig aandacht voor de socialiserende functie van het onderwijs en voor persoonsvorming"(p.28).

Net als de Onderwijsraad pleit Platform Onderwijs2032 ervoor dat naast output-indicatoren ook procesindicatoren een belangrijke rol spelen bij kwaliteitsbeoordeling door de Inspectie van het Onderwijs. "Toezicht moet niet alleen gaan over de resultaten die leerlingen op centrale examens en schoolexamens behalen, maar ook over de manier waarop scholen de kwaliteitszorg voor hun onderwijs ter hand nemen en hun onderwijsaanbod verantwoorden" (p. 57).

Het advies van de VO-raad over flexibilisering van het eindexamen (2018)
In het Sectorakkoord tussen het Ministerie van OCW en de VO-raad, daterend van 2014, is sprake van "meer flexibiliteit en ruimte in de inhoud en omvang van onderwijsprogramma's en de toetsing en examinering hiervan" (OCW en VO-raad, 2014 p. 8). Ook 'brede vorming' wordt in dit akkoord benadrukt: "Hoewel het behalen van een VO-diploma een belangrijk doel is en blijft voor het voortgezet onderwijs, biedt voortgezet onderwijs leerlingen zoveel meer dan een diploma of cijferlijst" (p. 11). In het advies wordt voorgesteld om het 'plusdocument' in te stellen als een mogelijkheid om brede vorming van leerlingen zichtbaar te maken. Net als de Onderwijsraad en Onderwijs2032 relativeert de VO-raad het belang van het centrale examen. Onderwijs is zoveel meer dan voorbereiding op het examen, wordt gesteld. De voorstellen van de VO-raad gaan echter in een andere richting: *flexibilisering* van het examen. Daarin verschilt de VO-raad van inzicht met de Onderwijsraad. In het advies van de Onderwijsraad uit 2015 'Maatwerk binnen wettelijke kaders' reageert de raad terughoudend over voorstellen tot flexibilisering van examens en eindtoetsen. "De Onderwijsraad adviseert de wettelijke kaders rond eindtoetsing niet te verruimen. In het huidige onderwijsstelsel vormt de eindtoetsing in het primair en voortgezet onderwijs een belangrijk ijkpunt" (p. 7). Zie ook de uitvoerige referentie aan dit Advies van de Onderwijsraad in hoofdstuk 3 van dit rapport.
Kennelijk was de VO-raad niet overtuigd van de argumentatie van de Onderwijsraad in 2015, want in 2018 komt de raad opnieuw met voorstellen tot herziening en flexibilisering van eindtoetsen en

examens. De notitie 'Examinering voortgezet onderwijs toe aan herijking', (VO-raad, 9 maart 2018) illustreert dit.
Uit het persbericht: "Er komen veel signalen uit het onderwijsveld dat de examensystematiek voor leerlingen en docenten knellend is. In de bovenbouw van het vo gaat bovenmatig veel tijd op aan de voorbereiding voor het examen, en het fenomeen examentrainingen heeft een hoge vlucht genomen. De nadruk verschuift zo van de inhoud van het onderwijs, naar hoe je een examen maakt ('teaching to the test'). Dit gaat ten koste van diepgang in het onderwijs en het leren van leerlingen. Ook ligt de focus in de examinering te eenzijdig op cognitieve ontwikkeling, terwijl het eigen maken van vaardigheden, socialisatie en persoonsvorming belangrijker worden in de maatschappij en het vervolgonderwijs. Dat zorgt voor uitholling van de waarde van het diploma".
Paul Rosenmöller, voorzitter van de VO-raad: "Waar de examinering het onderwijs zou moeten volgen, is het in de praktijk zo dat het het onderwijs stuurt. Dat leidt vooral in de bovenbouw tot een doorgeslagen focus op het leren voor het examen en het remt de ontwikkeling naar eigentijds onderwijs met meer aandacht voor brede vorming en maatwerk. Het examen lijkt een doel op zichzelf geworden, in plaats van een middel om in kaart te brengen of een leerling klaar is voor de vervolgstap. Een gezamenlijke herijking is nodig om de voorspellende waarde van het diploma voor succes in het vervolgonderwijs te versterken."
"Allereerst is nodig dat de examensystematiek, net als het onderwijs en de onderwijsorganisatie, meer flexibel wordt. Zo komen scholen die de mogelijkheid bieden om vakken in verschillend tempo of niveau te volgen met de huidige examensystematiek in de problemen. Daar moet wat de VO-raad betreft op korte termijn verandering in komen. Er moeten meer examenmomenten in het jaar komen, gezakte leerlingen moeten behaalde vakken kunnen behouden en er moet modulair geëxamineerd kunnen worden waardoor het bijvoorbeeld makkelijker wordt om vakken op een hoger niveau af te sluiten".

Naast de kritiek op een te smalle inhoud van toetsen en examens tamboereert de VO-raad sterk op de veronderstelling van "teaching to the test" en wordt opnieuw gepleit voor de flexibilisering, die in 2015 werd afgewezen door de Onderwijsraad, en die ook geen navolging vond in de Kamerbrief OCW over Toetsing en Examinering in het Voortgezet Onderwijs, gedateerd 28 november 2016.

5.3 Kritiek op basis van empirisch onderzoek naar negatieve bijeffecten van "accountability" en "high stakes testing"

Honingh en Ehren (2012) refereren aan onderzoeken, waarin de effecten van strategisch gedrag bij toetsing in het onderwijs zijn onderzocht. "Het simplificeren van prestaties en hanteren van eenzijdige meetbare uitkomsten kan zelfs leiden tot onwenselijke gevolgen zoals ritualisme, bureaucratisering en een eenvormig beeld van kwaliteit, zo blijkt uit onderzoek naar (bij)effecten van prestatiemeting (Behn, 2003; Bevan & Hood, 2006; De Bruijn, 2008; Van Thiel & Leeuw, 2002; Braithwaite, Makkai, & Braithwaite,2007)" (p. 69).
Een nadere analyse van de achtergrond van deze verschijnselen wordt gegeven door Koretz, (2005 en 2016) en Holcomb, Jennings en Koretz, 2012). In de Amerikaanse context is er sprake van 'high stakes testing' in het kader van het programma, No Child Left Behind. Koretz constateert dat er sprake is van 'test inflation' dat wil zeggen kunstmatig hoge toetsuitkomsten op lokale toetsen waarop scholen beoordeeld worden. Hij constateert dit door resultaten van dit soort 'high stakes tests' te vergelijken met resultaten op landelijke peilingen, waaraan voor scholen niet zulke grote belangen verbonden zijn.

'Test inflation', dat wil zeggen kunstmatig opgeschroefde toetsresultaten, ontstaat door gebrekkige inhoudsvaliditeit van toetsen, door strategisch gedrag van scholen en door 'teaching to the test'.

Het argument dat betrekking heeft op de inhoudsvaliditeit van toetsen is het meest fundamenteel. Toetsen worden geacht dekkend te zijn voor de onderwijsdoelstellingen en 'standaarden' waarvan ze de realisatie zouden moeten vaststellen. Koretz (2016) presenteert dit als een probleem van steekproeftrekking. Onderwijsdoelstellingen zou men zich kunnen voorstellen als een universum van inhoudsdomeinen, te demonstreren psychologische operaties en leertaken. Een inhoudsvalide toets wordt gedefinieerd als een representatieve steekproef uit de elementen uit het universum dat hoort bij een bepaalde doelstelling. Volgens Koretz benadert de praktijk van de constructie en jaarlijkse bijstelling van eindtoetsen zelden aan dit theoretische model. De selecties van items die tezamen een toets vormen zijn niet representatief en jaarlijkse bijstellingen zijn vaker oppervlakkige adaptaties van eerdere versies, dan dat er opnieuw een representatieve trekking van items plaatsvindt. Dit betekent dat het onderwijs en de toetsing daarvan structureel versmalt. Holcomb et al. (2012) zeggen hier het volgende over: "Under high-stakes conditions like those created by test-based accountability, educators face strong incentives to focus instruction on the specific content and format of tested material rather than the full domain of knowledge and skills represented in the state standards. *If these aspects of the test are predictable* and teachers focus on them, scores will become inflated, and inferences about mastery of the domain will be undermined" (p.12). Het effect van "score inflation" wordt verder versterkt door wat bekend staat als 'teaching to the test'. Leerkrachten besteden extra aandacht aan de inhouden, waarvan ze vermoeden dat deze deel uitmaken van het examen of eindtoets. Tevens wordt daarbij dan aangenomen dat dit ten koste gaat van andere inhouden, die ook van belang worden geacht. Ook wordt er geoefend op de vorm van toets-items en examenvragen; waarbij overigens de vraag is in hoeverre dit, onder bepaalde voorwaarden, niet als legitieme vorm van examenvoorbereiding gezien kan worden (Scheerens, 2017). Het strategisch gedrag van scholen dat het meest wordt aangehaald is het vertragen of terughouden van zwakke leerlingen, zodat de gemiddelde prestaties van de school hoger worden. In externe schoolevaluaties is dit tegen te gaan door ook gebruik te maken van rendementsindicatoren, zoals het percentage leerlingen dat slaagt voor het eindexamen zonder studievertraging te hebben opgelopen.

Versmalling van het onderwijs, door dit af te stemmen op de kwaliteitsmeting, komt niet alleen voor bij eindtoetsen, maar wordt soms ook verondersteld wanneer scholen zich voorbereiden op inspecties. Interessant is daarbij de kritiek die vroeger wel bestond op scholen die een systeem van zelfevaluatie uitzochten of zelf ontwikkelden dat erg leek op het vroegere toezichtkader van de inspectie, dat gebaseerd was op een aantal procesindicatoren van effectief onderwijs. Een interessante kwestie is ook de vraag of de inhoud van formatieve toetsen zou moeten worden afgestemd op 'summatieve' of eindtoetsen. Vanuit het gezichtspunt van curriculum alignment zou men dit willen stimuleren, maar door auteurs als Koretz wordt dit categorisch afgewezen, omdat het eenzijdigheid en 'versmalling' zou stimuleren.

Het is van belang om goed onder ogen te zien dat de resultaten van Koretz, cs gevonden zijn in de Amerikaanse context, waarin de prestatie druk en implicaties van scholen harder zijn dan in Nederland. Interessant is wat dit betreft wat Koretz zegt over alignment:

"Despite its benefits, alignment is not a guarantee of validity under high-stakes conditions. Even with superb alignment, the unavoidable incompleteness of tests makes them vulnerable to the inflationary effects of reallocation, of which alignment is a special case. Moreover, alignment offers no protection against the corrupting effects of coaching" (p.15).

Verder geeft Koretz een beschrijving van de toetspraktijk in Amerika, waaruit men de indruk krijgt dat de kwaliteit van de toetsen niet hoog is, en onderdoet voor de Nederlandse praktijk, die gevoed wordt door de grote psychometrische know-how bij het Cito en universitaire vakgroepen (vlg. ook Nusche, OECD, 2014).

5.4 Kritiek op basis van een veronderstelde "toets-gekte" en doorgeslagen rendement denken in Nederland

Op 19 oktober 2017 werd er door de Inspectie van het Onderwijs en het Ministerie van OCW een workshop georganiseerd onder de noemer 'Is meten wel weten?' en als subtitel 'Zijn we doorgeschoten in onze behoefte de kwaliteit van het leren uit te drukken in cijfers?' Van deze bijeenkomst verscheen begin 2018 een verslag dat op heldere wijze de stemming ten opzichte van toetsing, externe evaluatie, objectieve cijfers (waar het woord objectief consequent tussen aanhalingsteken is geplaatst) en examens is weergegeven. De inmiddels oude bekende, het verhaal van de smalle kijk op onderwijskwaliteit, keert ook in dit verslag weer terug, maar verder wordt ook vooral het gevoel gepeild, en weergegeven wat 'opiniemakers' (helaas niet nader gespecificeerd) er van vinden. Het verslag is een mooie samenvatting van de kritiek die ook in andere onderdelen van dit hoofdstuk is gedocumenteerd en zal daarom hieronder uitvoerig geciteerd worden.

Het verslag begint met een algemeen samenvattend beeld, waarbij de vraag gesteld wordt, of dit beeld inderdaad klopt.

"In vele toonaarden wordt in het publieke debat het volgende beeld geschetst: het onderwijs gaat gebukt onder het rendementsdenken; de behoefte om onderwijskwaliteit uit te drukken in getallen en die te ordenen tot lijstjes. Het is de dood in de pot, volgens de opiniemakers. Het leidt tot top down management; het heeft de leraar aangetast in zijn autonomie; de intrinsieke motivatie van kinderen verdwijnt als sneeuw voor de zon. Er wordt geschreven over grote 'prestatiedruk', '-pijn' zelfs. De grote boosdoener van dit alles is de overheid (met de inspectie als haar uitvoerder) met haar neo-liberalistische neiging tot beheersing en verantwoording. Een gebrek aan vertrouwen dus."

In een eerste reactie wordt opgemerkt dat er bij de gesprekspartners overeenstemming lijkt te zijn over het belang om de kwaliteit van het onderwijs te bewaken en daarbij cijfers te benutten. Rendementsgetallen en gemiddelden worden als nuttig gezien om aan te geven waar het goed of minder goed gaat. Er lijkt dus niet zozeer sprake van afwijzing van rendementsdenken, maar wel van 'doorgeschoten' rendementsdenken. Wat men daarmee bedoelt wordt uit de volgende citaten duidelijk. Er worden drie problemen gesignaleerd:

"*Probleem 1: Versmalling van het aanbod: alleen wat gemeten wordt, komt aan bod (*' teaching to the test')

Probleem 2: Doorgeschoten toets cultuur Leerlingen in het voortgezet onderwijs krijgen wel tot 180 proefwerken; hierdoor is een fijnmazig plaatsingsproces ontstaan waarin scholen primair op basis van rekenregels bepalen hoe de schoolloopbaan van een leerling eruit ziet. Een ander nadeel van de grote rol van toetsen is dat het invloed heeft op hoe en wat leerlingen leren. *"Ik leer vooral trucjes, ik haal goede cijfers omdat ik snap hoe proefwerken in elkaar zitten"*, aldus een leerling. Het zijn de toetsen die leerlingen aanzetten tot leren. Vooral in het voortgezet onderwijs. Het leren is hiermee extrinsiek gemotiveerd en volgens sommigen beklijft het geleerde dan maar slecht. Het leidt bovendien tot 'prestatiedruk' bij leerlingen". Men spreekt zelfs van 'prestatiepijn'.

Probleem 3: Het maakt scholen risicomijdend 'Neem maar liever het lichtere profiel, dat biedt meer kans op een goed examenresultaat.' Ten tweede maakt het hen behoudend in hun innovaties in het onderwijs. Alles is gericht op het centraal examen en het centraal examen dwingt hen tot een traditionele aanpak".

Vervolgens wordt de vraag gesteld wat de nadelige effecten van het rendementdenken in stand houdt. Dit zijn de antwoorden:

De afrekencultuur door overheid en media (en misverstanden daarover)
De inspectie heeft in het verleden te eenzijdig ingezet op rendementscijfers. De overheid heeft aangezet tot wantrouwen.

Behoefte aan informatie over kwaliteit
Sommige scholen vinden het belangrijk om goed op die lijsten naar voren te komen en sturen sterk op indicatoren als slaagpercentage. Zij hopen dat dit bijdraagt aan een betere concurrentiepositie.

Opwaartse druk
"Ouders en leerlingen vertelden ons dat cijfers en toetsen hen helpen om in de gaten kunnen houden of en hoe zij de schoolloopbaan moeten bijsturen. Ouders zetten scholen dus aan om veel toets- en rendementsgegevens te leveren. Voor de scholen aan de andere kant, bieden de 'objectieve' cijfers houvast om tegendruk te bieden aan deze opwaartse druk door ouders en leerlingen. Voor po-leraren om al vanaf groep zes de hoge verwachtingen van de ouders van Stijn te temperen. Voor vo-leraren om aan Sara duidelijk te maken dat Havo wel erg hoog gegrepen is."

Tot slot - sleutelrol voor leraren
Citaat leraar: "Ik breng nooit actuele gebeurtenissen in tijdens de les, dat kost me alleen maar tijd. Ik heb een examenprogramma af te werken."
Citaat leerling: "Mijn leraren duwen mij door de examenstof, daar leer ik niets van. Ik wil graag een leraar die interesse heeft voor wat ik denk over de stof."
"Volgens sommigen zijn leraren teveel verworden tot 'willoze' uitvoerders binnen een strak ingekaderd systeem."

Tenslotte bespreekt het verslag manieren om op een meer acceptabele wijze om te gaan met meten in het onderwijs; men gaat hierbij uit van de vraag: “Hoe het meten dienstbaar te maken aan goed onderwijs?”:

“Zie cijfers als startpunt voor een reflectief gesprek”
Om cijfers uit de sfeer van controle te halen is het bijvoorbeeld goed om leraren zelf hun kaders te laten maken waarin zij hun eigen rendementsgetallen bepalen. Schoolleiding en lerarenteams evalueren dan samen of zij hun eigen doelen hebben gehaald. Daarmee worden leraren eigenaar van de doelen en zijn cijfers niet langer van boven opgelegde kille getallen. Cijfers stimuleren dan juist de collectieve autonomie van leraren.

Streef naar een bredere verantwoording over kwaliteit

Neem de ruimte die er is
Scholen en teams hebben meer ruimte om het onderwijs in te vullen zoals zij willen dan zij ervaren. Nergens staat dat je als school 150 proefwerken moet afnemen, of dat je perse volgens een methode moet werken en dat je alleen maar de stof voor centraal examen moet aanbieden. Er kan verrassend veel. Lerarenteams kunnen in hoge mate zelf bepalen hoe zij het onderwijs willen inrichten. Aan de invulling van deze autonomie komen zij door gebrek aan tijd en werkdruk te weinig toe. Ook het centraal examen werkt dwingend, daarover hieronder meer.

Herzie het centraal examen
Het centraal examen heeft een groot terugslageffect op wat leerlingen leren, maar dat hoeft geen probleem te zijn. Zolang de inhouden die wij examineren dekkend zijn voor wat wij als maatschappij vinden dat leerlingen moeten kennen en kunnen, is het examen een goede waarborg. Op dit moment ervaren velen het terugslageffect als knellend. Het is te eenzijdig gericht op toetsbare kennis en het is volgens sommigen te veel. Bovendien dwingt het tot vakgericht onderwijs. Scholen die een vakdoorbrekend curriculum aanbieden komen dan in de problemen.

Herstel het schoolexamen in ere
Het schoolexamen geeft scholen de gelegenheid om eigen keuzes een plek te geven in het examen. Het is niet bedoeld als een oefenplaats voor het centraal examen. Scholen moeten deze ruimte (weer) nemen. Het schoolexamen is ook een goede de plek om 21e-eeuwse vaardigheden te belonen. Overheid, sectorraden en besturen kunnen scholen hierin actief stimuleren.

De volgende kanttekeningen zijn bij dit verslag te maken:
- De vraag of er meer objectieve evidentie bestaat voor de kritische waarderingen van de verschillende stakeholders, die zijn opgetekend, en die misschien vooral afkomstig zijn van niet nader geïdentificeerde opinieleiders.
- Ervaringen in relevante onderzoeksprojecten (Oomens, Exalto, De Jong, Scholten, Veldkamp, Janse, Scheerens, 2017; Scheerens & Exalto, 2017) over constructief gebruik van toetsen door scholen zijn moeilijk te rijmen met het beeld van leerkrachten die zich beklemd voelend door toetsen en examens. Het verplichte toets- en examenprogramma is redelijk beperkt (LVS, eindtoets en eindexamens) dus vraagt men zich af door wie men zich eigenlijk gedwongen voelt.
- De impliciete veroordeling van ‘rankings’ of ‘lijstjes’ die door de media gepubliceerd worden, terwijl hieraan toch, in zekere zin, een maatschappelijke behoefte bestaat.

- De neiging van de Inspectie van het Onderwijs om veronderstelde problemen bij externe summatieve evaluatie door te verwijzen naar de autonome scholen en meer formatieve evaluatie.
- De kennelijk bestaande aanname dat er een vakdoorbrekend curriculum nodig is, en dat dit een van de redenen is om het examen te herzien.

Verder blijkt uit recent onderzoek in opdracht van de VO-raad (Van der Ploeg & Weijers, 2018) dat de 'toetsgekte' mogelijk eerder te wijten is aan het toetsbeleid van de scholen dan van de landelijk verplichte toetsen. Dit onderzoek laat zien dat in de optiek van scholen vooral summatief gebruik van toetsen (cijfers geven, het fijnmazige plaatsingssyteem waar het inspectieverslag van spreekt) leerlingen motiveert om de toetsen te maken. Deze visie op schoolexaminering leidt tot veel toetsmomenten. Voor zover sprake is van toetsdruk of toetsgekte, levert het schoolexamen hier dus een belangrijke bijdrage aan. Bovendien, concluderen de onderzoeker, is hierdoor de kwaliteitsborging in termen van proces lastiger.
Ook uit opiniërende vragen aan schoolbestuurders, directies en docenten in dit onderzoek blijkt dat het schoolexamen veelal niet wordt gezien als een gelijkwaardig onderdeel aan het centraal examen, maar als een voorbereiding op het centraal examen. Dit laatste zien geënquêteerden als het belangrijkste doel van het schoolexamen. Belangrijker dan het examineren van onderdelen die vanuit de eigen onderwijsvisie relevant zijn of het toetsen van inhoud die in het centraal examen niet aan de orde komt.

Bevindingen focusgroep schoolleiders
Op basis van een focusgroep met schoolleiders uit het voortgezet onderwijs, die wij in het kader van dit onderzoek organiseerden op 20 november, 2018, concluderen we dat veel van de kritiek, die uit de hierboven geciteerde conferentie van de Inspectie en het Ministerie van OCW naar voren kwam, wordt bevestigd. Een kernpunt dat uit de focusgroep naar voren kwam was ambivalentie over de vele toetsen. Men vindt dat het toetsen is doorgeschoten, maar constateert tegelijkertijd dat de toetsen motiverend werken voor de leerlingen. Maar dit betreft dan extrinsieke motivatie, die minder gewaardeerd wordt dan intrinsieke motivatie. Volgens de deelnemers aan de focusgroepen zou dit moeten leiden tot herbezinning op het schoolexamen. Voorbereiding op het centraal examen kan een doel zijn, maar niet het hoofddoel. Veel overlap tussen schoolexamen en centraal examen dient te worden vermeden. Het centraal examen staat volgens de meeste deelnemers niet ter discussie. Wel zijn er veel voorstanders van een meer flexibele inrichting.

Misschien moet de oplossing worden gezocht in het stimuleren van het formatieve gebruik van de toetsen, door meer te investeren in terugkoppeling en uitleg van uitkomsten, en deze in te bedden in het onderwijsleerproces (McMilan & Nash, 2000). De wenselijkheden en mogelijkheden hiertoe zijn bestudeerd in onderzoek naar opbrengstgericht werken (Visscher en Ehren, 2011)[6]. Hierbij komt naar voren dat vaardigheden van leerkrachten om op een dergelijke formatieve wijze gebruik te maken van toetsen vaak tekortschiet, wat vraagt om professionele ontwikkeling en gerichte aanwijzingen bij de toetsen. In dit kader wordt wel gesproken over het versterken van de 'evaluation literacy' van docenten (Oomens, Veldkamp, & Scheerens, 2015; McMilan & Nash, 2000)).

6 Visscher, A. J., & Ehren, M. (2011). De eenvoud en complexiteit van opbrengstgericht werken. [Enschede]: [Universiteit Twente, Vakgroep Onderwijsorganisatie en -management

6 Het beoordelen en borgen van de kwaliteit van (studie)toetsen en examens

Piet Sanders, Arnold Brouwer en Anne Luc van der Vegt

Dit hoofdstuk gaat over het beoordelen van de kwaliteit van (studie)toetsen en examens met als focus de centrale examens van het voortgezet onderwijs. De Onderwijsraad gebruikt de term examens voor elke vorm van afsluitende of tussentijdse toetsing voor het vaststellen van leerresultaten met enig civiel effect. Van de vele functies die aan examens toegedicht worden, beschouwt de Onderwijsraad de kwalificerende functie als de belangrijkste. Bij de kwalificerende functie gaat het met name om het vaststellen van de kennis, vaardigheden en competenties die een student heeft opgedaan aan het eind van een opleiding.
Dit hoofdstuk bestaat uit drie paragrafen. De eerste paragraaf geeft een overzicht van de beoordelingssystemen die in Nederland gebruikt worden om de kwaliteit van toetsen te beoordelen (6.1). In annex 2 bij deze rapportage wordt het RCEC beoordelingssysteem voor de kwaliteit van toetsen en examens uitgebreid toegelicht. In paragraaf 6.2 beschrijven we hoe de kwaliteit van de schoolexamens wordt geborgd en kan worden geborgd.

6.1 Beoordelingssystemen voor de kwaliteit van toetsen

De Standards for Educational and Psychological testing worden sinds 1966 gepubliceerd door drie Amerikaanse instituten: de American Educational Research Association (AERA), de American Psychological Association (APA) en de National Council on Measurement in Education (NCME). De meest recente uitgave is die uit 2014 (APA, AERA & NCME, 2014). ‘The purpose of the *Standards* is to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretation of test scores for the intended test uses. Although such evaluations should depend heavily on professional judgment, the *Standards* provides a frame of reference to ensure that relevant issues are addressed.’
Het verschijnen van de *Standards* was de start voor de ontwikkeling van vele beoordelingssystemen van toetsen. In deze beoordelingssystemen worden de onderscheiden kwaliteitscriteria waaraan toetsen geacht worden te voldoen geoperationaliseerd. Sluijter, Hemker en Eggen (2018) geven een korte beschrijving van vijftien verschillende beoordelingssystemen: acht internationale en zeven Nederlandse beoordelingssystemen. Onderstaand overzicht van Nederlandse beoordelingssystemen is grotendeels ontleend maar niet identiek aan voornoemde publicatie.

1. *Beoordelingssysteem voor de kwaliteit van tests van de COTAN*
   Het beoordelingssysteem voor de kwaliteit van tests van de COTAN (Evers, Lucassen, Meijer, & Sijtsma, 2010) vindt zijn oorsprong in 1969 toen de Commissie Testaangelegenheden (COTAN), onderdeel van het Nederlands Instituut van Psychologen (NIP), begon met het publiceren van beoordelingen van tests. Het systeem is sindsdien verschillende malen herzien vanwege ontwikkelingen met betrekking tot toetstheorie en toetsconstructie.

2. *RCEC beoordelingssysteem voor de kwaliteit van studietoetsen en examens.*
   Het ‘RCEC beoordelingssysteem van de kwaliteit van studietoetsen en examens’ (Sanders, Van Dijk,

Eggen, Den Otter, & Veldkamp, 2016) is ontwikkeld om zowel de kwaliteit van studietoetsen als van (praktijk)examens te kunnen beoordelen. Het RCEC-beoordelingssysteem heeft de nodige raakvlakken met het systeem van de COTAN, maar is specifiek ontwikkeld vanuit het perspectief van toetsen en examens die in het onderwijs afgenomen worden.

3. *Beoordelingskaders van eindtoetsen en leerlingvolgsysteemtoetsen*
   Met ingang van het schooljaar 2014-2015 moeten alle schoolverlaters in het reguliere basisonderwijs een eindtoets maken die hun taal- en rekenvaardigheden meet. In 2017 konden scholen uit het basisonderwijs kiezen tussen de Centrale Eindtoets en vijf andere eindtoetsen. Die vijf eindtoetsen zijn door de Expertgroep PO, www.expertgroeptoetsenpo.nl, op kwaliteit beoordeeld aan de hand van het COTAN beoordelingssysteem. Daarnaast hanteert de Expertgroep PO een onderwijskundig en psychometrisch beoordelingskader voor leerlingvolgsysteemtoetsen.

4. *Regeling standaarden examenkwaliteit mbo*
   De beoordeling van de kwaliteit van de examens in het middelbaar beroepsonderwijs kent een complexe geschiedenis (Sluijter, 2016). In 2017 werd een norm voor valide exameninstrumenten opgesteld bestaande uit product, proces- en organisatie-eisen. Een exameninstrument moet tevens voldoen aan het geldende Onderzoekskader 2017 van de Inspectie van het Onderwijs.

5. *Kwaliteitscriteria voor competentie-assessmentprogramma's*
   Deze publicatie van Baartman (2008) geeft een overzicht van de criteria waar de examinering binnen competentiegericht onderwijs aan zouden moeten voldoen. Competentiegericht onderwijs is een term die een tijd in zwang is geweest om middelbaar en hoger beroepsonderwijs te beschrijven (zie ook paragraaf 3.5). Toetsing en examinering moet in dit soort onderwijs plaatsvinden in de vorm van een methodenmix (Straetmans & Sanders, 2001) of een competentieassessmentprogramma zoals Baartman dat beschrijft.

