Vergelijking van resultaten van Nederlandse leerlingen op de schriftelijke TIMSS-toets en de praktische TIMSS-toets in 1995 en 1999/2000

Pauline Vos
Rijksuniversiteit Groningen, Instituut voor Didactiek en Onderwijsvernieuwing, Groningen

Wilmad Kuiper
Universiteit Twente, Faculteit Gedragswetenschappen, Afdeling Curriculum, Enschede

Samenvatting
TIMSS, een internationaal vergelijkend onderzoek naar onderwijs in de exacte vakken, omvatte naast een schriftelijke toets een alternatieve toets waarin leerlingen uit leerjaar 2 van het voortgezet onderwijs in een practicumomgeving werden gebracht en werden getoetst op praktische vaardigheden als het doen van kleine wiskundige en natuurwetenschappelijke onderzoekjes. Volgens curriculumexperts sloot de toets goed aan bij de toepassingsgerichtheid van de basisvorming. In 1995 presteerden de Nederlandse leerlingen echter minder vaardig op deze toets dan verwacht. Zij kwamen in de internationale vergelijking nauwelijks uit boven het niveau van het internationale gemiddelde. Dit was opmerkelijk, omdat zij op de schriftelijke TIMSS-toets wél ruimschoots boven het internationale gemiddelde uitkwamen. De praktische TIMSS-toets werd in het voorjaar van 2000 herhaald. In dit artikel wordt hiervan verslag gedaan. Het blijkt dat de Nederlandse leerlingen in 2000 in wiskunde niet *en in science* wel *vooruit zijn gegaan qua toepassingsgerichte vaardigheden. Daarnaast blijken juist wiskundeleraren een positievere houding aan te nemen tegenover alternatieve, praktische toetsvormen. Ook wordt ingegaan op specifieke methodologische problemen bij het repliceren van de afname van praktische toetsen.*

1. Inleiding

De Third International Mathematics and Science Study (TIMSS) is een internationaal vergelijkend onderzoek naar het onderwijs in de exacte vakken. Het betreft hier de gebieden wiskunde en *science* (natuur/scheikunde, biologie en fysische aardrijkskunde). De studie wordt uitgevoerd onder auspiciën van de *International Association for the Evaluation of Educational Achievement* (IEA). TIMSS is uitgevoerd in 1995, 1999 en 2003. Dit artikel gaat over de afname in 1995 (TIMSS-95) en in 1999 (TIMSS-99) voor wat betreft leerjaar 1 en 2 van het voortgezet onderwijs (*Populatie 2*).

Via vragenlijsten werden gegevens verzameld over achtergrondkenmerken van leerlingen, leraren en scholen. Tevens werden met behulp van een schriftelijke toets prestaties van leerlingen gemeten. Aan TIMSS-95 Populatie 2 hebben 41 landen deelgenomen (Beaton, Martin e.a., 1996; Beaton, Mullis e.a., 1996). In 1999 is TIMSS voor het hoogste leerjaar van Populatie 2 herhaald, dus betreffende de leerlingen in de tweede klas van de basisvorming.

Aan TIMSS-99 hebben 38 landen deelgenomen. 26 landen, waaronder Nederland, participeerden zowel in TIMSS-95 als in TIMSS-99 (Martin, Mullis, e.a., 2000; Mullis, Martin, e.a., 2000). Het Nederlandse aandeel in beide studies is uitgevoerd door het OCTO van de Universiteit Twente.

Zowel in TIMSS-95 als in TIMSS-99 waren de scores van de Nederlandse populatie 2 leerlingen op de schriftelijke toets in de internationale vergelijking goed te noemen. Nederlandse leerlingen behoorden wat betreft wiskunde tot de *subtop*. Alleen de leerlingen in enkele Aziatische landen hadden een hogere score. Voor science was de score in de internationale vergelijking nog iets beter dan voor wiskunde. Slechts één land (Singapore) had in TIMSS-95 een betere score dan Nederland. De scores van de Nederlandse tweede klas leerlingen waren tussen 1995 en 1999 ook constant. In TIMSS-99 had zelfs geen enkel land een significant betere score dan Nederland (Singapore vertoonde namelijk een lichte achteruitgang ten opzichte van 1995). Nederland behoorde wat betreft science in 1999 met nog 16 andere landen tot de beste landen (Bos & Vos, 2000).

Voor beide vakgebieden lagen de scores van de Nederlandse leerlingen ruim boven de gemiddelde score van alle deelnemende landen (internationale gemiddelde). Dit was opmerkelijk omdat curriculumexperts oordeelden dat de internationale toets slechts in beperkte mate aansloot op het *beoogde curriculum* (de kerndoelen voor de basisvorming, c.q. de meest gebruikte methoden). Wiskunde-experts gaven aan dat slechts 71% van de wiskundige toetsitems paste bij het beoogde curriculum. Van de gecombineerde science-items uit de toets paste volgens de geconsulteerde curriculumexperts slechts 69% bij het beoogd curriculum. Enkele getoetste onderwerpen zoals algebra en scheikunde waren nog nauwelijks in het lesprogramma aan bod geweest. De wiskunde-experts hadden bezwaren tegen het toepassingsarme karakter van de sommen, terwijl de science-experts bezwaren hadden tegen de relatief grote hoeveelheid reproductieopgaven. Bovendien werd, met 75% multiple choice items, de vorm van de testitems minder geschikt geacht voor het toetsen van vaardigheden waaraan in de basisvorming belang wordt gehecht (Bos & Vos, 2000; Kuiper, Bos & Plomp, 1997, 1999).

In 1995 is, in aanvulling op de schriftelijke TIMSS-toets, in leerjaar 2 ook een praktische toets afgenomen. Dat is gebeurd in 20 landen, waaronder Nederland. Belangrijk argument voor het complementaire karakter van deze toets was dat met een schriftelijke toets vooral kennis en in mindere mate vaardigheden onderzocht kunnen worden. De TIMSS Praktische Vaardigheidstoets (PVT-95) is een toets bestaande uit vijf wiskundetaken (M-taken), vijf sciencetaken (S-taken) en twee gecombineerde wiskunde- en sciencetaken (G-taken) in een practicumomgeving. Leerlingen worden voorzien van manipulatieve materialen (klei, magneten, vouwblaadjes, plakband, enz) en eenvoudige meetapparatuur (thermometer, liniaal, weegschaal, enz). Zij worden getoetst met opdrachten als: het opzetten en uitvoeren van experimenten, het doen en beschrijven van waarnemingen, het rekenen met een zakrekenmachine, het zoeken naar regelmaat, en het noteren en interpreteren van meetgegevens (zie Appendix). In 1995 oordeelden drie wiskunde-experts (over de vijf M-taken en de twee G-taken) en even zovele natuur/scheikunde-experts (over de vijf S-taken en wederom de twee G-taken) dat de internationale praktische toets, gegeven de gerichtheid op het meten van toepassings-

Tabel 1. Oordeel Nederlandse curriculumexperts en resultaten van Nederlandse leerlingen op complementaire toetsen.

	Oordeel curriculumexperts*	Resultaat leerlingen**
TIMSS-95 (schriftelijk)	past matig	prestatie goed
PVT-95 (praktisch)	past goed	prestatie middelmatig

*) in het licht van de kerndoelen basisvorming.
**) in vergelijking tot leerlingen uit andere deelnemende landen.

gerichte vaardigheden, goed aansloot bij de kerndoelen van de basisvorming voor wiskunde en natuur/scheikunde (Kuiper, Bos & Plomp, 1997).
De scores van de Nederlandse leerlingen op de PVT-95, zowel voor science als voor wiskunde (inclusief de gecombineerde G-taken), onderscheidden zich echter nauwelijks van het internationale gemiddelde. Dit contrasteerde opvallend genoeg zowel met de prestaties op de schriftelijke toets (waarop wél een hoge internationale *ranking* werd bereikt) als met gegevens over de grote mate van geschiktheid van de toets in het licht van het beoogde curriculum voor de basisvorming. Een en ander staat weergegeven in Tabel 1. Het eerste contrast (verschil leerlingscores schriftelijke toets - praktische toets) is verticaal weergegeven in de laatste kolom. Het tweede contrast (verschil expertoordeel en leerlingscores schriftelijke toets) is horizontaal weergegeven in de onderste rij.

De vraag rees of deze contrasten konden worden toegeschreven aan de aandacht die in de basisvorming tot dan toe was besteed aan toepassingsgerichte vaardigheden in de exacte vakken. Wellicht volgde het onderzoeksmoment (1995) te kort op de formele invoering van de basisvorming (1993) en bleef de uitvoering van het nieuwe leerplan nog achter. Deze veronderstelling werd ook geschraagd door het lerarenoordeel over de PVT-95. In het kader van onderzoek naar het *uitgevoerd curriculum* hadden leraren wiskunde en natuur/scheikunde van de getoetste klassen de taken uit de toets voorgelegd gekregen. Zij oordeelden over het algemeen afwijzend op deze taken, in tegenstelling tot de curriculumexperts, hetgeen erop duidde dat niet alle getoetste inhouden en vaardigheden ook in de les aan de orde waren geweest (Bos & Kuiper, 1998).

