BIJLAGE

Deze bijlage is vooral beschrijvend van aard.[1] In de laatste paragraaf wordt aangegeven dat alertheid geboden is om ongewenst strategisch gedrag tegen te gaan. Voor het overige bestaat de beschrijving uit een schets van een aantal aspecten van internationale toetspraktijken, te weten:

1. Onderzoek naar effect centrale toetsing
2. Een overzicht van verplichte toetsen in een aantal Europese landen.
3. Twee goede praktijken met centrale toetsing: Noorwegen en Ontario.
4. Het publieke gebruik van toetsresultaten (accountability en strategisch gedrag)

1. Onderzoek naar effect centrale toetsing

Een deel van de bevindingen in de internationale onderzoeksliteratuur is gebaseerd op de effecten van centrale examens met een civiel effect. Hoewel de eindtoets PO niet het karakter heeft van een examen, kunnen deze bevindingen daarop wel van toepassing worden verklaard. De eindtoets heeft namelijk (samen met het schooladvies) consequenties voor het niveau waarop een leerling doorgaat in het voortgezet onderwijs. Dit prikkelt zowel de leerling als de school om zo goed mogelijk te presteren.

De context waarin (centrale) toetspraktijken uitgevoerd worden is belangrijk bij de weging van de internationale evidentie. Dit heeft bijvoorbeeld te maken met de autonomie van scholen en besturen, de aanwezigheid van een nationaal curriculum, de inrichting van het toezicht, de wijze van financiering en de verdeling van verantwoordelijkheden tussen centrale en decentrale overheden. Daardoor kan niet zonder meer worden gesteld dat centrale toetsing op zichzelf automatisch leidt tot betere prestaties. Zo constateert McKinsey & Company[2] dat toetsing als instrument om vorderingen van leerlingen continu te volgen, een van de hoofdkenmerken is die goed presterende stelsels gemeen hebben. Maar dan wel in combinatie met de volgende hoofdkenmerken:

a. Hoge kwaliteit en status van leraren.
b. Leraren worden goed opgeleid en bijgeschoold om kwalitatief goede instructie te geven (opbrengstgericht werken).
c. Systemen om kinderen met verschillende talenten en onderwijsbehoeften op een goed niveau te brengen.

Hierbij gaat het er ook om dat leraren en schoolleiders in hun klassen en scholen gebruik maken van assessment- en toetsresultaten en daarmee opbrengstgericht werken. Dit draagt bij aan een effectief stelsel. Die conclusie kan worden getrokken op basis van drie belangrijke overzichtsartikelen; Kluger & DeNisi (1996)[3], Black & William (1998)[4] en Hattie en Imperley (2007)[5]. De auteurs komen allen op basis van meta-analyses en literatuurstudies tot de bevinding dat feedback en gebruik van assessment- en toetsresultaten op klasniveau leiden tot effectiever onderwijs.

Dit neemt niet weg dat er empirisch onderzoek is, waarin sec het positieve effect van centrale eindtoetsen is aangetoond. Centrale eindtoetsen kunnen zorgen voor transparantie en verantwoording. Dit geeft scholen (leraren, schoolleiders en besturen) prikkels om zich sterker te concentreren op het verbeteren van prestaties van leerlingen. Centrale toetsen geven ook een duidelijk moment voor leerlingen om naartoe te werken. Sterke aanwijzingen voor positieve effecten van centrale toetsen op de onderwijskwaliteit worden gevonden in internationale empirische studies, waarin prestaties van leerlingen worden vergeleken in landen (of regio's) met en zonder centrale eindtoetsen.

[1] Een belangrijke bron voor deze beschrijving is het inventarisende onderzoek van Eurydice (2009): *National testing of Pupils in Europe: Objectives, Organisation and Use of Results.*
[2] Michael Barber en Mona Mourshed, 2007, *How the world's best-performing school come out on top*, London, McKinsey & Company.
[3] Kluger, A.N. en A. DeNisi, 1996, The Effects of Feedback Interventions on Performance: Historical Review, a Meta-Analysis and a Preliminary Feedback, In: *Psychological Bulletin*, vol. 119, pp. 254-284.
[4] Black, P. en D. William, 1998, Inside the black box: Raising standards through classroom assessment, *Phi Delta Kappan,* vol. 80*,* pp. 139-148.
[5] Hattie, J. and H. Timperley, 2007, The power of feedback, In: *Review of Educational Research*, Vol. 77, No. 1, 81-112.

