SCHOOLZELFEVALUATIE

HET EVALUEREN VAN EN DOOR SCHOLEN

Henk A. Moelands

Samenstelling promotiecommissie

Voorzitter/secretaris	prof. dr. H.W.A.M. Coonen
Promotoren	prof. dr. C.A.W. Glas
	prof. dr. P.F. Sanders
Leden	prof. dr. R.J. Bosker (Rijksuniversiteit Groningen)
	prof. dr. F.J.G. Janssens (Universiteit Twente)
	prof. dr. G.W. Meijnen (Universiteit van Amsterdam)
	prof. dr. J. Scheerens (Universiteit Twente)

ISBN 90-5834-333-2
Omslag: Marianne Brouwer
Druk: Print Partners Ipskamp B.V., Enschede

SCHOOLZELFEVALUATIE

HET EVALUEREN VAN EN DOOR SCHOLEN

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van rector magnificus,
prof. dr. W.H.M. Zijm,
volgens het besluit van het College van Promoties
in het openbaar te verdedigen
op donderdag 3 maart om 13.15 uur

door

Henricus Antonius Moelands
geboren op 6 november 1951
te Tilburg

Dit proefschrift is goedgekeurd door de promotoren prof. dr. C.A.W. Glas en prof. dr. P.F. Sanders.

Voorwoord

Kwaliteit van onderwijs is een onderwerp dat reeds een groot aantal jaren in de belangstelling staat. Van scholen wordt verwacht dat zij een actief beleid voeren om deze kwaliteit zo optimaal mogelijk te laten zijn. Veel is reeds over dit onderwerp geschreven en vele instrumenten zijn inmiddels ontwikkeld. Ook het onderhavige proefschrift heeft de kwaliteit van onderwijs als onderwerp. In het proefschrift wordt een systeem beschreven, EVADOS genoemd, dat scholen helpt bij het vergaren van informatie waarmee zij een beter beeld kunnen krijgen van de kwaliteit van het door hen geboden onderwijs. Het is aan de scholen om te bepalen hoe zij met deze informatie omgaan.

Toen ik mijn doctoraalexamen aan de Universiteit van Utrecht afsloot, gaf de voorzitter van de examencommissie mij mee dat als ik dacht aan een promotie het nu de tijd zou zijn. Hij doelde toen niet zozeer op het jaartal waarin wij ons toen bevonden, maar meer op de omgeving waar ik werkte: het Cito. Volgens hem zou het Cito met zijn vele expertises de perfecte omgeving zijn om te promoveren. Wat had hij dat goed ingeschat en hoe goed heb ik dat mogen ervaren. Niet alleen vanwege de vele expertises waarnaar hij verwees, maar zeker ook vanwege de grote collegialiteit die het Cito zo kenmerkt, met voor mij de afdeling Psychometrische Onderzoek en Kenniscentrum (POK) als het summum. Niet alleen omdat deze afdeling reeds vele jaren mijn thuisbasis is binnen het instituut, maar vooral vanwege de collega´s die daar werken. Een inspirerende omgeving met collega´s die gevraagd en ongevraagd steun geven. Ik ben hen daar zeer erkentelijk voor. Drie naaste collega´s ben ik vooral erkentelijk en hun namen wil ik dan ook graag vermelden. Als eerste Marius Ouborg met wie ik samen het project startte dat uiteindelijk uitmondde in dit proefschrift. Samen hebben we vele discussies gevoerd en vele werkzaamheden verricht. Ronald Engelen die onmisbaar was tijdens de eerste jaren van het project en zorgdroeg dat de voor mij wat lastigere analyses tot een goed einde

kwamen. Ook gedurende de andere jaren kon ik steeds bij hem aankloppen. Tot slot Anton Béguin die mij (be)geleid heeft op het terrein van de multilevel analyses en die steeds bereid was mij verder te helpen en in die zin voor mij een grote waarde had.

Bijzonder erkentelijk ben ik ook mijn werkgever, het Cito, voor de faciliteiten die mij geboden zijn bij het schrijven en afronden van mijn proefschrift. Een zeer speciale plaats daarin neemt Piet Sanders, hoofd van de afdeling Psychometrische Onderzoek en Kenniscentrum, in. Niet alleen als één van mijn promotoren, maar vooral als stimulator voor mij om het proefschrift te produceren. Steeds was Piet bereid mee te denken en teksten te becommentariëren. Gevraagd en zeker ook ongevraagd reikte hij ideeën of informatie aan. De voorzitter van de examencommissie te Utrecht moet met name hem in gedachten hebben gehad toen hij zijn uitspraak deed.
Ook mijn tweede promotor Cees Glas ben ik zeer erkentelijk. Hoewel Cees niet van het begin af aan bij de werkzaamheden betrokken was, waardeer ik zijn bereidheid om ook als promotor op te treden. Ook zijn opmerkingen na het doorlezen van de manuscripten hebben een positief effect gehad op het eindresultaat.

Tot slot bedank ik graag die personen die vaak niet op de voorgrond treden, maar zo belangrijk zijn in het eindtraject van een proefschrift. Als eerste wil ik mijn collega's van de afdeling DTP noemen, die mede ervoor gezorgd hebben dat het drukken van het proefschrift en dat wat daarvoor nodig is, tot een goed einde is gekomen. Ik wil met name Marianne Brouwer noemen, die het omslagontwerp voor haar rekening heeft genomen.
Nadat het manuscript van het proefschrift in concept klaar is, moet er nog een mooi geheel van gemaakt worden. Mijn grootste steun en toeverlaat daarbij was toch zeker Floor Berentsen, assistente van de afdeling POK. Voor al wat georganiseerd en geregeld moest worden om het proefschrift zover te krijgen zoals het nu voorligt, verdient zij de credits.

Inhoud

1 Inleiding

In de negentiger jaren van de vorige eeuw heeft het ministerie van OCenW een beleid ontwikkeld dat gericht is op het versterken van het vermogen van scholen om zelfstandig werk te maken van kwaliteitsverbetering. Een van de aandachtspunten van dit beleid betreft het bevorderen van kwaliteitszorg binnen scholen. In de ogen van de beleidsmakers dienen scholen een beleid te ontwikkelen dat gericht is op het bewaken en verbeteren van de kwaliteit van hun onderwijs.