In het kader van een tussentijdse evaluatie van de invoering van basisvorming werd het een maatschappelijk en wetenschappelijk relevante vraag geacht of herhaalde afname van de PVT zou resulteren in prestaties vergelijkbaar met die uit 1995. Daarnaast lag de vraag voor de hand of vakexperts en docenten tot een soortgelijk oordeel zouden komen over de geschiktheid van de praktische toets als in 1995 het geval was. Aldus werd besloten tot herhaalde afname van de PVT teneinde trendgegevens te verschaffen over praktische vaardigheden van leerlingen in de exacte vakken binnen de basisvorming. Tevens zouden de PVT-taken aan experts en leraren voorgelegd worden om inzicht te geven in trends in hun oordeel en in de samenhang hiervan met de verschillen in prestaties op de schriftelijke respectievelijk de praktische toets. Aldus zouden data worden vergaard op het niveau van het beoogd curriculum (de experts), het uitgevoerd curriculum (de leraren) en het bereikte curriculum (de leerlingen).
De herhaling van de praktische toets vond plaats in het voorjaar van 2000 (NWO-PROO project 411-203-04), vijf jaar na de eerste afname. In navolging

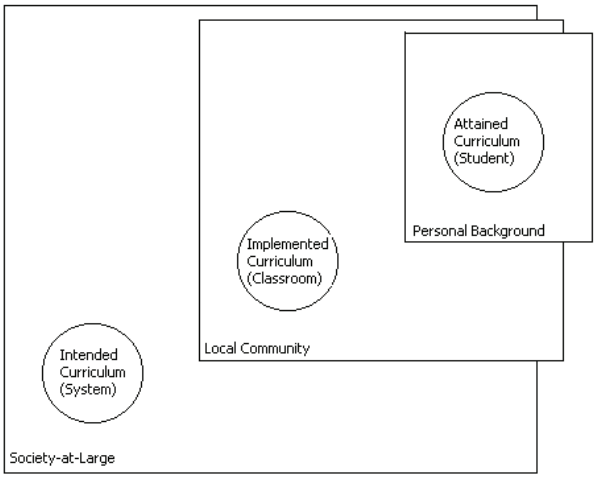
van TIMSS-95 en TIMSS-99 als aanduiding voor de schriftelijke toetsing, spreken we van PVT-95 en PVT-2000 voor de praktische toetsing. De volgende onderzoeksvragen stonden centraal.

- In hoeverre resulteert herhaalde afname van de praktische toets in leerjaar 2 van het voortgezet onderwijs in dezelfde prestaties? (vergelijking PVT-95 met PVT-2000)
- In hoeverre resulteert afname van de PVT-2000 in combinatie met de herhaalde afname van de schriftelijke toets TIMSS-99 in dezelfde verschillen in prestaties op PVT-95 en TIMSS-95? Ofwel: is er een trend in het verticale contrast uit tabel 1?
- In hoeverre zijn eventuele verschillen in prestaties op de PVT-2000 terug te voeren op de geschiktheid van deze toets in het licht van de kerndoelen voor de exacte vakken in de basisvorming (beoogd curriculum) en het feitelijk gegeven onderwijs in die vakken (uitgevoerd curriculum)?

In Vos (2002) zijn bovenstaande vragen ook onderzocht, maar dan alleen met betrekking tot de wiskundige delen van de toetsing. In het voorliggende artikel zal echter gerapporteerd worden over zowel wiskunde als science.

2. Conceptueel kader

Het doel van TIMSS is om middels internationale vergelijking een kennisbasis te leveren voor het onderwijs in de exacte vakken ten behoeve van beleidsmakers, curriculumonderzoekers en -specialisten. Het Nederlandse aandeel in TIMSS wordt uitgevoerd tegen de achtergrond van een conceptueel kader waarin drie verschijningsvormen van curricula voor de exacte vakken worden onderscheiden (cf. Kuiper, Bos & Plomp, 1997; Robitaille & Maxwell, 1996; Robitalle e.a., 1993): het beoogde, uitgevoerde en gerealiseerde curriculum, als geïllustreerd in figuur 1. Het curriculum wordt opgevat als een verzameling wiskunde- en science-inhouden in termen van begrippen, procesvaardigheden en houdingen. Het beoogde curriculum verwijst naar kerndoelen, examenprogramma en methoden. Het uitgevoerde curriculum verwijst naar het feitelijke



Figuur 1. IEA's curriculumniveaus (Robitaille, 1993).

onderwijsaanbod op school- en vooral klassenniveau. Het gerealiseerde curriculum verwijst naar de opbrengst van het onderwijs in de exacte vakken in termen van verworven kennis, vaardigheden en houdingen.
Met de schriftelijke en de praktische TIMSS-toetsen worden *kennis en vaardigheden* getoetst. De PVT is binnen TIMSS ontwikkeld vanuit een visie waarin natuurwetenschappelijk onderwijs zich baseert op een samenhang tussen processen en inhoud (procedurele en declaratieve kennis). Centraal staan open opdrachten om verklaringen voor onderzochte verschijnselen te geven. Een practicum is daarmee niet langer een ondersteunende illustratie van concepten (als *instap* vóór of als *demonstratie* ter rechtvaardiging van de theorie), maar een zelfstandige, toetsbare onderwijsactiviteit. Van de leerlingen wordt planmatig handelen gevraagd, in tegenstelling tot de "kookboek"-practica waarin aan de leerlingen stap-voor-stap handelingen worden voorgeschreven (Kind, 1999). Ervaringen met praktische tests zijn van recente datum (Doran & Tamir, 1992; Linn & Burton, 1994; Martin, Mullis, e.a., 2000; Shavelson, Baxter & Xiaohong, 1993) en de betrouwbaarheid in een internationaal vergelijkende context is vooralsnog enigszins problematisch (Shavelson, Baxter & Xiaohong, 1993; Zuzovsky 1999).

Internationale onderwijsinnovaties waarin leerlingen een plaats krijgen als betrokkenen die hun kennis construeren vanuit contextrijke ervaringen, hebben in TIMSS geleid tot de PVT en in Nederland tot de basisvorming. In de laatste staan de begrippen *toepassing*, *vaardigheid* en *samenhang* (TVS) centraal (Van Luyn, 1998). Met deze trefwoorden werd in 1993 een verplicht leerplan voor alle leerlingen in de onderbouw van het Nederlands voortgezet onderwijs ingevoerd. Naast invoering van de nieuwe vakken techniek en verzorging, werden leerplannen van bestaande vakken veranderd. Daarnaast kwam scheikunde (door samenvoeging met natuurkunde) voor het eerst op de lessentabel voor de tweede klassen. Uit het oude leerplan wiskunde werd een aantal onderwerpen vervangen door andere, met als leidraad dat het geleerde niet alleen in het latere leven zinvol moest zijn, maar ook op het moment van leren (Kok, Meeder, Wijers & Van Dormolen, 1992). Daarmee werd aangesloten op de wens van *vermaatschappelijking* van de leerinhouden, waarmee deze sterker dan voorheen uitgaan van levensechte, voor de leerlingen herkenbare situaties. Daarnaast streefde men tevens naar een *verwetenschappelijking* van het onderwijs waarin een belangrijke plaats is ingeruimd voor strategisch leren denken, problemen leren oplossen, en het aanleren van fundamenten van de natuurwetenschappelijke methode. De basisvorming moest naast vernieuwde lesinhouden samengaan met een *didactiek van authentiek leren*, die echter in de eerste jaren na de invoering nog slechts mondjesmaat uit de verf kwam (Doolaard, Cremers-van Wees & Bosker, 1999; Inspectie van het Onderwijs, 1999; Roelofs, Franssen, Houtveen & Lagerwij, 1999). Met name het aanbieden van activerend, onderzoeksgericht werk aan de leerlingen schoot tekort. Daarmee sluit de PVT niet alleen goed aan bij het *beoogde leerplan* van de basisvorming, maar voorziet het tevens als exemplarische onderwijsactiviteit in een leemte van het *uitgevoerde curriculum*, geschikt voor gebruik naast de bestaande methodes.

Tabel 2. Curriculumniveaus en bijbehorende operationalisaties.

Curriculum-niveau	Betrokkene	Operationalisatie
Beoogd	curriculumexpert	oordeel of taken passen bij kerndoelen
Uitgevoerd	vakleraar	oordeel of taken passen bij lesstof van de toetsklas
Gerealiseerd	leerling	scores op taken

3. Opzet van het onderzoek

Instrumentarium, steekproefopzet en methoden van dataverzameling en -verwerking volgden met het oog op vergelijkbaarheid alle procedures voor de dataverzameling uit 1995. Gegeven de eerdere afname werd voor de PVT-2000 afgezien van een *field trial*. In vergelijking met 1995 kampte het onderzoeksteam echter met een gebrek aan assistenten (studenten aan de universitaire lerarenopleiding in de exacte vakken). Dit leidde tot onverwachte capaciteitsproblemen in de personele bezetting en tot enkele aangepaste procedures in de dataverwerking. Waar van toepassing, zullen deze hieronder toegelicht worden.

Net als in 1995 zijn in 2000 experts en leraren bevraagd op geschiktheid van de PVT in het licht van het beoogde respectievelijk uitgevoerde curriculum. Met deze consultatie werd de operationalisatie van het conceptueel kader gecompleteerd (zie tabel 2).