Zo vergeleek Bishop[6] internationale toetsscores tussen landen met en zonder centraal examen. Hij vindt daarbij significante effecten[7] op kernvakken als wiskunde en natuurkunde.

Woessmann[8] vergeleek individuele prestaties van 260.000 leerlingen uit 39 landen die meedoen aan TIMMS[9] (13-jarigen) geanalyseerd. Toetsscores van individuele leerlingen werden daarbij gerelateerd aan persoonlijke achtergrondkenmerken, inzet van financiële middelen en allerlei institutionele kenmerken van onderwijssystemen, waaronder de aanwezigheid van centrale eindtoetsen. Hij vond statistisch significante positieve effecten van centrale eindtoetsen op de toetsscores op wiskunde en science. Een vergelijkbare analyse is door Fuchs en Woessmann[10] uitgevoerd op de PISA-toetsen in 32 landen, waaronder 28 OECD-landen. In deze studie worden positieve effecten van centrale eindtoetsen gevonden op alle toetsen.

Jürges et al.[11] maken gebruik van regionale variatie in schoolwetgeving in het Duitse voortgezet onderwijs om het effect van centrale eindtoetsen op onderwijsprestaties te identificeren. Hiervoor kijken ze naar de effecten op de TIMMS toetsscores op wiskunde en science. In alle Duitse provincies zijn centrale eindtoetsen voor wiskunde, maar in de meeste provincies is geen centrale eindtoets voor science. Zij vinden significant positieve effecten van centrale eindtoetsen op de prestaties voor science ten opzichte van provincie die daarvoor geen centrale toetsen afnemen.

Een andere observatie[12] is dat autonomie van scholen een sterker positief effect heeft op onderwijskwaliteit in landen met centrale toetsen en examens. Het idee hierachter is dat autonomie vooral werkt wanneer dit gekoppeld wordt aan verantwoording. Voor verantwoording zijn heldere, absolute normen een vereiste. Woessmann noemt centrale toetsen en examens in die zien ook wel de 'munteenheid' van het onderwijssysteem.

2. O*verzicht aantal lidstaten Europese Unie met verplichte testen voor leerlingen in het basisonderwijs en de onderbouw voortgezet onderwijs per leerjaar en vakken*

	Grade	Vakken	Doel
Belgie FR	6	Frans, wiskunde, natuur, geschiedenis, aardrijkskunde	Summatief[13]
Denemarken	2,4,6,8	7 vakken (afwisselend)	Monitor scholen en systeem, formatief[14]
Duitsland	3	Duits, rekenen	Monitor scholen en systeem
	8	Duits, wiskunde, Engels	Monitor scholen en systeem
Engeland	2	Rekenen, lezen, schrijven	Formatief
	5	Engels, rekenen, science	Monitor scholen en systeem,
Frankrijk	2	Frans, rekenen	Monitor scholen en systeem
	5	Frans wiskunde	Monitor scholen en systeem
	9	7 vakken	Summatief
Italie	9	Italiaans, wiskunde, science, Engels	Summatief
Luxemburg	3	Duits, rekenen	Formatief
	6	Duits, Frans, rekenen	Summatief
Noorwegen	5 en 8	Noors, wiskunde, Engels	Monitor scholen en systeem
Portugal	4 en 9	Portugees, wiskunde	In grade 4: monitor scholen en systeem; in