6 *Kwaliteitspiramide voor eigentijds toetsen en beoordelen*
   De kwaliteitspiramide voor eigentijds toetsen en beoordelen is ontwikkeld door Joosten-ten Brinke (2011) en richt zich op het borgen van de kwaliteit van toetsing en examinering binnen een instelling op verschillende met elkaar samenhangende niveaus: toetsonderdelen, toetsen, toetsprogramma's, toetsbeleid en de toetsbekwaamheid en toetsorganisatie.

7 *Checklist toetskwaliteit Eerste hulp bij toetsen*
   Meer gericht op docenten in het voortgezet onderwijs is de publicatie Eerste hulp bij toetsen (Maassen & Den Otter, 2014). Op basis van herkenbare vragen en opmerkingen van leerlingen worden aspecten van de kwaliteit van toetsen beschreven en geïllustreerd in de vorm van een checklist. Deze checklist is gebaseerd op basis van internationale richtlijnen, waaronder die voor testgebruik (International Test Commission, 2013), testbeveiliging (International Test Commission, 2014) en computergestuurde en internettesten (International Test Commission, 2005).
   Wat in het overzicht van Sluijter e.a. (2018) ontbreekt is het accreditatiekader van de Nederlands Vlaamse Accreditatie Organisatie (NVAO). In dit accreditatiekader voor het beoordelen van de kwaliteit van hogeronderwijsopleidingen en -instellingen, krijgt de toetsing speciale aandacht. Eén van de standaarden heeft betrekken op de adequaatheid van het systeem van toetsing, met daarin aandacht voor de validiteit, de betrouwbaarheid en de onafhankelijkheid van de beoordeling.

Gedetailleerde beschrijving van een beoordelingssysteem voor de kwaliteit van toetsen en examens
Dit hoofdstuk heeft betrekking op de door de overheid verplichte studietoetsen en examens die met de eerste vier voornoemde beoordelingssystemen en het NVAO accreditatiekader beoordeeld worden. De beoordelingen van deze beoordelingssystemen hebben consequenties voor (onderwijs)organisaties die toetsen en examens gebruiken. Indien examens van onvoldoende kwaliteit zijn, kunnen organisaties zelfs hun examenlicentie of accreditatie kwijtraken.
In Annex 2 van dit rapport wordt een nadere beschrijving gegeven van het RCEC beoordelingssysteem voor de kwaliteit van studietoetsen en examens. De redenen hiervoor zijn de volgende. Als je de verschillende systemen beschouwt, valt op dat er een grote overlap bestaat voor wat betreft de criteria die van belang geacht worden voor het bepalen van de kwaliteit van examens. Het RCEC systeem is specifiek ontwikkeld voor het beoordelen van toetsen en examens. Het RCEC beoordelingssysteem verschilt van het beoordelingssysteem van de COTAN dat bedoeld is voor het beoordelen van de kwaliteit van psychologische tests en studietoetsen, maar niet voor het beoordelen van de kwaliteit van examens. Andere hiervoor genoemde beoordelingssystemen zijn specifiek ontwikkeld voor het beoordelen van (praktijk)examens die afgenomen worden in het middelbaar beroepsonderwijs, de beoordeling van toetsing en examinering in het hoger onderwijs of de beoordeling van leerlingvolgtoetsen. Een andere belangrijke reden is dat met het RCEC beoordelingssysteem niet alleen schriftelijke toetsen bestaande uit gesloten vragen, open vragen, (praktijk)examens en computertoetsen beoordeeld kunnen worden maar dat er ook versies beschikbaar zijn voor het beoordelen van mondelinge examens en simulatorexamens.
De bespreking van het RCEC beoordelingssysteem in Annex 2 is grotendeels ontleend maar niet identiek aan Sanders (2016). Bij het criterium ‘representativiteit’ wordt dieper ingegaan op het aspect van de alignment tussen toets/examen en het curriculum/onderwijsdoelen. Het RCEC beoordelingssysteem onderscheidt zes criteria. Elk van de criteria wordt kort geïntroduceerd en de basisvragen die bij elk criterium worden gesteld, worden kort besproken.

6.2 Borging van de kwaliteit van het schoolexamen

In het voortgezet onderwijs weegt het schoolexamen even zwaar als het centraal examen bij het bepalen van het eindcijfer. De borging van de kwaliteit van het schoolexamen (SE) ligt op decentraal niveau; het is in de eerste plaats de verantwoordelijkheid van het schoolbestuur. Wel zijn er op landelijk niveau instrumenten ontwikkeld, die de school helpen een verantwoord SE samen te stellen en de kwaliteit daarvan te bewaken. Hieronder bespreken we deze instrumenten.

Richtlijnen en handreikingen

Eindtermen – Zoals we al beschreven in hoofdstuk 3, zijn de eindtermen die jaarlijks per vak worden opgesteld, onder verantwoordelijkheid van het Ministerie van OCW, leidend voor het opstellen van het schoolexamen. In het Eindexamenbesluit VO is vastgelegd dat het Ministerie een examenprogramma opstelt, waarin de examenstof staat omschreven, en bepaalt wat valt onder het CE en wat onder het SE (artikel 7). Wat betreft de tijdsplanning: het SE dient te zijn afgerond voordat het CE aanvangt (artikel 32).
De eindtermen voor het CE zijn gespecificeerd in syllabi, die bedoeld zijn om de constructie van de examens inhoudelijk aan te sturen en leraren te ondersteunen bij de examenvoorbereiding. Voor het SE

zijn er geen syllabi die voorschrijvend zijn, maar ontwikkelt SLO handreikingen die richtinggevend kunnen zijn bij het vormgeven van schoolexamens.
Bij het maken van toetsen hebben scholen de ruimte om eigen inhoudelijke keuzes te maken. De school stelt een programma van toetsing en afsluiting (PTA) op. Daarin wordt de inhoud van de examentoetsen vastgelegd, de toetsvorm en de normering.

Richtlijnen examenreglement en PTA – De school is verplicht een examenreglement en programma van toetsen en afsluiting (PTA) op te stellen. In het examenreglement doet de school in elk geval uitspraken over:

- maatregelen bij onregelmatigheden tijdens onderdelen van het eindexamen en de toepassing ervan;
- regels over de organisatie van onderdelen van het eindexamen;
- gang van zaken tijdens onderdelen van het eindexamen;
- herkansing van het schoolexamen en wanneer herkansing mogelijk is;
- samenstelling en het adres van de commissie van beroep.

Het PTA heeft in elk geval betrekking op:

- onderdelen van het examenprogramma in het schoolexamen;
- inhoud van het schoolexamen;
- wijze waarop het schoolexamen plaatsvindt;
- tijdvakken waarbinnen de toetsen van het schoolexamen beginnen;
- herkansing en de wijze van herkansing van het schoolexamen;
- regels waarop het cijfer voor het schoolexamen tot stand komt.

Handreikingen SLO – Per examenvak heeft SLO handreikingen geschreven voor de invulling van de schoolexamens. Deze helpen de school om, uitgaande van de eindtermen, inhoud en vorm van de toets te bepalen. Deze handreikingen zijn beschikbaar voor vmbo, havo en vwo. De handreikingen bevatten suggesties, maar geen voorschriften. Het is aan de school om te bepalen of de suggesties worden opgevolgd. Ook geeft SLO op zijn website informatie over de borging van de kwaliteit.

Checklist VO-raad – Ten behoeve van schoolleiding en examensecretarissen heeft de VO-raad een checklist opgesteld, waarmee kan worden nagegaan of de processen rondom de examinering goed zijn ingericht en goed worden uitgevoerd. Bovendien heeft de VO-raad een zelfevaluatie-instrument ontwikkeld. Dit is een hulpmiddel bij het borgen van de kwaliteit van de examens. De zelfevaluatie heeft onder meer betrekking op:

- Visie en verantwoordelijkheden;
- Regels en procedures;
- Kwaliteit van PTA en toetsen.

Bevordering kwaliteit van schoolexamens
Uit verscheidene studies blijkt dat docenten zich te veel beperken tot toetsopgaven die vragen om reproductie van kennis. Andere cognitieve processen, zoals toepassen van kennis of integratie met andere kennis, worden relatief weinig getoetst. Docenten ontwikkelen relatief weinig toetsitems zelf en putten vaak uit methodegebonden toetsen. Onderzoek laat zien dat docenten effectief getraind kunnen

worden in het maken van evenwichtige toetsen, als ze een professionaliseringsprogramma volgen waarin aandacht wordt besteed aan:

1) reflectie op de overtuigingen van docenten met betrekking tot de doelen van het onderwijs,
2) afstemming van deze doelen op instructie en toetsing,
3) gezamenlijke oefening op basis van nieuwe kennis en vaardigheden (Bijsterbosch, 2018).

Bij het ontwikkelen van toetsen die een beroep doen op verschillende cognitieve processen, zijn taxonomieën bruikbaar. Zeer bekend is de taxonomie van Bloom, later gereviseerd door Anderson en Krathwohl (2001), ontwikkeld als algemeen model voor de doelstellingen van het leerproces. De cognitieve niveaus in het gereviseerde model zijn:

- kennisreproductie;
- inzicht;
- toepassing;
- analyse;
- creatie/synthese;
- evaluatie.

Door SLO is een checklist ontwikkeld, waarmee docenten leerdoelen kunnen formuleren en beheersingsniveaus vaststellen. Docenten kunnen hier ook cursussen in volgen. Naast de taxonomie van Bloom zijn er ook andere taxonomieën. Romiszowski (1981) maakt bijvoorbeeld onderscheid in vier niveaus kennis en vier niveaus vaardigheden. In het Nederlandse onderwijs wordt met name de RTTI-taxonomie veel gebruikt. Deze afkorting verwijst naar een taxonomie met vier typen vragen: R=reproductievragen; T1=trainingsgerichte toepassingsvragen; T2=transfergerichte toepassingsvragen; I=Inzicht & Innovatievragen (Drost & Verra, 2015). De ontwikkelaars geven ook scholing in toetsbeleid en kwaliteitszorg.

Onderzoek naar kwaliteit schoolexaminering

Tien jaar geleden heeft het Cito onderzoek gedaan naar de vakinhoudelijke en toetstechnische kwaliteit van de schoolexamens in het voortgezet onderwijs. De belangrijkste conclusie was dat de kwaliteit op de overgrote meerderheid van de scholen in orde is. Niettemin stellen de beoordelaars vast dat er aanzienlijke kwaliteitsverschillen bestaan tussen de schoolexamens. Tekortkomingen van sommige schoolexamens: onevenredige accenten op verschillende vakonderdelen, te veel gebruik maken van methodebonden toetsen, te hoge of te lage moeilijkheidsgraad, ontbreken van schoolinterne kwaliteitscontrole op de toetsconstructie, te soepele of te strenge beoordeling. Het Cito beveelt onder meer aan om scholen te verplichten tot meer transparantie over de examenpraktijken.

Een actueel beeld van de mate waarin scholen aandacht besteden aan het borgen van de schoolexaminering wordt gegeven in recent onderzoek in opdracht van de VO-raad (Van der Ploeg & Weijers, 2018). Dit onderzoek is uitgevoerd naar aanleiding van de problemen met de examenpraktijk op twee vmbo-scholen in Maastricht. Doel van het onderzoek was duidelijk te maken in hoeverre bij scholen bekend is wat de wettelijke kaders en richtlijnen zijn en hoe het proces rondom het schoolexamen is ingericht. Dan gaat het om totstandkoming van het PTA en het naleven van procedures. De focus van dit onderzoek is dus een andere van het Cito-onderzoek van 2008.
De onderzoekers stellen vast dat procesmatige kwaliteitsborging nauwelijks voorkomt. Verbeteringen vinden veelal ad hoc plaats, naar aanleiding van geconstateerde problemen. Voor veel scholen is dit voldoende voor borging van de proceskwaliteit, maar er blijven zeker risico's over, bijvoorbeeld rondom het werken conform PTA en reglement of rondom afname van examens.

Kwaliteitscontrole wordt bemoeilijkt door de veelheid aan toetsen. Naast vaste schoolexamenweken werken scholen met extra toetsonderdelen, zoals schriftelijke overhoringen of huiswerk. Hierop is nauwelijks sprake van controle. De grote variatie en beperkte standaardisatie van schoolexamenonderdelen die buiten de toetsweken vallen, bemoeilijken de procesmatige kwaliteitsborging.

.

7 Innovaties in toetsen en examens in het VO

Bernard Veldkamp

In de laatste jaren zijn er veel vernieuwingen geweest binnen de toets-technologie die hun impact hebben gehad op het voortgezet onderwijs. De meest bekende vorm van toetsen is de individuele toets die schriftelijk afgenomen wordt, maar de grootschalige beschikbaarheid van computers, tablets en internet binnen de scholen hebben de technische mogelijkheden voor innovatieve vormen assessment sterk vergroot. In dit hoofdstuk wordt een aantal ontwikkelingen besproken en geëvalueerd binnen de context van het voortgezet onderwijs. Allereerst wordt ingegaan op gepersonaliseerd toetsen (7.1), vervolgens komt online toetsing aan de orde (7.2), daarna volgt authentiek toetsen (7.3) en tenslotte komt het gebruik van big data voor assessment aan de orde (7.4).

7.1 Gepersonaliseerd leren

Binnen het voortgezet onderwijs is veel aandacht voor gepersonaliseerd leren. De achterliggende gedachte is dat binnen klassen grote verschillen kunnen bestaan in niveau tussen de leerlingen en dat de effectiviteit, maar ook de aantrekkelijkheid, van het onderwijs sterk verhoogd kan worden door het niveau van het onderwijs aan te passen aan het niveau van de leerling. Een vergelijkbare gedachte ligt ten grondslag aan gepersonaliseerd toetsen. Deze gepersonaliseerde toetsen worden vaak digitaal afgenomen, zowel online als stand alone. In de literatuur wordt voor digitale gepersonaliseerde toetsen vaak de term adaptief toetsen (Wainer e.a., 2000) gebruikt. Toetsen worden gepersonaliseerd door het niveau, oftewel de moeilijkheid, van de toets aan te passen aan het beheersingsniveau van de kandidaat. Op deze manier krijgt de kandidaat geen vragen die veel te moeilijk of veel te makkelijk zijn. Daardoor raakt de leerling niet gefrustreerd of verveeld. Bovendien kan met behulp van psychometrie, meer specifiek met behulp van item response theorie (IRT) modellen, aangetoond worden dat adaptieve toetsen veel efficiënter meten, waardoor de toetslengte gereduceerd kan worden met 25%-50%. De combinatie van korte toetsen, digitale afname en toegenomen motivatie bij de leerlingen hebben adaptief toetsen populair gemaakt in het onderwijs. Een belangrijke opmerking hierbij is wel dat de toetsen gepersonaliseerd worden wat de moeilijkheid betreft, maar niet wat de inhoud aangaat. De inhoud van de toetsen blijft vergelijkbaar doordat ze allemaal voldoen aan de randvoorwaarden die in een syllabus of een toetsmatrijs aan de toetsen zijn gesteld.
Door gebruik te maken van IRT kan aangetoond worden dat vragen de meeste informatie opleveren over een kandidaat als de moeilijkheid van de vraag overeenkomt met diens vaardigheidsniveau. Intuïtief is duidelijk dat veel te makkelijke vragen bijna altijd correct beantwoord zullen worden en veel te moeilijke vragen bijna altijd incorrect. Een correct antwoord op een heel makkelijke vraag of een incorrect antwoord op een veel te moeilijke vraag levert dan ook weinig informatie op over wat nu precies het beheersingsniveau van de leerling is. In een kalibratiestudie wordt van te voren geschat wat de moeilijkheid van de verschillende vragen is en hoeveel informatie ze bij de verschillende beheersings-niveaus leveren. Geavanceerde algoritmes bepalen vervolgens, tijdens de toetsafname, welke vragen het meest geschikt zijn voor een specifieke leerling.
Deze algoritmes voor computer adaptief toetsen (CAT) werken op de volgende manier. Voorafgaand aan de toets wordt een beginschatting gemaakt van het niveau van de leerling. Dit kan op verschillende manieren. De begin schatting van het niveau van de leerling kan gelijk gesteld worden aan het

gemiddelde van de populatie of van de klas waar de leerling in zit. De docent kan een inschatting geven van het niveau van de leerling. Of het gemiddelde van prestaties op voorafgaande toetsen kan worden gebruikt. De tweede stap van het algoritme selecteert bij deze beginschatting één of meerdere vragen. De derde stap is de afname van deze vragen bij de leerling. Als de leerling de vragen heeft beantwoord, wordt, tijdens de vierde stap, het antwoord automatisch gescoord. Met behulp van psychometrie wordt vervolgens, in de vijfde stap, een schatting gemaakt van het niveau van de kandidaat. In stap zes wordt gekeken of de toets al kan stoppen of dat er nog meer vragen afgenomen moeten worden. Hiervoor zijn twee soorten stopregels beschikbaar. Voor sommige toetsen is van te voren vastgesteld hoeveel vragen er afgenomen moeten worden (fixed-length). Het grote voordeel van deze stopregel is dat alle leerlingen een even lange toets maken, wat de vergelijkbaarheid vergroot. Een tweede stopregel kijkt naar de nauwkeurigheid van de schatting van het beheersingsniveau (variable-length). Op het moment dat de nauwkeurigheid van de schatting hoog genoeg is, of als blijkt dat het afnemen van extra vragen niet leidt tot een nauwkeurigere schatting van het niveau, dan beëindigd de CAT. Met deze stopregel wordt optimaal gebruik gemaakt van de mogelijkheden om efficiënt te toetsen. De lengte van de toetsen zal alleen variëren voor de leerling in de klas. Als de toets nog niet voldoet aan de stopregels, dan gaan we weer naar stap twee van het algoritme. Bij de huidige schatting van het niveau worden weer één of meerdere vragen geselecteerd, die worden vervolgens afgenomen, automatische gescoord, de schatting van het niveau wordt ge-update en met de stopregel wordt gekeken of de toetsafname kan eindigen of dat er nog extra vragen nodig zijn. Dit proces wordt herhaald tot de toets stopt.
In de praktijk komt adaptief toetsen voor op verschillende manieren. Om optimaal gebruik te maken van de psychometrische mogelijkheden van CAT, zou na afname van elke vraag een nieuwe schatting gemaakt moeten worden van het niveau van de leerling en zou vervolgens uit alle beschikbare vragen de beste vraag geselecteerd moeten worden. Dit is alleen niet altijd haalbaar. Om te kunnen voldoen aan de eisen die in syllabi of toetsmatrijzen aan de inhoud van toetsen worden gesteld, worden randvoorwaarden opgelegd, waarmee het algoritme rekening moet houden bij het selecteren van vragen (van der Linden, 2005). Om te voorkomen dat algoritmes steeds dezelfde groep vragen selecteren, die snel bekend kunnen worden bij de leerlingen, worden restricties opgelegd aan hoe vaak een vraag geselecteerd mag worden (van der Linden en Veldkamp, 2004). Tenslotte wordt er ook geregeld gekozen voor multi-stage testing, een vorm van adaptief toetsen waarbij van te voren sets van vragen samengesteld worden. Nadat alle vragen uit zo'n set zijn beantwoord door de leerling, wordt een inschatting gemaakt van het niveau en krijgt de leerling een vervolg-set die van het zelfde niveau, moeilijker of makkelijker is. Een groot voordeel van deze vorm van adaptief toetsen is dat deze sets van te voren gereviewd kunnen worden door docenten en toetsexperts.
De verzameling vragen waar het algoritme uit kan selecteren wordt een itembank genoemd. Deze itembank is een database waarin voor elke vraag de vraag, de mogelijke antwoorden, het correcte antwoord, de psychometrische eigenschappen en hoe vaak de vraag is afgenomen worden vastgelegd. Een vuistregel die vaak gehanteerd wordt voor de grootte van de itembank is twaalf keer de lengte van de toets. De itembank moet geregeld ververst worden om te voorkomen dat de vragen uitlekken en bekend worden bij de leerlingen.
Een groot voordeel van gepersonaliseerd toetsen is dat de toetsen afgenomen kunnen worden op het moment dat de leerling er aan toe is. Als toetsen slechts gebruikt worden om de voortgang van de leerling te meten, dan kan dit veel voordelen opleveren. Er kan recht gedaan worden aan verschillen in snelheid van werken en het geeft meer flexibiliteit aan de leerling. Op het moment dat de toets gebruikt wordt voor beslissingen met betrekking tot de overgang naar een volgend leerjaar, of bij een examen, speelt alleen het risico van het uitlekken van de vragen. De leerlingen die het eerst de toets maken onthouden de vragen en spelen die door aan hun medeleerlingen. De uitgelekte vragen geven

vervolgens een vertekend beeld van het beheersingsniveau. In de beginjaren van CAT was dit een groot probleem. Het is meerdere keren gebeurd dat de inhoud van een itembank binnen enkele dagen verspreid werd via internet, soms zelfs in georganiseerd verband. Dit compromitteerde de validiteit van de scores. Om deze problemen te voorkomen, wordt de periode waarin de toets afgenomen wordt vaak beperkt en wordt gebruik gemaakt van item-exposure algoritmes om het risico van uitlekken te minimaliseren.
Tenslotte zijn er twee beperkingen/nadelen van adaptief toetsen. Het eerste punt betreft de kosten. Alhoewel elke toets in principe adaptief afgenomen kan worden, ongeacht de grootte van de itembank, wordt in de praktijk toch vaak de vuistregel gehanteerd dat het aantal vragen in de itembank ongeveer gelijk moet zijn aan twaalf keer de testlengte. Al deze items moeten ontwikkeld worden en ze moeten van te voren gepre-test zijn om de moeilijkheid en de andere psychometrische eigenschappen te schatten. Het ontwikkelen van een dergelijke itembank is daarom vrij kostbaar in vergelijking met het ontwikkelen van een gewone toets. Daarnaast is er een digitale infrastructuur en software nodig om de toetsen af te kunnen nemen. De kwaliteit van de digitale infrastructuur binnen het voortgezet onderwijs verbetert gelukkig snel. Steeds meer leerlingen krijgen de beschikking over chrome books of tablets en de software die nodig is, is steeds vaker open source of tegen een geringe vergoeding te krijgen. Ook wordt er veel onderzoek gedaan naar het geautomatiseerd schrijven van vragen en het optimaliseren van het pré-test proces om de kosten van het ontwikkelen van een item bank te verkleinen.
Een tweede punt betreft de noodzaak van het automatisch scoren van de antwoorden. Om het algoritme in staat te stellen om tijdens de toetsafname een schatting te maken van het beheersingsniveau, is het nodig dat de computer op een betrouwbare manier kan beoordelen of een antwoord correct is en in welke mate of dat het incorrect is. Het gevolg hiervan is dat adaptieve toetsen hoofdzakelijk gebruik maken van multiple-choice vragen met één correct antwoord of van vragen met een kort éénduidig antwoord. Deze vraagvormen kennen hun beperkingen. Het is een grote uitdaging om hogere-ordevaardigheden op deze manier te meten. Er komt weliswaar steeds meer technologie beschikbaar om tekst te interpreteren, maar op dit moment is dit nog een beperking van gepersonaliseerd toetsen.

7.2 Digitale toetsing

Een tweede ontwikkeling heeft te maken met digitale toetsing. De toegenomen beschikbaarheid van computers en tablets heeft de mogelijkheden voor digitale toetsing sterk vergroot. Digitale toetsing kent een groot aantal voordelen. In de vorige paragraaf is ingegaan op de mogelijkheid voor gepersonaliseerde assessment, maar daarnaast zijn er nog andere voordelen.
Het eerste voordeel betreft het gebruik van digitale media. Daarbij kan gedacht worden aan video's, geluidsfragmenten, chats, interactieve formats en digitale hulpbronnen. Door op een efficiënte manier gebruik te maken van de mogelijkheden van digitale media kan de aantrekkelijkheid van de vragen en het realistisch gehalte van de toets sterk worden verhoogd. Een belangrijke overweging bij het ontwikkelen van mediarijke toetsen blijft wel dat goed nagedacht moet blijven worden over de functionaliteit van de media om te voorkomen dat de media afleiden van het doel om te komen tot een betere meting van het beheersingsniveau. Als de toetsen afgenomen worden in een online omgeving, dan zijn de mogelijkheden voor het gebruik van digitale hulpbronnen bijna onbeperkt. Omdat onderlinge vergelijkbaarheid een belangrijk aspect is bij de kwaliteit van toetsing, worden toetsen vaak afgenomen in een afgesloten omgeving en wordt er veel aandacht besteed aan de vormgeving van de toets binnen verschillende browsers en toetsomgevingen.

Een tweede voordeel betreft de mogelijkheid voor het geven van feedback binnen een digitale leeromgeving. Bij feedback wordt onderscheid gemaakt tussen drie stappen. Wat is het leerdoel (feed-up), waar sta je nu (feed-back) en wat zijn de volgende stappen in het leerproces (feed-forward). Binnen de tweede stap spelen toetsen een belangrijke rol. Ze geven veel informatie over het beheersingsniveau van de leerlingen en geven ook informatie over de onderwerpen van de leerstof die niet worden beheerst, zowel op individueel als op groepsniveau. Op deze manier zijn digitale toetsen een belangrijk instrument met het oog op opportunities to learn. Feedback heeft het grootste effect op leerresultaten als de periode tussen het beantwoorden van de vragen en het krijgen van feedback kort is. Binnen digitale toetsing kunnen leerlingen direct feedback krijgen op hun antwoorden. Dat kan in de vorm van hints of aanwijzingen als het antwoord niet helemaal correct is, in de vorm van het correcte antwoord, of in de vorm van uitleg over hoe de leerling het antwoord zou kunnen verbeteren. Ook hierbij geldt dat de antwoorden automatisch gescoord moeten kunnen worden om gebruik te maken van deze voordelen van digitale toetsing. Als de toets afgenomen wordt in een online omgeving, dan biedt dit ook de mogelijkheden voor peer-feedback. Binnen het voortgezet onderwijs wordt er nog niet veel van deze mogelijkheid gebruik gemaakt, maar steeds meer online leeromgevingen bieden hiervoor mogelijkheden en bij MOOCs [7]is deze vorm van feedback vrij gebruikelijk. Naast feedback aan de leerlingen, geeft de toets ook feedback aan de docent en de schoolleider. Het helpt hen uitgaande van de leerdoelen invulling te geven in de volgende stappen van het leerproces.
Een derde voordeel gaat over analyse van de toetsresultaten. Een toets- en itemanalyse geeft inzicht in de psychometrische kwaliteit van de toets. Bij een digitale toetsafname kan de toets- en itemanalyse geautomatiseerd uitgevoerd worden. In een online omgeving waarbij leerlingen van verschillende scholen de toets maken, kunnen de gegevens gecombineerd worden, wat tot een hogere betrouwbaarheid leidt. Ook kunnen de toetsresultaten van verschillende leerlingen, klassen en scholen onderling worden vergeleken, wat meer inzicht geeft in de leerprestaties.
Een belangrijk aandachtspunt bij digitale toetsing en online toetsing zijn de beveiliging en de technische aspecten van de toetsafname. Dit betreft zowel de beveiliging van de toets tegen pogingen om op ongeoorloofde manier toegang te krijgen tot de vragen, maar ook beveiliging van de toetsresultaten en beveiliging van de gegevens van de school en de leerlingen. Daarnaast is het van belang dat alle leerlingen op een vergelijkbare manier toegang hebben tot de toets. Daarbij spelen technische aspecten een rol die te maken hebben met verschillende besturingssystemen, versies van software en verschillen in apparatuur. Ook is het van belang om aandacht te hebben voor leerlingen met visuele, auditieve of andere beperkingen.

7.3 Authentieke toetsing

Een derde ontwikkeling betreft de toegenomen aandacht voor authentieke toetsing. Authentieke toetsing wordt vooral toegepast bij het toetsen van complexe vaardigheden. Binnen het beroepsonderwijs worden verschillende taken getoetst door leerlingen deze taken uit te laten voeren in een realistische setting, bijvoorbeeld op de werkvloer of in een simulatie. Ook binnen het voortgezet onderwijs krijgt deze vorm van assessment steeds meer aandacht, bijvoorbeeld bij het meten van 21e eeuwse vaardigheden zoals informatievaardigheden, samenwerken, creatief denken, mediawijsheid of probleem oplossen. Aan deze vaardigheden wordt steeds meer gewerkt in het voortgezet onderwijs.