Het onderzoek is uitgevoerd binnen een random selectie van 50 van de in totaal 126 scholen die in het voorjaar van 1999 hadden deelgenomen aan de schriftelijke TIMSS-99-toets (Bos & Vos, 2000). Per deelnemende school is aselect één tweede klas geselecteerd waarin de schriftelijke toets is afgenomen. Het was niet mogelijk de praktische toets af te nemen bij dezelfde leerlingen die in het voorjaar van 1999 deelgenomen hadden aan de schriftelijke toets, simpelweg omdat die leerlingen al in leerjaar 3 zaten. In plaats daarvan werd voor elk van de 50 scholen één, naar schooltype identieke, tweede klas geselecteerd. De toets is afgenomen op 27 scholen (een respons van 54%). De steekproefomvang is kleiner dan in 1995, toen 49 scholen deelnamen (een respons van 98%). In 1995 is er echter tot driemaal toe bij non-respons een vergelijkbare vervangende school aangezocht (de respons vóór vervanging was toen 36%). De vervangingsmethode is in 2000 niet uitgevoerd.

In tabel 3 is de verdeling van schooltypes van de klassen in de steekproef

Tabel 3. Schooltype van scholen in de steekproef (procentueel).

Schooltypecombinaties	Percentage in 1995 (n=49)	Percentage in 2000 (n=27)
Ivbo	-	4
Vbo	18	22
Vbo/mavo	14	15
Mavo	8	11
Mavo/havo/vwo	35	-
Vbo/mavo/havo/vwo	21	-
Mavo/havo	-	19
Havo	-	4
Havo/vwo	-	11
Vwo	4	15
Totaal	100	100

van 1995 respectievelijk 2000 weergegeven. Duidelijk is dat de breedte van dakpanklassen is afgenomen. Was in 1995 nog ongeveer de helft van de deelnemende klassen een combinatie van drie of meer schooltypes (mavo/havo/vwo of vbo/mavo/havo/vwo), in 2000 zijn alleen klassen met één of twee schooltypes aanwezig. Dit weerspiegelt de ontwikkeling van vervroeging van schooltypekeuze van leerlingen in de basisvorming, zoals ook al door anderen is gerapporteerd (Doolaard, Cremers-van Wees & Bosker, 1999; Inspectie voor het Onderwijs, 1999).

Per geselecteerde tweede klas namen negen aselect gekozen leerlingen, namelijk de eerste negen van de alfabetische klassenlijst, deel aan de toets. Eén school hield zich niet aan deze randomisering en is uit de analyse verwijderd. Aan de PVT-2000 hebben derhalve 234 (=26x9) leerlingen deelgenomen.

In tabel 4a/b staat de verdeling van de getoetste leerlingen naar schooltype en sekse voor de toetsing in 1995 en 2000. De sekseverdeling is evenredig. Voor de indeling in schooltypen zijn, net als bij de rapportage van de schriftelijke toetsing (Bos & Vos, 2000), twee dichotome categorieën vbo/mavo en havo/vwo gemaakt. Uit de tabel blijkt dat er meer vbo/mavo-leerlingen dan havo/vwo-leerlingen zijn getoetst. Tevens blijkt dat in de dataset van 2000 in havo/vwo-klassen een lichte oververtegenwoordiging van meisjes is en in de vbo/mavo-klassen een oververtegenwoordiging van jongens. De oorzaak hiervan kan aan toeval worden toegeschreven omdat beide verdelingen ruim binnen de 95% betrouwbaarheidsmarges van een 50-50-verdeling liggen.

Tegelijk met de toets is een vragenlijst aan de leraren wiskunde en natuur/scheikunde van de betreffende toetsklassen afgenomen. Het responspercentage hierop (zie tabel 5) was voor de wiskundeleraren in 2000 ruimschoots hoger dan in 1995, terwijl deze voor de natuur/scheikundeleraren nagenoeg gelijk bleef. Daarnaast werden vijf curriculumexperts (drie voor wiskunde, twee voor natuur/scheikunde) geraadpleegd. De betrokken personen waren werk-

Tabel 4a. Verdeling getoetste leerlingen naar schooltypecombinatie en sekse in PVT-95 (Kuiper e.a. 1997).

	vbo/mavo	havo/vwo	Totaal (kolom-%)
Jongens	115	84	199 (46%)
Meisjes	144	94	238 (54%)
Totaal (rij-%)	259 (59%)	178 (41%)	437 (100%)

Tabel 4b. Verdeling getoetste leerlingen naar schooltypecombinatie en sekse in PVT-2000.

	vbo/mavo	havo/vwo	Totaal (kolom-%)
Jongens	71	46	117 (50%)
Meisjes	54	62	116 (50%)
Onbekend	1	-	1 (0%)
Totaal (rij-%)	126 (54%)	108 (46%)	234 (100%)

Tabel 5. Deelnemende leraren in PVT-95 en PVT-2000.

	1995		1999	
	Aantal	Respons	Aantal	Respons
Natuur/scheikunde	26	53 %	15	56 %
Wiskunde	20	41 %	20	74 %

zaam bij de landelijke Pedagogische Centra, het Freudenthal Instituut, lerarenopleidingen en het CITO.

Het onderzoeksinstrumentarium bestond allereerst uit de TIMSS praktische vaardigheidstoets. Deze toets bestaat uit twaalf taken: vijf wiskundetaken (M1 t/m M5), vijf science-taken (S1 t/m S5) en twee gecombineerde, vakoverschrijdende taken (G1 en G2). De taken worden in de vorm van een stationspracticum getoetst. Zes taken vergen een toetstijd van 30 minuten en zijn elk ondergebracht in één station. De overige zes taken vergen 15 minuten en zijn in tweetallen samengebracht tot één station. Aldus zijn er in totaal negen stations (voor negen leerlingen). In de appendix zijn korte omschrijvingen van de twaalf taken gegeven. De toetsafname duurt 90 minuten, waarin elke geselecteerde leerling individueel drie stations aandoet. Aan elk station wordt één taak gedurende 30 minuten of twee taken á 15 minuten uitgevoerd. Aldus volbrengt elke leerling drie à vijf van de twaalf taken. Om te voorkomen dat zich tijdens de toetsafname taakinteractie-effecten voordoen (waarbij leerlingen bij de ene taak ervaringen opdoen die hen kunnen helpen bij een volgende taak) zijn rotatieschema's opgesteld waardoor de leerlingen geheel eigen trajecten langs de stations afleggen (zie Harmon e.a., 1997).

Bij elke taak krijgen de leerlingen een werkblad met instructies en opdrachten, die uit het Engels zijn vertaald. Voor de beantwoording is op de werkbladen veel lege ruimte uitgespaard (Kuiper, Bos & Plomp, 1997). De (deel)opdrachten bestaan uit korte antwoord vragen (*reken met de rekenmachine de volgende vermenigvuldigingen uit:* 34x34, 334x334, 3334x3334), onderzoeksvragen die een uitgebreider antwoord vragen (*beschrijf wat je allemaal hebt gedaan om uit te zoeken welke magneet sterker is*), kennisvragen (*waarom denk je dat je hartslag op deze manier veranderde?)* en procedurele vragen (*als je de proef anders hebt moeten uitvoeren dan je van plan was, beschrijf dan wat je hebt veranderd*). In de (deel)opdrachten is het dilemma verwerkt dat zowel leerlingen met ruime alsook leerlingen met geringe zelfstandige onderzoeksvaardigheden aangestuurd moeten worden (Garden, 1999).

De toets is in 2000 afgenomen door een getrainde toetsleider, die beschikte over een mobiele set practicummaterialen. Op elke school werden daarmee dezelfde toetsomstandigheden gecreëerd. De toetsleider was tevens belast met de codering van de antwoorden volgens de internationale richtlijnen.

Het onderzoeksinstrumentarium bestond verder uit een vragenlijst voor leraren en een vragenlijst voor vakexperts. Beide vragenlijsten bevatten zeven van de twaalf toetstaken waarover een geschiktheidoordeel werd gevraagd. De vragenlijst voor de wiskunde-experts en -leraren bevatte de vijf M-taken en de twee G-taken; de vragenlijst voor de natuur/scheikunde-experts en -leraren bevatte de vijf S-taken en wederom de twee G-taken. Aan de vragenlijst voor de wiskunde-experts waren bovendien nog twee S-taken (Batterijen en Elastiekje) toegevoegd, om te meten of deze taken ook zouden passen binnen

toepassingsgericht wiskundeonderwijs (met de respectieve onderwerpen combinatoriek en grafische extrapolatie).
Aan de leraren werd gevraagd te beoordelen of zij de stof uit de taken onderwezen hadden, c.q. of zij de taken geschikt achtten voor opname in een toets over de tot dan toe behandelde stof c.q. vaardigheden (uitgevoerd curriculum, *opportunity to learn*). Laatstgenoemde maat is gebaseerd op onderzoek van De Haan (1992). De vraag aan de leraren wiskunde en natuur/scheikunde bij elk taakonderdeel was tweevoudig en luidde letterlijk:

1. Is de inhoud/vaardigheid onderwezen aan de toetsklas?
2. Er vanuit gaande, onafhankelijk van uw antwoord op vraag 1, dat u een praktische vaardigheidstoets zou moeten afnemen, zou u dit taakonderdeel opnemen?

De beantwoording is een indicatie voor het gegeven onderwijs, en of de leerlingen de gelegenheid hebben gekregen de betreffende kennis en vaardigheden aan te leren. Het tweede deel in de vraagstelling is een indicatie in hoeverre de leraar opname van de betreffende vraag mogelijk acht, als de leerlingen voldoende aansluitende *bagage* hebben meegekregen.