Over het bevorderen van de kwaliteitszorg binnen scholen heeft het ministerie afspraken gemaakt met de verschillende overkoepelende onderwijsorganisaties. Deze afspraken zijn neergelegd in het zogeheten Schevenings Beraad (1993/1994, p. 19), waarin wordt gesteld: '…De school is, binnen het raamwerk van de in de wetgeving bepaalde eisen van deugdelijkheid, verantwoordelijk voor de kwaliteit van haar onderwijs. De school voert een actief beleid van kwaliteitszorg. Zij kiest methoden en instrumenten en stelt vast welke resultaten worden geboekt en welke bijstellingen nodig zijn'.
Kwaliteitszorg is met andere woorden een belangrijk onderdeel van de strategische beleidsvorming en planning van scholen met als belangrijke aandachtspunten de optimalisering van het onderwijsresultaat en het kunnen afleggen van verantwoording.

Willen scholen een actief beleid van kwaliteitszorg voeren, dan zullen zij gegevens moeten verzamelen en evalueren om na te kunnen gaan of het gegeven onderwijs leidt tot de beoogde resultaten. Om de verschillende vormen van evalueren op het niveau van de school te typeren, hanteert Scheerens (1996a) drie tegenstellingen:
- kwantitatief versus kwalitatief,
- intern versus extern en
- proces- versus productgericht.

Bij de kwantitatieve evaluatie ligt de nadruk op validiteit en betrouwbaarheid van de gegevensverzameling. Kwalitatieve evaluatie gaat minder uit van eisen ten aanzien van validiteit en betrouwbaarheid, maar richt zich meer op de inhoudelijke kant van de informatie. Als voorbeeld van kwalitatieve evaluatie geeft Scheerens het vragen naar de tevredenheid over de examenresultaten, terwijl bij kwantitatieve evaluatie de examenresultaten zelf voorop staan.
Externe evaluatie heeft het grote voordeel van grotere objectiviteit en onpartijdigheid. Bij deze vorm van evaluatie zijn vaak actoren van buiten de school betrokken die de kwaliteit van het onderwijs beoordelen. Voorbeelden zijn de evaluatie door de inspectie, waarbij deze op een systematische wijze bepaalde aspecten van de kwaliteit van zowel het onderwijs in individuele scholen als het onderwijs in het gehele land beoordeelt, en onderlinge kwaliteitstoetsing tussen scholen in de vorm van wederzijdse visitatie. Ook cohort-onderzoeken in het basis- en voortgezet onderwijs kunnen een basis vormen voor externe evaluaties.
Bij interne evaluatie is meestal sprake van schoolzelfevaluatie. De school is daarbij niet alleen degene die de evaluatie initieert, maar ook degene die deze uitvoert, al of niet ondersteund door bepaalde (eventueel extern ontwikkelde) procedures en/of hulpmiddelen.
Productevaluatie richt zich met name op de effecten of 'producten' van het primaire proces in de organisatie. Procesevaluatie daarentegen heeft - zoals Scheerens (1996a, p. 15) dat noemt - 'transformatieprocessen of ondersteunende condities binnen organisaties als aandachtspunt'. Hoewel Scheerens de voorkeur geeft aan productevaluatie, stelt hij dat procesevaluatie als 'hulpmiddel' bij productevaluatie van grote betekenis is. Informatie over het (onderwijs)proces kan mogelijk meer inzicht verschaffen in de achtergronden van bepaalde mee- of tegenvallende resultaten.

De ontwikkeling van een procedure om de kwaliteit van een school te evalueren aan de hand van resultaten op toetsen is het onderwerp van dit proefschrift. De te ontwikkelen procedure wordt gemakshalve aangeduid met de term EVADOS, dat staat voor evalueren van en door scholen. In termen van de door Scheerens gegeven indeling in typen evaluaties, kan EVADOS gekenschetst worden als een procedure die bijdraagt aan een kwantitatieve, productgerichte en interne evalu-

atie van het onderwijs. Kwantitatief omdat de te verzamelen gegevens zoveel mogelijk moeten voldoen aan eisen ten aanzien van validiteit en betrouwbaarheid. Productgericht omdat in de procedure de leerresultaten van leerlingen bij de beoordeling van de kwaliteit van het onderwijs het uitgangspunt zijn. Intern gericht ten slotte omdat de procedure (in eerste instantie) bedoeld is voor de scholen zelf. Dit laatste wordt tot uitdrukking gebracht door het hanteren van de term 'schoolzelfevaluatie'. In de procedure wordt voor een deel ook aandacht besteed aan procesevaluatie, maar dan vooral om - aansluitend bij de opvatting van Scheerens - meer inzicht te krijgen in bepaalde mee- of tegenvallende resultaten.

Kwaliteit van het onderwijs

EVADOS heeft tot doel een bijdrage te leveren aan de bewaking en zonodig verbetering van de kwaliteit van het onderwijs. In navolging van De Groot (1983) zou kwaliteit omschreven kunnen worden als 'de mate waarin doelstellingen en functies van het onderwijs worden gerealiseerd'. In een in dit kader relevant vergaderstuk (Tweede Kamer, 1994-1995) maakt het ministerie van OCenW een onderscheid tussen kwaliteitszorg en kwaliteitsverbetering. Kwaliteitszorg wordt daarbij omschreven als 'alle activiteiten die erop gericht zijn om informatie te verzamelen over de kwaliteit, alsmede over factoren die van invloed zijn op de kwaliteit' (p. 5). Op schoolniveau kan kwaliteitszorg bijvoorbeeld neerkomen op het opsporen van sterke en zwakke plekken in het functioneren van een school en het registreren van de effecten van schoolbeleid. Op landelijk niveau heeft kwaliteitszorg tot doel informatie te verzamelen over sterke en zwakke plekken in het onderwijsbestel, het functioneren van dat bestel, het volgen van relevante maatschappelijke ontwikkelingen die van invloed kunnen zijn op dat bestel en het evalueren van de effecten van overheidsbeleid. Onder kwaliteitsverbetering verstaat het ministerie 'alle beleid dat gericht is op verbetering, zoals het formuleren van doelen, het vertalen van doelen in bijpassende methoden en instrumenten, het bepalen van de wijze waarop de kwaliteitszorg ingericht wordt, het bepalen van gewenste bijstelling op grond van in de kwaliteitszorg verzamelde informatie, alsmede het organiseren van het besluitvormingsproces in deze' (p. 5). Ook hierbij kan een onderscheid gemaakt

worden tussen het kwaliteitsbeleid van de school en het kwaliteitsbeleid van de overheid.

Voor een omschrijving van het begrip kwaliteit wordt in dit proefschrift aangesloten bij de opvatting van Scheerens die kwaliteit opvat als het geheel van wenselijke hoedanigheden, zoals deugdelijkheid, gezondheid of effectiviteit van organisaties (Scheerens, 1996a). Kwaliteitszorg omschrijft hij vervolgens als 'de actieve gerichtheid, zo men wil het 'beleid' van de organisatie om die wenselijke hoedanigheden ook inderdaad te manifesteren' (p. 12). In zijn optiek vraagt kwaliteitszorg om een actieve benadering, waarin 'kwaliteit' als een probleem wordt gezien en er actief gestreefd wordt naar registratie, handhaving of liefst verbetering van kwaliteit.
In het begrip 'kwaliteitszorg' liggen volgens Scheerens (1996a) de volgende vier deelprocessen besloten:

- meten en registreren;
- beoordelen en evalueren;
- kwaliteitsbewaking of kwaliteitsverbetering;
- organisatie van de kwaliteitszorg.