7 Massive Open Online Courses

Deze vaardigheden worden op dit moment vooral gemeten aan de hand van zelfperceptie. Dergelijke meetinstrumenten lopen tegen validiteitsproblemen aan omdat deze geen directe kennis en vaardigheden meten maar het zelfvertrouwen van leerlingen in kaart brengen. De scores zijn namelijk afhankelijk van de mate waarin de leerling vaardig is om de eigen digitale vaardigheid te beoordelen terwijl uit onderzoek blijkt dat de verschillen tussen zelfrapportage en daadwerkelijke vaardigheidsniveau vaak groot zijn (Allayar, 2011). Door authentieke taken voor te leggen, kan een veel betere indruk gekregen worden van het beheersingsniveau (Aesseart et al., 2014). Door leerlingen bijvoorbeeld toegang te geven tot een selectie van websites over een specifiek onderwerp en ze een aantal vragen over dat onderwerp voor te leggen, kan veel meer inzicht verkregen worden in de mate waarin ze in staat zijn om bruikbare informatie op internet te verzamelen (Heiting, 2018).
Bij het ontwikkelen van authentieke toetsen speelt wel de paradox van realisme en vergelijkbaarheid. Aan de ene kant wordt geprobeerd om de verschillende toets taken in een zo realistisch mogelijk setting en idealiter in de praktijk plaats te laten vinden. Op die manier kan het meeste inzicht verkregen worden in de beheersing door een leerling van bepaalde taken of vaardigheden. Aan de andere kant is het van belang dat de vragen of toetstaken die de leerlingen voorgelegd krijgen, wel onderling vergelijkbaar zijn om tot een eerlijke toetsing te komen. De oplossing die daarom vaak gekozen wordt is dat de vragen of taken in een gecontroleerde omgeving of binnen een simulatie uitgevoerd moeten worden.
Deze authentieke toetsen brengen nog wel een aantal uitdagingen met zich mee op psychometrisch vlak. Hoe kan gegarandeerd worden dat de scoring op een valide en betrouwbare manier plaatsvindt? Standaard meetmodellen uit de klassieke testtheorie of de item response theorie zijn vaak niet toepasbaar omdat die niet ontwikkeld zijn voor het meten van deze complexe vaardigheden. Ook kost het afnemen van authentieke toetsen vaak veel tijd en is het relatief duur om toetsen in een realistische omgeving af te nemen. Het gevolg is vaak dat de toets slechts een gering aantal vragen of taken kent. Nu wordt daarom nog vaak met beoordelaars gewerkt en is de score gebaseerd op een gemiddelde van hun beoordelingen. Deze manier van scoren is echter duur en tijdrovend. Er zijn inmiddels een aantal experimenten gedaan met het gebruik van Educational Design (Mislevy) en Bayesiaanse netwerken (de Klerk) om authentieke toetsen geautomatiseerd te scoren, maar deze oplossingen zijn nog niet op grote schaal beschikbaar binnen het voortgezet onderwijs.

7.4 Big data en toetsing

Tegenwoordig worden we overspoeld met data. Data van logfiles, klantenkaarten, social media, smart phones, sensoren, internet, en bijvoorbeeld smart watches. Deze data kunnen gecombineerd en geanalyseerd worden om tot nieuwe inzichten te komen over menselijk gedrag. De omvang van de data is enorm, zodat ook wel gesproken wordt van big data. Big data karakteriseert zich door de 3 V's (Gartner, 2018): Volume, Variety en Velocity. Volume slaat op de omvang van de data, al wordt het ook geregeld geïnterpreteerd als wat er allemaal met de data gedaan kan worden. Variety benadrukt meer de diversiteit van de data. De data is vaak ongestructureerd en kan niet in standaard databases opgeslagen worden. Velocity slaat tenslotte op de snelheid waarmee nieuwe data binnenkomen en/of opgevraagd worden.
In het onderwijs worden tegenwoordig grote hoeveelheden data vastgelegd. Veldkamp et al. (2017) onderscheiden een aantal verschillende databronnen. Zij beschrijven data afkomstig uit leerlingvolgsystemen, centrale toetsen, cohortonderzoek, (internationale) surveys, onderwijsinspectie, lesevaluaties, elektronische leeromgevingen, MOOCs, verslagen van (docent)vergaderingen, jaarverslagen, registraties, financiële stukken en overige ongestructureerde data. De individuele

bronnen voldoen meestal niet aan alle kenmerken van big data (volume, variety, and velocity) en één specifieke databron is meestal niet direct geschikt is voor het vinden van nieuwe verbanden tussen variabelen die nuttige informatie opleveren over het leerproces (educational data mining), maar een combinatie van deze bronnen biedt mogelijkheden. Tot nu toe is educational data mining hoofdzakelijk gebruikt om de individuele ontwikkeling van leerlingen te volgen en te onderzoeken welke variabelen daaraan gerelateerd zijn.
De vraag is of en hoe al deze data gebruikt kan worden voor toetsing. Technisch gezien is er veel mogelijk. Op basis van gegevens die over een leerling zijn verzameld, kan met behulp van machine learning een nauwkeurige voorspelling gemaakt worden van het beheersingsniveau. Toetsing kan daarmee volledig geïntegreerd met het leerproces plaatsvinden. Het directe gevolg is minder examenstress, minder piekbelasting in examentijd en efficiënt gebruik van de beschikbare data.
Om een dergelijk model te ontwikkelen is een dataset nodig met de beschikbare data uit het leerproces en een nauwkeurige meting van het beheersingsniveau. Deze data wordt opgesplitst in een trainingset en een testset. Vaak wordt hiervoor een randomverdeling van 70%/30% gehanteerd. 70% van de data wordt gebruikt om het model te trainen. 30% wordt gebruikt om te testen hoe goed het model werkt.
De ruwe data uit het leerproces kan allerlei vormen hebben. Met behulp van feature extraction wordt de ruwe data omgezet in variabelen die gebruikt kunnen worden om het model te bouwen. Daarna kunnen machine learning algoritmes, zoals support vector machines, neurale netwerken, naïve bayes of regression trees gebruikt worden om het beheersingsniveau te voorspellen op basis van deze variabelen (Friedman, Hastie & Tibshirani, 2001). Om te onderzoeken hoe goed het model in de praktijk werkt, wordt het toegepast op de testset. Bij een betrouwbaar model verschillen de prestaties nauwelijks tussen de trainingset en de testset. Om deze stappen uit te voeren kan gebruik gemaakt worden van standaard software zoals Cran R of Python.
Alhoewel zowel de data als de software beschikbaar is, wordt big data nog maar zelden gebruikt voor toetsing. Daarvoor zijn een aantal technische, juridische en ethische oorzaken. Op technisch vlak speelt mee dat big data analytics gegevens van een groot aantal leerlingen gebruikt om het beheersingsniveau van een individu te voorspellen. Alhoewel dit op populatieniveau tot goede resultaten leidt, wil dit nog niet zeggen dat dit ook voor elk individu geldt. Daarnaast worden bij deze methoden ook allerlei proxies gebruikt in plaats van de werkelijke prestaties. Een bekend voorbeeld is het automatisch beoordelen van essays met natural language processing technieken. Door te kijken naar, onder andere, de woorden die gebruikt zijn, de complexiteit van de zinnen en bijvoorbeeld de zinslengte wordt een cijfer toegekend. Juridisch gezien speelt daarnaast dat de Algemene verordening gegevensbescherming (AVG) sinds mei 2018 allerlei beperkingen oplegt aan het gebruik van data. Uitgangspunten binnen de AVG zijn de principes van rechtmatigheid, zorgvuldigheid, transparantie, vertrouwelijkheid, integriteit en dataminimalisatie. Door deze verordening wordt het risico op profiling (discriminatie op basis van data) kleiner, maar het gevolg is ook dat er minder data beschikbaar is voor analyses. Er is toestemming nodig van de leerlingen en soms van hun ouders om de leerlingdata te gebruiken voor beoordeling en daarnaast speelt doelbinding een rol, die vastlegt dat data alleen gebruikt mogen worden voor het doel waarvoor ze verzameld zijn. Dit houdt in dat al vóór de data verzameld worden, duidelijk moet zijn dat ze gebruikt gaan worden voor toetsing. Op ethisch gebied, tenslotte, spelen ook verschillende zaken. In hoeverre is het wenselijk om data uit het verleden te gebruiken om een model te ontwikkelen waarmee leerlingen worden getoetst? Mag je de data van de ene leerling gebruiken om een model voor andere leerlingen te ontwikkelen? Mag je een leerling beoordelen op basis van meer gegevens dat zijn/haar eigen prestatie?
Veldkamp et al (2017) beschrijven dat scholen kritisch staan tegenover de mogelijkheid om big data analytics in te zetten. Scholen, toetsleveranciers en softwareproducenten gebruiken graag data om te

monitoren en processen te verbeteren, om leerprestaties, studiesucces of uitval te voorspellen of om maatregelen te nemen om het onderwijs te verbeteren. De bereidwilligheid om data te delen is echter laag. Best-practises zijn nog te weinig voorhanden en de computersystemen zijn te divers om grootschalige data-uitwisseling op eenvoudige wijze vorm te geven.

8 De balans van bevestiging en kritiek bij het functioneren van examens en eindtoetsen

Jaap Scheerens

8.1 Inleiding

In dit hoofdstuk maken we de balans op over het functioneren van examens en eindtoetsen in Nederland. Dit gebeurt op basis van een afweging van de veronderstelde positieve bijdragen en de geleverde kritiek. Vervolgens wordt stilgestaan bij verschillende perspectieven om te kijken naar consolidatie dan wel verandering van het stelsel van toetsen en examens. Daarbij wordt onder meer stilgestaan bij inzichten over de 'maakbaarheid' van onderwijsstelstelsels; verbetering ten opzichte van verschillende facetten van onderwijskwaliteit en 'haalbare hefbomen' voor verbetering. Tenslotte worden de contouren van een beleidsgericht onderzoeksprogramma geschetst, die uitmonden in ideeën voor nader onderzoek van de examen- en toetsproblematiek. De bedoeling van dit onderzoek is de evidence-base voor toekomstig beleid ten aanzien van examens en eindtoetsen te versterken. Hoofdpunten zijn een nadere 'fact check' van zowel de geleverde kritiek als de positieve waardering van de bestaande verworvenheden en onderbouwing van verbeteringsstrategieën.

8.2 De balans van positieve verworvenheden en kritiek

Ondersteuning

Examens en eindtoetsen zijn op te vatten als bepalend voor de institutionele structuur van onderwijsstelsels. Met 'institutioneel' wordt in dit verband bedoeld dat ze de voornaamste interne spelregels en de communicatie over en weer met de ruimere maatschappelijke context vastleggen. Men zou dit ook kunnen uitdrukken door te spreken van de functie van examens voor de kwaliteitsborging van het onderwijssysteem. Daarnaast kan er een kwaliteitsbevorderende functie van examens en eindtoetsen worden onderkend, in die zin dat de werking ervan het onderwijs op een hoger peil brengt. Er wordt verwezen naar empirisch onderzoek dat deze werking ondersteunt.
Kwaliteitsborging manifesteert zich in de volgende functies:

- Eindtoetsing faciliteert doorstroom naar vervolgonderwijs;
- Eindtoetsing in het voortgezet onderwijs garandeert het civiel effect van diploma's;
- Eindtoetsing garandeert een gedeelde brede vorming met minimaal een basisniveau.

Examens en eindtoetsen zijn met deze functies een belangrijke pijler van onderwijsstelsels. Het wijzigen van examens is daarmee een gewichtige kwestie waarvoor men een gedegen evidence-based onderbouwing zou wensen. Het is de vraag of bijvoorbeeld de recente voorstellen van de VO-raad tot flexibilisering en differentiatie van examens aan deze eis voldoen (zie ook de stellingname hierover van de Onderwijsraad, anno 2015).

De gedachte dat eindtoetsing en examens ook tot kwaliteitsbevordering kunnen leiden is ingegeven door onderzoek waaruit bleek dat landen met een op standaarden gebaseerd examen beter scoren op internationale toetsen dan landen zonder examens. Hoewel deze uitkomsten niet helemaal

onomstreden zijn (zo zijn er bijvoorbeeld landen die geen examen hebben maar wel hoog scoren, zoals Vlaanderen), zijn er wel redenen die dit aannemelijk maken. Als mechanismen worden genoemd:

- Het stimuleren van de extrinsieke motivatie van leerlingen, docenten en scholen, door belangen die in het spel zijn bij de uitkomsten.
- Het stimuleren van leerprocessen, op basis van instrumentele feedback, het aanduiden van sterke en zwakke punten in het functioneren op macro en micro niveau.
- Het bieden van richting aan het onderwijs in het kader van 'curriculum alignment', in het bijzonder de goede aansluiting tussen nationale standaarden (eindtermen/referentieniveaus) enerzijds en de inhoud van examens en eindtoetsen anderzijds.

Het laatste punt, het gezichtspunt van curriculum alignment wordt van speciaal belang geacht, omdat het de discussie over teaching to the test (een van de meest gehoorde kritieken op examens en eindtoetsen) in een ander daglicht plaatst.

Kritiek

De titel van het Advies van de Onderwijsraad uit 2013 'Een smalle kijk op onderwijskwaliteit' vat een belangrijke lijn van kritiek samen. Deze lijn van kritiek is niet alleen afkomstig van de Onderwijsraad, maar wordt ook aangetroffen in beleidsnota's van de VO-raad, en recente publicaties van de Inspectie van het Onderwijs. In de terminologie van de hierboven besproken kernfuncties van onderwijs komt de kritiek erop neer dat er een te groot accent ligt op de kwalificatiefunctie, en dat dit ten koste gaat van socialisering en persoonlijke ontwikkeling. Het vigerende stelsel van eindtoetsen en examens versterkt deze tendens, omdat kennis en vaardigheden (kwalificatie) beter meetbaar worden geacht, daardoor de examens sterk bepalen, waarmee op zijn beurt het onderwijs zich ook vooral op kennis en cognitieve vaardigheden zou richten. De Onderwijsraad vindt dat socialisering en persoonlijkheidsvorming meer aandacht moeten krijgen, maar pleit niet direct voor vermindering op cognitief vlak. Het Platform Onderwijs2032 doet dit wel, en wil het cognitieve curriculum afslanken. Verder zien alle genoemde instanties teaching to the test als een groot probleem. Hierdoor krijgt de 'versmalling' van het onderwijs een extra stimulans. In Amerikaans onderzoek naar neveneffecten van strikt accountabilitybeleid, wordt een meer fundamentele analyse gegeven van de werking van teaching to test. Uitganspunt is dat toetsen en examens altijd beperkte en gebrekkige representaties zijn van wenselijke onderwijsopbrengsten, zoals aangegeven in onderwijsdoelstellingen. Vooral In situaties waarin er sprake is van grote consequenties van goede toetsuitkomsten zouden scholen een sterke prikkel ervaren om leerlingen te trainen voor het examen, ten koste van andere belangrijke onderwijsdoelen. Een tweede lijn van kritiek is dat er teveel en te rigide getoetst wordt. De Inspectie gaf hier onlangs in een conferentie een bloemlezing van, waarbij veel aandacht was voor toetsstress bij leerlingen en leerkrachten die zich 'beklemd' voelen door de overheid.

Op de onderbouwing en de uitwerking van deze kritiek is veel aan te merken. Aangezien het niet alleen om vrijblijvende beschouwingen gaat, maar er ook toegewerkt wordt naar drastische veranderingen in het stelsel van examens en toetsen, is het van belang een 'fact check' te doen op de assumpties die worden gedaan en ook de basisconcepten kritisch te bekijken. Tornen aan examens moet per definitie gezien worden als een stelselherziening. Voor zover de afspraken over op evidentie gebaseerd beleid, gedaan door de Commissie Dijsselbloem, nog gelden is er alle aanleiding voor grondige empirische en analytische voorstudies. Enkele voorbeelden zijn: hoe representatief zijn de negatieve houdingen tegenover toetsen die worden gesignaleerd werkelijk? Is er een feitelijke basis voor de aanname dat de huidige kernvakken zoveel tijd in beslag nemen dat er geen ruimte is om aan bredere kennis en vorming

aandacht te geven? Kritiek op het verschijnsel 'teaching to the test' heeft voor een deel een technische basis, namelijk gebrekkige inhoudsvaliditeit; waarom lijkt bij voorbaat te worden uitgesloten dat er ook technische oplossingen zijn voor dit probleem? Maar ook de ambivalente houding van het onderwijsveld ten opzichte van toetsen vraagt om nadere analyse en onderzoek. Het lijkt erop dat de overladenheid in het toetsen door de scholen zelf veroorzaakt wordt (Van der Ploeg & Weijers, 2018). Het aantal verplichte eindtoetsen, volgtoetsen en examens in het primair en secundair onderwijs is zeer beperkt. Opmerkelijk is de spanning tussen 'summatief' of 'beslissend' gebruik van toetsen, terwijl in het debat vooral formatief gebruik van toetsen wordt gepropageerd. Zou het mogelijk zijn werkbare combinaties toe te passen, waarin zowel het beoordelende moment (het krijgen van cijfers) als het formatieve gebruik van toetsen te combineren zijn?

8.3 Fundamentele kwesties bij het veranderen van het stelsel van toetsen en examens

Verschillende motieven en ambities tot verandering

In de kritiek op toetsen en examens, maar ook in de beschrijving van de positieve functies van examens is het begrip onderwijskwaliteit centraal gesteld. Wat soms expliciet is voorgesteld (Platform Onderwijs 2034) of anders 'in de lucht hangt' (nota's van de Onderwijsraad) is dat de examens zouden moeten veranderen om de kwaliteit van het onderwijs op peil te houden of te verhogen. Met behulp van het model van onderwijskwaliteit dat in hoofdstuk 3 is besproken, kunnen we beter aangegeven welke facetten van onderwijskwaliteit in het geding zijn als het gaat om 'betere' examens. We kijken daarbij dan specifiek naar het verhogen van de responsiviteit en de verbetering van de effectiviteit van het onderwijsstelsel.

1. Vergroten responsiviteit

Een voor de hand liggend motief om de examens te veranderen is dat de eindtermen van het onderwijs niet meer goed aansluiten bij de eisen van vervolgonderwijs en de maatschappij. Daarnaast is eventueel modernisering te noemen als motief om veranderingen aan te brengen. Een ander motief dat zeker meespeelt zijn allerlei overwegingen over verbetering van processen op basis van veranderende onderwijskundige of pedagogische voorkeuren, die eventueel *geen* relatie hebben met verbeterde opbrengsten. Deze 'proceskwaliteit' heeft geen plaats in het effectiviteitsmodel, omdat kwaliteit uiteindelijk moet zijn verankerd in verbetering van opbrengsten. Een dramatische implicatie zou zijn om de koppeling tussen processen en opbrengsten los te laten, het examen geheel of gedeeltelijk te laten vervallen en er een procesevaluatie voor in de plaats te stellen. Dit zou neveneffecten kunnen hebben in de vorm van bureaucratisering (standard operating procedures, ISO-achtige kwaliteitscontrole op procedures) en *goal displacement* (middelen worden tot doel verheven). In dit verband is vermeldenswaardig dat in het kader van het project Curriculum.nu, voor een deel van de bepleitte 'brede vaardigheden', met name in het domein van sociale en emotionele vaardigheden, wordt gesuggereerd hiervoor alleen het aanbod te specificeren en geen eindtermen. De implicatie hiervan zou zijn dat er op deze gebieden alleen procesevaluatie kan plaatsvinden.[8]

[8] Curriculum.nu (December, 2018) Verslag expertbijeenkomst 'curriculum en toetsing' dd 6-09-2018

Het perspectief van verbetering van de responsiviteit heeft net zoveel te maken met het nagaan of vigerende doelstellingen zouden moeten worden bijgesteld als met de inhoud van examens en eindtoetsen. Volgens proactief planningsdenken komen eerst de doelstellingen en pas daarna de instrumenten om de realisatie ervan te meten. In onderwijsstelsels met een goed gefundeerd stelsel van examens en eindtoetsen en minder ontwikkelde doelstellingsspecificaties zou men de pragmatische keuze kunnen maken om te proberen de wenselijk geachte veranderingen meteen op het niveau van examenprogramma's en toetsmatrijzen uit te drukken. De situatie in Nederland lijkt wel een beetje op deze situatie van gefundeerde examens en weinig heldere doelstellingen. Echter, in een recente studie (Scheerens en Exalto, 2017) bleek dat voor basisvakken als taal en rekenen de vigerende eindtermen en zeker de referentieniveaus veelal als een acceptabele basis werden gewaardeerd. Tegelijkertijd heeft de OECD in twee evaluaties van het Nederlandse onderwijs het belang van een meer expliciet doelstellingskader onderstreept. In die situatie verdient het misschien aanbeveling om beide operaties, expliciteren van doelstellingskader en aanpassing van toetsen en examens, te combineren.
Ook op dit vlak zouden heldere keuzes vooraf van belang zijn. Geheel volgens de afspraak tot evidence-based onderwijsbeleid zou er een systematisch en representatief empirisch analytisch doelstellings-onderzoek kunnen worden ondernomen om het gewenste verbeterde doelstellingskader te onderbouwen (vlg. Stroomberg, 1977). Efficiëntie zou een belangrijke overweging moeten zijn bij dit soort vooronderzoek. Daarom zou overwogen kunnen worden om bij de verandering uit te gaan van bestaande examens en eindtoetsen en wijzigingen aan te brengen in de vorm van aanvulling door eventueel nieuwe sub-domeinen van items en eliminatie van wat minder relevant wordt geacht.

2. Vergroten effectiviteit

Een tweede motief om de examens bij te stellen is vergroting van de effectiviteit. De vraag is hier of examens en eindtoetsen de potentie waarmaken om positief bij te dragen aan de opbrengsten van het onderwijs. Aangrijpingspunten tot verbetering zijn hier de mechanismen die de positieve invloed van examens kunnen verklaren: stimuleren van extrinsieke motivatie, feedback ten behoeve van leerprocessen, en stimulering van curriculum alignment. Vanuit dit gezichtspunt kan gedacht worden aan een aantal gebieden, dat voor optimalisering en technische verbetering in aanmerking komt. Het is overigens interessant om te constateren dat de geleverde kritiek op examens voorbij lijkt te gaan aan de mogelijkheid van onderbenutting van technische mogelijkheden en sub-optimale uitvoering. Het zoeken naar oplossingen in de sfeer van technische verbetering staat niet op de agenda. Ook de door onderzoek deels onderbouwde insteek dat het stelsel van toetsen en examens als een 'hefboom' voor verbetering van onderwijsprestaties kan worden gezien ontbreekt in het Nederlandse debat. Mogelijkheden tot verbetering kunnen voor elk van de genoemde mechanismen (motivatie, feedback en alignment) worden geïnventariseerd.
Als het gaat om motivationele aspecten lijkt een milde prestatiedruk te prefereren boven een situatie van enerzijds vrijblijvendheid en anderzijds een regime met sterke sancties en beloningen. Hoewel de percepties hiervan verschillen (vergelijk het uitvoerig geciteerde verslag van de Inspectie uit 2017) lijkt het Nederlandse accountability-regime noch te slap, noch te hard. Nader onderzoek over de waardering en perceptie van het stelsel kan hierover meer duidelijkheid brengen.
Wat betreft de feedback functie zou kunnen worden nagegaan of op een aantal aspecten verbeteringen nodig zijn: het stimuleren van een criterium gerichte interpretatie van eindtoetsen en volgsystemen, het optimaliseren van het diagnostisch karakter van toetsen en het vergroten van de mogelijkheid tot adaptief toetsen.
Voor wat betreft de verbetering van curriculum alignment, zou in de eerste plaats gekeken moeten worden naar wat is omschreven als 'horizontale alignment', de aansluiting tussen (nationale)

onderwijsdoelstellingen en eindtoetsen en examens. Dit thema is op te vatten als analyse en onderzoek naar de inhoudsvaliditeit van de desbetreffende instrumenten. Een specifiek aandachtspunt hierbij zou de relatie tussen imperfecte dekking van doelstellingen en 'test inflation' (vgl het werk van Koretz) kunnen zijn. Binnen een context van beleid dat zowel het doelstellingskader zou willen herconstrueren als de examens bijstellen of updaten, zou gezocht kunnen worden naar een gecombineerd ontwerp.
Een ander onderzoeksthema in het kader van alignment is legitieme toetsvoorbereiding. Al in 1985 werd door Van der Linde toepassing van item-banking voorgesteld om betere toetsvoorbereiding mogelijk te maken (Van der Linde, 1985). Op basis van itembanken kunnen oefentoetsen worden samengesteld die inhoudelijk de examenstof dekken zonder examen items te dupliceren.
Wat betreft een eventuele optimalisering van 'verticale alignment' speelt de organisatorisch bestuurlijke opbouw van het Nederlandse onderwijsstelsel een grote rol. Verticale curriculum alignment betreft de aansluiting van verschillende curriculum componenten als doelstellingen, leergangen, leerboeken, schoolwerkplannen en formatieve toetsen, tot en met het dekking van de leerdoelen in de lessen ('content covered', of 'opportunity to learn'). Een optie is om verticale alignment over te laten aan het spel van vrije krachten van autonome scholen en marktgerichte providers; een andere optie is om te denken aan enigerlei vorm van centrale monitoring van de desbetreffende koppelingen en aansluitingen. Een en ander hangt nauw samen met de verhouding tussen standaardisatie en autonomie in het Nederlandse onderwijs. We zullen hier in de volgende paragraaf verder bij stilstaan.

Het spanningsveld tussen standaardisatie en autonomie
In voorafgaande hoofdstukken zijn examens en eindtoetsen gezien als structurele voorzieningen voor kwaliteitsborging en kwaliteitsverbetering. Het gaat hierbij om de standaardisering van de output van onderwijs. De optimalisering ervan is in de Nederlandse situatie begrensd door de hoge mate van autonomie van scholen, een invloedrijk middenveld van raden en een teruglopende centrale bestuurskracht. Bronneman (2011) merkt in dit verband op dat de invloed van het Ministerie van OCW gedurende de laatste 20 jaar is afgenomen en dat het overdragen van zeggenschap naar schoolbesturen niet tot meer autonomie van leerkrachten heeft geleid.

Scheerens (2016, p. 371- 373), bespreekt de bestuurlijke context en vergelijkt 'het accountability scenario' qua potentieel tot effectiviteitsverhoging met andere scenario's.
Kijkend naar het Nederlandse op kwaliteit gerichte onderwijsbeleid vallen twee scenario's op: het good governance scenario, gebaseerd op autonomie van de school, en het accountability- of verantwoordingsscenario, met voorzieningen rondom evaluatie en beoordeling. Autonomie zit in het DNA van het Nederlandse onderwijssysteem. De basis is het principe van onderwijsvrijheid, waarmee confessionele scholen dezelfde rechten kregen als openbare scholen. Onderwijsorganisaties en belangenverenigingen, zoals vakbonden of verenigingen van schoolbesturen, waren georganiseerd rondom de zuilen: een katholieke zuil, een protestante zuil en een neutrale zuil. Samen vormden deze organisaties een sterk invloedsblok, dat, volgens sommigen, corporatieve karakteristieken had (Leune, 1983). Deze sterke compenserende macht van onderwijsorganisaties bestaat ook vandaag de dag in gemoderniseerde vorm, waarbij de zuilorganisaties zijn vervangen door de raden voor primair en voortgezet onderwijs (PO- en VO-raad). Deze structuur drukt een algemene aversie uit tegen 'staatspedagogiek' en heeft gevolgen voor kwaliteitsbeleid en de rol van de onderwijsondersteunende structuur. Hervormings- en innovatiebeleid moeten zoveel mogelijk 'bottom up' zijn. Opmerkelijk is dat in het recente advies van de Onderwijsraad over examinering en toetsing de verantwoordelijkheid van

de centrale overheid voor het formuleren van eindtermen, in de context van curriculumvernieuwing, juist weer benadrukt wordt[9]. De raad neem stelling tegen de vermenging van het vaststellen van doelstellingen en uitwerking voor het onderwijs, welke hij signaleert in de optiek binnen het project Curriculum.nu

Onderwijsondersteunende organisaties werken niet als technische structuur die de implementatie van landelijk beleid kunnen faciliteren, zelfs wanneer deze binnen het kader van beleidsdoelen werken zoals vastgesteld in convenanten tussen het Rijk en het onderwijs. Zij worden gecontroleerd door scholen of door een netwerk van scholen die onder het zelfde bestuur vallen.

De relatief hoge score van Nederland op internationaal vergelijkende toetsen wordt wel uitgelegd als een gevolg van de grote autonomie van scholen. Internationaal onderzoek levert echter geen overtuigend bewijs voor de veronderstelling dat autonomie een oorzaak is van hoge onderwijsprestaties (Vgl. Maslowski, Scheerens & Luyten, 2007).