Aan de experts is gevraagd of de taken passen bij het beoogd curriculum, als omschreven in de kerndoelen voor de basisvorming en de daarop gebaseerde meest gebruikte methoden.

4. Dataverwerking en betrouwbaarheid

In TIMSS worden open antwoorden uit de leerlingentoetsen gecodeerd met een dubbelcijferige code. Het eerste cijfer van de code wordt gebruikt voor de mate van correctheid (3, 2, 1 of 0 punten). De codes 20, 21, 22 en 23 leveren dus alle dezelfde correctheidsscore van 2 punten. Het tweede cijfer in de code is een aanduiding van de gangbare fouten en/of de gebruikte oplossingsstrategie (de diagnostische score). Codeurs werden voorafgaand aan de toetsafname getraind.

Volgens de TIMSS-procedures dient een vooraf vastgesteld percentage van het leerlingwerk door twee afzonderlijke codeurs te worden gecodeerd. In 1995 werd 10% van de leerlingantwoorden dubbel gecodeerd. De overeenstemming van de codes op de antwoorden van de toets in 1995 werd gecheckt op de correctheidscode (eerste cijfer uit de code gelijk) en op de diagnostische code (beide cijfers van de code gelijk). Een percentage van 100% betekent totale overeenstemming tussen de twee codeurs in het coderen van de leerlingantwoorden.

In de internationale vergelijking bleek Nederland een lage intercodeursovereenstemming op de leerlingantwoorden op de PVT-95 te hebben. Lag het internationaal gemiddelde op een overeenstemming over de gehele toets op 91%, in Nederland lag dit op 82% (Harmon e.a., 1997). In het uitzonderlijke geval van één item (een onderdeel uit taak S4 Elastiekje) gaven de twee Nederlandse codeurs slechts op 52% van de leerlingantwoorden een overeenkomstige correctheidscode. Hieruit blijkt dat in 1995 de betrouwbaarheid van de score op sommige onderdelen van taken enigszins twijfelachtig was. In TIMSS zijn echter geen drempels aangebracht voor een vereiste minimale codeursovereenstemming.

Bij de PVT-2000 was slechts één codeur beschikbaar. Het betrof een ervaren leraar wiskunde (o.a. diverse examenklassen), die al eerder door het internationale TIMSS Study Center getraind was in het coderen van de schrif-

Tabel 6. Betrouwbaarheid (Cronbach alpha) en vergelijkbaarheid (χ^2-toets) van leerlingscores per taak.

	Items in de taak	Alpha in 1995	Alpha in 2000	χ^2-toets op scores van 1995-2000
M1 Dobbelsteen	6	0.50	0.64	77
M2 Rekenmachine	7	0.71	0.68	99
M3 Vouwen	4	0.83	0.76	53
M4 Bocht	8	0.59	0.62	100
M5 Inpakken	3	0.61	0.65	28
S1 Hartslag	4	0.61	0.59	33
S2 Magneten	2	0.65	0.36	0
S3 Batterijen	4	0.54	0.68	18
S4 Elastiekje	7	0.58	0.39	0
S5 Oplossingen	7	0.63	0.63	57
G1 Schaduwen	6	0.64	0.61	1
G2 Klei	8	0.85	0.78	0

telijke toets (TIMSS-99). Deze codeur onderging een training van één dag in het coderen van de PVT-2000. Om een trend te kunnen meten, moesten de data zowel betrouwbaar als vergelijkbaar zijn. Cronbach's alpha berekening voor de betrouwbaarheid bracht een aantal problemen aan het licht voor de taken S2 (Magneten) en S4 (Elastiekje), met alpha's van 0,36 respectievelijk 0,39 (zie tabel 6). Een waarde groter dan 0,6 zou acceptabel zijn.

Daarnaast moesten de codes uit 1995 en 2000 vergelijkbaar zijn. Weliswaar waren de toetscondities zoveel mogelijk identiek gehouden, maar toch bleken er kleine afwijkingen in de gebruikte apparatuur. Bij taak S2 (Magneten) waren in 1995 twee totaal verschillende magneten gebruikt (niet alleen verschillend in kracht, maar ook in afmeting), terwijl in 2000 de magneten meer met elkaar overeenkwamen. Bij taak G1 (Schaduwen) was in tegenstelling tot 1995, in 2000 de zaklamp deels afgeplakt waardoor de schaduwranden scherper werden. Bij taak G2 (Klei) was in 1995 een fragiele, metalen weegschaal gebruikt, terwijl in 2000 een robuuste, handzame plastic weegschaal gehanteerd werd.

Om te onderzoeken of de verschillende toetsomstandigheden gevolgen hadden voor de scores, is onderzoek gedaan naar de verdelingen van de leerlingantwoorden in beide metingen. Indien de omstandigheden ongeveer gelijk zijn, de codeurs ongeveer gelijk handelen, en de prestaties van de leerlingen tussen 1995 en 2000 licht gaan verschuiven, dan kunnen frequenties van de antwoordcodes niet volledig verschillen. Met een χ^2-toets kan worden vergeleken of twee dataverzamelingen ongeveer dezelfde verdeling hebben. De χ^2-toets op de codes van de leerlingantwoorden uit 1995 en 2000 toonde voor vier taken duidelijk aan, dat de codes uit 1995 en 2000 wezenlijk verschilden, zoals blijkt uit de laatste kolom van tabel 6. Hier is een waarde groter dan 5 acceptabel. Onder de vier problematische taken waren de drie hierboven genoemde.

Ook de vergelijkbaarheid van de codes van taak S4 (Elastiekje) bleek onzeker, hoewel zich hier geen verschillen in de toetsomstandigheden hadden voorgedaan. Het codeerschema bleek echter te vaag en was in 1995 en 2000 verschillend geïnterpreteerd. Het probleem deed zich voor bij ongeveer 10% van de leerlingen die tijdens de toets hun meting van de uitrekking van het

elastiekje fingeerden. Zij maakten een grafiek waarin elk ringetje het elastiekje precies ½ cm oprekte. Dit gaf een lineaire, *kaarsrechte* grafiek. Een ware meting zou daarentegen schommelingen te zien geven, met stapjes variërend tussen 0,3 cm en 0,6 cm per toegevoegd ringetje. De grafiek vertoont in de praktijk afwijkingen van een rechte lijn. In het internationale codeerschema van taak S4 (Elastiekje) staat: maximaal 3 punten toekennen indien aan twee criteria is voldaan:

1. de leerling heeft tenminste vijf metingen genoteerd;
2. de data vertonen een toename.

In 1995 werd het hoogste puntenaantal toegekend omdat er vijf *meetpunten* in een grafiek waren uitgezet. In 2000 werd er op dit onderdeel veel strenger geoordeeld omdat er geen metingen waren verricht. Daarmee bleken de data voor deze taak niet vergelijkbaar.

Het bovenstaande overwegende is besloten vier taken niet in de trendmeting van de leerlingprestaties mee te nemen. Na verwijdering van de taken S2, S4, G1 en G2 bleven nog acht taken over. Voor de internationale vergelijking uit 1995 bleek deze ingreep slechts geringe veranderingen op te leveren.

In tabel 7 zijn de 19 landen gerangschikt, die in 1995 zowel aan de schriftelijke als aan de praktische toets deelnamen. Voor de TIMSS-95 score is een sommatie gemaakt van de wiskunde- en de science-score, zoals berekend door het TIMSS International Study Centre, op basis van IRT en *plausible values* (Beaton, Martin e.a., 1996; Beaton, Mullis e.a., 1996). Nederland staat bij de schriftelijke TIMSS-95-toets redelijk hoog genoteerd. Nederland maakt hier deel uit van een groep landen, die onderling niet significant in score verschillen, maar wel ruimschoots boven het internationale gemiddelde uitkomen. Deze *subtop* in de TIMSS-95 kolom van tabel 7 bestaat uit de landen van Nederland tot en met Zwitserland (plaatsen 3 t/m 6).

Tabel 7. Rangschikking en score van 19 landen in TIMSS-95 en PVT-95 (bij 12, resp. 8 taken).