Voor het meten en registreren van de kwaliteitszorg dient geoperationaliseerd te worden op basis waarvan de kwaliteit bepaald gaat worden. Voor het beoordelen van deze kwaliteit dient een maatstaf ontwikkeld te worden die precies aangeeft wanneer een gewenste standaard bereikt is. Kwaliteitsbewaking heeft betrekking op alles wat men doet met de resultaten van kwaliteitsregistratie en -evaluatie. Bij kwaliteitsbewaking is er sprake van 'vinger aan de pols houden'. Kwaliteitsverbetering geeft aan dat men de resultaten van registratie en beoordeling wil benutten om (het functioneren van) de organisatie te verbeteren en te veranderen, waarbij beide genoemde oriëntaties zowel vanuit een extern controleperspectief (verantwoording) als vanuit een intern managementperspectief (verbetering) kunnen plaatsvinden. De organisatie van de kwaliteitszorg ten slotte heeft betrekking op de mate waarin een organisatie toegerust is om aan kwaliteitszorg te doen.

In EVADOS is als operationalisatie voor de kwaliteit gekozen voor de resultaten van leerlingen op de toetsen uit het Cito-leerlingvolgsysteem. Als maatstaf voor gewenste standaarden is gekozen voor een intern en een extern referentiekader. Wat betreft het derde deelproces richt EVADOS zich niet op het externe controleperspectief, maar op het interne managementperspectief. Op deze drie deelprocessen zal in dit proefschrift nog uitgebreid ingegaan worden. De organisatie van de kwaliteitszorg komt niet aan bod.

Naar een procedure voor kwaliteitszorg

Cremers, Rekveld, Brandsma en Bosker (1995) hebben 31 instrumenten beschreven die scholen kunnen gebruiken voor kwaliteitszorg. Hun studie maakt duidelijk dat er verschillende systemen van zelfevaluatie te onderscheiden zijn, en dat elk systeem zijn eigen kenmerken heeft, zijn eigen eisen stelt en zijn eigen mogelijkheden heeft. Volgens Bosker en Scheerens (1995) dient de evaluatie van de kwaliteit van een school in eerste instantie gebaseerd te zijn op de prestaties van de leerlingen. Zij geven aan dat bij procedures voor schoolzelfevaluatie die gericht zijn op procesfactoren het gevaar van 'goal-displacement' bestaat, dat wil zeggen dat deze factoren als evaluatiedoel gezien gaan worden in plaats van mogelijke verklarende factoren voor behaalde leerprestaties. Scheerens (1989) hanteert in dit verband zelfs de uitdrukking 'operatie gelukt, patiënt overleden'. Dat leerlingprestaties van belang zijn, vermeldt ook Creemers (1995, p. 5) 'At the moment there are a lot of descriptions of different projects in education available, touching stories about what happened in practice, good relations built by change agents in schools, outside of schools, more inclusion of parents, a better playground, and so on. But the question remains: did factors at the school level and the classroom level that we know from research and theory, really induce better schools or better results?… One of the main issues of school effectiveness research that should also be found in school improvement is that it is ultimately about results, student achievement'.

Uit de studie van Cremers e.a. (1995) zijn twee redenen af te leiden om een procedure voor schoolzelfevaluatie te ontwikkelen.

In de eerste plaats blijkt dat er voor scholen nagenoeg geen instrumenten of procedures beschikbaar zijn die zich concentreren op leerprestaties. En leerprestaties zijn, zoals onder andere Bosker en Scheerens en ook Creemers aangeven, erg belangrijk voor het bepalen van de kwaliteit van het onderwijs. In de tweede plaats blijkt dat voorzover er instrumenten of procedures beschikbaar zijn, deze zich richten op het niveau van de individuele leerling en niet op het niveau van de klas of de school. Voor een school is het in het kader van de verantwoordelijkheid voor haar eigen onderwijs belangrijk zicht te krijgen op de ontwikkeling van de school als geheel. Steeds zal de school periodiek geïnformeerd moeten worden over hoe zij ervoor staat, of genomen beleidsbeslissingen tot het gewenste effect hebben geleid en of de ontwikkelingen van groepen van leerlingen verloopt zoals verwacht en bijvoorbeeld geen afwijkend patroon vertoont in vergelijking met de ontwikkeling van vergelijkbare groepen in voorgaande jaren. Op basis van deze informatie kan een school desgewenst acties ondernemen. EVADOS dient scholen van de hiervoor genoemde informatie te voorzien en hen zo te ondersteunen bij het verder vormgeven van kwaliteitszorg.

Functionaliteit van de te ontwikkelen procedure

Voogt (1995, p. F3325-4) omschrijft zelfevaluatie (zij gebruikt daarvoor het begrip ‘schooldiagnose’) als ‘elke systematische procedure van dataverzameling en -analyse in en voor en door scholen over het actuele functioneren met de bedoeling prioriteiten te stellen en besluiten te nemen over schoolverbeteringen’. Centraal in deze omschrijving staat het komen tot schoolverbeteringen. Schoolverbeteringen kunnen betrekking hebben op het functioneren en het welbevinden van diverse geledingen binnen de school, op het tot stand brengen van positieve attitudeveranderingen bij leerlingen en op het verwerven van relevante kennis, inzichten en vaardigheden door de leerlingen. Een school kan ook een aantal specifieke doelen nastreven zoals het aanbieden van een zo breed mogelijk vakkenpakket, kindvriendelijk lesgeven of het terugdringen van het aantal kinderen dat spijbelt. Alle keuzes die gemaakt worden zullen in ieder geval moeten aansluiten bij de kaders die van overheidswege opgelegd worden. Deze kaders laten zich in algemene termen omschrijven als: brede vorming, het

bevorderen van een ononderbroken ontwikkelingsgang, maatwerk voor iedere leerling, het voorbereiden van jeugdigen op het spelen van een verantwoorde rol in de multiculturele samenleving en het bestrijden van achterstanden. De invulling van deze algemene doelen is onder invloed van maatschappelijke ontwikkelingen voortdurend onderwerp van bezinning en dialoog.