Er is daarentegen wel empirisch bewijs voor de hypothese dat de *combinatie* van grote zelfstandigheid op processen en centrale controle op de outcomes effectief is (Woessmann, et al, 2009, OECD, 2010, 2013). Deze combinatie benadrukt de principes van new public management, waarbij verwacht wordt dat de voordelen van krachtige, professionele autonomie bij werkprocessen en stevige controle op outcomes de beste resultaten leveren. In de Nederlandse situatie is deze combinatie van autonomie en standaardisatie van opbrengsten aan de orde, en dus een goede kandidaat als verklaring van de relatief goede prestaties.
Een deel van de kritiek en weerstand tegen examens en objectieve eindtoetsen heeft te maken met de machtsstrijd tussen ‘centraal’ en ‘decentraal’. De lijn die al aangegeven is in het citaat van Bronneman (2011) heeft zich verder doorgezet. Zeker wanneer men de ‘opiniemakers’ volgt die de Onderwijsinspectie (2017) ten tonele voert. Ondanks de teruglopende invloed van het centrale niveau zouden scholen en docenten zich beklemd voelen door de verplichte toetsen en examens. Zoals eerder aangegeven, zou het de moeite waard zijn om hier ‘fact checks’ toe te passen door middel van representatief opinieonderzoek en een analyse van de vrije ruimte die scholen hebben, naast de kernvakken die getoetst en geëxamineerd worden.

8.4 Naar een beleidsgericht onderzoeksprogramma gericht op fundamentele vragen rondom examens en toetsing

Zoals eerder betoogd moet de verandering van examens en eindtoetsen opgevat worden als een stelselherziening. Conform de afspraken die zijn gemaakt naar aanleiding van het rapport van de Commissie Dijsselbloem (2008) zouden stelselhervormingen alleen kunnen plaatsvinden op basis van een evidence-based benadering. Kort gezegd komt een op evidentie gebaseerde aanpak neer op een grondige ex ante evaluatie van het beoogde beleid, de assumpties waarop het gebaseerd is en de haalbaarheid, alsmede op afspraken over een ex post evaluatie van het programma. Hieronder wordt een aantal thema’s besproken dat een plaats heeft in een ‘ex ante’ evaluatie van voornemens om

[9] Onderwijsraad (2018) Toets wijzer: Naar een eigen(tijdse) wijze van toetsen en examineren. Den Haag: Onderwijsraad.

examen en eindtoetsen te veranderen. Samen vormen deze thema's de contouren van een beleidsgericht onderzoeksprogramma.

We gaan uit van drie soorten onderzoeksvragen:

a) *Vragen die betrekking hebben op een relatief eenvoudige 'fact check' op enkele aannames die in het debat over examens en toetsen tot nu toe gedaan zijn.*
b) *Vragen die als 'pijlers van evidentie' gelden bij de kritiek op de bestaande situatie en vigerende gedachten over hervorming van examens en eindtoetsen. Als zodanig beschouwen we in de eerste plaats empirisch analytisch doelstellingsonderzoek, en in de tweede plaats onderzoek naar de onderwijsbaarheid en meetbaarheid van '21st century skills'*
c) *Vragen over 'curriculum alignment' als een geheel van maatregelen dat de onderwijskwaliteit-bevorderende functie van examens en eindtoetsen kan verhogen.*

Ad a) De feitelijke basis voor enkele aannamen in het debat

<u>Beklemde docenten; gestreste leerlingen?</u>
De indruk wordt gewekt (vgl. Inspectie van het Onderwijs, 2017) dat scholen en docenten zich 'bekneld voelen' door de verplichtingen die de overheid oplegt wat betreft examens en eindtoetsen. De onderbouwing die hiervoor wordt gegeven is dat 'opiniemakers' dit vinden. Het is wenselijk dat deze opvatting wordt gecheckt door middel van een representatieve opiniepeiling onder relevante doelgroepen, ouders, docenten, schoolleiders en schoolbestuurders. Het is van belang dat bij deze peiling een goed onderscheid wordt gemaakt tussen examens en eindtoetsen die daadwerkelijk door de overheid verplicht worden en toetsen die door de scholen zelf gekozen zijn, dan wel deel uitmaken van methoden of onderwijsleerpakketten. Een verwant thema is de veronderstelde 'toetsstress' bij leerlingen. Naar dit thema lijkt nog nauwelijks onderzoek te zijn gedaan, en zou de mogelijkheid van een systematische peiling nagegaan kunnen worden. Verder is een nadere search naar relevant onderzoek op dit terrein nuttig.

<u>Cognitieve 'verenging' en vrije ruimte in het curriculum</u>
In de beschouwingen over 'een smalle kijk op onderwijskwaliteit' wordt de indruk gewekt dat er een te groot accent ligt op kernvakken en andere schoolvakken die worden geëxamineerd en dat er onvoldoende ruimte is voor sociale en persoonlijke vorming. Met andere woorden: de sturende werking van het examen zou een beletsel zijn voor scholen om voldoende tijd te besteden aan de 'brede ontwikkeling' van leerlingen.
Het is de moeite waard om het waarheidsgehalte van deze stelling nader te onderzoeken, en hierbij eventueel ook internationale vergelijkingen te betrekken. In de analyse is eventueel de enorme groei van particuliere bijscholing en examentraining te betrekken, die er op zouden kunnen wijzen dat er juist te weinig tijd wordt besteed aan basisvakken.

<u>Wat is de meerwaarde van vakkenintegratie?</u>
Vakkenintegratie is een curriculumvraagstuk dat relevant is voor toekomstig onderzoek, omdat het consequenties kan hebben voor 'verticale alignment', de aansluiting tussen curriculum en examen. In de vorige paragraaf hebben we betoogd dat alignment een aangrijpingspunt is bij het benutten van examens om de effectiviteit van het onderwijs te vergroten.
In het advies van het Platform Onderwijs2032 wordt gepleit voor vakkenintegratie. Ook in andere bijdragen aan het debat wordt aangenomen dat vakkenintegratie een goede zaak is (Onderwijsinspectie,

2017). Vakkenintegratie is een zeer forse ingreep, die uiteraard ook grote implicaties voor het examen heeft. De argumentatie die het Platform geeft is als volgt:
"Om het onderwijs meer betekenis voor leerlingen te geven, stelt het Platform voor die kennis in drie leerdomeinen te clusteren: Mens &Maatschappij, Natuur & Technologie, Taal & Cultuur. Leerlingen maken zich de kennis van die domeinen op een diepgaande manier eigen: niet van alles een beetje, maar meer van minder. Ze leren kennis uit verschillende vakken met elkaar in verband te brengen aan de hand van maatschappelijke vraagstukken. Scholen brengen hun leerlingen behalve kennis ook vakoverstijgende vaardigheden bij, die eveneens tot de vaste basis behoren. Het gaat om leervaardigheden, creëren, kritisch denken, probleemoplossend vermogen en samenwerken" (p. 8-9). Hier staan opvattingen tegenover die de waarde van de verschillende leervakken juist benadrukken, onder meer vanuit het gezichtspunt van aansluiting bij het vervolgonderwijs, terwijl duidelijk is dat de genoemde vaardigheden ook geoefend kunnen worden in relatie tot vakgebonden inhouden. Een literatuurstudie die argumenten en de onderbouwing daarvan nader documenteert zou wenselijk zijn om na te gaan of er voldoende basis is voor een dergelijke ingrijpende herstructurering. Overigens kan vakkenintegratie ook anders plaatsvinden dan door het combineren van inhoudelijke vakgebieden. In het kader van de bepleite 'brede vaardigheden' en 'skills' kunnen bepaalde vaardigheden wellicht in meerdere vakgebieden worden gestimuleerd (voorbeelden zijn systematische probleemoplossing, en motivatieaspecten bij feedback op prestaties).

Ad b)"pijlers van evidentie" bij de vigerende plannen tot hervorming van examens en eindtoetsen

Empirisch analytisch doelstellingsonderzoek

Systematisch onderzoek naar de identificatie, legitimering en uitwerking van onderwijsdoelstellingen is node gemist in de onderbouwing van de curriculumplannen van Platform Onderwijs2032 (vgl. Scheerens, 2016). Het belang van onderwijsdoelstellingen komt in dit rapport op meerdere fronten duidelijk naar voren, onder meer als kern van het gepresenteerde model van onderwijskwaliteit, maar ook als basis voor de uitwerking van toets- en examenprogramma's. Opnieuw is hier het begrip 'alignment' van belang. Een goede aansluiting tussen doelen en examens helpt om curriculaire focus aan te brengen, hetgeen de opportunity to learn vergroot (zie paragraaf 4.4).
Als aanvulling op de vigerende procedure om verder te komen met curriculumvernieuwing in het kader van Curriculum.nu en die, naar verluidt, vooral uit consultatie van het onderwijsveld bestaat, zou een representatief opgezet, empirisch analytisch doelstellingsonderzoek verder recht doen aan het belang van deze onderneming. De methodologie voor dit soort onderzoek is beschikbaar (vgl. Stroomberg, 1977).

De meetbaarheid en onderwijsbaarheid van 21ste -eeuwse vaardigheden

De stelling dat niet-cognitieve componenten een belangrijker plaats zouden moeten hebben in het onderwijs is een van de belangrijkste achterliggende motieven in het debat over zowel doelstellingen en curriculum vernieuwing, als over eventuele bijstelling van examens en eindtoetsen.
Vanuit onderwijskundige optiek is dit niet direct een nieuw thema. Zo maakte Eric de Corte in 1973 het onderscheid tussen inhoudelijke en formele onderwijsdoelstellingen. Bij de formele doelstellingen ging het met name om cognitieve processen die op meerdere vakgebieden relevant waren (toenmalige voorbeelden waren leren samenwerken en 'leren leren'). Vervolgens ontstond er op het gebied van beroepsonderwijs belangstelling voor bredere competenties, die wel werden gezien als combinaties van cognitieve, affectieve en conatieve facetten (Kuijpers, 2003). Ook de OECD had een inbreng in enkele publicaties over zogenoemde 'key competencies' (Rychen and Salganic, 2003)). Geleidelijk deed

internationaal de term '21st century skills' opgang. Inmiddels was het concept onder invloed van vooral bijdragen uit de VS in de sfeer van 'emotionele intelligentie', 'positieve psychologie' en 'karakter ontwikkeling' verbreed.
Naast meetbaarheid is de onderwijsbaarheid van de niet-cognitieve attributen in het geding. Globaal zijn er twee benaderingen: specifieke cursussen om dit soort vaardigheden bij te brengen en het geven van meer expliciete aandacht aan deze facetten in 'normale' vakgerichte onderwijsleersituaties. Het gaat daarbij dan bijvoorbeeld om sociale facetten van samenwerking bij groepstaken, het pedagogisch omgaan met emotionele kanten van prestatie-feedback, het bespreken van normen en waarden bij gedragsproblemen en reflectie op interculturele verschillen (vlg. J. Scheerens: *Informal Learning for Active Citizenship at School*, 2009). Burgerschapskunde is inmiddels ingevoerd, en is ook evalueerbaar als onderwijsopbrengst, omdat er een duidelijk cognitieve component aan zit.
De gegevensbasis over effecten van programma's tot karaktervorming is beperkt en niet overtuigend (Scheerens, 2017). Uit schooleffectiviteitsonderzoek, waarin geprobeerd is niet-cognitieve effecten te meten, blijkt telkens dat de effecten die de school erop heeft aanmerkelijk kleiner zijn dan bij de cognitieve opbrengsten (vgl. bv Timmermans, 2015). Bovendien speelt de nature/nurture-discussie als het gaat om – niet alleen de onderwijsbaarheid –maar überhaupt de beïnvloedbaarheid van factoren die in de buurt komen van persoonlijkheidskenmerken. Samengevat zou vooronderzoek wenselijk zijn, alvorens te besluiten op welke wijze men het curriculum en de examens zou willen bijstellen in de richting van 21st century skills.

Ad c) Vragen over 'curriculum alignment' als een geheel van maatregelen dat de onderwijskwaliteit bevorderende functie van examens en eindtoetsen kan verhogen.

Onderzoeksthema's in verband met de functie van toetsen en examens in de context van curriculum alignment zijn:

- 'Horizontale alignment'; de inhoudsvaliditeit, c.q. doelstellingsvaliditeit van examens en toetsen
- Toets- en examentechnische innovaties.
  Een nadere uitwerking van de ontwikkelingen die in de Hoofdstuk 6 en 7 van dit rapport zijn behandeld.
- Haalbaarheid van gecombineerde constructie van een doelstellingskader en een programma van examens en eindtoetsen.
  Bestuurlijk-organisatorische randvoorwaarden in Nederland, voor zover centrale monitoring van verticale alignment zou kunnen worden overwogen.
- Internationaal vergelijkend onderzoek naar curriculum alignment met een aantal voor Nederland interessante onderwijsstelsels.

Bijlage 1 Referenties

American Psychological Association, American Educational Research Association & National Council on Measurement in Education (2004). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition).* New York: Longman

Argyris, C., & Schön, D.A. (1978). *Organizational learning: a theory of action perspective.* Reading, Massachusetts: Addison- Wesley Publishing Company.

Baartman, L. (2008). *'Assessing the assessment'. Development and use of quality criteria for Competence Assessment Programmes*. Proefschrift: Dutch Interuniversity Centre for Educational Research.

Baker, R (2010). Data mining for education. In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education* (3rd edition). Oxford, UK: Elsevier.

Ball, S.J., Mac Maguire, M., and Braun, A.. (2012) *How Schools Do Policy: Policy Enactments in Secondary Schools*. New York: Routledge.

Behn, R.D. (2003). *Why Measure Performance? Different Purposes Require Different Measures.* Public Administration Review, 63(5), 586-606.

Bevan, G. & Hood, C. (2006). *What's measured is what matters: Targets and gaming in the English Public Health Care system.* Public Administration, 84(3), 517–538.

Bevan, G. & Hood, C. (2006). What's measured is what matters: Targets and gaming in the English Public Health Care system. *Public Administration,84*(3), 517–538.

Biesta, G. (2012a). "The future of teacher education on: Evidence, competence or wisdom?" in *Research on Steiner Education* 3:1(pp. 8-21).

Biesta, G. (2012) *Goed onderwijs en de cultuur van het meten.* Den Haag: Bool/Lemma.

Bishop, J. (1997). *The effect of National Standards and Curriculum-Based Exams on Achievement*. Cornell University. Center for Advanced Human Resource Studies.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to Improve Learning*. New York: McGraw-Hill
Tyler, R.W. (1949). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago.

Bol, T., Witsche, J. Van der Werfhorst, H.G., & Dronkers, J. (2014) Curricular Tracking and Central Examinations: Counterbalancing the Impact of Social Background on Student Achievement in 36 Countries (2014$0 *Social Forces* 92(4) 1545–1572, June 2014
doi: 10.1093/sf/sou003 Downloaded from https://academic.oup.com/sf/article-abstract/92/4/1545/2235815.

Boom, Stefan (2003) A continuing story: Casustoetsing in het eindexamen geschiedenis, *Kleio, 44-1*, januari 2003, 19-25.

Borghans, L., Velden, R. van der, Büchner, C., Coenen, J., & Meng, C. (2007). *Het meten van onderwijskwaliteit en de effecten van recente onderwijsvernieuwingen*. Deelonderzoek uitgevoerd door Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA). Maastricht: Universiteit Maastricht, ROA.

Bosker, R.J., & Scheerens, J. (1995). A self-evaluation procedure for schools using multilevel modelling. *Tijdschrift voor Onderwijsresearch*, 20(2), 154-164.

Braithwaite, J., Makkai, T., & Braithwaite, V. (2007). *Regulating Aged Care; Ritualism and the New Pyramid*. Cheltenham: Edward Elgar.

Bronneman-Helmers, R, (2011), *Overheid en onderwijsbestel Beleidsvorming rond het Nederlandse onderwijsstelsel, (1990-2010)* The Hague:, SCP-publication 2011-31.

Bruijn, H. de (2008). *Managers en Professionals; Over management als probleem en oplossing.* Den Haag: Sdu Uitgevers bv.

Carnoy, M., Elmore, R., & Siskin, L. (2003) (Eds.). *The New Accountability. High Schools and High-Stakes Testing*. New York/London: Routledge Falmer.

Case, B. and Zucker, S. (2005) Horizontal and Vertical Alignment. *Presentation Beijing, China at the China–US Conference on Alignment of Assessments and Instruction.* Pearson Policy Report.

Causa, O., & Chapuis, C. (2009). *Equity in Student Achievement Across OESO Countries: An investigation of the role of policies*. OESO Economics Department Working Papers, No. 708. OESO Publishing. Doi: 101787/223056645650.

Chibulka, J.G, & Derlin, R.L. (1995). State educational performance reporting policies in the U.S.: accountability's many faces. *International Journal of Educational Research, 23*(6), 479-492.

De Klerk, S., Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & education,* 85, 23-34.

Downing, S.M., & Haladyna, T.M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Drost, M. & Verra, P. (2015) *Handboek RTTI – Opbrengstgericht werken met RTTI.* Bodegraven: Uitgeverijplus.

Dulmers, R.J. (1988) (Red.). *Marketing voor scholen*. Alphen aan den Rijn: Samsom.

Elmore, R.F. and Associates,. (1990) *Restructuring School; The Next Generation of Educational Reform.* San Francisco: Jossey Bass.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, S. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests.* Amsterdam: NIP. [www.psynip.nl: https://www.psynip.nl/wp-content/uploads/2016/07/COTAN-Beoordelingssysteem2010.pdf].

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York, NY, USA:: Springer series in statistics.

Gartner (2018). *What Is Big Data? - Gartner IT Glossary - Big Data*. Gartner IT Glossary (retrieved on 07/06/2018).

Glas, G.V. (1972) The many faces of accountability. *Phi Delta Kappan,* 53, 635-659.

Groot, A.D. de (1986). Is de kwaliteit van het onderwijs te beoordelen? In: A.D. de Groot, *Begrip van evalueren*. Den Haag: VUGA.

Hanushek, E.A., & Woessmann, L. (2005). *Does Educational Tracking Affect Performance and Inequality? Differences- in-Differences Evidence across Countries*. NBER Working Paper, No. 11124.

Hanushek, E.A., & Woessmann, L. (2009). *Do better schools lead to more growth?* Cognitive Skills, Economic Outcomes, and Causation. NBER, Cambridge, MA, WP 14633, National Bureau of Economic Research (January).

Hattie, J. (2009). *Visible Learning*. Abingdon: Routledge.

Heiting, M.C. (2018). Unpublished doctoral thesis.

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation, 163-189*. Greenwich, CT: Information Age Publishing.

Honingh, M.E. ; Ehren, M.C.M (2012) Onderwijstoezicht in een polycentrisch sturingsmodel; Dilemma's bij het vaststellen en verbeteren van de onderwijskwaliteit. *Bestuurskunde*, vol. 21, iss. 4, (2012), pp. 64-72.

Inspectie van het Onderwijs (2017) *"Is meten wel weten?" met als subtitel "Zijn we doorgeschoten in onze behoefte de kwaliteit van het leren uit te drukken in cijfers?"* Verslag studiemiddag op 19 oktober, 2019. Utrecht: Inspectie van het Onderwijs.

International Test Commission (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. [www.intestcom.org: https://www.intestcom.org/files/guideline_computer_based_testing.pdf].

International Test Commission (2013). *ITC Guidelines on Test Use. Version 1.2.* [www.intestcom.org: https://www.intestcom.org/files/guideline_test_use.pdf].

International Test Commission (2014). *International Guidelines on the Security of Tests, Examinations, and Other Assessments.* [www.intestcom.org: https://www.intestcom.org/files/guideline_test_security.pdf].

Janssens, F.J.G. (2011). The impact of the publication of school performance indicators in the Netherlands. *International Journal of Educational Policies*, 5(2), 55-73.

Joosten-ten Brinke, D. (2011). *Eigentijds toetsen en beoordelen*. Lectorale rede. Tilburg: Fontys Lerarenopleiding Tilburg.

Kane, Th.,J., Kerri, K.A., & Pianta, R. (2014) *Designing teacher evaluation systems. New guidance from the Measures of Effective Teaching Project.* San Francisco: Jossey Bass.

Kleij, F.M. van der (2013). *Computer-based feedback in formative assessment*. Proefschrift. Enschede: Universiteit Twente.

Kuijpers, M.A.C.T. (2003) *Loopbaanontwikkeling. Onderzoek naar "Competenties".* Enschede, Twente University Press.

Little, Leeuw, A.C.J. de (1982). *Organisaties, management, analyse, ontwerp en verandering. Een systeemvisie*. Assen: Van Gorcum.

Looney, J. W. (2011) *Alignment in complex education systems: achieving balance and coherence*. OECD Education Working Paper No. 64.

Louis, K. S., Dretzke, B. &Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement, 21*(3), 315 -336.

Maassen, N., & Den Otter, D. (2014). *Eerste hulp bij toetsen: grip op toetskwaliteit*. Enschede: RCEC. https://www.nro.nl/wpcontent/uploads/2014/12/RCEC_Kwaliteit_Toets_Checklist_2014.pdf.

McMillan, J. H., & Nash, S. (2000). *Teacher Classroom Assessment and Grading Practices Decision Making.* Paper presented at the 2000 NCME Annual Meeting, April 27, New Orleans.

Maslowski, R. ; Scheerens, J. ; Luyten, H. / The effect of school autonomy and school internal decentralization on students' reading literacy. In: *School Effectiveness and School Improvement.* 2007 ; Vol. 18. pp. 303 – 334.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4), 6-20.

Ministerie van OCW (2007). *Krachtig meesterschap*. Kwaliteitsagenda voor het opleiden van leraren 2008-2011. Den Haag: Ministerie OCW.

Ministerie van OCW (2007). *Tekenen voor Kwaliteit*. Kwaliteitsagenda Voortgezet Onderwijs. Den Haag: Ministerie OCW.

Ministerie van OCW (2007). *Scholen voor morgen*. Kwaliteitsagenda PO. Den Haag: Ministerie OCW.

Ministerie van OCW (2008). *Werken aan vakmanschap*. Strategische agenda Beroepsonderwijs en Volwasseneneducatie 2008-2011. Den Haag: Ministerie OCW.

Ministerie van OCW (2009). *Bestel in beeld 2008*. Den Haag: Ministerie van OCW.

Mourshed, M., Chijioke, C., and Barber, M. (2010). *How the world's most improved school systems keep getting better.* McKinsey and Company.

Nusche, D., Braun, H., Halasz, G., & Santiago, P. (2014). *OECD reviews of evaluation and assessment in education: Netherlands 2014*. Paris, France: OECD.

OECD (2010). *Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States*. Paris: OECD.

OECD (2011) *Quality time for students; Learning in and out of schools.* Paris; OECD Publishing.

OESO (2007). *PISA 2006. Science Competencies for Tomorrow's World.* Paris: OESO.

Onderwijsraad (2007*) Sturen van vernieuwende onderwijspraktijken 19 november 2007* | Verkenning Den Haag: Onderwijsraad.

Onderwijsraad (2015*) Maatwerk binnen wettelijke kaders: eindtoetsing als ijkpunt voor het funderend onderwijs* 13 november 2015 | Advies. Den Haag: Onderwijsraad.

Onderwijsraad (2016*) Een ander perspectief op professionele ruimte in het onderwijs*. 27 september 2016 | Advies. Den Haag: Onderwijsraad.

Onderwijsraad (2013) *Een smalle kijk op onderwijskwaliteit*. 4 november 2013 | Advies. Den Haag: Onderwijsraad.

Onderwijsraad (2016**)** *De volle breedte van onderwijskwaliteit.* 10 mei 2016 | Advies Den Haag: Onderwijsraad.

Oomens, M., Veldkamp, B & Scheerens, J. (2015) *Marktverkenning formatieve evaluatie.* Utrecht: Oberon.

Oomens, M., Exalto, R., De Jong, A., Scholten, F., Veldkamp, B., Janse, R., Scheerens, J. (2017) *Marktonderzoek formatief evalueren. Een onderzoek naar vraag en aanbod*. Utrecht: Oberon.

Operation Education (2017) Waarom een Centraal eindexamen?. https://operation.education/dossier-waarom-centraal-eindexamen/

Pannecoucke, I. (2005). *Gewikt en gewogen: Schoolkeuze tussen principe en feiten*. Antwerpen: Universiteit van Antwerpen, OASES – Onderzoeksgroep Armoede, Sociale Uitsluiting en de Stad.

Peschar, J.L., & Wesselingh, A.A. (1985). *Onderwijssociologie: een inleiding*. Groningen: Wolters-Noordhoff.

Pirsig, R.M. (1999). *Zen and the art of motorcycle maintenance*. London: Vintage.

Platform Onderwijs2032 (2016) *Eindadvies: Ons onderwijs2032*. Rijksoverheid. https://www.rijksoverheid.nl/documenten/rapporten/2016/01/23/eindadvies-platform-onderwijs2032-ons-onderwijs2032.

Python software corporation (2018). *Python.* https://www.python.org.

Rand News Release (July 25, 2000). *Explaining achievement gains in North Carolina and Texas*.

Rapport "Commissie Dijsselbloem" Parlementair Onderzoek Onderwijsvernieuwingen (2008) https://www.parlement.com/id/vhnnmt7mtyqi/parlementair_onderzoek.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for statistical computing, Vienna, Austria. URL: http://www.R-project.org/

Rychen D., and Salganik, L. H. (2003) *Key Competencies for a Successful Life and a Well-Functioning Society.* Hogrefe Publishing.

Sandahl, J. (2015) *Citizenship Education in Upper Secondary School: Subject Knowledge and Citizenship Education in History and Social Science Education*. Stockhom: University of Stockholm.

Sandahl, J. (2015) *Social studies as socialisation, qualification and subjectification* Paper prepared for Panel "Citizens still in School: Motivations for Political Participation", chaired by Julie Ane Ødegaard and Trond Solhaug, ECPR's General Conference (Montreal), 26-29 August 2015.)

Sanders, P.F., Dijk, P. van, Eggen, T., Otter, D. den, & Veldkamp, B. (2016). *RCEC Beoordelingssysteem voor de kwaliteit van studietoetsen en examens.* Enschede: RCEC. Online: www.rcec.nl/Beoordelingssysteem.

Sanders, P.F. (Red.) (2013). *Toetsen op School*. Arnhem: Cito. Online: www.toetsenopschool.nl.

Sanders, P.F. (Red.) (2016). *Toetsen op School: Hoger onderwijs*. Arnhem: Cito. Online: www.toetsenopschool.nl.

Sanders, P., Brouwer, A., & Veldkamp, B. (2017*). Onderzoek naar de inhoudsvaliditeit van een tweetal centrale examens*. Enschede: RCEC.

Scheerens, J. (1983). *Evaluatie-onderzoek en beleid: methodologische en organisatorische aspecten* (Vol. 68): Stichting voor Onderzoek van het Onderwijs.

Scheerens, J. (2013). *Educational evaluation and assessment in the Netherlands. Addendum to the 2012 report*. Retrieved from http://www.oecd.org/education/school/Netherlands%20CBR%20Update.pdf.

Scheerens, J. (2009). *Het innoverend vermogen van de onderwijssector en de rol van de ondersteuningsstructuur*. Notitie en presentatie in opdracht van het Ministerie van OCW. Enschede: Vakgroep Onderwijsorganisatie en –management. Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). *Review and Meta-Analyses of School and Teaching Effectiveness*. Enschede: University of Twente, Department of Educational Organization and Management. 374 pp.

Scheerens, J. Luyten, H., and Van Ravens, J. (2011) *Perspectives on educational quality. Illustrative outcomes on primary and secondary schooling in the Netherlands.* Dordrecht, Heidelberg, New-York, London: Springer.

Scheerens, J. (2016) *Educational effectiveness and Ineffectiveness. A critical review of the knowledge base.* Dordrecht, Heidelberg, New-York, London: Springer.

Scheerens, J. en Exalto, R., (2017). *"Teaching to/from the test". Een verkennende studie naar het in lijn brengen van doelen, toetsen, curriculum en onderwijsaanbod*. Utrecht: Oberon.

Scheerens, J. (2017) ed. *Opportunity to learn, curriculum alignment and test preparation. A review study. Dordrecht, Heidelberg, New York, London: Springer.*

Scheerens, J. (2014) ed. *Effectiveness of time investments in education. Insights from a review and meta-analysis.* Dordrecht, Heidelberg, New-York London: Springer.

Scheerens, J. (2009) *Informal Learning for Active Citizenship at School*, Dordrecht, Heidelberg, New-York, London: Springer.

Sireci, S.G. (1998a). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299-321.

Sireci, S.G. (1998b). The construct of content validity. *Social indicators Research*, 45, 83 – 117.

Sireci, S.G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26, 100 – 107.

Sluijter, C. (2016). De kwaliteit van toetsen en examens in het middelbaar beroepsonderwijs. In P.F. Sanders (Red.) *Toetsen op School. Middelbaar Beroepsonderwijs*. (pp. 102-116). Arnhem: Cito. Online: www.toetsenopschool.nl.