TIMSS-95 wiskunde + science			PVT-95 (12 taken)			PVT-95 (8 taken)		
	Land	Score		land	score		Land	score
1	Singapore	1250	1	Singapore	71	1	Singapore	70
2	Tsjechië	1138	2	Engeland	67	2	Engeland	66
3	**Nederland**	**1101**	3	Zwitserland	65	3	Roemenië	64
4	Slovenië	1101	4	Australië	65	4	Australië	64
5	Australië	1075	5	Zweden	64	5	Zwitserland	63
6	Zwitserland	1067	6	Schotland	62	6	Zweden	60
7	Canada	1058	7	Noorwegen	62	7	Noorwegen	60
8	Engeland	1058	8	Roemenië	62	8	Schotland	60
9	Zweden	1054	9	Tsjechië	61	9	Slovenië	60
10	USA	1036	10	Slovenië	61	**10**	**Nederland**	**60**
11	Nw Zeeland	1033	11	Canada	60	11	Tsjechië	59
12	Noorwegen	1030	12	Nw Zeeland	60	12	Nw Zeeland	58
13	Schotland	1015	**13**	**Nederland**	**60**	13	Canada	58
14	Spanje	1004	14	USA	55	14	USA	53
15	Roemenië	968	15	Spanje	54	15	Spanje	52
16	Cyprus	937	16	Iran	52	16	Iran	50
17	Portugal	934	17	Portugal	47	17	Portugal	45
18	Iran	898	18	Cyprus	46	18	Cyprus	42
19	Colombia	796	19	Colombia	39	19	Colombia	34
	Intl.gemidd	1029		Intl. gemidd	59		Intl. gemidd	57

Voor wat betreft de PVT-95 berekend over de voltallige twaalf toetstaken is er een groep van acht landen op de plaatsen 6 t/m 13 waarvan de scores onderling weinig verschillen. Nederland maakt deel uit van deze *middenmootgroep*. De gemiddelde score van elk van dit achttal landen ligt net boven het internationaal gemiddelde. De *ranking* is hierbij vertekenend omdat de gemiddelde scores dicht bij elkaar liggen. Afhankelijk van de selectie van een deelverzameling van taken veranderen de gemiddelde scores en de onderlinge rangvolgorde van deze acht landen enigszins, evenals de onderlinge afstanden. Bij de scoreberekening van de PVT-95 over twaalf dan wel acht taken (met weglating van de vier taken S2, S4, G1 en G2) komt Nederland iets hoger in de rangvolgorde, maar zit nog steeds in dezelfde *middenmootgroep* van acht landen. Weglating van de vier taken voor de trendanalyse heeft dus geen gevolgen voor de stelling dat de gemiddelde score van de Nederlandse leerlingen op de wiskunde- en science-taken in vergelijking tot de gemiddelde score op de schriftelijke toets *middelmatig* is. Alleen indien de vijf wiskundetaken uit de toets wordt geselecteerd, zou Nederland in de *subtop* komen, net als op de schriftelijke toets (Vos, 2002). De Nederlandse leerlingen scoorden in de internationale vergelijking namelijk relatief minder op de science-taken, inclusief de gecombineerde taken, dan op de wiskundetaken.

In het vervolg van dit artikel zullen voor de trendvergelijking van de toetsresultaten dus slechts acht taken (vijf wiskundetaken, drie science-taken) beschouwd worden. Voor de analyse van andere dan trendvergelijkingen zijn wel alle twaalf taken geschikt, zoals bijvoorbeeld voor de sekseverschillen in 1995 en de sekseverschillen in 2000. Ook voor beschouwingen over het beoogd en het uitgevoerd curriculum zijn alle twaalf taken geschikt, aangezien men op deze curriculumniveaus niet gehinderd werd door gewijzigde apparatuur of codeursinterpretaties.

5. Resultaten

Eerste onderzoeksvraag

Analyse van de prestaties op de praktische TIMSS-toets in 1995 respectievelijk 2000 leert, dat er op de vijf wiskundetaken sprake is van een *nultrend*. Op de drie science-taken is er wel een significante verbetering, waardoor de gemiddelde score over *alle* acht taken ook een significant verschil maakt, bij een tweezijdige 95% betrouwbaarheidsinterval.

Op de PVT-95 scoorden de Nederlandse leerlingen middelmatig in de internationale vergelijking, met name op de science-taken, en de prestaties op de toets van 2000 laten in dit opzicht een significante verbetering zien. Met het resultaat van 2000 zouden de Nederlandse leerlingen in de *subtop* van de PVT-95 rangschikking komen (posities 3-5). Hierbij dient echter aangetekend te worden dat niet bekend is of leerlingen in andere landen zich in vijf jaar tijd ook op praktische vaardigheden hebben verbeterd, aangezien de replicatie van PVT-95 alleen in Nederland heeft plaatsgevonden.

De praktische vaardigheden van Nederlandse leerlingen lijken in de basisvorming dus te zijn toegenomen, zij het niet op wiskundig vlak. In tabel 8 staan de gemiddelde percentages correct per taak, uitgesplitst naar vbo/mavo en havo/vwo. Op de wiskundetaken zien we opvallend constante cijfers. Op de science-taken is over de gehele linie een vooruitgang te zien. Waar dit significant is, is dit gemarkeerd. De significant hogere totaalscore is enerzijds toe te

Tabel 8. Gemiddelde score percentage correct per practicumtaak, vergelijking 1995-2000, naar schooltypecombinatie

	Vbo/mavo		Havo/vwo		Totaal	
	1995 n=259	2000 n=126	1995 n=178	2000 n=108	1995 n=437	2000 n=234
M1 Dobbelstenen	73	67	83	82	77	74
M2 Rekenmachine	54	52	73	68	62	59
M3 Vouwen	67	72	80	83	73	77
M4 Bocht	63	66	74	72	68	69
M5 Inpakken	46	49	61	60	52	54
Subtotaal wiskunde	*61*	*62*	*74*	*73*	*66*	*67*
S1 Hartslag	37	44	57	59	45	51
S3 Batterijen	59	64	71	77	64	70
S5 Oplossingen	47	58*	58	63	51	60*
Subtotaal science (3 taken)	*48*	*56**	*62*	*68*	*53*	*62**
Totaal (8 taken)	**56**	**59**	**70**	**71**	**61**[1]	**64***

*) significant trendverschil

[1]) Het verschil met de score van Nederland, zoals in tabel 7 is weergegeven, is toe te schrijven aan afronding en aan de afwijkende Nederlandse populatie (met n=435) die voor de internationale vergelijking werd geselecteerd.

schrijven aan de vbo/mavo-leerlingen, anderzijds aan een betere prestatie van álle leerlingen op taak S5 (Oplossingen). Het betreft een scheikundige taak en het is mogelijk dat de vertraagde integratie van scheikunde in het lesprogramma van de tweede klassen maakt, dat de leerlingen met deze scheikundige taak in 2000 beter overweg konden dan in 1995.

Zoals te verwachten is, zijn de verschillen tussen vbo/mavo en havo/vwo aanzienlijk. Op alle taken, met uitzondering van M5 (Inpakken) en S5 (Oplossingen), is dit verschil significant. De verschillen tussen de schooltypecombinaties lijken in 2000 licht (maar niet significant) verminderd ten opzichte van de verschillen in 1995.

Omdat in de steekproef de meisjes in havo/vwo en jongens in vbo/mavo oververtegenwoordigd zijn, is geen vergelijking tussen jongens en meisjes opgenomen. Wel worden in tabel 9a/b de sekseverschillen per schooltypecombinatie apart gerapporteerd. Hieraan zijn toegevoegd de resultaten voor de vier taken, die niet geschikt zijn voor de trendvergelijking. Voor de sekseverschillen binnen beide metingen zijn deze echter wel bruikbaar omdat de seksen in 1995 en in 2000 onder dezelfde condities meededen.
Waar zich significante verschillen voordoen, is dit met een asterisk (*) gemarkeerd bij de hoogst scorende groep. Er zijn enkele opmerkelijke verschuivingen waarneembaar. Het significante sekseverschil op de totaalscore voor vbo/mavo uit 1995 is verdwenen. Voor zover zich in 1995 significante verschillen op afzonderlijke taken voordeden, waren deze telkens in het nadeel van de meisjes. In 2000 behoeft dit beeld bijstelling: op de taak waarop een significant verschil aanwezig is (M3 Vouwen), is dit in het voordeel van de meisjes. Ook op enkele andere taken tonen zij zich licht beter. Op het havo/vwo is het sekseverschil licht toegenomen, wederom in het voordeel van de meisjes, hoewel het verschil niet significant is.

Tweede onderzoeksvraag
Op de schriftelijke TIMSS-toets scoorden de Nederlandse leerlingen in zowel

Tabel 9a. Gemiddelde score correct in 1995 per practicumtaak, vergelijking naar sekse.

	Vbo/mavo			Havo/vwo		
	m n=144	j n=115	totaal n=259	m n=94	j n=84	totaal n=176
M1 Dobbelstenen	75	72	73	83	83	83
M2 Rekenmachine	56	51	54	80*	66	73
M3 Vouwen	64	71	67	85	74	80
M4 Bocht	61	66	63	74	75	74
M5 Inpakken	45	46	46	57	67	61
Subtotaal wiskunde	*60*	*61*	*61*	*73*	*74*	*74*
S1 Hartslag	31	43*	37	57	57	57
S2 Magneten	95	85	91	100	100	100
S3 Batterijen	52	67*	59	71	72	71
S4 Elastiekje	65	65	65	83	73	78
S5 Oplossingen	45	50	47	54	62*	58
Subtotaal science	*57*	*61*	*58*	*71*	*71*	*71*
G1 Schaduwen	26	31	28	52	54	53
G2 Klei	44	42	43	49	52	50
Subtotaal gecomb. taken	*35*	*36*	*36*	*50*	*54*	*52*
Totaal (alle 12 taken)	*53*	*56**	*54*	*68*	*68*	*68*

*) significant sekseverschil

Tabel 9b. Gemiddelde score correct in 2000 per practicumtaak, vergelijking naar sekse

	Vbo/mavo			Havo/vwo		
	m n=54	j n=71	totaal n=126	m n=62	j n=46	totaal n=108
M1 Dobbelstenen	72	63	67	87*	74	82
M2 Rekenmachine	53	50	52	71	61	68
M3 Vouwen	85*	67	72	85	81	83
M4 Bocht	63	69	66	75	69	72
M5 Inpakken	49	48	49	60	60	60
Subtotaal wiskunde	*63*	*61*	*62*	*75*	*71*	*73*
S1 Hartslag	44	44	44	62	56	59
S2 Magneten	79	74	76	79	100*	86
S3 Batterijen	61	66	64	80	73	77
S4 Elastiekje	61	62	62	70	68	70
S5 Oplossingen	61	56	58	64	62	63
Subtotaal science	*61*	*60*	*61*	*72*	*69*	*71*
G1 Schaduwen	38	39	39	55	63	59
G2 Klei	63	65	64	78	73	75
Subtotaal gecomb. taken	*49*	*52*	*51*	*65*	*65*	*65*
Totaal (alle 12 taken)	*59*	*59*	*59*	*72*	*69*	*71*

*) significant sekseverschil

Tabel 10. Trendvergelijking gemiddeld percentage correct, per vak en toetstype.