Hoewel bij EVADOS de nadruk gelegd wordt op leerresultaten, die gezien worden als de belangrijkste indicatie voor de kwaliteit van het geboden onderwijs, kan uit de behaalde leerresultaten niet rechtstreeks de kwaliteit van het onderwijs worden afgeleid. De omstandigheden waaronder dat onderwijs heeft plaatsgevonden kunnen namelijk zo wezenlijk verschillen, dat het niet juist zou zijn een uitspraak te doen over de kwaliteit van het onderwijs wanneer deze alleen gebaseerd zou zijn op leerresultaten. Zo is het goed mogelijk dat de kwaliteit van het onderwijsleerproces op een school met lagere leerresultaten hoger is dan de kwaliteit van het onderwijs op een school met hogere leerresultaten. EVADOS moet scholen dan ook in staat stellen bij de beoordeling van de leerresultaten rekening te houden met de omstandigheden waaronder het onderwijs heeft plaatsgevonden. Voor zover mogelijk moet bij de weergave van de resultaten rekening gehouden (kunnen) worden met een aantal relevante input- en procesgegevens. Scholen kunnen zich met deze gegevens beter spiegelen aan voor hen vergelijkbare groepen, omdat mogelijke invloeden van een aantal factoren, zoals bijvoorbeeld de samenstelling van de populatie op de school, verdisconteerd zijn, om op basis van de resultaten gerichtere (beleids-)acties te kunnen ondernemen.

Schoolzelfevaluatie vraagt van scholen periodiek gegevens te verzamelen over:
- de resultaten van het onderwijs aan (groepen van) leerlingen;
- het onderwijsaanbod;
- proceskenmerken;
- leerlingkenmerken.

EVADOS stelt scholen in staat om op basis van deze gegevens conclusies te trekken over de mate waarin men tevreden is met de gevonden resultaten. Afhankelijk van het resultaat kan een school besluiten maatregelen ter

verbetering van de kwaliteit te nemen.

Moelands en Sanders (1996) onderscheiden drie categorieën onderwijskundige meetinstrumenten. De eerste categorie betreft meetinstrumenten die tot doel hebben de kwaliteit van een onderwijssysteem te beoordelen. De tweede categorie betreft meetinstrumenten die tot doel hebben de kwaliteit van een leerling te beoordelen. De derde categorie ten slotte betreft meetinstrumenten die tot doel hebben de kwaliteit van het onderwijsleerproces te beoordelen.
EVADOS kan zowel bij de eerste als de derde categorie ingedeeld worden. Indien de procedure tot doel heeft de kwaliteit van het onderwijs op klas-, groeps- of schoolniveau (meso-niveau) te beoordelen en scholen van informatie te voorzien voor het nemen van (beleids)beslissingen, dan behoort EVADOS tot de eerste categorie. EVADOS is met name hiervoor bedoeld. Ingeval er sprake is van het leveren van informatie over de kwaliteit van het onderwijsleerproces aan docenten op basis waarvan zij kunnen besluiten het onderwijs aan hun leerlingen anders in te richten, is EVADOS een voorbeeld van de derde categorie.

De ontwikkeling van een procedure voor schoolzelfevaluatie vraagt om keuzes. Deze keuzes worden hierna kort toegelicht. In het vervolg van dit proefschrift komen de keuzes uitvoerig aan bod.

1 Herhaalde metingen in de tijd

Bij de ontwikkeling van EVADOS is ervoor gekozen de leerresultaten in de tijd te kunnen volgen. Het vaststellen van de kwaliteit mag niet een éénmalige activiteit zijn, maar dient in meerdere jaren plaats te vinden. Bovendien dienen de resultaten uit de verschillende jaren met elkaar vergeleken te kunnen worden. Deze keuze vraagt om meetinstrumenten die vergelijkingen in de tijd mogelijk maken.

Voor de ontwikkeling van EVADOS is gekozen voor de toetsen uit het Cito-leerlingvolgsysteem, die de mogelijkheid bieden om vergelijkingen in de tijd en ook over leerjaren heen mogelijk te maken. Door toepassing van itemresponsmodellen zijn voor een aantal leerstofonderdelen meetschalen ontwikkeld waar-

mee vorderingen van individuele leerlingen in de tijd gevolgd kunnen worden (Gillijns & Moelands, 1992). Met behulp van deze schalen kan dan zichtbaar gemaakt worden hoeveel een leerling in een bepaalde periode vooruit is gegaan ten opzichte van eerdere meetmomenten.

Na het afnemen van de toetsen en het registreren van de vorderingen van de individuele leerlingen kan een docent, door het aggregeren van deze leerlinggegevens, informatie krijgen over het functioneren van groepen van leerlingen (Moelands & Ouborg, 1995). Door dit voor een aantal jaren te doen, ontstaat een overzicht van de gemiddelde schaalscores die in de achtereenvolgende jaren door de groepen zijn behaald.

Uitspraken over de kwaliteit van het onderwijs krijgen pas hun waarde als deze gelegd kunnen worden naast een standaard (een referentiekader). Wat als referentiekader dienst gaat doen zal vastgesteld moeten worden. Bij EVADOS is ervoor gekozen scholen in staat te stellen hun prestaties te vergelijken met hun eigen prestaties in voorgaande jaren, en met de prestaties van andere scholen. In het laatste geval is het wel van belang dat vastgesteld gaat worden op basis waarvan die vergelijking gaat plaatsvinden.

2 Toegevoegde waarde

De toegevoegde waarde informeert een school over haar bijdrage aan het onderwijs. Een belangrijk aspect bij de toegevoegde waarde betreft de leerwinst: het verschil tussen voor- en nameting. Door als outputmeting te kiezen voor de toetsen van het Cito-leerlingvolgsysteem hoeven aparte voormetingen niet plaats te vinden. De resultaten op deze toetsen zijn immers in de tijd vergelijkbaar, waardoor steeds informatie over het (start-)niveau aan het begin van een (nieuwe) periode bekend is. Bij de interpretatie van de leerwinst zijn de omstandigheden waaronder het onderwijs plaatsvond en kenmerken van leerlingen aan wie het onderwijs is gegeven belangrijk, omdat deze twee factoren mede het te behalen resultaat beïnvloeden. De schoolresultaten dienen dan ook gecorrigeerd te worden voor achtergrondvariabelen op leerling-, klas- en schoolniveau die bepaalde

uitkomsten zouden kunnen verklaren. Door deze correcties krijgt een school informatie over haar bijdrage aan de leerprestaties.