Sluijter, C. Hemker, B, & Eggen, T. (2018). *Beoordelen van de kwaliteit van toetsen en examens. Deel 1: Systemen en criteria*. Arnhem: Cito.
Online: www.toetswijzer.nl.

Straetmans G., & Sanders P.(2001). *Beoordelen van competenties van docenten*. Den Haag: Programmamanagement EPS/HBO-raad.

Stroomberg, H. P. (1977) *Communale rekendoelen. Een empirisch onderzoek naar doelstellingen van het rekenonderwijs*. Amsterdam: RITP (Academisch Proefschrift).

Timmermans, A. (2012) *Value added in educational accountability.* Groningen: University of Groningen (doctoral thesis).

Thiel, S. van, & Leeuw, F.L. (2002). The performances Paradox in the Public sector. *Public Performance & Management Review, 25*(3), 267-281.

UNESCO (2004). *The Quality Imperative*. EFA Global Monitoring Report 2005. Paris: Unesco.

Van der Linden, W. J. (2006). *Linear models for optimal test design*. Springer Science & Business Media.

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.

VO-raad: *Examinering voortgezet onderwijs toe aan herijking*. 29 maart 2018.

Van de Werfhorst, H. G. (2011). Selectie en differentiatie in het Nederlandse onderwijsbestel: gelijkheid, burgerschap en onderwijsexpansie in vergelijkend perspectief. *Pedagogische Studiën*, *88*(4), 283-297.

Van de Werfhorst, H. G., & Mijs, J. J. B. (2010). Achievement inequality and the institutional structure of educational systems: a comparative perspective. *Annual review of sociology*, *36*, 407-42.

Van de Werfhorst, H., Dronkers, J., Karsten, S., Van der Velden, R., & Webbink, D. (2011) *Educational systems and four central functions of education.* Research program funded by NWO Programming Council for Educational Research (PROO) 2011-2015.

Van der Ploeg, S. & Weijers, S. (2018). *Kwaliteitsborging schoolexaminering voortgezet onderwijs.* Utrecht: Oberon.

Veldkamp B.P. (2016. De inhoud en constructie van toetsen. In Sanders, P.F. (Red.), *Toetsen op School: Hoger onderwijs* (pp. 21 – 30). Arnhem: Cito.
Online: www.toetsenopschool.nl.

Veldkamp, B., Schildkamp, K., Keijsers, M., Visscher, A., & de Jong, T. (2017). *Verkenning data-gedreven onderwijsonderzoek in Nederland.* Universiteit Twente, The Netherlands.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer.* Routledge.

Webb, N. L. (1997). *Research monograph No. 6. Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.

William (2013) *Introduction to Sociology* is Rice University.
https://opentextbc.ca/introductiontosociology/chapter/chapter16-education/

Woessmann, L., Luedemann, E., Schuetz, G., & West, M.R. (2009). *School Accountability, Autonomy and Choice around the World*. Cheltenham, UK/ Northampton, MA, USA: Edward Elgar.

Woessmann, L., Luedemann, E., Schuetz, G., & West, M.R. (2009). *School Accountability, Autonomy and Choice around the World*. Cheltenham, UK/ Northampton, MA, USA: Edward Elgar.

Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives,* 30, 3-32.

Woessmann, L. (2018). Central exit exams improve student outcomes. External school leaving exams raise student achievement and improve how grades are understood in the labor market, IZA *World of Labor* January 2018: 419.

Annex 1: Verantwoording opzet en onderzoeksmethoden

Anne Luc van der Vegt en Jaap Scheerens

In deze verantwoording maken we duidelijk op welke manier de onderzoeksvragen worden beantwoord en welke onderzoeksmethoden hiervoor zijn gehanteerd.

Beantwoording van de onderzoeksvragen
Het onderzoek geeft antwoord op drie typen onderzoeksvragen: 1) vragen over de typering en functies van toetsing en examinering, 2) vragen over basiskenmerken van het Nederlands examenstelsel, ook in internationaal perspectief, 3) vragen over fundamentele thema's. Hieronder volgt een volledige weergave van de onderzoeksvragen uit het onderzoeksvoorstel, op basis waarvan NRO subsidie heeft verstrekt. Per onderzoeksvraag geven we aan in welke hoofdstukken deze wordt beantwoord.

1. *Onderzoeksvragen over de typering en functies van toetsing en examinering*
   a) *Vanuit welke basisoptiek moeten examens geanalyseerd worden?*
      Hierbij worden als invalhoeken niet alleen kwaliteitsborging maar ook kwaliteitsverbetering gekozen.
   b) *Welke kernfuncties heeft toetsing en examinering, en welke evidentie is relevant om te kunnen nagaan hoe deze functioneren?*
      Als uitgangspunt worden de volgende functies onderscheiden:
      1) de legalistische functie van examens als formalisering van het civiel effect van het met goed gevolg doorlopen hebben van onderwijsprogramma's;
      2) de evaluatieve en publieke rekenschapsfunctie van examens (Engelse term: accountability);
      3) de kwaliteitverhogende werking van examens;
      Ten aanzien van deze laatste functie wordt aandacht besteed aan drie mechanismen: 1) de motiverende functie voor actoren, met name leerlingen, 2) het stimuleren van leer- en verbeterprocessen op basis van feedback op basis van examenresultaten, 3) het bieden van curriculaire focus. De derde "functie" is een relatief nieuw gezichtspunt om naar examens te kijken, zeer relevant in discussies waarin examens niet zelden als "belemmerend" worden neergezet.

De beantwoording van vraag 1 klinkt door in het hele rapport. Vraag 1a) en 1b) zijn daarbij nauw met elkaar verbonden. Dit wordt duidelijk in hoofdstuk 4, waar kwaliteitsborging en kwaliteitsverbetering worden besproken als kernfuncties van toetsing en examinering. Daarna volgt een bespreking van de kritiek op de kwaliteitsbevorderende werking van toetsen en examens (hoofdstuk 5). Het *borgen* van kwaliteit staat vervolgens centraal in hoofdstuk 6; de mogelijkheden tot *verbetering* van kwaliteit, door innovaties in toetsen en examens, in hoofdstuk 7.

2. *Onderzoeksvragen over basiskenmerken van het Nederlandse examenstelsel, ook in internationaal vergelijkend perspectief*
   a) *Wat zijn mijlpalen in de historische ontwikkeling van toetsing en examinering?*
   b) *Wat zijn de relevante criteria om het Nederlandse examenstelsel beschrijvend te typeren?* Mede in verband met internationale vergelijking wordt gekeken naar *dekking* (hoeveel onderwijsprogramma's en onderwijsniveaus worden gecoverd), *besturingsniveau* (welke instantie gaat over de examens, gemengde systemen, school en centraal examen, ideeën over policentrische kwaliteitsbewaking), en *methodologische typering*, kwantitatieve toetsing,

kwalitatieve methoden, inbedding in een ruimere onderwijs evaluatieve context, met name ook verband met meer formatieve evaluatie. Een aandachtspunt hierbij is het onderscheid tussen centraal examen (CE) en schoolexamen (SE) en de relatie tussen beide.

c) *Hoe is de waardering van het Nederlandse examenstelsel?*
Hierbij gaat het om waardering door betrokkenen, verrichte evaluaties (om door de Inspectie), en reviews door gezaghebbende internationale organisaties (OECD, Europese Commissie). Ook zal hierbij gekeken worden naar beschikbare gegevens over prestatiedruk en examenstress bij leerlingen.

Het eerste hoofdstuk geeft beknopt de historische ontwikkeling weer van het examen in Nederland en geeft daarmee antwoord op vraag 2a.
Wat betreft vraag 2b: in de hoofdstukken 2 en 3 wordt inzicht gegeven in de criteria waaraan de kwaliteit van het onderwijs kan worden afgemeten, vanuit verschillende maatschappelijke functies en verschillende belanghebbenden. Die criteria zijn van belang bij het beschrijven van het Nederlandse examenstelsel, in vergelijking met andere Europese landen. De internationale vergelijking wordt gemaakt in Annex 3. Onderscheid tussen schoolexamen (SE) en centraal examen (CE) komt aan de orde bij de bespreking van de geschiedenis van het examen (hoofdstuk 1), de kwaliteit van SE en CE vanuit het oogpunt van verschillende belanghebbenden (hoofdstuk 3.5), kritiek op het examen (hoofdstuk 5) en kwaliteitsborging en verbetering van examens (6.1 over centrale examens en 6.2 over schoolexamens).
Vraag 2c wordt deels beantwoord in hoofdstuk 1, waar internationaal onderzoek naar het Nederlandse stelsel, van o.a. de OECD, wordt besproken. Ook hoofdstuk 4 is relevant voor de beantwoording van deze vraag, met name de analyse van mechanismen als accountability, curriculaire focus en alignment, die de kwaliteitsbevorderende werking van examens verklaren. Ook hierbij worden internationale studies besproken (zie met name paragraaf 4.4 en 4.5).

3. *Onderzoeksvragen over fundamentele thema's*
Fundamentele thema's in relatie tot toetsing en examinering zijn a) de kwaliteitbevorderende functie, b) de houding binnen het onderwijs ten opzichte van toetsing en examinering, c) verandering en modernisering van het examen.
   a) *Welke theoretische en empirische onderbouwing is er te geven van de kwaliteitbevorderende functie van examens (te midden van andere vormen van onderwijsevaluatie en assessment)?*
   Hierbij kan kwaliteit worden afgemeten aan o.a. externe kwaliteitsbeoordeling (Inspectie), waardering van stakeholders (ouders en leerlingen), aansluiting op vervolgonderwijs.
   b) *Welke examen- en toetskenmerken zijn van belang om de kwaliteitbevorderende werking te stimuleren?*
   Denk aan inhouds- en criteriumvaliditeit van toetsen en examens, mogelijkheid tot adaptief toetsen, facilitering van "legitieme" toets voorbereiding; "meta-datering" van toetsen en examens met het oog op te realiseren cognitieve operaties.
   c) *Welke plaats hebben examens in een goede curriculaire "uitlijning" (alignment) van onderwijsstelsels, en welke mogelijkheden zijn er voor aansluiting met meer formatieve toetsen en volgsystemen?*
   d) *Welke factoren spelen een rol bij acceptatie en weerstand tegenover toetsen en examens en welke mogelijkheden zijn er om positieve betrokkenheid te vergroten?*
   Bij weerstand kan worden gedacht aan prestatiedruk en examenstress. Bij het vergroten van acceptatie aan ondersteuning van feedback, en "feedforward" en aansluiting bij evaluatie en feedback als integraal onderdeel van effectieve instructie en het professionele repertoire van leerkrachten.
   e) *Welke posities zijn er in de huidige discussie over verandering en modernisering van het examen?* (vgl. het debat rondom "Onderwijs2032", discussienota's van de Onderwijsraad over

onderwijskwaliteit en het advies van de Onderwijsraad dat in het voorjaar van 2018 wordt verwacht);

f) *Welke waarden en normatieve posities spelen daarbij een rol?* (bijvoorbeeld gelijkheid van kansen);
Hoe "evidence based" zijn de belangrijkste stellingnames? (zie ook Onderwijsraad, 2014[10] en 2017[11]).
g) *Wat zijn belangrijke thema's voor verder onderzoek in een programmalijn met betrekking tot het examenstelsel?*
Hierbij wordt onder meer gedacht aan empirische internationaal vergelijkend onderzoek, empirisch doelstellingsonderzoek in het kader van "Onderwijs2032" en fundamenteel onderzoek naar de toetsbaarheid en "onderwijsbaarheid" van "21ste -eeuwse vaardigheden.

De fundamentele vragen over toetsing en examinering (onderzoeksvraag 3) vormen een rode draad door het hele rapport.
Vraag 3a: verklaringen voor de kwaliteitsbevorderende functie van examens worden besproken in hoofdstuk 4.
Vooral hoofdstuk 7 is relevant voor het antwoord op vraag 3b. Hierin worden relevante examen- en toetskenmerken genoemd en ingegaan op mogelijkheden om aan de hand hiervan toetsen en examens te vernieuwen.
Alignment (vraag 3c) is een onderwerp dat uitvoerig aan de orde komt in paragraaf 4.4 en 4.5.
Acceptatie en weerstand tegenover toetsing (vraag 3d) komt met name aan bod in hoofdstuk 5, over de huidige kritiek op examens en toetsing. In datzelfde hoofdstuk gaan we in op verschillende posities ten aanzien van modernisering van het examen, onder meer van Onderwijs2032 en de Onderwijsraad (vraag 3e). De positie van de Onderwijsraad wordt ook besproken in hoofdstuk 2 en 4.
Tenslotte is hoofdstuk 8 gewijd aan de thema's voor verder onderzoek. Hier wordt, zoals gespecificeerd in de onderzoeksvraag, onder meer aandacht besteed aan doelstellingenonderzoek en aan curriculumvraagstukken, vanwege de verbondenheid tussen curriculum en examens, die besloten ligt in het begrip 'alignment'.

Verantwoording onderzoeksmethoden
Hieronder beschrijven we op welke manier het onderzoek is uitgevoerd en waar er afwijkingen zijn van de subsidieaanvrage.

Desk research
Dit onderdeel is uitgevoerd conform de subsidieaanvraag. Er is gebruik gemaakt van:
- wetenschappelijke literatuur, zowel Nederlands als internationale literatuur zoals beschreven in subsidieaanvraag;
- documentatie over het Nederlandse onderwijsbestel, zoals in beschreven in subsidieaanvraag;
- beleidsdocumenten, zoals adviezen van Onderwijsraad, Onderwijs2032, VO-raad, rapportages over relevant beleidsonderzoek;
- opiniërende artikelen, op basis van een literatuursearch in vakbladen.

Opiniërend onderzoek naar de waardering van het examenstelsel door belanghebbenden

10 https://www.onderwijsraad.nl/publicaties/2014/toegevoegde-waarde/item7107
11 https://www.onderwijsraad.nl/upload/documents/publicaties/volledig/Werkprogramma-2018.pdf

In deze studie is ervoor gekozen de waardering voor het examenstelsel te bespreken in twee focusgroepen, één met schoolleiders en één met toets- en examenexperts (zie ook annex 4). Er is nog niet een breed opgezet kwantitatief opinie-onderzoek gehouden onder vertegenwoordigers van scholen voor voortgezet onderwijs, vervolgonderwijs, leerlingen en ouders.
De belangrijkste reden daarvoor is de vraagstelling voor een kwantitatief opinie-onderzoek nauw luistert. Met een vraagstelling die onvoldoende aansluit bij de onderwijspraktijk bestaat het risico dat de resultaten niet valide zijn. Deskresearch leverde onvoldoende aanknopingspunten op voor de constructie van vragenlijsten. De vragenlijsten van het Cito voor het meten van de waardering van examens achtten we voor dit onderzoek niet goed bruikbaar.
Het leek ons daarom wenselijk om in gesprekken met schoolleiders en toets- en examenexperts het spectrum van mogelijke opvattingen te verkennen, zodat deze input in vervolgonderzoek benut kan worden bij het formuleren van vragen en stellingen voor een grootschaliger opinie-onderzoek.
Verder wordt in de rapportage verwezen naar de bevindingen van het recente onderzoek in opdracht van de VO-raad, waarin onder meer is gevraagd naar de mening over het schoolexamen in relatie tot het centraal examen.

Internationaal vergelijkend onderzoek: internationaal panel en case study Italië
De case study is uitgevoerd door middel van enkele bezoeken aan Italië door de hoofdauteur van dit rapport. Tijdens deze bezoeken zijn interviews gehouden met vertegenwoordigers van INVALSI, de centrale instantie voor toetsontwikkeling in Italië en het Italiaanse Ministerie van OCW. Verder is gebruik gemaakt van documentatie die het Italiaanse bestel en systeem van nationale toetsen beschrijft. De case beschrijving is opgenomen als onderdeel van annex 3 bij dit rapport: ‘Examinations in an international perspective’.
Verder is een beschrijving gemaakt van Vlaanderen en Zweden, op basis van documentatie over die het toetssysteem in deze landen beschrijft. Ook deze beschrijvingen zijn opgenomen als onderdeel van annex 3 bij dit rapport: ‘Examinations in an international perspective’.

Bespreking rapport met panel van relevante respondenten
De eerder genoemde focusgroep van toets- en examenexperts is ook benut voor een bespreking van het conceptrapport. Aan de focusgroep werd deelgenomen door vertegenwoordigers van CvTE, SLO, Cito, Docentplus en de leden van het onderzoeksconsortium: Oberon, Universiteit Twente en RCEC.

Annex 2: RCEC beoordelingssysteem voor de kwaliteit van studietoetsen en examens

Piet Sanders en Arnold Brouwer

Het Research Center voor Examinering en Certificering (RCEC) heeft een systeem ontwikkeld voor de beoordeling van de kwaliteit van toetsen en examens die in het onderwijs gebruikt worden. Het beoordelingssysteem hanteert zes criteria:
1 Doel en gebruik
2 Kwaliteit van toets- en examenmateriaal
3 Representativiteit
4 Betrouwbaarheid
5 Standaardbepaling en normhandhaving
6 Afname en beveiliging

Elk criterium wordt beoordeeld door middel van vragen die als 'onvoldoende' (O, score 1), 'voldoende' (V, score 2) en 'goed' (G, score 3) beoordeeld en gescoord worden. Op basis van (de som van) de scores op de onderscheiden vragen wordt een criterium beoordeeld als 'goed', 'voldoende' of 'onvoldoende'. Hierna presenteren we de volledige informatie van het criterium 'Doel en gebruik'. De presentatie van de andere criteria beperkt zich tot een korte inleiding en de vragen die bij het betreffende criterium gesteld worden. Het beoordelingssysteem wordt steeds verder ontwikkeld; de meest recente versie is op te vragen via de website van het RCEC, zie www.rcec.nl > Beoordelingssysteem.

Criterium 1: Doel en gebruik
Bij de beoordeling van dit criterium wordt het doel en het gebruik van toetsen en examens beoordeeld. Is duidelijk 'wat' we toetsen of examineren en 'waarom' we dat doen? Het doel van toetsen en examens is te beoordelen of kandidaten over de vereiste kennis, vaardigheden of houdingen beschikken. Het gebruik van toetsen en examens betreft de beslissingen die op basis van de door de kandidaten behaalde resultaten op toetsen en examens genomen worden. Merk op dat in het RCEC-beoordelingssysteem onder examens ook praktijkexamens verstaan worden.
Basisvraag 1.1: Is aangegeven wat de doelgroep(en) van de toets of het examen is (zijn)?
Bij onvoldoende beoordeling kan men de twee andere vragen van dit criterium overslaan en doorgaan met criterium 2
Basisvraag 1.2: Is het meetdoel van de toets of het examen beschreven?
Bij onvoldoende beoordeling kan men basisvraag 3 van dit criterium overslaan en doorgaan met criterium 2
Basisvraag 1.3: Is aangegeven wat het gebruiksdoel van de toets of het examen is?
Aanwijzingen bij basisvraag 1.1: Is aangegeven wat de doelgroep(en) van de toets of het examen is (zijn)?
In ieder geval moet de opleiding benoemd zijn waarvoor de toets of het examen wordt ingezet. Maar ook het aangeven van de leeftijd, het beroep, het opleidingsniveau of de relevante voorkennis van kandidaten zijn mogelijkheden om de doelgroep te definiëren. Deze informatie kan onder andere belangrijk zijn bij het beoordelen van de inhoud van de toets of het examen zoals het taalgebruik en de gehanteerde normen of cesuren.

Aanwijzingen bij basisvraag 1.2: Is aangegeven wat het meetdoel van de toets of het examen is?
Een toets of examen moet vaststellen wat kandidaten na afloop van een onderwijstraject wel en niet beheersen. Wat de kandidaten geacht worden te beheersen, kan onder andere aangegeven worden als:
• de beheersing van een bepaald construct (bijvoorbeeld 'leesvaardigheid');
• de beheersing van een exameneenheid van een examenprogramma (bijvoorbeeld de exameneenheid 'havo-examen wiskunde');
• de beheersing van een kerntaak, beroepstaak of werkproces (bijvoorbeeld uit een mbo kwalificatiedossier);
• de beheersing van een competentie (bijvoorbeeld 'analyseren' van een assistent-drogist).

Een toets of examen die een construct of een competentie meet, zal een gedetailleerde beschrijving met voorbeelden moeten geven van de theorie waarop het bedoelde construct of de competentie gebaseerd is. Dit impliceert dat niet volstaan kan worden met 'deze toets meet het construct leesvaardigheid' of 'dit examen meet de competentie analyseren'. Indien de toets of het examen niet expliciet naar een construct of competentie verwijst, zullen de onderwerpen, kerntaken of exameneenheden die in de toets of het examen aan de orde komen, beschreven moeten worden. In voorkomende gevallen kan volstaan worden met een verwijzing naar relevante brondocumenten. Bij deze vraag is het van belang dat de relevantie van de inhoud van de toets of het examen voor het beoogde doel aannemelijk gemaakt wordt. Daarbij kan bijvoorbeeld een toetsmatrijs van de toets of het examen goede diensten bewijzen.
Aanwijzingen bij basisvraag 1.3: Is aangegeven wat het gebruiksdoel van de toets of het examen is? Een toets of examen kan gebruikt worden voor:
• Selectie.
Afhankelijk van het toets- of examenresultaat wordt een leerling wel of niet toegelaten tot een opleiding (bijvoorbeeld toelating tot een numerus fixus opleiding).
• Classificatie
Afhankelijk van het toets- of examenresultaat volgen leerlingen verschillende onderwijsprogramma's die tot verschillende diploma's of certificaten leiden (bijvoorbeeld de eindtoets(en) in het basisonderwijs).
• Plaatsing
Afhankelijk van het toets- of examenresultaat volgen leerlingen verschillende onderwijsprogramma's die tot hetzelfde certificaat of diploma leiden (bijvoorbeeld een zelfbeoordeling ten behoeve van de BOL- of BBL-leerweg in het mbo).
• Certificering of diplomering
Afhankelijk van het toets- of examenresultaat wordt wel of niet een diploma of certificaat verstrekt (bijvoorbeeld de examens in het voortgezet onderwijs).
• Monitoring
Afhankelijk van het toetsresultaat wordt vastgesteld of de leerling al of niet voortgang heeft geboekt (bijvoorbeeld het gebruik van leerlingvolgsysteemtoetsen in het basisonderwijs).
Een ander gebruiksdoel van toetsen die in het onderwijs opgang doet, betreft de drie benaderingen van formatief assessment:
• data-based decision making (DBDM), in de Nederlandse literatuur opbrengstgericht werken (OGW) genoemd;
• assessment for learning (AfL), in de Nederlandse literatuur ook wel toetsen of evaluatie van het leren genoemd;
• diagnostische toetsen (DT).

Voor meer informatie over formatief toetsen verwijzen we naar het proefschrift van Van der Kleij (2013), te downloaden op de website van het RCEC, www.rcec.nl, of naar de het marktonderzoek formatief evalueren van Oomens e.a. (2017).

Criterium 2: Toets- of examenmateriaal

Bij de beoordeling van dit criterium wordt een onderscheid gemaakt tussen toetsen en examens die met behulp van de computer worden afgenomen, en toetsen en examens die schriftelijk worden afgenomen en uit gesloten en/of open vragen bestaan, en (praktijk)examens die uit één of meer praktijkopdrachten bestaan. Ook kunnen toetsen en examens uit combinaties van gesloten vragen, open vragen en praktijkopdrachten bestaan. Bij gesloten of (meervoudige) meerkeuzevragen moet de kandidaat het goede antwoord (of goede antwoorden) selecteren, bij open vragen het goede antwoord (of goede antwoorden) formuleren en bij een (praktijk) examen de praktijkopdrachten uitvoeren.
Er worden bij dit criterium twee basisvragen gesteld. In het algemeen geldt dat om de score(s) op een toets of examen zinvol te kunnen interpreteren, de toets of het examen zodanig dient te zijn afgenomen en gescoord dat andere, niet-beoogde factoren geen invloed kunnen uitoefenen op de totstandkoming van de score(s). Zo dient bijvoorbeeld de afname en instructie dusdanig gestandaardiseerd te zijn dat de invloed van variatie in instructie of verschil in de toets- of examenleider of van de afnamesituatie op de score is geëlimineerd of in ieder geval binnen de grenzen van het mogelijke is beperkt. Ook moet de scoring zo objectief mogelijk zijn. Voor de beoordeling van een toets of examen die schriftelijk afgenomen wordt, dient men te beginnen met vraag 2.1, voor een computertoets of -examen met vraag 2.8. Indien er van een toets of examen zowel een schriftelijke versie als een computerversie bestaat, dient de kwaliteit van het toets- en examenmateriaal van beide versies te worden beoordeeld.
Basisvraag 2.1: Zijn de vragen of opdrachten gestandaardiseerd?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.
Basisvraag 2.2.a: Is er sprake van een geautomatiseerd of objectief scoringssysteem? en/of
2.2.b: Als de scoring door beoordelaars gebeurt, is dan het beoordelingsvoorschrift volledig en duidelijk?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.
2.3: Zijn de vragen of praktijkopdrachten, het toets- of examenboekje, de antwoordschalen en/of het antwoordformulier zodanig ontworpen dat fouten bij de invulling voorkomen worden?
2.4: Is het scoringssysteem zodanig ontworpen en beschreven dat fouten bij de scoring voorkomen worden?
2.5: Is de instructie voor de kandidaat volledig en duidelijk?
2.6: Zijn de vragen of opdrachten correct geformuleerd?
2.7: Hoe is de kwaliteit van het toets- of examenmateriaal?

Voor computertoetsen gelden de volgende vragen:
Basisvraag 2.8 Zijn de vragen gestandaardiseerd?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.
Basisvraag 2.9: Is er sprake van een geautomatiseerd of objectief scoringssysteem?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.
2.10: Is de software zodanig ontworpen dat fouten door onjuist gebruik voorkomen worden?
2.11: Is de instructie voor de kandidaat volledig en duidelijk?

2.12: Zijn de vragen correct geformuleerd?
2.13: Hoe is de kwaliteit van de vormgeving van de gebruikersinterface?
Op basis van deze vragen wordt vastgesteld of het toets- of examenmateriaal van onvoldoende, voldoende of goede kwaliteit is. Hieraan zou nog een vraag toegevoegd kunnen worden over hoe er omgegaan wordt met fouten en onduidelijkheden in het examen. Is hiervoor een procedure afgesproken en wordt die ook toegepast?