	Science		Wiskunde	
	TIMSS (schriftelijk)	PVT (3 taken)	TIMSS (schriftelijk)	PVT (5 taken)
1995	64	53	63	66
1999/2000	62	62*	65	67

*) significant trendverschil

1995 als 1999 internationaal gezien goed (Mullis, Martin e.a., 2000; Martin, Mullis e.a., 2000; Bos & Vos, 2000). In tabel 10 zijn de gemiddelde p-waarden over alle items van de twee vakgebieden in beide metingen weergegeven. Zowel bij wiskunde als bij science is geen significante verandering in de score op de schriftelijke toets te zien. Bos en Vos (2000) wijten de geringe afname in de sciencescore aan de vervanging van toetsitems uit 1995 waarbij een aantal abstractere science-items aan de toets van 1999 waren toegevoegd (op het gebied van *milieu* en *aard van de wetenschappen*).

Voor de praktische toets worden hier wederom alleen de acht taken beschouwd, die geschikt geacht werden voor de trendvergelijking. Wat betreft wiskunde is het beeld vergelijkbaar met de schriftelijke toets: er is een lichte, niet-significante toename op de gemiddelde score over de vijf wiskundetaken. De schriftelijke en de praktische toets geven daarmee hetzelfde beeld: een *nultrend*. De gemiddelde toename op de drie science-taken uit de praktische toets is echter opmerkelijk, en valt op in vergelijking tot de constante score op de science-items uit de schriftelijke toets. De leerlingen in leerjaar 2 hebben zeven jaar na de formele invoering van de basisvorming duidelijk een grotere vaardigheid verworven op het gebied dat door deze drie taken (S1 Hartslag, S3 Batterijen, S5 Oplossingen) wordt bestreken, terwijl de meer theoretische kennis op hetzelfde niveau is gebleven. De drie taken omvatten zeer verschillende kennisgebieden, namelijk respectievelijk biologie, natuurkunde en scheikunde. Ze toetsen verschillende manieren van onderzoeksaanpak, namelijk waarnemingen doen bij één veranderende variabele (S1 en S5) en combinatorisch probleemoplossen (S3). Hieruit kan worden afgeleid, dat de leerwinst in de basisvorming te vinden is op het gebied van het strategisch denken en het toepassen van fundamenten van de natuurwetenschappelijke methode.

Derde onderzoeksvraag

Hieronder worden de oordelen van de experts en de leraren weergegeven over de geschiktheid van de taken uit de praktische vaardigheidstoets in het licht van de kerndoelen respectievelijk het gegeven onderwijs.

De oordelen met betrekking tot het *beoogde curriculum* zijn in tabel 11 weergegeven als gemiddelden over zowel de experts als de onderdelen van een taak. Een cijfer 100 betekent, dat álle experts álle onderdelen uit een taak als geschikt beoordeelden. Een cijfer 50 kan betekenen dat alle experts eenstemmig de helft van de onderdelen uit een taak geschikt vonden. Het kan ook betekenen dat de helft van de experts álle onderdelen van een taak geschikt en de andere helft van de experts alle onderdelen óngeschikt vond.

De wiskunde-experts oordeelden in 2000, evenals in 1995 toen hetzelfde instrument werd gehanteerd, zéér positief tot redelijk positief over de wiskundetaken. Hoewel sommige onderdelen niet letterlijk tot de beoogde leerstof behoorden, pasten zij meestal wel naar de aard van het breed geformuleerde

Tabel 11. Oordeel vakexperts 2000 over geschiktheid van subonderdelen, per taak.

Taak (aantal onderdelen)	Natuur/scheikunde-experts	Wiskunde-experts
M1 Dobbelstenen (5)	-	100
M2 Rekenmachine (6)	-	75
M3 Vouwen (4)	-	63
M4 Bocht (6)	-	83
M5 Inpakken (3)	-	100
S1 Hartslag (3)	50	-
S2 Magneten (2)	100	-
S3 Batterijen (4)	88	13
S4 Elastiekje (6)	92	100
S5 Oplossingen (6)	67	-
G1 Schaduwen (6)	42	58
G2 Klei (3)	33	17

algemene kerndoel: *een wiskundige werkhouding ontwikkelen waarbij systematisch en methodisch werken, generaliseren, kritisch beoordelen van gegevens en uitkomsten alsmede het creatief bedenken van oplossingen aan de orde komen*. Ook de natuur/scheikunde-experts oordeelden zéér positief tot redelijk positief over vier van de vijf science-taken. De taak met de laagste beoordeling (S1 Hartslag) kreeg de aantekening dat deze eerder onder biologie valt.

Over de gecombineerde (vakoverschrijdende) taken werd het volgende geoordeeld. Taak G1 (Schaduwen) werd door de wiskunde- en natuur/scheikunde-experts als middelmatig, respectievelijk minder geschikt geacht. Taak G2 (Klei) leverde een behoorlijk afwijzend oordeel op van zowel de natuur/scheikunde- als de wiskunde-experts. Eén wiskunde-expert gaf echter aan, dat deze taak wel geschikt was als toetsing van niet-routinematig probleemoplossen.

Omdat twee science-taken een wiskundige component bevatten (taak S2 Batterijen bevatte combinatoriek, taak S4 Elastiekje bevatte grafiek tekenen en extrapoleren), werden deze als vakoverschrijdende taken in 2000 ook aan de wiskunde-experts voorgelegd. Zij oordeelden bijna unaniem, dat taak S2 níet en taak S4 wél tot het beoogd curriculum van hun vak behoorde. Hieruit blijkt de betrekkelijkheid van de door het International TIMSS Study Center in Boston gehanteerde definities over welke PVT-taken een typische science-, wiskunde- dan wel vakoverschrijdende taak zijn (Harmon e.a., 1997).

Kortom, negen van de twaalf taken kregen een positief oordeel in het licht van de kerndoelen voor wiskunde en natuur/scheikunde voor de basisvorming, waarbij een tiende taak (S1 Hartslag) mogelijkerwijs ook geschikt zou zijn voor biologie. Het oordeel van de experts in 2000 verschilde slechts in onderdelen van het oordeel uit 1995 en ondersteunt wederom de stelling dat de PVT goed past bij het beoogde curriculum in de basisvorming.

De vraag aan de leraren was tweeledig: enerzijds werd gevraagd naar het *onderwezen* zijn van de inhoud van de taken, en anderzijds werd gevraagd naar mogelijke *opname* van een taakonderdeel in een zelf gemaakte toets. In tabel 12a/b zijn de gemiddelde percentages van het antwoord “ja” per taakon-

Tabel 12a. Oordeel wiskundeleraren over taakonderdelen van zeven wiskundetaken (M1 t/m M5, G1 en G2) uit de PVT.

	Vbo/mavo		Havo/vwo	
	1995 n=12	2000 n=11	1995 n=7	2000 n=9
Onderwezen	32	54	41	58
In eigen toets opnemen	50	79	39	72

Tabel 12b. Oordeel natuur/scheikundeleraren over taakonderdelen van zeven science-taken (S1 t/m S5, G1, G2) uit de PVT.

	Vbo/mavo		Havo/vwo	
	1995 n=17	2000 n=8	1995 n=8	2000 n=7
Onderwezen	55	48	57	46
In eigen toets opnemen	66	63	74	81

derdeel op beide vragen weergegeven over álle taken die aan de leraren werden voorgelegd. Voor de wiskundeleraren zijn dit de vijf M-taken en de twee G-taken. Voor de natuur/scheikundeleraren zijn dit de vijf S-taken en de twee G-taken. De antwoorden zijn uitgesplitst naar schooltypecombinatie en onderzoeksjaar. Door de kleine aantallen leraren is voorzichtigheid geboden bij het trekken van conclusies.

Uit tabel 12a/b blijken verschillende zaken. Over het algemeen zijn de leraren die de moeite namen om de vragenlijst in te vullen, niet onverdeeld positief over de toets. Over de gehele linie zijn de onderwerpen door ongeveer de helft van deze leraren onderwezen. Kijken we naar de trend, dan is er bij de wiskundeleraren een positieve trend zichtbaar en bij de natuur/scheikundeleraren géén trend (een licht negatieve trend). Bij wiskunde zouden de inhouden dus in 2000 méér aan de orde zijn gekomen dan in 1995, en bij science in gelijke mate. Dit staat in contrast met de prestaties van de leerlingen. Immers, de prestaties op de wiskundetaken zijn niet gestegen, die op de science-taken juist wel.