3 Procesindicatoren

Als een school constateert dat de resultaten van haar onderwijs achterblijven bij de verwachtingen, dan dient zij de mogelijkheid te hebben om gerichte maatregelen ter verbetering van de resultaten te nemen. De school zal moeten nagaan wat mogelijke oorzaken voor het achterblijven van de resultaten kunnen zijn. Het spreekt voor zich dat, gegeven de complexiteit van het onderwijs, het lang niet altijd mogelijk is precies dé oorzaak of dé oorzaken aan te geven. Uit het schooleffectiviteitsonderzoek zijn factoren bekend die van invloed kunnen zijn op de resultaten van het onderwijs. Genoemd worden bijvoorbeeld de betrokkenheid van de schoolleiding en de sfeer binnen de school. Ook de wijze waarop de leerkracht zijn lessen organiseert en de mate waarin hij regelmatig de vorderingen van zijn leerlingen nagaat, worden als belangrijk gezien. In dit proefschrift zal met name op één procesindicator (Toets Curriculum Overlap) nader ingegaan worden.

4 Toets Curriculum Overlap

Voor het beoordelen van de resultaten van leerlingen op het door scholen gebruikte toetsinstrumentarium, zijn twee vragen relevant.

1 Sluiten de opgaven van het gebruikte instrumentarium in voldoende mate aan op het geboden onderwijs?
2 In hoeverre is het geboden onderwijs afgestemd op de kerndoelen, zoals deze gelden voor het basisonderwijs?

De eerste vraag is de meest wezenlijke. Als de gebruikte toetsen niet aansluiten op het geboden onderwijs (geïmplementeerd curriculum), dan geven de resultaten van de leerlingen op de toetsen, ongeacht welke doelstellingen nagestreefd worden, geen of onvoldoende informatie over de kwaliteit van dat onderwijs. Voor een (maatschappelijke) waardering van het geboden onderwijs is een extern referentiekader nodig, waaraan de tweede vraagstelling refereert. Voor het basisonderwijs kunnen hiervoor de in 1998 vastgestelde kerndoelen gelden (Ministerie van Onderwijs Cultuur en Wetenschappen, 1998). Voor het kunnen

doen van een uitspraak over de kwaliteit van het onderwijs is het van belang zicht te hebben op de inhoudsvaliditeit van het door scholen gebruikte meetinstrument ten opzichte van de kerndoelen. Indien mogelijk niet alleen aan het einde van het basisonderwijs (groep acht), maar ook tijdens de andere leerjaren. Indien een school een signaal krijgt dat haar onderwijs inhoudelijk niet leidt tot de kerndoelen, dan is dat een reden voor overleg binnen de school dat zou kunnen leiden tot bijstellingen van het onderwijsprogramma.

5 Geen extra belasting voor scholen

Voor de implementatie en het gebruik van EVADOS is het van belang dat scholen zo min mogelijk belast worden. Waar mogelijk dient het verzamelen van extra gegevens voor het doen aan zelfevaluatie tot een minimum beperkt te blijven. De mogelijkheden hiervoor zijn binnen het basisonderwijs aanwezig. Scholen maken gebruik van schooladministratiepakketten waarin veel gegevens zijn opgeslagen of opgeslagen kunnen worden. Ook zijn er diverse computerprogramma's beschikbaar waarin leerresultaten ingevoerd kunnen worden. Door de relevante gegevens uit deze pakketten met elkaar in verband te brengen, krijgen scholen zonder al te veel (extra) inspanningen, gegevens voor de evaluatie van hun onderwijs ter beschikking. Indien nog andere, aanvullende informatie nodig is, zal deze op een zo gebruiksvriendelijke wijze verzameld moeten worden.

Inhoud van dit proefschrift

Dit proefschrift gaat over het ontwikkelen van een procedure voor schoolzelfevaluatie. Er wordt een systematiek beschreven waarmee scholen de kwaliteit van het door hen geboden onderwijs kunnen evalueren. Vanuit dit oogpunt kan het proefschrift getypeerd worden als een ontwerp van een procedure voor schoolzelfevaluatie. De in dit proefschrift beschreven werkzaamheden hebben voor een deel plaatsgevonden als onderdeel van een vierjarig (1995-1999) gezamenlijk project (ZEBO-project) van het Instituut voor Toetsontwikkeling (Cito), het Instituut voor Leerplanontwikkeling (SLO) en het Onderzoekscentrum voor Toegepaste Onderwijskunde (OCTO). Het project had tot doel een instrument voor schoolzelfevaluatie te ontwikkelen. Elk instituut leverde vanuit zijn exper-

tise een bijdrage aan het ZEBO-project. Voor de SLO betekende dit het ontwikkelen van een instrument waarmee het onderwijsaanbod kwalitatief beoordeeld kan worden. Het OCTO ontwikkelde een instrument voor het meten van school- en klaskenmerken. Het Cito droeg zorg voor de koppeling van leerlingachtergrondgegevens met de resultaten van leerlingen op de toetsen uit het Cito-leerlingvolgsysteem en het informeren van scholen over de ontwikkeling van groepen van leerlingen in de tijd. Ook het ontwikkelen van een instrument Toets Curriculum Overlap behoorde tot de taakstelling van het Cito. Dit proefschrift betreft de Cito-bijdrage aan het ZEBO-project. Telkens als in dit proefschrift verwezen wordt naar EVADOS, heeft dit betrekking op het Cito-aandeel in het ZEBO-project.

In hoofdstuk twee van dit proefschrift wordt het conceptueel model voor schoolzelfevaluatie dat als basis diende voor de ontwikkeling van EVADOS besproken. Op basis van literatuur over schooleffectiviteit en schoolverbetering wordt aangegeven welke input- en procesfactoren relevant zijn als mogelijke verklarende factoren voor een verandering in leerresultaten.

In hoofdstuk drie komt de functionaliteit van EVADOS aan de orde. Bovendien wordt nader ingegaan op de wijze waarop de procedure scholen kan voorzien van interne en externe referentiegegevens. Ook zal in dit hoofdstuk het begrip 'toegevoegde waarde' uitgewerkt worden.

Hoofdstuk vier is gewijd aan de componenten input, proces en output. In dit hoofdstuk zal een in veel scholen gebruikt schooladministratiepakket als informatiebron voor inputgegevens besproken worden. Ook komt in dit hoofdstuk een uitwerking van de procesvariabele Toets Curriculum Overlap aan de orde. Ten slotte vindt in dit hoofdstuk een bespreking plaats van de toetsen uit het Cito-leerlingvolgsysteem en zal op het belang van deze toetsen voor EVADOS ingegaan worden.