[6] Bishop, J.H., 1997, The Effect of National Standards and Curriculum-Based Examinations on Achievement, *American Economic Review*, vol. 87, pp. 260-264.
[7] De meeste effecten lopen uiteen van 0.10 tot 0.20 standaarddeviatie. Dat zijn geen kleine effecten. Ter vergelijking, het effect op leerprestaties van een goede tegenover een gemiddelde leraar ligt in dezelfde orde van grootte.
[8] Woessmann, L, 2003, Schooling resources, educational institutions and student performance: The international evidence, *Oxford Bulletin of Economics and Statistics*, vol 65, nr. 2, pp. 117-170
[9] Trends in International Mathematics and Science Study
[10] Fuchs, T. and L.Woessmann, 2007, What accounts for international differences in primary schools learning across countries, Mimeo, IFO Institute for Economic Research at the University of Munich.
[11] Jürges et al., 2005, The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany, *Journal of the European Economic Association* vol. 3, nr. 5., pp 1134-1155.
[12] Woessmann, 2003 en Fuchs en Woessmann, 2007.
[13] Een summatieve toets vindt plaats aan het einde van een leerproces, meestal om een eindoordeel uit te spreken of te selecteren.
[14] Een formatieve toets dient om het leerproces, indien nodig, in de gewenste richting bij te sturen.

			grade 9: summatief
Spanje	4 en 8	Taal, wiskunde, science ICT, burgerschap, cultuur	Monitor scholen en systeem (per regio)
Zweden	5 en 9	Zweeds, Engels, Wiskunde	In grade 5: formatief; in grade 9: summatief

(Grade 2 komt overeen met groep 4 basisonderwijs, grade 4 met groep 6, grade 6 met groep 8, grade 9 komt overeen met klas 3 voortgezet onderwijs etc.)

Landelijke toetsen voor alle scholen en leerlingen zijn er bijvoorbeeld in Australië, Denemarken, Zweden en Engeland. Australië heeft sinds eind vorige eeuw een jaarlijks, nationaal testprogramma voor taal en rekenen inclusief nationale benchmarks (in alle staten). Afname is in de grades 3, 5, 7 en 9. Testen zijn bedoeld om vast te stellen of leerlingen voldoen aan de nationale benchmarks en leraren te informeren of er verbeteringen in het onderwijs nodig zijn.

Een interessant voorbeeld vormt Engeland dat aanvankelijk veel nationale toetsen had ingevoerd en daaraan vergaande consequenties voor scholen verbond. Dit ging ten koste van het draagvlak onder scholen en om die reden hebben de Engelsen hun toetspraktijk aangepast. Het aantal (centrale) toetsmomenten nu is teruggebracht tot drie (op 7-, 11- en 16-jarige leeftijd). De zevenjarigen worden door hun leraar getoetst hun beheersing van de moedertaal en rekenen en op elfjarige leeftijd worden de leerlingen landelijk getoetst op moedertaal, rekenen-wiskunde en science. De resultaten van deze verplichte landelijke toetsen worden door Ofsted (de Engelse inspectie) gebruikt om de scholen te beoordelen. Door te kijken naar de verschillen tussen twee toetsmomenten spreekt men ook over de zogenaamde 'value added' (vergelijkbaar met ons streven om leerwinst in beeld te brengen). Door contextvariabelen als etniciteit mee te nemen spreekt men ook over 'contextual value added' (vergelijkbaar met toegevoegde waarde). Al deze informatie wordt ook door scholen en schoolbesturen gebruikt om de eigen opbrengsten te evalueren.

Duitsland en de VS organiseren hun landelijke toetsen op het niveau van hun (deel)staten. In de VS moeten staten de ontwikkeling van hun leerlingen in taal en rekenen-wiskunde in kaart brengen; die toetsen zijn gekoppeld aan bepaalde standaarden. Landen en regio's als Finland en Vlaanderen (met een goede positie in PISA[15]) hebben geen nationale toetsen voor alle leerlingen. Ze doen wel jaarlijks, steekproefsgewijs onderzoek met toetsen onder leerlingen van ongeveer 12 jaar oud. Daarmee evalueren zij periodiek het niveau van het nationale onderwijs op één of meer vakken. De functie van deze toetsen is beperkt tot een instrument voor het ontwikkelen van nationaal onderwijsbeleid en wordt op het niveau van het stelsel niet gebruikt voor de beoordeling van de opbrengsten van scholen.

3. Twee goede praktijken met centrale toetsen

Noorwegen

Noorwegen is, mede door tegenvallende PISA cijfers in 2000, gestart met een verbeterprogramma, waarin opbrengstgericht werken aan de hand van centrale toetsen een belangrijk plaats in neemt. In 2003 is het National Quality Assessment System ingevoerd, dat zich richt op kwaliteitsontwikkeling in het basis- en het voortgezet onderwijs. Naast informatie over leeropbrengsten, waardering van de leeromgeving, schoolsucces en hulpmiddelen, bevat dit systeem ook nationale toetsen.