Wat betreft opname van de taken in een zelf gemaakte toets is bij de wiskundeleraren wederom een positieve trend te zien en bij de natuur/scheikundeleraren geen trend. Het lijkt erop dat de bevraagde wiskundeleraren een positievere attitude hebben gekregen ten opzichte van de alternatieve toetsvorm (en dan vooral de vbo/mavo-leraren), terwijl de natuur/scheikundeleraren (en dan vooral de havo/vwo-leraren) in 1995 al redelijk positief stonden tegenover deze vorm van toetsing en weinig van mening zijn veranderd. Dat in 2000 de natuur/scheikundeleraren in vbo/mavo het minst positief staan tegenover de praktische toets geeft te denken. Juist leerlingen van dit schooltype hebben veel baat bij praktische ervaringen.

Om het oordeel van de leraren in verband te brengen met het oordeel van de vakexperts, zijn in tabel 13a/b dezelfde berekeningen nog eens uitgevoerd, maar nu separaat voor die taken die de curriculumexperts goed vonden passen bij het beoogde curriculum voor de basisvorming en voor de taken die niet goed passen bij het beoogde curriculum. Voor wiskunde zijn dit enerzijds de taken M1 t/m M5, anderzijds de taken G1 en G2. Voor science zijn dit enerzijds de taken S2 t/m S5 en anderzijds de biologische taak S1 (Hartslag) en de taken G1 en G2. In elke cel staan dus twee cijfers: één voor het leraaroor-

Tabel 13a. Oordeel wiskundeleraren over taakonderdelen die wel/niet bij beoogd curriculum passen (M1 t/m M5, resp G1 en G2) uit de PVT.

	Vbo/mavo		Havo/vwo	
	1995	2000	1995	2000
	n=12	n=11	n=7	n=9
Onderwezen	37/19	58/42	43/38	69/29
In eigen toets opnemen	53/43	80/76	47/24	78/59

Tabel 13b. Oordeel natuur/scheikundeleraren over taakonderdelen die wel/niet bij beoogd curriculum passen (S2 t/m S5, resp S1-G1-G2) uit de PVT.

	Vbo/mavo		Havo/vwo	
	1995	2000	1995	2000
	n=17	n=8	n=8	n=7
Onderwezen	61/48	51/43	58/56	48/44
In eigen toets opnemen	73/57	65/61	80/67	88/71

deel over de taken die volgend de curriculumexperts wél passen bij het beoogde curriculum, en één voor het oordeel over de taken die níet passen.
Wat opvalt, is dat in alle gevallen de oordelen over de taken die wél passen bij het beoogde curriculum hoger zijn dan de vergelijkbare getallen over de taken die niet geschikt worden bevonden. Het verschil tussen passend en niet-passend bij de kerndoelen is voor de leraren van belang in hun oordeel. De wiskundeleraren (zowel vbo/mavo als havo/vwo) lijken dit onderscheid op het gebied van *onderwezen* zijn van de taken duidelijker te maken dan de natuur/scheikundeleraren, zowel in 1995 als in 2000.

Voor wat betreft het opnemen van de taken in een toets lijken de havo/vwo-leraren (zowel wiskunde als natuur/scheikunde) het onderscheid tussen wel/niet passend bij de kerndoelen meer te maken dan de vbo/mavo-leraren. Een duidelijke trend in de leraaroordelen is niet goed waarneembaar, met uitzondering van de havo/vwo-wiskundeleraren die aangeven dat de taken in 2000 meer onderwezen zijn dan in 1995 en bij wie de bereidheid tot opname van de taken in een toets toegenomen lijkt te zijn.

De beantwoording van de derde onderzoeksvraag wordt enigszins gehinderd door de lage lerarenrespons. Met enig voorbehoud ontstaat het volgende beeld. Wat betreft wiskunde passen de vijf M-taken uit de PVT zowel in 1995 als in 2000 onveranderd goed bij het beoogde curriculum. Maar in 1995 geeft slechts ongeveer 1/3 van de leraren tegenover ruim de helft van de leraren in 2000 aan dat de inhouden en vaardigheden in hun lessen onderwezen zijn. Ook is er een toename te constateren in het lerarenoordeel om taken in een eigen toets op te nemen. Daarmee blijft het uitgevoerde curriculum echter nog steeds achter op het beoogde curriculum, en alleen bij de wiskundeleraren in havo/vwo lijkt de discrepantie in vijf jaar tijd minder groot te zijn geworden. De vertaling hiervan in verbeterde leerlingprestaties (het gerealiseerde curriculum) laat echter op zich wachten. Op de wiskundetaken zijn deze in vijf jaar tijd niet toegenomen.

Vier van de vijf S-taken passen goed bij het beoogde natuur/scheikunde-curriculum (de vijfde taak, S1, is een biologie-taak). De oordelen van de natuur/scheikundeleraren vertonen echter geen duidelijke trend. Zowel in 1995 als in 2000 geeft ongeveer de helft van alle leraren aan dat de inhouden onderwezen zijn. Een positief geschiktheidsoordeel over het mogelijk opnemen van de taken in een eigen toets wordt in beide metingen gegeven door onge-

veer 2/3 van de vbo/mavo-leraren en door een grote meerderheid (ongeveer 4/5) van de havo/vwo-leraren. Het lijkt er daarmee op dat de natuur/scheikundeleraren niet meer of minder zijn gaan aansluiten bij het beoogde curriculum; waarschijnlijk organiseerden zij in het verleden ook al practica. De leerlingen zijn echter wel beter gaan presteren op de science-taken. Maar het leggen van een verband tussen enerzijds de leerlingprestaties en anderzijds de leraaroordelen is moeilijk, aangezien van de vier science-taken die passen bij het beoogde natuur/scheikundecurriculum er twee (S2 Magneten en S4 Elastiekje) niet geschikt geacht werden om mee te nemen in de trendanalyse van de leerlingprestaties. Op de overige twee taken (S3 Batterijen en S5 Oplossingen) zijn de leerlingprestaties licht (niet significant) toegenomen.

6. Conclusies

De algemene internationale doelstelling van TIMSS is een vergelijking van verschillen tussen landen in de opbrengst van het onderwijs in de exacte vakken, een analyse van sterke en zwakke punten en van systeem-, school-, klas- en leerlingkenmerken die daarmee samenhangen, en het op nationaal niveau daar waar nodig formuleren van beleidsaanbevelingen voor de verbetering van het onderwijs in de exacte vakken (cf. Kuiper, Bos & Plomp, 1997).

In eerder onder IEA-vlag uitgevoerd internationaal vergelijkend onderzoek naar de kwaliteit en opbrengst van het onderwijs in de exacte vakken zijn prestaties van leerlingen alleen gemeten met behulp van schriftelijke toetsen. In TIMSS is met deze traditie gebroken door in leerjaar 2 van het voortgezet onderwijs een praktische vaardigheidstoets af te nemen in aanvulling op de internationale schriftelijke wiskunde- en sciencetoets. Deelname van Nederland aan deze *performance assessment* werd van groot belang geacht in het licht van de invoering van basisvorming en de tussentijdse evaluatie daarvan. Immers, door ook praktische vaardigheden mee te nemen in de opbrengstmeting werd meer recht gedaan aan het beoogde vernieuwde karakter van het onderwijs in de exacte vakken in de basisvorming.

De resultaten van de afname van de internationale praktische vaardigheidstoets in het voorjaar van 1995 en de replicatie ervan in 2000 leggen een (vooralsnog) relatief zwakke kant bloot van het onderwijs in de exacte vakken in de onderbouw van het voortgezet onderwijs. Immers, de praktische toets paste naar het inzicht van vakexperts weliswaar op hoofdlijnen redelijk tot goed bij de kerndoelen, maar het oordeel van leraren over de geschiktheid van de toets alsmede de prestaties van leerlingen op de toets vielen in 1995 minder gunstig uit. In 2000 blijken zowel de wiskundeleraren als de natuur/scheikundeleraren in hun oordelen over de toetstaken weinig te zijn opgeschoven. De inhoud van de taken is slechts bij ongeveer de helft van de leraren aan bod gekomen. De bereidheid om de taken in een toets op te nemen is alleen bij de wiskundeleraren toegenomen. Dit vertaalt zich vooralsnog niet in verbeterde prestaties op de wiskundetaken. Wel is er een verbeterde prestatie van de leerlingen op de drie science-taken geconstateerd. Hoewel er geen verband met de leraaroordelen lijkt te zijn, hebben leerlingen een grotere vaardigheid verkregen in strategisch denken en toepassen van fundamenten van de natuurwetenschappelijke methode. De geconstateerde sekseverschillen (ten nadele van de meisjes) uit 1995 zijn niet opnieuw waargenomen. Daarentegen blijken de meisjes op enkele taken significant beter te scoren dan de jongens.

Daarnaast blijkt uit dit onderzoek dat het gebruik van een praktische toets als trendmeting en in internationaal vergelijkend verband nog de nodige verbeteringen behoeft op het gebied van betrouwbaarheid en vergelijkbaarheid. Het is gebleken dat consistentie van toetsomstandigheden en codeursinterpretaties nauw luistert. Kleine afwijkingen in apparatuur tijdens de toets kunnen leiden tot bijvoorbeeld makkelijker meetbare onderzoekswaarnemingen, waardoor leerlingen meer tijd overhouden voor resterende deelopdrachten en systematisch hogere scores behalen. Dergelijke afwijkingen geven te denken voor de internationale vergelijkbaarheid, omdat kleine apparatuurverschillen ook tussen landen voorkomen. Hoewel het International TIMSS Study Centre duidelijk geprobeerd heeft met strakke richtlijnen de verschillen te minimaliseren, blijken er nog zeker punten voor verbetering vatbaar te zijn. Het voorliggende onderzoek wijst echter ook uit dat het *aantal* taken in de PVT voldoende groot is voor het enigszins tegen elkaar uitmiddelen van meetfouten. Bij weglating van een aantal minder betrouwbare taken uit de toets blijkt vooral de *ranking* te veranderen, maar de onderlinge verschillen tussen landen blijven ongeveer eenzelfde beeld te geven.