Voor het doen van uitspraken over de kwaliteit van het onderwijs dienen de relevante inputfactoren, procesfactoren en outputgegevens aan elkaar gerelateerd

te worden. Een aantal van deze factoren heeft betrekking op het niveau van de leerling, bijvoorbeeld zijn sociaal-economische status. Een aantal factoren, zoals de betrokkenheid van de schoolleiding, heeft betrekking op het niveau van de school. Bij de interpretatie van leerresultaten dient niet alleen rekening gehouden te worden met relevante input- en procesfactoren, maar ook met de niveaus waarop zij betrekking hebben. Is het een factor die hoort bij de leerling, bij de klas, de school of zelfs bij meerdere niveaus? Om met de effecten van deze factoren op de diverse niveaus rekening te houden en om betrouwbare en valide uitspraken te kunnen doen, zullen statistische modellen gehanteerd moeten worden. Multilevel modellen bieden die mogelijkheid (Bosker & Scheerens, 1995). In hoofdstuk vijf wordt hierop ingegaan.

Een belangrijke indicator in dit proefschrift voor de kwaliteit van het onderwijs zijn de resultaten van leerlingen op toetsen. In hoofdstuk 6 wordt aangegeven hoe de ontwikkeling van groepen van leerlingen in de tijd op basis van hun resultaten op toetsen in beeld gebracht kan worden. Ook worden de in hoofdstuk 5 te bespreken multilevel modellen toegepast. Aangegeven zal worden wat voor informatie over de kwaliteit van het onderwijs de toepassing van deze modellen voor scholen oplevert.

Hoofdstuk zeven ten slotte richt zich op de (nabije) toekomst. Om EVADOS te gebruiken voor uitspraken op beleidsniveau zullen (landelijke) referentiegegevens verzameld moeten worden. Gegeven de aard van EVADOS zullen deze gegevens frequent verzameld en geanalyseerd moeten worden. Ook de terugrapportage van de bevindingen dient frequent plaats te vinden. Hierbij speelt datacommunicatie een belangrijke rol. Er dient gezocht te worden naar een systeem waarmee op een efficiënte en vooral ook gebruiksvriendelijke wijze data getransporteerd kunnen worden. Dit onderwerp zal met name in hoofdstuk zeven aan de orde komen. Ook het nut van het oprichten van een Expertisecentrum voor Kwaliteitszorg komt aan bod. Met een dergelijk centrum kan - met gebruikmaking van de verworvenheden waarvan in dit proefschrift verslag wordt gedaan - een permanent monitorsysteem opgezet worden, waarbij de doelgroep zich niet hoeft te beperken tot scholen. Ook gemeenten, samenwerkingsverban-

den en andere organisaties die op enigerlei wijze betrokken zijn bij het onderwijs zouden van de diensten van een dergelijk centrum gebruik kunnen maken.

2 Het CIPO-model als uitgangspunt voor school-zelfevaluatie

In de inleiding is zelfevaluatie (schooldiagnose) omschreven als ‘elke systematische procedure van dataverzameling en data-analyse in en voor en door scholen over het actuele functioneren met de bedoeling prioriteiten te stellen en besluiten te nemen over schoolverbeteringen’ (Voogt, 1995). Bij de bespreking van deze omschrijving is aangegeven dat deze verbeteringen betrekking kunnen hebben op deelaspecten als het functioneren en het welbevinden van diverse geledingen binnen de school, op het tot stand brengen van attitudeveranderingen bij leerlingen en op het verwerven van kennis, inzichten en vaardigheden door de leerlingen. Ook specifieke doelen kunnen onderwerp van verbeteringen zijn, zoals het aanbieden van een zo breed mogelijk vakkenpakket, kindvriendelijk lesgeven of het terugdringen van het aantal kinderen dat spijbelt. In dit proefschrift heeft schoolzelfevaluatie betrekking op de school als geheel en niet op afzonderlijke elementen als schoolleiding, leerklimaat en methoden. Dat neemt niet weg dat vastgesteld moet worden welke elementen (indicatoren) bijdragen aan het verkrijgen van een goed beeld van de school. Voor zover wenselijk en noodzakelijk zullen keuzes gemaakt moeten worden. Voor het maken van deze keuzes is een beschrijvingskader van de school ‘als geheel’ nodig, op basis waarvan een oordeel over de kwaliteit van een school uitgesproken kan worden. Bovendien biedt een dergelijk kader een basis om scholen met elkaar te vergelijken.

2.1 Effectiviteitsmodellen

Schooleffectiviteit betreft de wijze waarop de school als organisatie zijn doelen bereikt. Quinn en Rohrbaugh (1983), aangehaald door Scheerens (1996b), onderscheiden de volgende vier uit de organisatieliteratuur afgeleide effectiviteitsmodellen: 'het human-relationsmodel', 'het interprocesmodel', 'het opensysteemmodel' en 'het rationele-doelmodel'. Elk model hanteert een verschillend gezichtspunt bij het bepalen van de kwaliteit van een organisatie. Zo hecht bijvoorbeeld het 'human-relationsmodel' veel waarde aan het onderwerp arbeidssatisfactie, terwijl het 'rationele-doelmodel' vooral kijkt naar onderwijsopbrengsten.

In dit proefschrift staan onderwijsopbrengsten in de vorm van leerlingresultaten op toetsen uit het Cito-LVS centraal, een keuze die aansluit bij het rationele-doelmodel. Belangrijke begrippen binnen dit model zijn productiviteit en efficiëntie. Vanuit dit model zou gezocht moeten worden naar indicatoren die scholen optimaal helpen de productiviteit en efficiëntie van hun organisatie vast te stellen en zonodig te verbeteren.

Aan het ontwikkelen van EVADOS ligt een beschrijvingsmodel ten grondslag. De factoren uit dit beschrijvingsmodel, alsmede het niveau waarop deze factoren werkzaam zijn, zijn afgeleid uit de resultaten van het schooleffectiviteitsonderzoek. In het vervolg van deze paragraaf zal allereerst een kort historisch overzicht van dit type onderzoek gegeven worden en daarna zal nader ingegaan worden op het te hanteren beschrijvingsmodel. Voor het overzicht is vrijelijk gebruik gemaakt van publicaties van Scheerens (1989), Brandsma (1993) en Van Petegem (1994, 1997).

2.2 Het schooleffectiviteitsonderzoek

De oorsprong van het schooleffectiviteitsonderzoek is gelegen in een vijftal onderzoeksstromingen (Scheerens, 1989):

- onderzoek naar (on)gelijkheid van kansen;
- economisch onderzoek naar onderwijsproductiefuncties;
- evaluatie van compensatieprogramma's;
- onderzoek naar effectieve scholen en evaluatie van schoolverbeteringsprogramma's;
- onderzoek naar de effectiviteit van leerkrachten en instructieprocessen.