Sinds 2007 is er een systeem met veel aandacht voor opbrengstgericht werken. Er wordt in grade 5 (vergelijkbaar met groep 7) en grade 8 (vergelijkbaar met klas 2 in het voortgezet onderwijs) getoetst op drie basisvaardigheden (Noors, Engels en rekenen-wiskunde). Daarnaast worden deze drie vakken geëxamineerd aan het eind van het lager secundair onderwijs. De toetsen bieden inzicht in de vorderingen van leerlingen ten aanzien van het bereiken van referentieniveaus en vormen een bestanddeel van de kwaliteitsgegevens over scholen. Zij worden benut voor twee doelen: het meten van leervorderingen bij leerlingen als hulpmiddel voor het onderwijs in de klas en het bieden van inzicht in de onderwijskwaliteit van het Noorse stelsel. Het draagvlak onder leraren is groot, omdat de overheid het primaat heeft gelegd bij het eerste, formatieve, doel.

De invoering van deze centrale toetsen, in combinatie met een forse investering in de kwaliteit van leraren en schoolleiders, heeft er mede toe geleid dat Noorwegen nu een duidelijke stijgende lijn laat zien voor de resultaten bij leesvaardigheid (een lichte stijging op PIRLS[16]) en rekenen-

[15] Programma for International Student Assesment
[16] Progress in International Reading Literacy Study

wiskunde en 'science' (een forse verbetering op TIMMS). Daarmee is Noorwegen een voorbeeld van een praktijk waarbij centrale toetsing als onderdeel van een bredere verbeterstrategie tot resultaten leidt.

Ontario, Canada

Canada doet vaak met enkele deelstaten mee aan internationale vergelijkingen. In het midden van de jaren negentig van de vorige eeuw vroegen publiek en ouders om een betere beoordeling van de resultaten van leerlingen en scholen. In 1996 heeft de regering van Ontario de 'Education Quality and Accountability Office' (EQAO) opgericht dat tot taak kreeg toetsen te ontwikkelen en beschikbaar te stellen voor onder meer leerlingen van de onderbouw en bovenbouw van de basisschool. Ontario heeft er toen voor gekozen om alle leerlingen te toetsen, omdat dat volgens de provincie de beste manier is om het onderwijs te verbeteren, op het niveau van de leerling, van de klas, van de school, van het schoolbestuur en van de provincie als geheel; landelijke steekproeven alleen zouden daarin tekortschieten.

Het doel van toetsen is om de ontwikkeling van leerlingen te stimuleren en om de verantwoording in het publieke scholen te versterken. Alle leerlingen nemen in principe deel aan deze provinciale toetsen.

Alle leerlingen op publieke scholen maken provinciale taal- en rekentoetsen in de onderbouw en bovenbouw van de basisschool; de toetsen weerspiegelen de kernvaardigheden voor taal en rekenen, zoals die zijn beschreven in het 'curriculum van Ontario'. Leerlingen, leraren, scholen en bestuurders ontvangen een toetsrapport dat past bij hun rol. Met de resultaten leggen scholen en schoolbesturen niet alleen verantwoording af aan ouders en publiek, maar zij gebruiken de uitkomsten ook om het onderwijs en de prestaties verder te verbeteren. Schooldirecties en leraren gebruiken de rapportage van het EQAO over de toetsresultaten voor de eigen schoolverbetering, waarvoor de doelen en acties onderdeel vormen van het jaarlijkse *School Improvement Plan*. Complementair daaraan gebruiken de onderwijsbesturen de analyse van de toetsresultaten voor de doelen en acties op bestuursniveau tot verbetering van het onderwijs. Deze maken onderdeel van het jaarlijkse *Board Improvement Plan*. Tot slot gebruikt het ministerie van onderwijs de gegevens om sterke en zwakke punten in het onderwijsstelsel te benoemen en zo nodig actie te ondernemen. De *Board Improvement Plans* vormen daarbij de basis voor het inhoudelijk gesprek tussen het ministerie en schoolbesturen.