Correspondentie over dit artikel aan Pauline Vos, Rijksuniversiteit Groningen, Instituut voor Didactiek en Onderwijsvernieuwing, Nijenborgh 4, 9747 AG Groningen. Email: f.p.vos@fwn.rug.nl.

English summary

Comparison of Dutch student achievements on the TIMSS written test and the TIMSS performance assessment in 1995 and 1999/2000

TIMSS, an international comparative study on mathematics and science achievement, consisted of a written test and an optional practical test (performance assessment). In 1995 in the Netherlands, both tests were administered to students in grade Secondary 2. The practical test focused on the assessment of practical mathematics and science skills, like designing and conducting a small-scale experiment. Consulted curriculum experts considered the practical test appropriate to the application-oriented attainment targets for mathematics and the science subjects in lower secondary education. However, student performances on the practical test were poorer than expected, contrary to their achievement on the written test.

The practical test was repeated in spring 2000, the design and outcomes of which are reported in this article. From the repeat it appears that (i) there is improvement in achievement in practical science skills only, and (ii) mathematics teachers have developed a more positive attitude towards more practical modes of assessment.

Literatuur

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Kelly, D. L. & Smith T. A. (1996). *Mathematics achievement in the middle school years. IEA's Third International Mathematics and Science Study*. Boston: Boston College.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzales, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years. IEA's Third International Mathematics and Science Study.* Boston: Boston College.

Bos, K. Tj. & Kuiper, W. A. J. M. (1998). Praktische vaardigheden internationaal vergeleken. *NVOX, 23*, 366-372.

Bos, K. Tj., Kuiper, W. A. J. M. & Plomp, Tj. (1999). Student performance and curricular appropriateness in the Netherlands. *Studies in Educational Evaluation, 25*, 269-276.

Bos, K. Tj. & Vos, F. P. (2000). *Nederland in TIMSS-1999, exacte vakken in leerjaar 2 van het voortgezet onderwijs.* Enschede: Universiteit Twente.

Bos, K. Tj., Kuiper, W. A. J. M. & Plomp, Tj. (2001). TIMSS results of Dutch grade 8 students in international perspective: Performance assessment and written test. *Studies in Educational Evaluation, 27,* 79-94.

Doolaard, S., Cremers-van Wees, L. M. C. M. & Bosker, R. J. (1999). *Basisvorming in 1996; beschrijving en vergelijking met de periode voor invoering*. Enschede: Universiteit Twente.

Doran R. L. & Tamir, P. (Eds.) (1992). An international assessment of science practical skills. *Studies in Educational Evaluation, 18*, 1-102.

Garden, R. (1999). Development of TIMSS performance assessment tasks. *Studies in Educational Evaluation, 25*, 217-241.

Haan, D. M. de (1992). *Measuring Test Curriculum Overlap.* Enschede: Universiteit Twente.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzales, E. J. & Orpwood, G. (1997). *Performance assessment in IEA's Third International Mathematics and Science Study.* Boston: Boston College.

Inspectie van het Onderwijs (1999). *Werk aan de basis. Evaluatie van de basisvorming.* Den Haag: SDU.

Kind, P. M. (1999). Performance assessment in science, what are we measuring? *Studies in Educational Evaluation, 25*, 179-194.

Kok, D., Meeder, M., Wijers, M. & Dormolen, J. van (1992). *Wiskunde 12-16, een boek voor docenten.* Utrecht/Enschede: Freudenthal Instituut/SLO.

Kuiper, W. A. J. M., Bos, K. Tj. & Plomp, Tj. (1997). *Wiskunde en de natuurwetenschappelijke vakken in leerjaar 1 en 2 van het voortgezet onderwijs. Nederlands aandeel in TIMSS populatie 2.* Enschede: Universiteit Twente.

Kuiper, W. A. J. M., Bos, K. Tj. & Plomp, Tj. (1999). Mathematics achievement in the Netherlands and appropriateness of the TIMSS mathematics test. *Educational Research and Evaluation, 5* (2), 85-104.

Linn, R. L. & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational measurement: Issues and Practice, 13*, 5-15.

Luyn, J. van (1998). *Basisvorming: de basis van het studiehuis.* Den Haag: PMVO.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J. & Smith, T. A. (2000). *TIMSS 1999 international mathematiccs Report, Findings from IEA's repeat of the Third International Mathematics and Science Study at the eighth grade.* Boston, MA: Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., Garden, R. A. & O'Connor, K. M., (2000). *TIMSS 1999 international science report, Findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth grade.* Boston, MA: Boston College.

Roelofs, E. C., Franssen, H. A. M., Houtveen A. A. M. & Lagerweij, N. A. J. (1999). Een dieptestudie naar authentiek leren in de basisvorming. Docentgedrag, methodengebruik en leerlingpercepties. *Pedagogische Studien, 76* (4), 258-272.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton E. & Nicol, C (1993). *Curriculum frameworks for mathematics and science. TIMSS Monograph nr. 1*. Vancouver: Pacific Educational Press.

Robitaille, D. F. & Maxwell, B. (1996). In D. F. Robitaille & R. A. Garden (Eds.), *Research questions and study design. TIMSS Monograph No. 2* (pp. 34-43). Vancouver: Pacific Educational Press.

Shavelson, R. J., Baxter G. P. & Xiaohong, G. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Vos, F. P. (2002). *Like an ocean liner changing course: The grade 8 mathematics curriculum in the Netherlands 1995-2000*. Enschede: Universiteit Twente.

Zuzovsky, R. & Harmon, M. (Eds) (1999). TIMSS performance assessment. *Studies in Educational Evaluation, 25,* 173-323.

Zuzovsky, R. (1999). Problematic aspects of the scoring of the TIMSS practical performance assessment: some examples. *Studies in Educational Evaluation, 25*, 315-323.

Appendix

De twaalf taken uit de PVT-2000 omvatten activiteiten als het doen van proefjes met behulp van eenvoudige practicummaterialen, het opzetten en uitvoeren van experimenten, het doen en beschrijven van waarnemingen, het uitvoeren van berekeningen (o.a. met een zakrekenmachine), het noteren van meetresultaten in een tabel, het interpreteren van gegevens uit een tabel, en het geven van verklaringen voor waargenomen verschijnselen (cf. Bos & Kuiper, 1998; Kuiper, Bos & Plomp, 1997).

M1 - Dobbelstenen
Een transformatieregel toepassen op de ogen van een dobbelsteen (even ogen: plus 2, oneven ogen: min 1); waarnemingen doen (40 keer gooien), noteren en frequenties verklaren.

M2 - Rekenmachine
Regelmaat in getalpatronen ontdekken; ontbinden in factoren.

M3 - Vouwen en knippen
Met een schaar in een gevouwen blaadje gewenste vormen maken; een vouwplan tekenen.

M4 - De bocht om
Rechthoeken als schaalmodellen voor meubels ontwerpen; interpretatie van mogelijke meubelvormen; rechthoekjes testen of ze bij een gegeven model van een tweedimensionale gang geschoven kunnen worden; een regel ontdekken voor afmetingen van rechthoeken die wel/niet door de gang passen.

M5 - Inpakken
Ruimtelijke schetsen maken van doosjes waarin 4 pingpong-ballen passen; bijbehorende uitslagen schetsen; één doosje met correcte maten ontwerpen.

S1 - Hartslag
Door lichamelijke inspanning (een trapje op-en-af lopen) versnelling van hartslag vaststellen: metingen doen, weergeven en verklaren.

S2 - Magneten
Determinatie van de relatieve kracht van twee magneten (welke van de twee is sterker?). Metingen verwoorden.

S3 - Batterijen
Vier gelijke batterijen zijn gegeven en een zaklamp waarin twee batterijen tegelijk getest kunnen worden. Met de mededeling dat er twee lege batterijen zijn, is de determinatie hiervan gevraagd. Oplossingsstrategie weergeven.

S4 - Elastiekje
Onderzoek naar de uitrekking van een elastiekje, waaraan ringetjes worden gehangen. Meten en weergegeven van metingen. Extrapolatie voor nog twee extra ringetjes die niet aanwezig zijn.

S5 - Oplossingen
Onderzoeken of bepaalde bruistabletten in heet water sneller oplossen dan in koud water. Metingen doen, waarnemingen weergeven, verklaren.

G1 - Schaduwen
Onderzoek naar een relatie van de afstanden tussen lamp, voorwerp en projectiescherm indien de schaduw een vaste maat heeft (schaduw is tweemaal zo breed als het voorwerp). Metingen doen, weergeven, verklaren.

G2 - Klei
Bij een balansweegschaal een oplossingsstrategie verzinnen om bij twee gegeven gewichten (20g en 50g) eenheden van respectievelijk 10g, 15g en 35g af te wegen ("broodjes klei"). Verwoorden van strategie.