Twee voorbeelden van het onderzoek naar (on)gelijkheid van kansen zijn de onderzoeken van Coleman, Campbell, Hobson, McPartland, Mood, Weifeld en York (1966) en Jencks , Smith, Acland, Bane, Cohen, Gintis, Heys en Michelson (1972). In hun studie onderzochten zij de relatie tussen een aantal schoolfactoren of -kenmerken en leerprestaties. Zij concludeerden dat het milieu van herkomst van de leerling een belangrijkere determinant van onderwijsuitkomsten was dan de school die de leerling bezocht. Met andere woorden: niet de invloed van de school op leerlingprestaties is van belang, maar de gezinssituatie en de sociale herkomst van leerlingen. Blijkbaar is het onderwijs niet opgewassen tegen de maatschappelijke achtergrond van leerlingen en heeft het wat dat betreft nauwelijks invloed op hun schoolloopbaan.

De tweede onderzoeksstroming vindt zijn oorsprong in het economisch georiënteerde onderzoek naar onderwijsproduktiefuncties. In deze benadering staat de input-outputrelatie centraal. De school zelf wordt als 'black box' gezien. In deze stroming staat de vraag 'welke inputkenmerken hebben effect op de outputkenmerken' centraal. Daarbij wordt voor een aantal interveniërende factoren zoals sociaal economische status en intelligentie gecorrigeerd. Dit type onderzoek concentreerde zich op inputkenmerken als: leerkracht/leerlingratio, vooropleiding en ervaring van leerkrachten, het salaris van de leerkrachten en de uitgaven

per leerling. Uit deze economische georiënteerde onderzoeken blijkt dat de inputvariabelen slechts een gering effect hebben op de output.

De derde onderzoeksstroming concentreert zich op het voorkómen van leerachterstand bij kinderen uit sociaal minder weerbare milieus, vergelijk het Nederlands project 'Onderwijs en Sociaal Milieu' (OSM) uit de zeventiger jaren.

De vierde onderzoeksstroming kan beschouwd worden als de stroming die het meest de kern van het schooleffectiviteitsonderzoek raakt. De vier overige onderzoeksstromingen richten zich niet zozeer op de effectiviteit van de school als geheel, of op de effectiviteit van schoolkenmerken. Het weerleggen van de conclusie dat scholen er niet zoveel toe doen (vergelijk Coleman e.a. en Jencks e.a) is een belangrijke inspiratiebron geweest voor deze onderzoeksstroom. Getracht wordt de 'black box' van de school open te breken door schoolkenmerken te onderzoeken die te maken hebben met organisatie, vormgeving en inhoud van het gebeuren op de school.

De vijfde stroming ten slotte richt zich op (het effectief gedrag van) de leerkrachten in de klas. Zoals in de jaren zestig en zeventig allerlei persoonlijkheidskenmerken van leerlingen centraal stonden (vergelijk Coleman e.a. en Jencks e.a.), staat in deze vijfde stroming het leerkrachtgedrag centraal (proces-productstudies). Deze onderzoeken leidden tot de conclusie dat leerkrachtgedrag wel degelijk van invloed is op de prestaties van leerlingen, hetgeen in tegenstelling was met de bevindingen uit bijvoorbeeld het onderzoek van Coleman e.a. (1966).

De onderzoeken van Coleman e.a. en Jencks e.a. ondervonden veel kritiek. Niet alleen input- en outputkenmerken, maar ook andere variabelen, zoals proceskenmerken, moesten volgens de critici meegenomen worden. Een voorbeeld van een input-proces-outputbenadering in Groot-Brittanië is het onderzoek 'Fifteen Thousand Hours' van Rutter, Maughan, Mortimore, Ouston en Smith (1979). Uit dit onderzoek bleek dat tussen scholen met vergelijkbare leerlingpopulaties duidelijke verschillen bestonden. De conclusie was dat scholen in kwaliteit

verschillen en dat het voor leerlingen wel degelijk uitmaakt welke school zij bezoeken. Blijkbaar kunnen verschillen tussen leerlingen, in tegenstelling tot datgene wat Coleman e.a. en Jencks e.a. in hun onderzoeken concludeerden, voor een deel wel verklaard worden door schoolkenmerken. Voorbeelden van schoolkenmerken zijn: het doceergedrag van leerkrachten, het gebruik van beloning en straf en de stabiliteit van het docententeam. Ook in de Verenigde Staten kwamen Brookover, Beady, Flodd, Schweitzer en Wisenbaker (1979) tot de bevinding dat de school er toe doet.

Ofschoon er veel kritiek geleverd is op de methodologie van voornoemde onderzoeken (zie Van Petegem, 1997, verwijzend naar Scheerens, 1989), is het belang ervan groot. Immers, als het mogelijk is de verschillen in effectiviteit tussen scholen te verklaren met - manipuleerbare - schoolkenmerken, dan biedt dat wellicht mogelijkheden de effectiviteit van scholen te verhogen.

Uit de uitkomsten van het schooleffectiviteitsonderzoek in de jaren zestig en zeventig komen vijf factoren naar voren die positief samenhangen met leerresultaten. Deze factoren die bekend staan als het vijffactorenmodel van Edmonds (1979) zijn (in willekeurige volgorde):

- onderwijskundig leiderschap;
- nadruk op basisvaardigheden;
- hoge verwachtingen van de prestaties van leerlingen;
- een ordelijk en veilig klimaat;
- frequente evaluatie van de vorderingen van leerlingen.

Uit toepassing van deze vijf factoren in grootschalige schoolverbeteringsprojecten blijkt dat er geen effect is op leerprestaties op langere termijn, waaruit geconcludeerd zou mogen worden dat het vijffactorenmodel wellicht te eenvoudig is en niet beantwoordt aan de complexiteit van de onderwijswerkelijkheid (Van Petegem, 1997).

Belangrijk in het kader van het onderzoek naar de effectiviteit van scholen is het onderzoek 'School Matters' van Mortimore, Sammons, Stoll, Lewis en Ecob (1988). In hun studie gebruikten zij multilevel modellen om de afzonderlijke

effecten van de school en de klas op zowel individueel als op groepsniveau te onderzoeken. Ook hun onderzoek bevestigt dat de school er toe doet en een grote bijdrage levert aan de progressie van leerlingen. Bovendien, zo blijkt uit hun onderzoek, is de ene school effectiever voor de ene groep leerlingen dan voor de andere. Zij concluderen dat er twaalf factoren zijn die van invloed zijn op de effectiviteit van scholen.