In 2011 bleek de leesvaardigheid van leerlingen in Ontario hoger dan in 2001 en 2006. Bovendien liggen deze resultaten iets hoger dan in Nederland. Voor rekenen-wiskunde en science geldt dat de resultaten in het basisonderwijs in Ontario fors verbeterd zijn ten opzichte van midden jaren negentig.

Bishop[17] heeft een analyse gepresenteerd waarin hij gebruik maakt van variatie tussen verschillende provincies in Canada. Canada heeft een gemengd systeem, waarin sommige provincies centrale toetsen kennen en sommige niet. Analyses op schoolniveau laten zien dat scholen met een centrale eindtoets significant beter presteren op zowel wiskunde als science. Deze resultaten geven aan dat onderwijsprestaties kunnen worden verbeterd door de disciplinerende werking van centrale eindtoetsen.

4. Publiek gebruik van toetsresultaten (accountability en strategisch gedrag)

Accountability

Het gebruik van toetsresultaten voor de verantwoording over de prestaties van scholen is vooral onderzocht in de Verenigde Staten, maar er is ook enige evidentie uit landenstudies. In een vergelijking van de prestaties van leerlingen op internationale toetsen vinden Woessmann et al.[18] dat de prestaties hoger zijn in landen waarin beoordelingen van individuele scholen worden vergeleken met prestaties in het district of op landelijk niveau.

Tweederde van de EU-landen gebruikt landelijke toetsen, waarvan de resultaten op school- en landelijk niveau worden geaggregeerd, als onderdeel van verantwoording over de prestaties. Sommige landen, zoals Engeland en Schotland, verplichten of stimuleren scholen de resultaten te gebruiken in hun interne verantwoording over de kwaliteitsanalyse. Dit gegeven wordt vervolgens

[17] Zie noot 6.
[18] Woessmann, L. et al., 2009, *School accountability, autonomy, and choice around the world,* Cheltenham, UK and Northampton, MA.

ook ingebracht in de dialoog met de externe toezichthouder. In bijvoorbeeld Engeland en Zweden spelen de toetsresultaten een rol in het toezicht op de scholen. In een aantal landen (waaronder Denemarken, Engeland, Zweden en IJsland) worden de toetsresultaten openbaar gemaakt. Soms gaat dat in de vorm van ruwe data (Zweden), soms gecorrigeerd voor achtergrondgegevens (IJsland) en soms een combinatie van die twee (Engeland). In een aantal landen (zoals Denemarken, Schotland, Finland en Noorwegen) worden schoolresultaten ook gebruikt door lokale autoriteiten. De meeste landen gebruiken de schoolprestaties ook om op nationaal niveau te rapporteren over de kwaliteit van het onderwijsstelsel en om kwaliteitsbeleid te voeren.

Het verbinden van verantwoording aan toetsresultaten blijkt een positief effect te hebben op de prestaties van scholen. Drie studies voor de Verenigde Staten (Hanushek en Raymond[19], Jacob[20] en Dee en Jacob[21]) wijzen er op dat de invoering van 'school accountability' een positief effect heeft op leerprestaties. Hanushek en Raymond onderzochten of het publiceren van de prestaties van scholen als zodanig hiervoor verantwoordelijk was, maar ze vonden geen afzonderlijk significant effect hiervan op leerprestaties. Het effect werd eerder veroorzaakt door gevolgen te verbinden aan de prestaties van scholen in de vorm van sancties. Een uitgebreidere bespreking van deze en andere studies naar effecten van accountability-systemen is te vinden in De Wolf en Janssens.[22]

Strategisch gedrag
Wat verder opvalt is dat er in sommige landen een intensief debat gaande is over de rol en functie van centrale, gestandaardiseerde toetsen. Dit debat wordt heviger naarmate leraren en scholen verantwoordelijk worden gehouden voor de toetsresultaten door daar rechtspositionele en financiële gevolgen aan te verbinden. Onder de 'dreiging' van directe en vergaande gevolgen kan het risico van (ongewenst) strategisch gedrag toenemen. Dit gedrag is dan geen direct uitvloeisel van centrale toetsing zelf, maar van de gevolgen die verbonden worden aan de verantwoording van de resultaten.