Criterium 3: Representativiteit
Bij dit criterium wordt de representativiteit van de toets of het examen beoordeeld. Representativiteit heeft betrekking op zowel de inhoud als moeilijkheidsgraad van de toets of het examen. De inhoud van toetsen en examens is gebaseerd op wat een kandidaat wordt onderwezen. Deze leerdoelen worden, afhankelijk van de onderwijssector, geformuleerd als kerndoelen, eindtermen, basiskwalificaties, kerntaken of competenties. Omdat deze doelen nog te algemeen zijn om er toetsen of examens op te kunnen baseren, dienen ze uitgewerkt te worden tot toetsbare leerdoelen. Indelingsschema's of taxonomieën voor menselijk presteren, meestal aangeduid met toetsmatrijzen, vormen hierbij een nuttig hulpmiddel om de beoogde leerdoelen uit te werken tot toetsbare leerdoelen. Omdat de meeste studietoetsen en examens eerder directe metingen van menselijk gedrag zijn (en veelal ook door gebruikers als zodanig opgevat worden) dan metingen van constructen of competenties, wordt bij dit criterium prioriteit gegeven aan de inhoud van de toets of het examen. Mochten bij sommige toetsen of examens wel constructen of competenties worden gemeten, en als daar in de verantwoording van de toets of het examen bewijzen voor worden aangevoerd, dan zal de beoordelaar van de toets of examen de beoordeling op die bewijzen baseren. Zie hierover ook hoofdstuk 4 van Toetsen op School (Sanders, 2013), te downloaden op www.toetsenopschool.nl. Behalve dat de inhoud van de toets of het examen de leerdoelen dient te representeren, dient de moeilijkheidsgraad van de vragen of praktijkopdrachten, en dus de toets of het examen, ook afgestemd te zijn op de beoogde doelgroep. Dit betekent in de praktijk dat het merendeel van de vragen of praktijkopdrachten niet te moeilijk of te makkelijk mag zijn voor de kandidaten.
Behalve dat de inhoud van de toets of het examen de leerdoelen dient te representeren, dient de moeilijkheidsgraad van de vragen of praktijkopdrachten, en dus de toets of het examen, ook afgestemd te zijn op de beoogde doelgroep. Dit betekent in de praktijk dat het merendeel van de vragen of praktijkopdrachten niet te moeilijk of te makkelijk mag zijn voor de kandidaten.
Basisvraag 3.1: Is de toetsmatrijs, het examenprogramma, examenplan, competentieprofiel of de operationalisatie van het construct een adequate representatie van het meetdoel?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 4
Basisvraag 3.2 Is de moeilijkheidsgraad van de vragen of de praktijkopdrachten afgestemd op de beoogde doelgroep?
In het RCEC beoordelingssysteem kan representativiteit opgevat worden als de alignment tussen de toets of het examen en het curriculum/de leerdoelen.
Om bedoelde alignment te onderzoeken is door Sanders, Brouwer en Veldkamp (2017) een onderzoekskader ontwikkeld om de representativiteit, ook wel inhoudsvaliditeit genoemd, van examens te beoordelen. Het onderzoekskader is gebaseerd op publicaties van Sireci (1998a, 1998b) en Sireci en Faulkner-Bond (2014). Het onderzoekskader voor het beoordelen van de representativiteit of inhoudsvaliditeit van een toets of examen onderscheidt de volgende vier aspecten:
1. domein definitie,
2. domein representatie

3. domein relevantie
4. adequate toetsconstructieprocedure

Domein definitie verwijst naar hoe het 'construct', bijvoorbeeld de vaardigheid vwo wiskunde A, operationeel gedefinieerd wordt. Wat betreft de examens uit het voortgezet onderwijs komt die operationele definitie neer op het geven van (a) gedetailleerde beschrijvingen van de inhoud van het domein en de cognitieve vaardigheden waar een beroep wordt gedaan, en (b) toetspecificaties die zowel de specifieke inhoudscategorieën of inhoudsgebieden als de cognitieve niveaus benoemen. Evaluatie van de domein definitie houdt in dat de overeenstemming onderzocht wordt tussen de operationele definitie van het examen en de heersende opvattingen over het domein van vakdeskundigen uit het werkveld. Voor dit onderzoek worden veelal onafhankelijke vakdeskundigen ingezet. De mate waarin belangrijke aspecten van het construct of het curriculum niet voorkomen in de toetsspecificaties is een belangrijk criterium voor de beoordeling van de domein definitie.
Domein representatie verwijst naar de mate waarin een toets het domein zoals dat gedefinieerd is door de toetsspecificaties adequaat representeert en meet. Evaluatie van de domein representatie vereist de inzet van vakdeskundigen die de items/vragen/opdrachten bekijken en beoordelen. Het is de taak van de vakdeskundigen om te bepalen of de items in voldoende mate het beoogde domein representeren.
Domein relevantie betreft de mate waarin elk item van een examen relevant is voor het beoogde domein. Items die belangrijke aspecten van het domein meten, zouden hoge beoordelingen voor domein representatie moeten krijgen en items die minder belangrijke aspecten meten lage beoordelingen. Ook hier zou men aan vakdeskundigen moeten vragen de relevantie van items voor bepaalde toetsspecificaties te beoordelen en deze beoordelingen zou men dan binnen elke inhoudscategorie kunnen aggregeren om de domein representatie te bepalen.
Adequate toetsconstructieprocedures zullen de inhoudsvaliditeit van examens bevorderen door bij alle onderscheiden stappen van het toetsconstructieproces kwaliteitscontrole te doen plaatsvinden. In de loop der jaren zijn er vele toetsconstructieprocedures voorgesteld die veelal grote overeenkomsten vertonen. Een van de bekendste toetsconstructieprocedures is het twaalf stappenplan dat beschreven staat in het eerste hoofdstuk van Downing en Haladyna (2006). In navolging hierop introduceerde Veldkamp (2016) een tien stappenplan ten behoeve van de (centrale) examens zoals wij die in Nederland kennen.
Het onderzoekskader is tot nu toegepast bij twee centrale examens (Sanders, Brouwer, & Veldkamp, 2017) en lijkt geschikt voor het beoordelen van de representativiteit of inhoudsvaliditeit van centrale examens.
Een andere aanpak om de representativiteit of inhoudsvaliditeit te beoordelen, wordt gebruikt bij de beoordeling van het beoordelingsaspect 'Dekking' van de mbo examens, www.valideringexamens.nl. Aan 'Dekking' worden de volgende producteisen gesteld:

- Er is aangetoond welk onderdeel(en) van de kwalificatie het exameninstrument dekt. Hiertoe is een relatie gelegd tussen de kwalificatie-eisen en het exameninstrument.
- Voor de kwalificatie is er een herleidbaar examenresultaat per kerntaak:
    - Indien een exameninstrument een deel van een kerntaak dekt, dient de verhouding tot de andere exameninstrumenten, die samen de gehele kerntaak dekken, inzichtelijk te zijn.
    - Indien een exameninstrument kerntaakoverstijgend is, dient inzichtelijk te zijn hoe tot de examenresultaten per kerntaak wordt gekomen.
- Het exameninstrument examineert niet meer dan in de kwalificatie beschreven staat.
- De examenvorm sluit aan bij de te beoordelen inhoud.

- De context en de inhoud van het exameninstrument doen recht aan complexiteit, verantwoordelijkheid en zelfstandigheid op het niveau van een beginnend beroepsbeoefenaar.
- De inhoud (opdrachten, vragen) van het exameninstrument is voor één uitleg vatbaar, is passend bij de kwalificatie-eisen en ligt op het niveau van de beginnend beroepsbeoefenaar. Het taalgebruik is helder en past bij de doelgroep.
- Wanneer het exameninstrument uit meerdere examenonderdelen bestaat, past de omvang van het onderdeel (omvang in tijd en energie die in het betreffende onderdeel gestoken moet worden) bij het belang van de kwalificatie-eisen.
- Het exameninstrument dekt de eisen van de kwalificatie. In de verantwoording bij het exameninstrument is aangegeven of, en waarom, dat middels het examineren van alle kwalificatie-eisen of middels representatieve examinering van de kwalificatie-eisen wordt gerealiseerd.
- Als er representatief wordt geëxamineerd dan gelden de volgende criteria:
  ○ de opvatting over het beroep / de beroepen waarvoor wordt geëxamineerd is afgestemd met het beroepenveld. De inhoud van het beroep is hierin leiden
- Als er representatief wordt geëxamineerd dan gelden de volgende criteria:
  - de opvatting over het beroep / de beroepen waarvoor wordt geëxamineerd is afgestemd met het beroepenveld. De inhoud van het beroep is hierin leidend;
  - op basis van de opvatting over het beroep / de beroepen en de afstemming met het beroepenveld worden keuzes voor de examinering gemaakt;
  - indien sprake is van een keuze, dient deze representatief te zijn. Een representatieve keuze betekent dat het examen een juiste indicatie geeft van het (toekomstige) beroep wat betreft niveau, inhoud en complexiteit. Gemotiveerd wordt:
    - welke kwalificatie-eisen in ieder geval worden geëxamineerd. Examens of examenonderdelen hiervoor worden door alle deelnemers uitgevoerd;
    - welke overige keuzes zijn gemaakt voor de te examineren kwalificatie-eisen.
  - in sommige gevallen kunnen kwalificatie-eisen middels een steekproef worden geëxamineerd. Voor de steekproef gelden de volgende voorwaarden:
    - de kwalificatie-eisen waaruit een steekproef wordt getrokken zijn gelijkwaardig;
    - de steekproef kan niet uit kerntaken getrokken worden. De kwalificatie-eisen, waaruit een steekproef getrokken kan worden, zijn: werkprocessen, kennisaspecten, vaardigheden, gedragingen, situaties, handelingen of beroepscontexten;
    - de steekproef is onvoorspelbaar voor de deelnemer;
    - de omvang van de steekproef voor deze kwalificatie-eisen geeft de juiste indicatie van het (toekomstige) beroep wat betreft niveau, inhoud en complexiteit. Dit bepaalt het aantal examenonderdelen. Alleen in het geval de selectie van het aantal onderdelen voor de steekproef volkomen arbitrair is, kan minimaal ⅓ als richtlijn worden aangehouden.

Deze twee sets met vragen kunnen worden gebruikt om de representativiteit of inhoudsvaliditeit van examens te beoordelen.

Criterium 4: Betrouwbaarheid

Bij dit criterium wordt de betrouwbaarheid van (de scores van) een toets of examen beoordeeld. Bij betrouwbaarheid gaat het om de vraag of we vertrouwen kunnen hebben in de scores die kandidaten op een toets of examen behalen. De betrouwbaarheid is te kwantificeren met een betrouwbaarheidscoëfficiënt, het percentage misclassificaties en de standaardmeetfout. De

betrouwbaarheidscoëfficiënt heeft een ondergrens van 0,0 en een bovengrens van 1,0. Een hoge betrouwbaarheidscoëfficiënt geeft aan dat we vertrouwen hebben in de betrouwbaarheid van het examen. We mogen dan verwachten dat indien de kandidaten twee keer hetzelfde examen zouden maken, zij een vergelijkbare score behalen. Onder misclassificaties verstaan we het aantal kandidaten dat als gevolg van de onbetrouwbaarheid van de toets of het examen ten onrechte gezakt en ten onrechte geslaagd is. Bij een hoge betrouwbaarheid is het aantal misclassificaties gering. Met de standaardmeetfout kunnen we de vraag naar de betrouwbaarheid van de scores van individuele kandidaten beantwoorden, oftewel welke andere score had een kandidaat ook op de toets of het examen kunnen behalen. Bij een hoge betrouwbaarheid ligt de score van het eerste examen erg dicht bij de score op het tweede examen. Als een examen met behulp van item response theorie wordt gescoord, dan kan naar de betrouwbaarheid van de geschatte latente vaardigheid gekeken worden in plaats van naar de betrouwbaarheid van de testscore.
Voor meer informatie over het berekenen en interpreteren van de betrouwbaarheid van een examen verwijzen we naar hoofdstuk 3 van Toetsen op School (Sanders, 2013) te downloaden op www.toetsenopschool.nl.
Basisvraag 4.1: Zijn of worden betrouwbaarheidsgegevens verstrekt?
Bij onvoldoende beoordeling van deze vraag kan men direct doorgaan naar criterium 5.
4.2: Zijn of worden de betrouwbaarheidsgegevens correct berekend?
4.3: Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets of het examen genomen worden?

Criterium 5: Standaardbepaling en normhandhaving
Bij dit criterium beoordelen we hoe de normen van toetsen of examens bepaald zijn en hoe de normen van vergelijkbare of parallelle toetsen of examens gehandhaafd zijn. Normeren kunnen we onderscheiden in relatief en absoluut normeren. Het doel van relatief normeren is het opstellen van relatieve normen. Kenmerkend voor relatieve normen is dat de score van een kandidaat vergeleken wordt met de scores die andere kandidaten van dezelfde doelgroep behaald hebben. Relatieve normen worden vooral gebruikt bij psychologische tests en studietoetsen die door testuitgevers op de markt worden gebracht. Het doel van absoluut normeren is het opstellen van absolute normen. Kenmerkend voor absolute normen is dat de score van een kandidaat vergeleken wordt met een score die als standaard, norm of cesuur aangeduid wordt.
De bespreking hier beperken we tot absoluut normeren. Het doel van absoluut normeren is het bepalen van absolute normen/standaarden/cesuren. De term standaard wordt gebruikt als afkorting voor prestatiestandaard of beheersingsstandaard. De meest bekende standaard of cesuur van een toets of examen is de score die de grens vormt tussen een voldoende of slagen en een onvoldoende of zakken. Dat betekent dat kandidaten met een score gelijk aan of boven de cesuur een voldoende behalen en dus geslaagd zijn voor de toets of het examen en dat kandidaten met een score onder de cesuur een onvoldoende behalen en dus gezakt zijn. Het is echter ook mogelijk om voor een toets of examen meer dan één standaard of cesuur te bepalen. In dat geval worden de kandidaten niet in twee categorieën (gezakt of geslaagd) maar in meerdere categorieën, bijvoorbeeld de cijfers 1,0 tot en met 10,0 onderverdeeld. Methoden voor standaardbepaling kunnen onderscheiden worden in methoden die gebaseerd zijn op de beoordeling van de vragen/opgaven/opdrachten van een toets of examen, bijvoorbeeld de methode van Angoff, en methoden die gebaseerd zijn op de beoordeling van de kandidaten die een examen maken, bijvoorbeeld de methode van de contrasterende groepen. Voor meer informatie zie hoofdstuk 9 van Toetsen op School (Sanders, 2013) te downloaden op www.toetsenopschool.nl. Het tweede doel van normeren bij toetsen of examens is het handhaven van

eenmaal bepaalde standaarden. De methoden om standaarden (normen) te handhaven worden meestal met normhandhaving aangeduid. Zie voor meer informatie op www.toetswijzer.nl > ToetsSpecials de ToetsSpecial over normering bij de centrale examens in het voortgezet onderwijs.

Voor toets/examen met absolute normen
5.1: Worden standaarden/normen/cesuren verstrekt?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 6.
5.2: Is de standaardbepaling correct bepaald?
5.2.a: Is de standaardbepalingsmethode op de juiste wijze uitgevoerd?
5.2.b: Zijn de beoordelaars/vakdeskundigen naar behoren geselecteerd en getraind?
5.2.c: Is er voldoende overeenstemming tussen de beoordelaars?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 6.

Voor toets/examen met relatieve normen
5.3 Wat is de kwaliteit van de verstrekte normen?
5.3.a: Zijn de normgroepen groot genoeg?
5.3.b: Zijn de normgroepen representatief?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 6.
5.4 Worden de betekenis en de beperkingen van de normschaal duidelijk gemaakt voor de gebruiker en is het type normschaal in overeenstemming met het doel van de toets of het examen?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 6.
5.5 Worden er gegevens verstrekt over:
5.5.a: Gemiddelden, standaardafwijkingen van de scoreverdeling?
5.5.b: Nauwkeurigheid van de meting en daarbij behorende intervallen (standaardmeetfout, standaardschattingsfout, testinformatie/ standaardfout)?

Voor toets/examen met absolute en relatieve normen
5.6: Worden normen/standaarden/cesuren gehandhaafd?
Bij onvoldoende beoordeling (1) van deze vraag kan men doorgaan met criterium 6
5.6.a: Is de methode op grond waarvan de norm(en) is (zijn) gehandhaafd correct?

Criterium 6: Afname en beveiliging
Om de afname van een toets of examen goed te doen verlopen, moet er informatie beschikbaar zijn die met name voor de surveillant bedoeld is (zie ook functieprofiel surveillant op www.nvexamens.nl). Deze informatie dient in overzichtelijke vorm, hetzij op papier hetzij digitaal, beschikbaar te zijn. Voor toetsen of examens die via de computer worden afgenomen, geldt dat tevens specifieke aanwijzingen moeten worden gegeven met betrekking tot de installatie en/of het opstarten en het gebruik van de toets of het examen. Soms is deze informatie gebundeld in een aparte installatiehandleiding. Een toets of examen dient ook ‘goed’ beveiligd te zijn, dat wil zeggen dat al het mogelijke gedaan moet worden om de toegang tot de toets of het examen, het toets- of examenmateriaal en de toets- of examenresultaten te beveiligen. Bij de beveiliging van toetsen of examens dienen we een onderscheid te maken tussen enerzijds computertoetsen of -examens en anderzijds schriftelijke toetsen en examens.

Basisvraag 6.1: Is er voor de surveillant voldoende informatie over de afname van de toets of het examen beschikbaar?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan.
6.1.a: Is de informatie (voor de surveillant) volledig en duidelijk?
6.1.b: Wordt de mate van deskundigheid die vereist is om de toets of het examen af te nemen vermeld?
Basisvraag 6.2: Is de toets of het examen voldoende beveiligd?
Bij onvoldoende beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan.
Extra vragen voor afname via computer
6.3: Wordt er informatie gegeven over de installatie van de computersoftware?
6.4: Wordt er informatie gegeven over de bediening en mogelijkheden van de software?
6.5: Zijn er voldoende mogelijkheden voor technische ondersteuning?

.

Annex 3: Examinations in an international perspective

Jaap Scheerens

Aims and scope of this paper

The purpose of this chapter is to provide an international context to the report on examinations and testing in the Netherlands. The chapter consists of a case study on Italy, and briefer sketches on education in Sweden and the Flemish part of Belgium. These countries are chosen, because they provide interesting patterns of similarity and differences, as compared to the Dutch educational system. The presentation describes three major system characteristics: examinations, educational testing and curriculum standardization. The choice of the first two characteristics follows the content of the report on the Netherlands. Curriculum standardization is chosen as a third descriptive characteristic, because it provides a most relevant context for the position of examinations and testing. More specifically it provides a context for the alignment of examinations and assessment in the broader educational context of the countries concerned.

A CASE STUDY OF EXAMINATIONS, TESTING AND CURRICULUM STANDARDIZATION IN ITALY
Overall description of elementary and lower secondary education

Primary education is compulsory, has an overall length of 5 years and is attended by pupils aged 6 to 11. Although they are two different levels of education, each with its own specificities, primary education and lower secondary education make up the first cycle of education which lasts a total of eight years. Comprehensive institutes group primary schools, lower secondary schools and pre-primary schools managed by a single school manager. The purpose of comprehensive institutes is to assure didactic continuity within the same cycle of education.
The aim of primary education is to provide pupils with basic learning and the basic tools of active citizenship. It helps pupils to understand the meaning of their own experiences.
Primary education is divided, for teaching purposes only, into a first year linked to pre-primary education, followed by a further two periods of two years each.

Secondary education is organized into the following levels:

- Lower level, called 'first-level secondary school' (*scuola secondaria di primo grado*)
- Upper level, called 'second cycle of education' (*secondo ciclo di istruzione*), which is made up of:
  - State-run general and vocational upper secondary school (*scuola secondaria di secondo grado*)
  - vocational education and training (*Istruzione e formazione professionale* - IFP*)* which are run at regional level

Lower secondary school is compulsory, it lasts for 3 years and is attended by pupils aged 11 to 14. Lower secondary school and compulsory primary school make up the first cycle of education which lasts eight years altogether. However, each portion of the first cycle has its own specificities.
Lower secondary school aims at fostering the ability to study autonomously and at strengthening the pupils' attitudes towards social interaction, at organizing and increasing knowledge and skills and at providing students with adequate instruments to continue their education and training activities

The first two years of the second cycle of secondary education and training are compulsory. Together with the eight compulsory years of the first cycle of education, they make up the 10 years of compulsory education (from 6 to 16 years of age) and can be undertaken at any of the State or regional second-cycle institutions.

Examinations

In Italy, at the level of primary and secondary schooling, there used to be three main examinations, one at primary school level (ISCED 1), one at lower secondary level (ISCED 2, middle school) and one at upper secondary level (ISCED 3). Apart from these examinations there are specific examinations for private schools, for art colleges and for students who have followed alternative learning routes (e.g in other countries).
The three main examinations, as described in the European Glossary of Education (Eurydice, 2004) were:
Esame di licenza elementare
Grammatical variants: Esami di licenza elementare
Level: ISCED 1
Explanatory note: Examination held by the school at the end of primary education *(scuola primaria)*. It includes two written tests and one oral test. It is taken before an examining board consisting of the class teachers and two others appointed by the teachers' council. It leads to the *Diploma di licenza elementare*.
It should be noted, however, that "final assessment" at the end of the primary school has no longer the status of an examination, the situation was changed in 2004 by decree *n. 59/2004.*The current situation is based on assessments by the school, at the end of the primary school period. This is described in a more recent document by EURYDICE: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-primary-education-23_en

The relevant passage is cited below:
"At the end of primary school, pupils receive a certification of the competences achieved. The certification of competences refers to the 'Student's profile' included in the National Guidelines for the curriculum that describes the subject and citizenship-related competences each pupil is expected to hold at the end of the first cycle of education. Moreover, the certification of competences must refer to the eight key competences for lifelong learning defined at European level (2006/962/EC) and take into account important competences developed by pupils through non-formal and informal learning. Competences are evaluated through a four-level scale, each level described with explanatory indicators. Schools (teachers) are in charge of drawing up the certificate. The Ministry has provided schools with the model valid nationwide for the certification of competences (annex A to the Ministerial Decree no. 742/2017)".

Esame di licenza media
Grammatical variants: Esami di licenza media
Level: ISCED 2
Explanatory note: Examination held by the school at the end of lower secondary education *(scuola media)* before an examining board. It includes three written tests in Italian, mathematics and one foreign language, as well as a test in all subjects. This examination leads to the *Diploma di licenza media*.

EURYDICE, 2018 https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-lower-secondary-education-18_ provides further description about the examination at the end of lower secondary education. This is cited below:
"The first-cycle State leaving examination takes place at the end of the third year of lower secondary school. It aims at verifying knowledge, skills and competences acquired by pupils at the end of the first cycle of education, according to the objectives established by the National guidelines for the curriculum.
Admission of candidates
Pupils access the final exam if they:

- attended at least 75% of the annual teaching time;
- did not incurred in disciplinary measures due to particularly serious events implying non admission to the next grade or to the final exam;
- participated in the external national standardized testing in Italian, Mathematics and English.

Pupils access the final examination even if they have not totally or partially met the expected learning attainment targets in one or more subjects (i.e. have marks lower than 6/10). However, the Class council can decide to refuse admission to the final examination to pupils with marks lower than 6/10 in one or more subjects. In this case, decisions must be taken by majority vote and be adequately motivated.
In case of admission to the final exam, the Class council assigns pupils an admission mark in tenths, taking into account pupils' overall learning process for the three-year period of lower secondary education. The admission mark can even correspond to a non-sufficient assessment, i.e. be less than 6/10, and counts in the average of the final examination mark. Before being admitted to the examinations the students have to do the INVALSI tests at this level (language, mathematics and English).
The examination board and the organization of the exam
An examination board is set up in every school. All teachers of all classes make up the board and the school manager chairs it.
The examination board is divided into sub-commissions, each corresponding to a class, which administer the examinations and carry out the assessment.
The examination consists of three written tests and an interview.
The exam takes place between the end of the school year and the 30th of June. Written tests are held in three, not necessarily consecutive, different days according to the calendar set by individual schools.
The examination board draws up the three written tests, in coherence with the competence development targets established in the National guidelines for the curriculum, and establishes the duration of each test, that must not exceed four hours, as well as the assessment criteria common to all sub-commissions.
Contents of the State examination
The three written tests cover the following subjects:

- Italian or other official instruction language
- mathematics
- foreign languages

The interview usually takes place after the three written tests" (and covers all school subjects)

"Pupils who pass the first-cycle leaving State examination, receive the first-cycle of education leaving Diploma (*Diploma conclusivo del primo ciclo di istruzione*).
The Diploma shows the student's personal data and the final mark obtained at the State examination, the length of studies, as well as the foreign language/s and the musical instrument tested at the examination. Each Diploma is given a progressive number shown alongside the year of printing. It also

shows the registration number assigned in the 'Register of Diplomas', which the School manager keeps under her/his responsibility, and the date of delivery. Diplomas are printed by the State printing office (*Istituto poligrafico dello Stato*), according to the model provided every year by the Ministry of Education and are delivered to schools through the Regional school offices (USRs).
Pupils who pass the final State examination receive also a certificate attesting the competences acquired at the end of the first cycle of education.
The certification of competences refers to the 'Student's profile' included in the National Guidelines for the curriculum that describes the competences each pupil is expected to hold at the end of the first cycle of education. Moreover, it must refer to the eight key competences for lifelong learning defined at European level (2006/962/EC). Finally, the certification of competences takes into account important competences developed through non-formal and informal learning. Competences are evaluated through a four-level scale, each level described with explanatory indicators. Schools (teachers) are in charge of drawing up this section of the certificate.
A specific section, drawn up by Invalsi, integrates the certification of competences with a description of attainment targets reached by pupils in the standardized national testing in Italian, mathematics and English.
The Ministry has provided schools with the **model** valid nationwide for the certification of competences (annex B to the Ministerial Decree no. 742/2017)".

Esame di Stato conclusivo dei corsi di studio di istruzione secondaria superior
Grammatical variants: Esami di Stato conclusive dei corsi di studio di istruzione secondaria superiore
Level: ISCED 3
Explanatory note: National examination which, since 1998/99, has replaced the examination held at the end of upper secondary education (in a *liceo* or *istituto*). It is held before an examining board appointed by the ministry. Pupils are allowed to enter the examination only if they have received satisfactory marks during their final year. It includes three written tests: one in Italian, one in a subject related to the particular course of study concerned, and a cross-curricular test. There is also one oral test. This examination leads to the *Diploma di superamento dell'esame di Stato conclusivo dei corsi*

The examinations at the end of the first cycle and at the end of the second cycle (upper secondary) represent the core of the Italian examinations. The two core examinations are structured in a similar way: entrance to the exam is based on internal school assessment and the examinations consist of written assignments and an oral examination.

The examinations for alternative learning routes are described as follows:
Esame integrativo
Country: Italy
Grammatical variants: Esami integrativi
Level: ISCED 1, 2, 3 and 5
Explanatory note: Examination held by the educational institution concerned for candidates who are either studying at any educational level in Italy or abroad and wish to enroll in the equivalent year in another type of institution of the same level in Italy, or who have completed the supplementary upper secondary course of study *(corso integrativo)*lasting 1 year (after having successfully completed

upper secondary studies lasting less than 5 years). The tests are devised by the educational institution concerned. Following successful completion of this examination, candidates receive a certificate with no specific title or indication of marks, which gives them access to tertiary education.

Educational testing

Assessment context at primary school level

Pupil assessment is both formative and summative and focuses on pupils' learning processes as well as on their overall learning outcomes. It should also be consistent with the learning objectives established in the Plan for the educational offer (PTOF) of each school, with the National guidelines for the curriculum and with the personalized plans. In the PTOF, the Teachers' assembly of each school also defines the methods and criteria for assuring that pupil assessment is equal, transparent and fair.
Daily, periodic and final assessment of the pupils in a class is the responsibility of the teachers of that class. Periodic assessment takes place at the end of each term. For assessment purposes, the school year is divided, into three-month or four-month terms, as established by each school. Final assessment takes place at the end of each school year. Pupils do not take final examinations at the end of primary school.
All the teachers for a given class participate in the assessment procedure (known as *scrutinio*): subject teachers, support teachers, teachers of Catholic religion or of the alternative activities depending on each pupil's choice and all other teachers who have carried out activities in the class.
Periodic and final evaluation of pupils' learning outcomes in each subject is expressed in numerical marks out of ten (from 0 to 10), corresponding to ten learning levels. The description of each pupil's learning processes and of the overall learning levels reached, supplements the numerical assessment.
A mark equal or higher than 6/10 means a sufficient attainment of the learning targets expected for the relevant level of education. A mark lower than 6/10 means that the attainment of learning targets is partially lacking. Marks lower than 6/10 do not affect pupil's progression to the next grade. Class teachers can decide for the non-admission to the following grade only in exceptional cases and by unanimous vote. Decision must be motivated.
Class teachers evaluate each pupil's behaviour through a synthetic report assessment. Pupils' behavior assessment refers to the development of citizenship competences.
Teachers of Catholic religion and teachers of the activities alternative to Catholic religion evaluate pupils through separate synthetic reports describing the interest shown in the subject and the results achieved.
The National Institute for the Evaluation of the Education System (*Istituto nazionale per la valutazione del sistema di istruzione e formazione* - INVALSI) carries out the external assessment of pupils. National standardized testing of pupils takes place during the second and fifth grades of primary school, within the month of May. National testing in the second grade aims at verifying pupils' learning attainment levels in Italian and Mathematics, while in the fifth grade covers also English. Tests in English must be consistent with the Common European framework of reference for languages.
Special dispositions apply for the assessment of pupils with special educational needs and of hospitalized pupils.
Pupils are admitted to the next grade of primary school and to the first year of lower secondary school even if they have not achieved the expected attainment targets, i.e. with marks lower than 6/10 in one or more subjects. At periodic or final assessment, the school warns in good times families of pupils with low marks and autonomously organizes specific measures and actions to help pupils improve their learning results.

Non-admission of a pupil to the next grade is only exceptional and decided by all class teachers upon unanimous agreement and specific motivations.
Pupils do not take exams at the end of primary school. In fact, primary school, together with lower secondary school, is part of one only school cycle called “first cycle of education” and the Italian Constitution establishes that final exams are held only at the end of each cycle of education. Source: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-primary-education-23_en
The INVALSI tests
Since 2008 all pupils in Italy sit for tests in reading, mathematics/arithmetic at the following levels. At grades 2 and 5 of primary school; at grade 8 and 10 of the middle school, and (from 2019 onwards) also at grade 13 of upper secondary education. EURYDICE: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-primary-education-23_en . Since 2017 tests in English are taken at grades 5, 8, and 10, and in 2019 also at the level of grade 13 (Aljello, 2017). Since the school year 2017-2018 the tests in grade 8, taken in a period before the Final State Examination (test in April) is compulsory for admission to the examination. The tests are based on the goals of the National Guidelines, (national curriculum) which, are prescriptive, and are anchored to them, i.e. each item is linked to a specific goal (ibid).
In the brochure titled “The INVALSI Tests according to INVALSI (2018) https://invalsi-areaprove.cineca.it/docs/2018/INVALSI_tests_according_to_INVALSI.pdf the purposes of testing are described. The need for testing is motivated on the basis of persisting problems of inequity in Italian education, and to facilitate innovation. “Using the same tests for everyone helps us to identify areas that need improvement” (ibid). The main rationale for testing is to discover weaknesses in learning, which can then be met by improvement oriented measures. Test results present one type of indicator that is part of the framework for school self-evaluation (RAV). Among possible aims of educational testing, school self- evaluation stands out in the Italian context. In the brochure a lot of care is taken to explain what the tests are **not** for: judging schools, judging teachers and judging students. “Obviously, the tests do not measure everything. Therefore, they are not used to assess either the pupil or the teacher, and they are only one of many components of a school’s self-assessment. But they often make it possible to see what is difficult to see by oneself, avoiding the risk of being self-referential” (ibid). The context of freedom of education is emphasized to point at the personal professional role that teachers have in optimizing instruction. Since the late nineteen nineties the autonomy of schools and teachers, as far as shaping and implementing the curriculum has increased. “The test results are an indication of the skill level reached, but they cannot explain the underlying reasons. Every situation, whether positive or negative, is determined by factors that only the teachers involved can identify”. “Therefore, the tests cannot tell us how to teach, which is a prerogative and a duty of teachers. However, they can indicate where the efforts should be concentrated. Developments in regulations, which over time have widened the possibility of choosing times, methods and places, has led to a gradual recognition and expansion of teaching freedom “(ibid).
The brochure also mentions that the tests are tapping more than just rote learning and memorization: ”In a changing world, schools must inevitably change as well. Whereas, in the past, people were mainly expected to carry out tasks designed by others, the people of today must be able to think and decide for themselves. Pupils must therefore study what they have always studied, but they must also get used to using the knowledge, linking it to other knowledge, and applying it to solve new problems. This is why the tests are not simply a memory exercise, but one of reasoning” (ibid)
The tests are processed by INVALSI and schools receive summary reports of the results. As indicators of the socio economic status of students are included, the summary reports also contain reference points, or benchmarks, in the sense of national averages, and comparison to schools with similar student

background characteristics. A start has been made with digital application of the tests. With computerized testing, new scoring techniques and reporting frameworks have been introduced, as well as adaptive use of the tests. Test results are also used for research purposes. Among others the application of value-added analyses allow for an estimate of the "school effect". At national level the tests results are used to compare the learning results between regions. The stunning differences in achievement between north and south are a strong motive to further search for underlying explanations.

Curriculum standardization

The issue of curriculum standardization addresses the degree to which school curricula are formally specified by the central level. This may take the form of explicit performance standards, for each subject, specification of educational objectives and sub- or intermediary objectives, mandatory subject matter sequences and even prescribed textbooks. As a matter of fact these are all considered as input facets of the curriculum. The question of whether systems have standard based examinations or aligned high stakes final tests is indicated with the term "output standardization".

Internationally comparative indicators on curriculum (input) standardization are not available, to our knowledge. Sometimes approximations are used that are based on OECD's Education at a Glance, or information based on the PISA background questionnaires. Van de Werfhorst, Elfers and Karsten (2015) use "the absence of school autonomy" as an indicator of input standardization. Absence of school autonomy is based on the absence of authority of the school in the choice of textbooks, the program of course offerings and the substances of courses. They place Italy on a scale from -1 to 2, at a level of about -.04, which means that Italy is relatively low on curriculum specification or input standardization from the center. This is another way of stating that the Italian system has relatively high school autonomy. On this scale Greece is at level 2, France at level 0 and the Netherlands at -.07.

Pisa 2012 (OECD 2013, p51) has an indicator for school autonomy that is based on the discretion of schools about the curriculum and assessment. The scale runs from -1.5 to +1.5 and is based on an index titled : "Index of school responsibility for curriculum and assessment". Here Italy scores positively at .04, and the Netherlands at 1.0.

A third indicator on standardization can be obtained from the 2004 edition of Education at a Glance (OECD, 2004, chapter D). This indicator stands for "Planning and Structures" and is defined as referring to decisions about: opening and closing of schools, creation or abolition of a grade level, design of programs of study, selection of a program of study taught in a particular school, choice of range of subjects in a particular school, definition of course content, setting of qualifying examinations for a certificate or diploma, and "credentialing": (examination content, marking and administration). Clearly this is a more general composite indicator containing elements of both input and output standardization. The indicator measures the percentage of decisions in this domain (planning and structures) that is taken by the school. On this indicator Italy scores close to 40% and the Netherlands 100%

Official documentation about the Italian national curriculum gives more "flesh and blood" to the international indicators on "input standardization".

General curriculum guidelines

The "Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo d'istruzione" (annali della pubblica istruzione numero speciale, 2012) describes the curriculum for Kindergarten,

primary school, and “middle school” (ISCED 3, lower secondary education.
http://www.indicazioninazionali.it/wp-content/uploads/2018/08/Indicazioni_Annali_Definitivo.pdf

The curriculum document describes the general pedagogical ideals of the Italian instructional system. Although the curriculum is concretely specified and prescriptive a lot of emphasis is given to the autonomy of teachers and schools.

The idea of a “core curriculum” is mentioned, in which only the most essential elements are prescribed. “ (13) “We are not expecting an attitude of mere “application” to these curricular guidelines, because that would not be in line with the principle of autonomous responsibility, but rather an open dialogue about the way the school should act, also considering skills of didactic innovation and the most effective management of learning”(13) “With the national curriculum guidelines the general goals, the learning goals and the aims for competency development are stipulated for each discipline and field of application.” (13)

The eight European Key Competencies are mentioned as a frame of reference for the Italian curriculum. The curriculum document contains a description of general characteristics of the curriculum, the overall organization and features, which reflect current thinking about education, including responsibilities for evaluation, inclusive education and school improvement.

Learning goals
“Learning goals are based on knowledge fields, more specific knowledge and skills that are seen as absolutely necessary to reach the general goals for competency development”.
The objectives are organized in thematic cores, and defined with respect to longer periods: the period of nursery school, the primary school period and the period of the first stage of secondary education.

Evaluation
“ The teachers are responsible for student evaluation, and documentary registration of this. They are also responsible for the choice of instruments within the framework of criteria that are considered in collegial consultation. The intermediary assessment and the periodic evaluations must be coherent with the objectives and targets that are stipulated in the curriculum guidelines and must be dedicated to the curriculum”.
Schools are also responsible for school self-evaluation, which should provide reflection on all facets of the internal school organization and the school curriculum, in order to enhance effectiveness and transparency and service to external accountability requests

Educational testing and school improvement
Schools are also responsible for school self-evaluation, which should provide reflection on all facets of the internal school organization and the school curriculum, in order to enhance effectiveness and transparency and service to external accountability requests.

Inclusive education
The curriculum document mentions a series of specific official decrees and guidelines about promoting educational opportunities and integration for disadvantaged learners and students from other cultures. The policy of inclusive education encompasses students with disadvantaged social background and students with special needs, due to intellectual and physical handicaps (p. 20).

Professional learning communities
"The professionality of teachers enriches itself in collaborative work, continuous professional development, didactic reflections and exchange with the world of science and culture. This enriching collaborative work is seen as being stimulated by educational leadership of the school directorship by stimulating pedagogical coordination" (p. 21)

Pedagogical and didactic specification

Characterization of instruction in three subsequent phases (nursery, elementary and lower secondary)

The nursery school hosts children of three to six. The activities that are to take place in this period of schooling are described in the guidelines. For this period of schooling "end terms" of knowledge, skills and attitudes are specified that are expected to be attained to be prepared for the elementary school period. Examples are:
"Recognize and express emotions, and awareness of desires and fears, become conscious of their own state of minds and those of others"
" Share experiences and games, learn to use material and resources with others, gradually learns to confront conflicts and respects behavioral rules in personal and public contexts"
"Demonstrates first capacities of logic reasoning, starts to learn coordination in space and time and orientation with respect to the world of symbols, representation by various media and technology"

"Express themselves in a personal way, with creativity and participation and with sensibility with respect to culture, language and experience"

The "first cycle of education" comprises primary school and lower secondary school (31)

The mission"(il senso") of the primary school is described as follows
" The school offers situations and context that enable students to understand the world and themselves, become knowledgeable about their body as a "good" that is worthwhile to take care of, find stimuli to develop analytic and crucial thinking, "learn to learn", cultivate their phantasy, and original thought, share possible ideas about reality, and reflect on the sense and consequences of choices that are made. The school stimulates the development of the necessary capacities to learn to read emotions, and manage them, by setting longer term goals, and trying to reach them. Simulates attitudes of celebrating to do things well, and to take care of the environment they live in, both the natural and the social environment" (31)
The school mission is further described in terms of "basic cultural literacy", citizenship and knowledge and attitudes concerning the Italian constitution.
"It is the specific task of the first cycle to stimulate basic alphabetization in languages and codes that make up the structure of our culture, broaden the horizon to other cultures with whom we are living together and with respect to the new media".
"This is about culture and social norms which include instruments like writing, arithmetic, languages and knowledge of the various disciplines, which should always be integrated"
" In the lower secondary school access to the various disciplines is realized in terms of points of view that concern reality, modes of knowing, interpretation and representation of the world"

"The competences that are acquired in the specific disciplines also serve to stimulate broader transversal competencies, which represent essential conditions for full personal development and for active participation in social life, and oriented at the values of living together and civic communities. Citizenship competencies are stimulated continuously in all learning activities, making use of opportunities that are offered in each of the disciplines"
Teaching the Italian language is seen as the most important instrument to stimulate communication and to provide access to knowledge. Therefore teaching language has high priority and is the responsibility of all teachers, including those specialized in other school subjects.

The learning context should meet the following requirements:
- A flexible use of the classrooms and other spaces in the school;
- The importance of the school library is underlined;
- Respect and appreciation of the experience and prior knowledge of the students;
- Providing a diversified teaching repertoire that respect differences between students;
- Stimulating discovery learning;
- Encouraging cooperative learning;
- Actualizing "learning to learn";
- Structuring the didactic approach by laboratory like situations.

Didactic specification in school subjects in primary and lower secondary education

The curriculum covers the following subjects:

Italian language
English language (or a second foreign language)
History
Geography
Mathematics
Science
Music
Art (*arte e immagine*)
Physical education
Technology
Catholic religion

The curriculum guidelines provide a summary description of the subject matter area, specify sub-domains and then provide general goals (traguardi) and more specific learning objectives, which are formulated at the end of the fifth year of primary education and the third year of secondary education.

Illustration of the didactic specification for mathematics (lower secondary education)

General goals (traguardi) for the end of elementary school
The student is confident in arithmetic treating rational numbers, masters the various ways of expressing and estimating the scale of numbers, and basic operations. Recognizes and names the forms of planes and spaces, their representation and makes connections with the relations between elements. Analyzes and interprets data representations to deduct measures of variance and for making decisions. Solves problems in various contexts on the basis of information.

Orients him/herself on probability assessments in situations of uncertainty.

Has developed a positive attitude towards mathematics, on the basis of significant experiences and has understood that mathematical instrument are useful in many real life situations.

Illustrations of Learning objectives mathematics, final year of lower secondary

Numbers

- Carries out additions, multiplications, fractions, classicifations and equations between known numbers (natural numbers, whole numbers, fractions and decimal numbers), if possible as mental arithmetic or by using the usual written algorithms.

Space and forms

- *Reproduce geometrical forms and designs, in appropriate ways, with precisions, and appropriate instruments (ruler, square, compass, goniometer, and goniometric software).*
-

Relations and functions

- Interpreting, constructing and transforming formula, which contain letters to express relationships and characteristics in a general way.
- Explore and solve problems y means of first degree equations

Data and predictions

- Representing data sets, also using spreadsheets. To confront data, in specific situations, with decisional options, making use of frequency distributions and relative frequencies. Choosing and applying central tendency measures (mean, median, arithmetical mean) when appropriate given the nature and characteristics of the available data. Knowing how to evaluate the variability of a data set, computing, for example, the variance.

In the document “Indicazione nazionali e nuove scenari”, (2018) the Ministry of Education provides an update on the Guidelines, published in 2012. The update places citizenship as the central theme of the curriculum, and gives guidelines and suggestions how instruction in the school subjects can be more closely dedicated to this overall aim of citizenship. https://www.orizzontescuola.it/wp-content/uploads/2018/02/Indicazioni-nazionali-e-nuovi-scenari.pdf

Discussion

The case-study provides a descriptive picture of the Italian educational system, which is concentrated on examinations, tests and curriculum standardization. As part of a report that is dedicated to current debates about examinations and testing in the Netherlands the key issue is what can be learned from a comparison between the two countries. The conclusions are tentative because of the limits of the pilot study on which the report is based. The conclusions are therefore to be seen as impressions that would require further corroboration.

Correspondence greater than differences

The case study report does no justice to the historical developments between educational systems. Still cultural traditions and deeply rooted governance styles between Italy and the Netherlands are different, and are expected to influence the current situation in a significant way. Where the current description shows many similarities, the way they are interpreted and implemented will tend to be shaped by the different developmental histories of schooling. So what are the apparent similarities?

Autonomy
The Italian curriculum guidelines put a lot of emphasis on the autonomy of schools and teachers to implement the curriculum guidelines. Also in the area of student assessment there is an important role for teachers and schools. Curriculum guidelines are interpreted by each school, which is expressed in the POF, a document which would be called a "school working plan" in the Netherlands.

Key competences and "citizenship"

Particularly in the recent (2018) update on the curriculum guidelines, a lot of emphasis is given on "citizenship" as an overall theme that cuts across the traditional disciplines. What is also emphasized is the idea of general subject related competencies, for example mathematical literacy, and meta-cognition "learning to learn".

Elaborate systems of examinations and tests

The description shows that Italy has an elaborate system of examinations and tests. The examinations are high stakes, provide diplomas and give access to follow up education. The tests are mainly low stakes, to be used for formative purposes and school self- evaluation. Although the examination and test systems in the two countries are different, they are comparable in terms of scope and frequency, which is considered high, in comparison to other (third) countries.

Modernization of schooling and instruction

Apart from modernization of content (see above) the Italian curriculum guidelines mention a series of desirable characteristics of the education process that appears quite international, and is similar to current trends in the Netherlands; examples are: collaborative learning between students, collegial consultation and collaboration between teachers, continuous professional development, use of ICT applications in learning, testing and teaching, ideals of school improvement. Inclusive education should be mentioned in particular.

Within these overriding similarities differences to a degree are expected to exist with respect to autonomy, where the Italian curricular guidelines are more detailed and explicit than in the Netherlands, and where the impression is that curriculum guidelines will tend to be followed with more fidelity than is the case in the Netherlands. A more pragmatic attitude in Italy with respect to key competences and citizenship education, where these are clearly placed as facets of subject matter related teaching. This issue is still more open in the Netherlands (where a complete curriculum overhaul is considered to give more space to 21st century skills). Another difference is that the examination system in Italy currently seems uncontested, whereas it is very much in discussion in the Netherlands,

where the tendency among important stake holders is to de-centralize examinations further. It should be noted, however, that the Italian examinations just have an external element for the subjects language, mathematics and English as a foreign language; all subjects are covered in oral examinations that are conducted by the schools. As a matter of fact the Italian examinations at secondary level have a mix of external structuring and school level specification that is comparable to the Dutch *school examination.*
With respect to testing the Italian tests are predominately low-stakes, while the Netherlands has a high stakes test at the end of primary school.

Further study needed on standardization and alignment

More input standardization, a centrally specified curriculum, could be seen as facilitating both output standardization and alignment between inputs and outputs. The international indicators on input standardization are relatively weak. The best contribution so far is the definition by Van de Werfhorst and Bol (2016), in which input standardization is defined as the lack of school autonomy in establishing which text books are to be used, as well as the choice of course- offering and the contents of courses. On this indicator Italy is slightly more standardized -.04 than the Netherlands, -.07. On the broader indicator of central regulation of planning and structures in education, Education at a Glance 2018, places Italy as much more centrally standardized (67% of all decisions in this domain taken at the central level) than the Netherlands, which scores 0% of decisions taken at the central level. Apparently the current state of the art on these indicators leaves much room for improvement. As indicators of curriculum alignment are not as yet available from international studies, further empirical work, would be needed in this area as well. In Italy the alignment between curriculum objectives and the contents of tests in language mathematics and English is assured in the test construction process (Aljello, 2017,p. 4). On another interesting facet of the alignment issue, the fact that the contents of examinations and high stakes tests may have a "backwash effect" on what is taught, I found a reference in the online journal Scuola 7, (Monti, 2018)[12] in which such a backwash effect was described as a first positive effect of the new test in English that was designed in line with the European quality Framework (QCER), implying that, because of this design instruction would become more focused on applications and higher cognitive skills.

[12] Come si è detto, le prove propongono testi che possono spaziare come argomento dalla scienza alle news, dalle curiosità ai temi sociali, e come tipologia testuale possono includere testi espositivi, narrativi, argomentativi, regolativi, continui e non continui. Per questo motivo la prima ricaduta positiva delle prove potrebbe essere quella di riportare il focus dell'insegnamento e della valutazione sulla lingua viva e reale della comunicazione, ed esporre gli studenti a diversi tipi di linguaggi, e non solo a quelli più tradizionalmente collegati ad un particolare indirizzo di studi.

Brief comparative notes on examinations, testing and curriculum standardization in Sweden and Belgium (Flanders)

Sweden

Compulsory education is provided through a single structure that corresponds to primary and lower secondary education (ISCED levels 1 and 2). Children start Year 1 at age 7 and complete compulsory school at age 16 (Year 9). All children between age 7 and age 16 attend school, which is free of charge. There is no tracking: everyone follows the same path and the same curriculum from Year 1 to Year 9.(OECD, 2015, p. 20)

Student performance on the Program for International Student Assessment (PISA) has declined dramatically, from near the OECD average in 2000 to significantly below the average in 2012. No other country participating in PISA saw a steeper decline than Sweden over that period (OECD, 2017, p 7).

"There is a lack of capacity and clarity in roles and responsibilities at various levels of the education administration, and local autonomy is not matched with adequate public accountability. These are key challenges for improving student performance. Lack of clarity and differing views on education priorities are diluting school improvement efforts and have led to cherry-picking of priorities at the local level. Unclear education priorities and a piecemeal approach to reform hinder the alignment, coherence and potential impact of reforms and policies. In addition, assessment and evaluation arrangements remain underdeveloped, leading to a lack in coherence and unreliable data and information on student Performance" (ibid).

Examinations

Sweden has no formal central examinations in primary and secondary education. It does have school leaving certificates. Assessment is fully taken care of by the schools. The school leaving certificates are described as follows in the European Glossary of Education (Eurydice, 2004):

Slutbetyg fran grundskolan

Leaving certificate awarded at the end of compulsory education (the 9-year*grundskola*) on the basis of pupil assessment throughout the final year. The results of the national examinations in Swedish (or Swedish as a second language), English and mathematics are taken into consideration, in order to ensure that marks are comparable nationally. The certificate indicates the subjects of the courses taken and marks obtained. Pupils with the minimum pass grades in Swedish, English and mathematics are admitted to upper secondary education.

Slutbetyg fran gymnasieskolan

Leaving certificate awarded at the end of upper secondary education (the 3-year *gymnasieskola*) on the basis of the marks obtained for each course. In order to facilitate standardization of pupil assessment, the results of national examinations in some subjects may be used. The

certificate indicates the subjects of the courses taken and marks obtained. It is a basic requirement for entry to tertiary education.

Educational testing

National tests are compulsory at the end of the Years 3, 6, and 9 in Swedish, Swedish as a second language and mathematics. In 2010, these summative tests for Year 9 students were expanded to include science. National assessments for Year 3 and Year 6 in Swedish/Swedish as a second language, mathematics and English (Year 6 only) are intended for diagnostic and formative purposes.(OECD, 2015, p. 22. The Swedish approach combines national standard-setting and central test development with a high degree of trust in school professionals to carry out evaluation and assessment. While key elements of evaluation and assessment are well established at student, teacher, school and system levels, challenges remain in aligning the different elements to ensure consistency and complementarity (OECD, 2011, p.5) The OECD report makes the following recommendations:

Increase the reliability of national assessments and building teacher capacity. As national assessments play a key role in Sweden's evaluation and assessment system, it is important that the results are reliable and nationally consistent. Currently, the national tests are scored locally by students' own teachers. A re-correction of national assessments showed that teacher grading was uneven. This raises concerns about fairness in grading and also reduces the adequacy of national test results as a measure of school and system performance. High quality training and professional development for effective assessment are essential to strengthen teachers' practices. External moderation can further help increase consistency and comparability of national test results. Options for doing this include having a second grader in addition to the students' own teachers, employing professionals for systematic external grading and/or moderation, or introducing a checking procedure by a competent authority or examination board (ibid, p.5).. the system lacks a reliable measure of learning outcomes to monitor if national learning goals are being achieved… (ibid p.6)

Develop a coherent framework for evaluation and assessment. "The well-detailed elements of evaluation and assessment currently do not link into a coherent framework. The development of a strategic plan or national framework for evaluation and assessment could help optimize alignment between the different components. It could provide an overview and reference for all actors working with evaluation and assessment in education, outline evaluation and assessment requirements at different levels, clarify responsibilities related to these requirements and map the range of tools that are available to optimize practices. It should be complemented by competency descriptions for those who carry evaluation responsibilities and be followed up by specific professional development opportunities" (Ibid).

Curriculum standardization

Curriculum standardization, as defined in previous sections, is defined by proxy indicators from OECD's Education at a Glance, or information based on the PISA background questionnaires. When using "the absence of school autonomy" as an indicator of input standardization, following Bol and Van de Werfhorst, (2016) Sweden is placed at a level of -.01 (on a scale that runs from -1 to +2) This means that Sweden is close to average on input standardization. As indicated before Italy is placed at a level of about -.04, and the Netherlands at -.07; Greece is at level 2.

Pisa 2012 (OECD 2013, p51) has an indicator for school autonomy that is based on the discretion schools have concerning the curriculum and assessment. The scale runs from -1.5 to +1.5 and is based on an index titled : "Index of school responsibility for curriculum and assessment". Here Sweden scores -

.04, indicating that school discretion over curriculum and assessment is slightly below average. As indicated before Italy scores positively at .04, and the Netherlands at 1.0, which means that their schools are more autonomous.
On the indicator “planning and structures” defined in Education at a Glance 2004, Sweden scores 0, which indicates that Swedish schools would have no autonomy whatsoever over de elements of this indicator that were defined in the section on Italy. In the 2018 version of Education at a Glance Sweden scored 83% of decisions taken in the domain of planning in structures being made by the Central level, Italy scored 67% on this indicator. By contrast the Netherlands scores 0% on this indicator (OECD, 2004, p 427, OECD 2018 p.420). For the Flemish community of Belgium 33% of decisions on planning and structures was taken at the central (state) level in 2018). Comparatively speaking Sweden would therefore be considered as high on curriculum standardization as compared to Italy, Flanders and the Netherlands.
The curriculum for compulsory education, valid nationwide, went through a reform in 2011. It specifies that all schools should base their work on the same fundamental values and ensure that all students embrace these values. Local planning must seek to give practical expression to the goals and guidelines for education set out in the Education Act, the curriculum and syllabi. The overall goals are expressed as knowledge, skills and attitudes that the students are to master during compulsory school. The choice of tools and methods are not regulated. Within Sweden’s decentralized steering of the school system, they are left to individual school organizers to determine. (OECD, 2017, 22)
Since a major administrative reform in the early 1990s, Sweden has one of the most decentralized education systems in the world, with its 290 municipalities in charge of organizing and operating school services. School leaders and teachers also have wide-reaching autonomy in deciding on study options, teaching materials and methods. The role of the national Government and agencies is to set curriculum goals and monitor outcomes rather than to focus on inputs and processes. In this highly decentralized context, evaluation and assessment are crucial to ensure that professionals get the information and feedback that they need to improve the quality of their work (OECD, 2011. p/7)

Belgium (Flemish part)

There is compulsory and free education for all children from age 6 to 18 in Belgium1. Although the vast majority of children pursue this in schools, some parents may choose home education for their children. The Flemish school system is stratified and schooling is organized in four main stages, with the first streaming of children into different types of education at the end of primary school:
• Primary education: children follow six years of primary education from age 6 to 12. There is also an offer of seven years of special primary education. At the end of primary education, children who achieve the objectives of the curriculum receive a certificate (curriculum objectives are based on the Flemish attainment targets).
• Secondary education (first stage): students follow two years of education from age 12 to 14. This is either in the “A stream” in which students must achieve the Flemish attainment targets, or in the “B stream” in which students pursue the Flemish developmental objectives. All children who have received a certificate for primary education are enrolled in the “A stream” (this was 84% of all children in 2010). Those who wish to pursue a vocational education or who have not achieved the primary education certificate enroll in the “B stream”. Upon completion of their first year in “B stream”, some students may be eligible to transfer to the first year of the “A stream”, if they wish to do so.

• Secondary education (second stage): from age 14 to 16, students attend one of four distinct types of secondary education (OECD, 2011, p. 16). Upper secondary education is diversified into general, technical, arts and vocational streams.

Flemish schools enjoy a high degree of autonomy and are free to develop their own educational policies, including curriculum, assessment, certification and any self-evaluation activities. However, in order to be able to award official qualifications or to receive funding, schools must meet certain conditions set by the Flemish authorities, including: following a core curriculum (attainment targets or developmental objectives according to the stage or type of education); and allowing the Flemish authorities to assure their quality (this is done via the Inspectorate). However, all schools belong to an educational network and may choose to use a curriculum and/or tests developed by the different umbrella organizations within these networks. (OECD, 2011, 11)

The quality of teaching and learning should be at the heart of self-evaluation and inspection activities. One way to promote a common understanding of the key factors influencing teaching and learning is to ensure that the goals, indicators and criteria used in both inspection and self-evaluation are sufficiently similar. If evaluation priorities are not clear, there is a risk that schools develop self-evaluation activities only to satisfy demands for external accountability during inspection.

The OECD examiners recommended an increase in the use of information (collected at either the school level or by the Flemish authorities) for both internal and external school evaluation. "There seems an urgent need to establish a protocol whereby schools provide on an annual basis selected data on student performance from their chosen monitoring systems. Currently, the Inspectorate lacks key information on the output part of the CIPO inspection framework, as it does not have regular performance information for schools on which to base its risk assessment. Further, objective performance data play a critical role in monitoring equity within an education system and further ways could be found to collect more information to support school evaluation". (OECD, 2011, p.12)

Examinations

Flanders has no formal central examinations. At the same time there is an elaborate system of certificates, for which school internal examinations are mentioned. The certificates are described as follows in the European Glossary of Education (Eurydice, 2004):

Getuigschrift basisonderwijs
Country: Belgium (Flemish Community)
Level: ISCED 1
Explanatory note: Certificate awarded to pupils who have satisfactorily achieved curricular objectives at the end of primary education lasting 6 years. Assessment for the certificate, which is carried out by the teachers collectively, includes written examinations set by the school. The certificate may also be awarded to pupils who pass the first year of general secondary education (A-stream) or the second year of vocational secondary education (B-stream). The certificate gives access to the first year of the first stage of secondary education, and may in certain cases be obtained by pupils in special primary education.

Getuigschrift (+)
Country: Belgium (Flemish Community)
Level: ISCED 2 and 3
Explanatory note: This certificate takes several forms, depending on the type of education received:
A certifi ate testifying to completion of the first stage of secondary education: it is known as the *getuigschrift eerste graad secundair onderwijs* and is awarded at the end of the first 2 years of secondary education. It entitles pupils to enter the first year of the second stage of secondary education leading to the *getuigschrift tweede graad secundair onderwijs*.
A certificate testifying to completion of the second stage of secondary education: it is known as the *getuigschrift tweede graad secundair onderwijs* and is awarded at the end of the second two-year stage of secondary education, to pupils who have already obtained the *getuigschrift eerste graad secundair onderwijs*. The *getuigschrift tweede graad secundair onderwijs*, which may replace the *Studiegetuigschrift*, enables pupils to enter the first year of the third stage of secondary education leading to the *Diploma secundair onderwijs*.
These certificates indicate the type of education concerned, and are awarded with the so-called *orienteringsattest* (*Orienteringsattest A* and *Orienteringsattest B*) and take into account when deciding whether pupils will be admitted to the next level of education. Pupils who do not pass the examinations held on completion of a school year (secondary education) receive an *Orienteringsattest C*
The *getuigschrift*, which is awarded to pupils at the end of the fi fth year of social and vocational education). These pupils are offered social an dvocational training with a view to securingtheir integration into a normal environment and work situation, and have to acquire all skills specifi ed in a training course chosen from a setof possible options. These skills are listed on the certifi cate which gives access to the corresponding profession or occupation, and entitles its holders to qualify for block release vocational training leading to a *Getuigschrift van alternerende beroepsopleiding*. In some cases, pupils may also obtain the certificate by successfully completing the block release training in form 3 of special secondary education.
The *getuigschrift van leertijd* (apprenticeship certificate), which is awarded at the end of the second stage of vocational secondary education including a simultaneous apprenticeship.
The written, oral or practical examinations are set by the individual school and cover all subjects studied. The name of the certificate is followed by reference to the course concerned.