Zo onderzocht Jacob[23] het accountability beleid dat in 1996-1997 in publieke scholen in Chicago is ingevoerd. Hierbij ging het om extern gebruik van toetsresultaten door het bevoegde schooldistrict met directe en vergaande consequenties voor de scholen. Scholen werden namelijk onder toezicht geplaatst als minder dan 15 procent van de leerlingen de nationale standaarden haalde. Als zich vervolgens geen verbetering voordeed werden docenten of schoolleiders herplaatst of ontslagen. Dat zijn vergaande gevolgen die we in Nederland niet kennen. Jacob vond weliswaar een aanzienlijke verbetering van toetscores in wiskunde en lezen, maar er was geen verbetering op toetsen die geen deel uitmaakten van het accountability beleid. Daarnaast werden meer leerlingen verwezen naar het speciaal onderwijs en bleven zij vaker zitten. Burgess et al.[24] onderzochten de effecten van accountability beleid in het Verenigd Koninkrijk. Dit beleid zette scholen aan om zich vooral in te spannen voor leerlingen waarvan ze verwachten dat die de standaard nog kunnen halen. De onderzoekers vonden dat de prestaties van leerlingen die de standaard waarschijnlijk niet konden halen slechter waren geworden na de introductie van het nieuwe beleid.

Het risico van dit soort negatieve effecten lijkt toe te nemen wanneer toetsresultaten worden gebruikt als belangrijke factor in beslissingen over (de hoogte van) de bekostiging van scholen of over rechtpositionele gevolgen voor leraren en schoolleiders. In het Nederlandse primair onderwijs is dat niet aan de orde. Er is geen sprake van prestatiebekostiging of ontslag van schoolleiders en leraren op basis van (de verantwoording over) toetsresultaten van leerlingen. De ontwikkeling naar opbrengstgericht werken en daarmee voor beter onderwijs aan alle leerlingen wordt ondersteund door een evenwichtig gebruik van toetsen. Daarom moeten we alert zijn op het optreden van ongewenste neveneffecten. Dit kan door:

[19] Eric A. Hanushek & Margaret E. Raymond, 2005. Does school accountability lead to improved student performance?, *Journal of Policy Analysis and Management*, vol. 24 (2), pp. 297-327.
[20] Jacob, B., 2005, Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools, *Journal of Public Economics*, vol. 89, no. 5-6, pp. 761-796.
[21] Dee, T., en B. Jacob, 2011, The impact of No Child Left Behind on student achievement, *Journal of Policy Analysis and Management*, vol. 30, no.3, pp. 418-446.
[22] Wolf, I.F. de en F.J.G. Janssens, 2007, Effects and side effects of school inspections and accountability in education: a review of empirical studies, *Oxford Review of Education,* vol. 33, no. 3, pp. 379-396.
[23] Jacob, B.A., 2002, Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools, NBER Working Paper 8968.
[24] Burges, S., C. Propper, H. Slater, en D. Wilson, 2005, Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools, CEPR Discussion Paper.

- Een inhoudelijk kwalitatief goede eindtoets die adequaat het beheersingsniveau meet van wat leerlingen qua kernvaardigheden aan het eind van het primair onderwijs worden geacht te kennen en kunnen. Het voorbereiden van leerlingen op de toets is daarmee onderdeel van het regulier onderwijsproces. De kern van het onderwijs is er immers voor een belangrijk deel op gericht om leerlingen te begeleiden naar het gewenste beheersingsniveau. Door de huidige, kwalitatief hooggewaardeerde Eindtoets Cito als basis te nemen voor de centrale eindtoets, wordt hiermee rekening gehouden.
- De beoordeling van de onderwijskwaliteit van een school niet alleen te baseren op de resultaten van centrale toetsing maar daarbij andere aspecten te blijven betrekken zoals kenmerken van leerlingen en het onderwijsleerproces.
- De effecten van invoering van de centrale eindtoets te monitoren en evalueren, niet alleen voor wat het betekent voor de leerlingen, de scholen en hun prestaties, maar ook voor het al dan niet optreden van ongewenst strategisch gedrag.