Educational testing

All student assessment and related certification at key stages of the schooling system is designed and conducted at the school level. Although, schools may seek support and may conduct some externally designed tests

The National Assessment Programme (this monitors only a sample of schools, but other schools can choose to administer parallel versions of the test) and some specific funding for schools include evaluation requirements.

Over the last decade, there has been increased importance given to indicators and output measures in evaluating educational quality (Ministry of Education and Training and the University of Antwerp, 2010). Among other factors, this is attributed to a general shift to focusing on output measurements in Europe

and internationally, as well as the availability of attainment targets and developmental objectives and their periodic assessment as part of the National Assessment Program (see Chapter 2). Further, it is argued that a focus on outputs is more cost-effective than the assessment of processes.

Importantly, in this context there is consensus among the Ministry of Education and Training and all stakeholders to avoid the "ranking or comparisons (of schools) based on numerical data" as it is perceived such comparisons would increase competition among schools and not improve quality (Ministry of Education and Training and the University of Antwerp, 2010). Indeed, there is no collection of comparable performance information on all schools in the Flemish Community. (ibid, 21)

The clarification that schools are responsible for systematically monitoring their quality is the logical progression in a series of initiatives by the Ministry of Education and Training to stimulate schools to conduct self-evaluations (see Chapters 2, 3 and 4). One of the major policy projects is "Strengthening internal quality care of schools". Notably, this has led to the development of parallel versions of the National Assessment Program (NAP): although the NAP is a periodical sample survey to monitor the implementation of the Flemish attainment targets at the system level, schools are now able to administer parallel versions of these tests as part of their own evaluation activities. All schools participating in the NAP – and now also schools participating in international student assessments – receive feedback reports for their own evaluation and improvement purposes. Further, a key mechanism has been to make special funding available via school participation in innovation projects or as part of the GOK policy (see above) subject to requirements for schools to evaluate in related areas.

Curriculum standardization

On Bol and Van de Werfhorst's indicators of input standardization and output standardization, Belgium (Fl) scores 0 for each one of them. This means that central control over curricular aspects in Flanders is median and that Flanders has no central examinations.
On OECD PISA's indicator on school autonomy on curriculum and assessment in the field of mathematics, Belgium has a position of -.08 on the scale that was used (from – 1.5 to + 1.5).
On the indicator "planning and structures" defined in Education at a Glance 2004, Belgium has a score of slightly over 40%, which means that in Belgium schools decide over about 40% of all decisions in the area of planning and structures. It should be noted, that this indicator was based on information about the French part of Belgium, so its validity for Flanders is doubtful. On the same indicator Flanders received a score of 67% in 2018, (OECD, 2018, p. 420)

There are two major requirements related to school evaluation: they must follow a core curriculum set by the Flemish authorities (attainment targets or developmental objectives depending on the stage or type of education); and they must allow the Flemish authorities to assure their quality (this is done via the Inspectorate) (OECD, 2011, 17)

Schools enjoy a high degree of autonomy in the Flemish Community. As part of the Belgian Constitution, "Freedom of education" gives the right to any natural or legal person to start a school. This "freedom of organization" allows each school to develop its own educational policies, including its own pedagogical plan, teaching methods, curriculum and timetables, as well as to appoint its own staff. Although, schools receiving public funding are required to operate within a regulatory framework, they still enjoy "considerable autonomy" (Ministry of Education and Training and the University of Antwerp, 2010).

At the same time, “Freedom of education” gives the right to all parents to choose a school for their child(ren). This means that schools are accountable to parents and lays the foundation for potentially strong competition among schools (ibid, p.18)

Final discussion on comparing examinations and testing in The Netherlands, Italy, Belgium (Fl) and Sweden

First of all, it should be emphasized that the comparative case-study had a very limited scope, and should be seen as a very preliminary pilot study, rather than a fully- fledged empirical study. Therefore the conclusions are tentative.

When comparing the four countries the following patterns could be discerned:
The Netherlands and Italy have central examinations and highly developed educational testing systems.
Sweden and Flanders have neither central examinations, nor highly developed educational testing systems.
So, it could be concluded that Italy and the Netherlands are high on “output standardization” and Sweden and Belgium (Fl) are low on output standardization. It should be noted, however, that the central part of final examinations at secondary education in the Netherlands is more extensive (in the sense of subjects covered) than in Italy.

When turning to pro-active curriculum structuring as a form of “input standardization” the situation is more complex. The following observations can be made:

- All four educational systems put a high premium on the autonomy of schools and teachers
- Yet, they score differently on proxy quantitative indicators of input standardizations with the Netherlands and Belgium scoring quite low, Italy somewhat higher and Sweden relatively high.
- Given the weakness of the quantitative indicators on input standardization these results should be taken with a lot of reservation. Still, for the two systems, on which actual curriculum documents were reviewed, the Netherlands and Italy, the above pattern was supported: Italy seems to have a more detailed prescriptive set of curriculum guidelines than the Netherlands.

Methodologically it could be the case that social desirability leads countries to downplay curriculum standardization. Emphasizing the value of autonomy in education has a high symbolic and political value: everyone likes it and not that much needs to be done because it is already there in the form of traditional professional autonomy of teachers. Currently emphasizing autonomy is more popular than ever (e.g. Schleicher, 2018). To try and penetrate the autonomy rhetoric it could be worthwhile to go back in time and study historical trends. An illuminating example is Sahlgren’s analysis of the Finnish educational miracle. In his book “Real Finish lessons” (2015) he explains early success on PISA by referring to a history of centralistic governance and traditional teaching. In the same vein centralistic trends in curriculum standardization might have strong historical backgrounds in Italy and Flanders. In the Netherlands school autonomy has a fairly long tradition, and examinations have always acted as the “countervailing” alternative option for central focus and control.

And what does all of this matter? From an educational effectiveness perspective there seem to be several alternative hypotheses on the table:

- Educational systems need a mixture of control and autonomy, where input and output standardization could replace one another.

- Educational systems need (a minimum level of) both input and output standardization for optimal performance.
- Educational systems should maximize autonomy and get rid of input and output standardization as much as possible.

Empirical research so far seems to favor the model of what was once presented as “new” public management, free process and control outcomes. See for example Woessmann’s results where the positive influence of a standardized examination is particularly evident in otherwise highly autonomous systems (Woessmann, 2018).

Finally, apart from new Finnish lessons, there is perhaps also a Swedish lesson. The Swedish system was traditionally quite centralistic. Next followed a period where rather radical decentralization took place. More recently they are considering some re-centralizations, also prompted by critical reviews by the OECD (OECD, 2011, OECD, 2017). Perhaps the Swedish lesson is that systems should be reluctant in upsetting traditional governance approaches and be very careful in drastically overhauling institutional features like standardized curricula and examinations.

References

Aljello, A. (2017) The Education system in Italy. The National Evaluation System. R0me: INVALSI

Clanchini Monti (2018) I test di inglese: una novità assoluta? Scuola7 (1 ottobre, 2018)

EURYDICE (2004) European Glossary on Education.(2004) Volume 1 – Second edition

Examinations, Qualifications and Titles. Brussels: EURYDICE, 2004

Indicazioni nazionali per il curricolo della scuola dell’infanzia e del primo ciclo d’istruzione

Annali della pubblica istruzione (2012)
http://www.indicazioninazionali.it/wp-content/uploads/2018/08/Indicazioni_Annali_Definitivo.pdf

Indicazioni-nazionali-e-nuovi-scenari ***Annali della pubblica istruzione*** (2018)
https://www.orizzontescuola.it/wp-content/uploads/2018/02/Indicazioni-nazionali-e-nuovi-scenari.pdf
Ministero dell'Istruzione dell'Università e della Ricerca (2018) Scuola secondaria di secondo grado

http://www.miur.gov.it/scuola-secondaria-di-secondo-grado

EURYDICE (2018) Assessment in Primary Education –Italy https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-primary-education-23_en Brussels: EYRYDICE
EURYDICE (2018) Assessment in General Lower Secondary Education - Italy

https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-lower-secondary-education-18_en Brussels: EYRYDICE

OECD (2011) School Evaluation in the Flemish Community of Belgium, *Reviews of Evaluation and Assessment in Education.* Paris: OECD Publishing *http://dx.doi.org/10.1787/9789264116726-en*

OECD (2011), *OECD Reviews of Evaluation and Assessment in Education: Sweden 2011,* OECD Publishing. *http://dx.doi.org/10.1787/9789264116610-en*

OECD (2015) Improving Schools in Sweden: An OECD Perspective. Paris: OECD

Sahlgren, G. H. (2015). Real finish lessons. The true story of an education superpower. Center for Policy Studies. Surrey.

Scheerens, J. Glas, C.A.W., Luyten, H., Jehangir, K. (2013) System level correlates of educational performance. A study based on PISA 2009 data. Enschede: University of Twente. Unpublished Manuscript

Schleicher, A (2018), *World Class: How to build a 21st-century school system*, Strong Performers and Successful Reformers in Education, OECD Publishing, Paris.
http://dx.doi.org/10.1787/4789264300002-en

Van de Werfhorst, H., Elfers, R., Karsten, S. ed. (2015)Onderwijsstelsels vergeleken, leren, werken en burgerschap. Meppel: Ten Brink

Woessmann, L. (2018). Central exit exams improve student outcomes. External school leaving exams raise student achievement and improve how grades are understood in the labor market, IZA *World of Labor* January 2018: 419.

Annex 4: Verslag van de bijeenkomsten van twee focusgroepen: schoolleiders en toets- en examenexperts

Anne Luc van der Vegt en Jaap Scheerens

Een samenvatting van de rapportage over de literatuurstudie is voorgelegd aan twee focusgroepen, op 20 november en 10 december 2018. De eerste focusgroep bestond uit tien schoolleiders voortgezet onderwijs, de tweede uit experts op het gebied van toetsing en examinering, aangevuld met vertegenwoordigers van onderzoeksconsortium: Oberon, Universiteit Twente en RCEC. De bijeenkomsten werden geleid door Jaap Scheerens en Anne Luc van der Vegt.

Deelnemers

Focusgroep schoolleiders
Rob Aarts, OMO sg, Helmond
Wiebe Brouwer, St. Vrije Scholen, Amsterdam
Hans Buijze, gepensioneerd rector
Marij Dings, Bonnefanten College, Maastricht,
Ajolt Elsackers, St. Maartenscollege, Den Haag
Wendy Groen, Wolffert Tweetalig, Rotterdam
Jeroen van Grunsven, Picasso Lyceum, Zoetermeer
Eduard Nagel, Bonaventuracollege, Leiden
Erik van der Vaart, Via Nova College, Utrecht
Johan Veenstra, Comenius College, Hilversum

Focusgroep toets- en examenexperts
Marinka Drost, Docentplus
Jan Kastelein, CvTE, sectormanager vmbo
Willem Rosier, SLO
Piet Sanders, RCEC
Hendrik Straat, Cito
Bernard Veldkamp, Universiteit Twente
Petra Verra, Docentplus

Doel en gespreksonderwerpen

Doel van de bespreking was om te reflecteren op de bevindingen van de onderzoekers en om input te verzamelen voor een opinieonderzoek onder het onderwijsveld, dat in vervolg op deze studie kan worden uitgevoerd.

Met de focusgroepen is gediscussieerd aan de hand over vier thema's:
1. De functie van het examen
2. Kritiek op het examen
3. Examen en curriculum
4. Centraal examen en schoolexamen

Het onderstaande verslag is een samenvatting van de opinies en argumenten die naar voren zijn gebracht. Per thema worden de standpunten binnen beide focusgroepen samengevat. Bij de schoolleiders doen we dit als reacties op enkele stellingen, waarop zij hebben gereageerd. Bij de focusgroep van toets- en examenexperts vond een meer 'open' discussie plaats, zonder stellingen.

1. Functie van het examen

Focusgroep schoolleiders

Stelling: Examens werken motiverend
De meeste schoolleiders vinden niet dat examens motiverend zijn voor leerlingen. Dat wil zeggen: met examens worden leerlingen *extrinsiek* gemotiveerd, maar niet *intrinsiek*. Het laatste is wenselijk, het eerste niet, is de mening van de meeste schoolleiders.
Gezien de hoeveelheid toetsen die scholen afnemen, lijken scholen er echter vanuit te gaan dat leerlingen niet zonder extrinsieke motivatie kunnen. 'Dat is het oude denken,' reageert één van de schoolleiders. 'Als je dat blijft zeggen, dan zal dat ook zo blijven. '
Volgens anderen is extrinsieke motivatie op zich niet verkeerd, mits leerlingen ook intrinsiek gemotiveerd zijn. 'Je hebt leerlingen die versnellen. Die hebben een vergelijkbare motivatie als een topsporter. Ze leggen hun examen eerder af en zijn daarvoor super gemotiveerd.'
Daarvoor is 'eigenaarschap' nodig bij leerlingen, vindt een andere schoolleider. Eigenaarschap is schaars tegenwoordig, maar kan bevorderd worden door meer differentiatie en maatwerk.

Stelling: Examens geven focus aan het onderwijs
Schoolleiders vinden dat dit het geval is. Maar die focus wordt niet door iedereen positief beoordeeld. 'Zodra je een summatieve toets hebt gehaald, dan vergeet je de stof ook zo snel mogelijk.'
Bij formatieve toetsen speelt dit probleem niet of minder, vinden enkele schoolleiders. Zij pleiten voor meer gebruik van formatieve toetsen. Daarbij helpt het om de toetsen 'kleiner' te maken, zodat je meer ruimte hebt om leerlingen te laten reflecteren op wat ze aan het doen zijn. Een schoolleider verwacht dat dit ook kan helpen bij het maken van een onderbouwde studiekeuze.
Volgens sommigen zou het examen juist nog meer focus aan het onderwijs kunnen geven. 'Als je je op het examen richt, hoef je zestig procent van de methode niet te behandelen. De tijd die je bespaart kun je aan andere zaken besteden.' Focus betekent dus niet dat *al* het onderwijs op het examen gericht moet zijn.
Een andere schoolleider bestrijdt dat dit type curriculaire focus wenselijk is. Vakinhoudelijke kennis is heel betrekkelijk. Uit een gesprek met oud-scholieren heeft hij geleerd dat zij – inmiddels afgestudeerd

– hadden willen leren samenwerken, verantwoordelijkheid dragen, met tijd om gaan, mediawijsheid, zelfreflectie, presenteren. De inhoud van de avo-vakken wordt door enkele schoolleiders veel te veel bepaald door wat wetenschappers belangrijk vinden, niet door wat mensen nodig hebben in de beroepspraktijk.

Focusgroep toets- en examenexperts

Vanuit RCEC wordt een kanttekening geplaatst bij de functie van examens als toegangspoort tot het vervolgonderwijs. Het slagen voor een examens leidt tot diplomering, maar dat hoeft niet altijd te betekenen dat de gediplomeerde wordt toegelaten tot vervolgonderwijs. Daarvoor kunnen aanvullende eisen gelden. Ook is te overwegen dat niet *alle* eisen voor een diploma hoeven gelden voor toelating tot vervolgonderwijs. Moet een wiskundige een voldoende hebben voor moderne vreemde talen? Misschien niet. Je kunt dus onderscheid maken tussen de functies van diplomering en toelating tot vervolgonderwijs.
Voor het beroepsonderwijs wordt als specifiek facet van het civiel effect genoemd: 'bescherming van de burger'. Je wilt voorkomen dat technici ongediplomeerd hun beroep uitoefenen. De burger zou daar nadeel van kunnen ondervinden.
De vertegenwoordiger van het CvTE vindt dat de waarde van het examen als indicator voor onderwijskwaliteit beperkt is. Dat leidt ertoe dat scholen zich teveel richten op de examenresultaten. Dit is een 'perverse prikkel'. Je zou eigenlijk het succes van leerlingen in het vervolgonderwijs moeten meenemen. Van de havo-gediplomeerden van sommige scholen switcht meer dan de helft in het eerste jaar van het hoger beroepsonderwijs. Meer aandacht voor LOB zou dit kunnen verhelpen. Maar die aandacht vertaalt zich niet in hogere examencijfers. Daarom is het onverstandig scholen alleen op die examenresultaten te beoordelen
Meer aandacht voor LOB kan er ook toe leiden dat leerlingen een bewuste keuze maken voor een geschikte beroepsopleiding, in plaats van na de mavo maar de havo te kiezen, omdat ze het nog niet weten. Te meer daar veel mavo-gediplomeerden vervolgens stranden op de havo.

2. Kritiek op het examen

Focusgroep schoolleiders

Stelling: Huidige examens leiden tot versmalling van het curriculum
Bijna alle schoolleiders zijn het (zeer) eens met deze stelling. Het centrale examen is het schoolexamen gaan 'domineren'. Dit komt onder meer doordat de Inspectie is gaan kijken naar het verschil tussen gemiddelde examencijfers op SE en CE. Ook al worden scholen hier niet meer op beoordeeld, in de beleving van scholen is dat risico er nog steeds. Om een te groot verschil te voorkomen, zorg je dat SE zoveel mogelijk op het CE lijkt. Maar dit heeft een keerzijde. Leuke, alternatieve toetsen verdwijnen. 'Het werkt zeer demotiverend voor docenten,' stelt een schoolleider vast.
Een andere schoolleider vindt da het CE niet tot versmalling hoeft te leiden. 'Dat doen we onszelf aan. Ik denk dat je met 30% van de stof die we nu allemaal doorploeteren met de examenleerlingen ook het examen haalt. Die 70% kan je dan ook aan andere dingen besteden.'
Een schoolleider: 'Wat ik veel interessanter vind dan het verschil tussen SE en CE is: waar sta ik met mijn centraal examen in de benchmark. Als dat ok is, denk ik: mijn docenten voldoen blijkbaar. Dan wil ik

graag ruimte behouden om impulsen aan mijn leerlingen te geven. Want dat motiveert leerlingen.' Anderen betwisten het belang van zo'n centrale benchmark.

De druk van presteren wordt niet alleen gevoeld in de bovenbouw. 'Het begint al in de brugklas. Daar stel je vast of een kind wel voldoende cijfers haalt om over te kunnen gaan. Terwijl je je ernstig moet afvragen of je wel toetst in het belang van het kind. Er zijn scholen die er over denken om leerlingen niet over te laten gaan, omdat die te weinig kans zou hebben om te gaan slagen. Dat heeft een drukkend effect op de resultaten van de school. De vraag is of het aanpassen van het systeem van centrale examinering effect heeft op de doorstroom binnen het onderwijs.

Stelling: De huidige examens zijn te weinig flexibel; meer maatwerk is wenselijk.

Veel schoolleiders vinden maatwerk wenselijk, maar op welke manier, dat is nog niet iedereen duidelijk. Een mogelijkheid is bijvoorbeeld dat vwo-leerlingen niet elk vak op vwo-niveau doen en toch een academische studie kunnen volgen. 'Wellicht zou je daar vervolgopleidingen bij moeten betrekken? Een studie geneeskunde zou misschien niet op elk vak VWO niveau nodig hebben.'

Het eindexamen is nu teveel een eindpunt in het denken, vinden sommigen. Je zou willen dat je de resultaten vervolgens weer toepast, dus formatief gebruikt. 'De zelfreflectie en het continu denken van 'wat ik nu aan het doen ben, levert mij dit nog op wat ik tot doel had gesteld?', dát wil je graag die jonge mensen meegeven. Dat is een eeuwig durend proces.'

Het belang van het examen als toelating tot vervolgonderwijs wordt ter discussie gesteld. De decentrale selectie speelt daarbij een rol. 'De hele matching is hierop gebaseerd, zoals in Utrecht. Als student doe je ervaring op, en docenten doen ervaring op met jou als student. Dat heeft voorspellende waarde dan je cijferlijst.'

Een andere schoolleider stelt een meer soepele overgang voor tussen vo en wo. 'Als je graag een bepaalde studie wilt doen in het vervolgonderwijs zou het mooi zijn als je al in een deel van het laatste jaar daar in kan groeien en kan ervaren of het wat voor jou is. En of je het juiste niveau voor instroom hebt.'

Focusgroep toets- en examenexperts

De meningen over de kritiek op het examen verschillen. Vertegenwoordiger RCEC vindt dat 'teaching to the test' op zich niet verkeerd is, als de test goed geconstrueerd is. 'Bij het staatsexamen gebeurt dat ook.' Andere experts vinden dat er een risico van versmalling is. Dit vloeit niet rechtstreeks voort uit de exameneisen, maar eerder uit de beoordeling van onderwijsresultaten door de Inspectie. Gesignaleerd wordt dat het systeem van SE en CE scholen de mogelijkheid biedt om het curriculum te verbreden, maar dat daar te weinig gebruik van wordt gemaakt. Scholen zorgen dat het SE niet te veel afwijkt van het CE, om te voorkomen dat de cijfers te veel verschillen.

Wat betreft flexibilisering door meerdere toetsmomenten voor het CE te introduceren: volgens CvTE is dit alleen mogelijk als het SE niet hoeft te worden afgerond voor leerlingen aan het CE beginnen. Anders vraagt de constructie van SE-toetsen veel te veel tijd. Afschaffen van het tweede tijdvak zou ook een mogelijkheid kunnen zijn.

Wat betreft flexibilisering door leerlingen op verschillende niveaus examen te laten doen: daarvoor is de verschillende lengte van opleidingen in het vo een obstakel. Een oplossing zou volgens CvTE kunnen zijn: voeg havo en vwo samen en maak voor elk twee niveaus. Per vak bepalen de leerlingen dan op welk niveau ze examen doen.
Vertegenwoordiger Docentplus vermoedt dat dit kan leiden tot een zwaardere rol van decentrale selectie door hoger onderwijs.

3. Examen en curriculum

Focusgroep schoolleiders

Stelling: Een goed examen meet de leeropbrengst van alle onderdelen van het curriculum.
Sommige schoolleiders hinken op twee gedachten. 'Volgens de logica zou je willen dat de aansluiting tussen het examen en de doelstelling optimaal is. Maar aan de andere kant zijn er een heleboel dingen die belangrijk zijn, die misschien toch niet te meten zijn.'
Anderen vindt dat het examen zich zou moeten beperken tot het toetsen van een kerncurriculum. Dit laat de school meer vrijheid voor het inrichten van de rest van het onderwijs.
Eén van de schoolleiders verwoordt het standpunt dat examen en curriculum niet naadloos op elkaar hoeven te passen. 'Je moet vaststellen of een leerling aan de eisen van een bepaald niveau heeft voldaan. Daarmee geef ik niet aan of het zwaarder of lichter moet zijn. Bij het examen moet je een bepaald niveau halen, als je dat niet haalt, heb je een probleem.'

Focusgroep Experts

Moet met het examen ook de functies van persoonsvorming en socialisering worden gemeten? Sommige experts zien daartoe mogelijkheden. Vertegenwoordiger van het Cito: 'In het SE heb je de ruimte om andere opdrachten te geven, meer socialisatie en persoonsvorming, meer praktijkopdrachten. Dan is het examen als geheel meer in balans.' Toch vindt het Cito niet dat het examen alle onderwijsdoelen moet afdekken. Sommige dingen moet je wel in het leerplan opnemen, zonder ze te examineren. Vanuit CvTE wordt dit onderschreven. 'Niet alles wat belangrijk is, is examineerbaar.'

4. Centraal examen en schoolexamen

Focusgroep schoolleiders

Stelling: Het schoolexamen is van belang als voorbereiding op het centraal examen
Schoolleiders onderschrijven dat dit zo is. Tegelijkertijd vinden ze het onwenselijk dat er veel overlap is tussen het CE en het SE. Hoe die overlap kan worden voorkomen is niet uitgebreid besproken.

Stelling: Het centraal examen zou zwaarder moeten wegen dan het schoolexamen.

Er zijn schoolleiders die het bespreekbaar vinden om het centraal examen af te schaffen. Decentrale selectie zou daarvan de plaats kunnen innemen. Het examen is de duidelijkste exponent van het summatieve toetsen. ‘We werken naar het eindpunt van de CE toe. Het summatieve denken wordt versterkt doordat scholen hun eigen onderwijs hiernaar inrichten. Het hoeft niet zo te zijn, maar gebeurt wel.’
Anderen zijn op zich niet tegen het CE. ‘Het hangt er volledig van af wat je gaat toetsen in het CE. Wat houdt het CE in de nieuwe vorm?’
Ten slotte zijn er ook schoolleiders die het CE wel willen handhaven maar het SE ter discussie stellen. ‘Wat mij betreft stoppen we met SE. De druk van het centraal examen via het SE, daar zou ik van af willen. Zoals deze er nu is met toezicht van de inspectie.’ Voor behoud van het CE pleit volgens enkele schoolleiders ook dat dit helpt voor het borgen van de kwaliteit.

Focusgroep toets- en examenexperts

Dat het SE te veel overlapt met het CE, onderschrijven de experts. Sommigen, bijvoorbeeld van Docentplus, vinden dat de richtlijnen voor het SE specifieker zouden moeten worden gemaakt. Anderen, bijvoorbeeld CvTE, vinden dat scholen juist meer ruimte moeten krijgen. Er is al zoveel ‘dichtgetimmerd’ in het onderwijs. In dat kader zou bijvoorbeeld de tweede correctie kunnen worden afgeschaft.
Door SLO wordt erop gewezen dat in syllabi al behoorlijk wat richting wordt gegeven aan het SE. Bovendien zijn er handreikingen voor de constructie van schoolexamens. Alleen worden ze niet optimaal gebruikt. Heel veel van wat in het CE wordt getoetst, komt ook in SE-toetsen terecht. We moeten dus op zoek gaan naar mechanismen om te bevorderen dat het SE meer schooleigen wordt.
Als voorbeeld van een effectieve werkwijze om overlap tussen SE en CE tegen te gaan, wordt door het CvTE de innovatie van de beroepsgerichte in het vmbo genoemd. Daar zijn nu 9 profielvakken (combi’s van metaal-, elektro- en installatietechniek) die centraal worden geëxamineerd en 250 keuzevakken, met een smalle vakinhoud. Voor de keuzevakken wordt alleen een schoolexamen afgelegd. De stof voor CE en SE is hierdoor disjunct, heeft geen gemeenschappelijke elementen.

Slotvraag: Wat zijn de belangrijkste uitdagingen voor de instandhouding, modernisering of vernieuwing van de examens in het voortgezet onderwijs?

Focusgroep schoolleiders

Over curriculumontwikkeling:
‘Het lastige van curriculumconstructie is dat het ook duurzaam en bestendig moet zijn. Maar het is na korte tijd weer zo anders dat je dit moet aanpassen. De kunst is om een basiscurriculum te maken wat duurzaam is en wat eigenlijk altijd handig is om in je bagage te hebben.’

Over aansluiting curriculum in verschillende schooltypen:

'Het havo- en vwo-curriculum sluit niet goed op elkaar aan. Als je op een ander niveau examen wilt doen dan wordt het je wel lastig gemaakt. Als je nu je herkansing ook niet haalt op een hoger niveau, mag je dan nog examen doen op je eigen niveau. Maar dat kan eigenlijk helemaal niet, omdat je dan nog andere examenonderwerpen moet gaan leren.
Je kan niet aan flexibilisering werken als de curricula en daarmee de examens niet op elkaar aansluiten.'
Als oplossing voor dit probleem wordt genoemd een modulaire opbouw van het hele curriculum, door de schooltypen heen.

Focusgroep toets- en examenexperts

De experts vinden unaniem dat het CE niet ter discussie zou moeten staan. Met het CE zou in ieder geval een kerncurriculum moeten worden geëxamineerd. Scholen hebben dan ruimte voor een eigen invulling van het SE en het curriculum.
Nader onderzoek wordt gevraagd naar het fenomeen van examenstress. Is deze te reduceren of is deze van alle tijden en plaatsen?
Ook de doorlopende leerlijn tussen po en vo vraagt aandacht. Het curriculum sluit niet op elkaar aan en de wijze van toetsen ook niet.

Oberon

Postbus 1423, 3500 BK Utrecht
t 030 230 60 90 | f 030 230 60 80
info@oberon.eu | www.oberon.eu

Utrecht, maart 2019