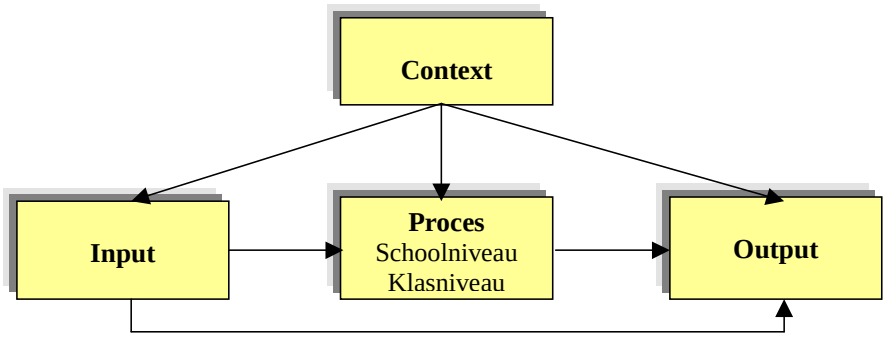
Creemers (1994) geeft aan dat inmiddels vele internationale onderzoeken hebben plaatsgevonden naar factoren die verschil zouden maken tussen effectieve en niet-effectieve scholen. Hij verwijst daarbij naar Levine en Lezotte (1990) die een overzicht geven van meer dan honderd factoren die op de een of andere wijze correleren met effectiviteit. In tegenstelling tot Mortimore e.a. (1988) die twaalf factoren vonden, geeft Creemers aan dat in twaalf Nederlandse onderzoeken minder dan twaalf factoren gevonden werden die een onderscheid maakten tussen effectieve en niet-effectieve scholen. Een aantal van deze factoren komt overeen met de eerder genoemde factoren uit het vijffactorenmodel van Edmonds, zoals: ordelijk klimaat, frequente evaluatie, prestatiegericht beleid, hoge verwachtingen en directe instructie.

2.3 Een beschrijvingskader voor schoolzelfevaluatie

Uit de hiervoor besproken resultaten van het schooleffectiviteitsonderzoek komt naar voren dat diverse factoren de effectiviteit van een school bepalen. De vraag is nu welke van deze factoren relevant zijn om op te nemen in een beschrijvingskader dat als basis kan dienen voor de ontwikkeling van een procedure voor schoolzelfevaluatie? Bij het weergeven van een dergelijk beschrijvingskader dient rekening gehouden te worden met de complexiteit van het onderwijs en de organisatieniveaus waarop de factoren werkzaam zijn.

2.3.1 De complexiteit van het onderwijs

Om de complexiteit van het onderwijs te beschrijven maakt men veelal gebruik van modellen. Een bekend onderwijsmodel is het CIPO-model, een acroniem voor Context, Input, Proces en Output. Figuur 2.1 geeft de onderlinge relatie aan tussen de componenten die door het model onderscheiden worden.



Figuur 2.1
Het CIPO-model

Uit het CIPO-model kan worden afgeleid dat de componenten context, input en proces mede bepalend zijn voor de bereikte of te bereiken output. Tot de component 'input' worden niet alleen leerlingkenmerken zoals het beginniveau van de leerling, hun motivatie en interesses gerekend, maar ook docentkenmerken. Ook beschikbare ruimtes en materiële voorzieningen, de beschikbaarheid van onderwijsmateriaal en van begeleidingsfaciliteiten behoren tot deze component. Voorbeelden van de component 'proces' zijn didactische werkvormen, leeractiviteiten, beschikbare lestijd, de mate waarin leerlingen in de gelegenheid zijn gesteld zich de lesstof eigen te maken en de organisatie van het onderwijs. Kortom alles wat op school- en klasniveau met het onderwijsproces en de wijze waarop leerlingen dat ervaren te maken heeft. De component 'output' kan op verschillende zaken betrekking hebben. Bijvoorbeeld in hoeverre geformuleerde kerndoelen of daarvan afgeleide tussendoelen bereikt zijn, op de uitstroom naar het

speciaal onderwijs en naar vormen van voortgezet onderwijs, en op het aantal drop-outs. Voorbeelden van de component ‘context’ zijn de demografische kenmerken van de wijk waar de school zich bevindt. Ook kenmerken als schoolgrootte en bestuursvorm als ook de gemeente waartoe een school behoort, zijn voorbeelden van de component ‘context’.
Het CIPO-model laat zien dat als men een uitspraak wil doen over de output van het onderwijs, men rekening moet houden met de andere componenten. De componenten ‘input’ en ‘proces’ zijn voor een school het meest manipuleerbaar en vormen om die reden voor een school een aangrijpingspunt om de output van het onderwijs te verbeteren. Natuurlijk is ook de component ‘context’ van belang. Ook deze component zal zeker de output van het onderwijs kunnen beïnvloeden. De school zal echter variabelen die tot de component ‘context’ gerekend kunnen worden in de regel niet kunnen beïnvloeden. Mocht een wijk vanwege de samenstelling van de populatie van invloed zijn op de output van het onderwijs, dan is dat voor een school een gegeven. Een school zal niet in staat zijn deze samenstelling in een voor haar positieve zin te beïnvloeden. En voor zover een school toch invloed wil uitoefenen op de samenstelling van haar populatie, dan zal zij bijvoorbeeld de toelatingsvoorwaarden moeten veranderen, wat een ‘inputvariabele’ is.

2.3.2 Organisatieniveaus

In een school kunnen drie organisatieniveaus onderscheiden worden: de leerling, de klas en de school, waarbij de leerling als het laagste niveau gezien wordt en de school als het hoogste. Leerlingen maken deel uit van een klas en een klas maakt op zijn beurt weer deel uit van een school. Men zou op deze wijze kunnen spreken van een hiërarchische structuur. Met deze niveaus moet voor het ontwikkelen van een procedure in het kader van schoolzelfevaluatie zo mogelijk rekening gehouden worden. Per niveau zijn variabelen te onderscheiden die de effectiviteit van het onderwijs kunnen beïnvloeden. Een voorbeeld op het niveau van de leerling is de startkwalificatie waarmee een leerling het onderwijs

aanvangt en op het niveau van de school de managementkwaliteiten van de schoolleider.
Uit onderzoek blijkt dat het effect van de verschillende variabelen op de diverse niveaus verschillend is. Zo verklaren met name leerlingkenmerken als intelligentie en sociaal economische status het grootste deel van de variantie in leerlingprestaties. Instructievariabelen die vrij direct ingrijpen op het leerproces van de leerlingen hebben een sterkere relatie met leerprestaties dan verder verwijderde schoolkenmerken als schoolklimaat en stijl van leiderschap (Scheerens, 1989). Het voorgaande betekent overigens niet dat de variabelen op hoger niveau niet van belang zouden zijn. Deze variabelen leveren ook een bijdrage aan de resultaten of zijn voorwaardenscheppend voor datgene wat plaatsvindt op de lagere niveaus (Creemers, 1994). Zo zijn afspraken op schoolniveau mede bepalend voor dat wat zich in de klas afspeelt.

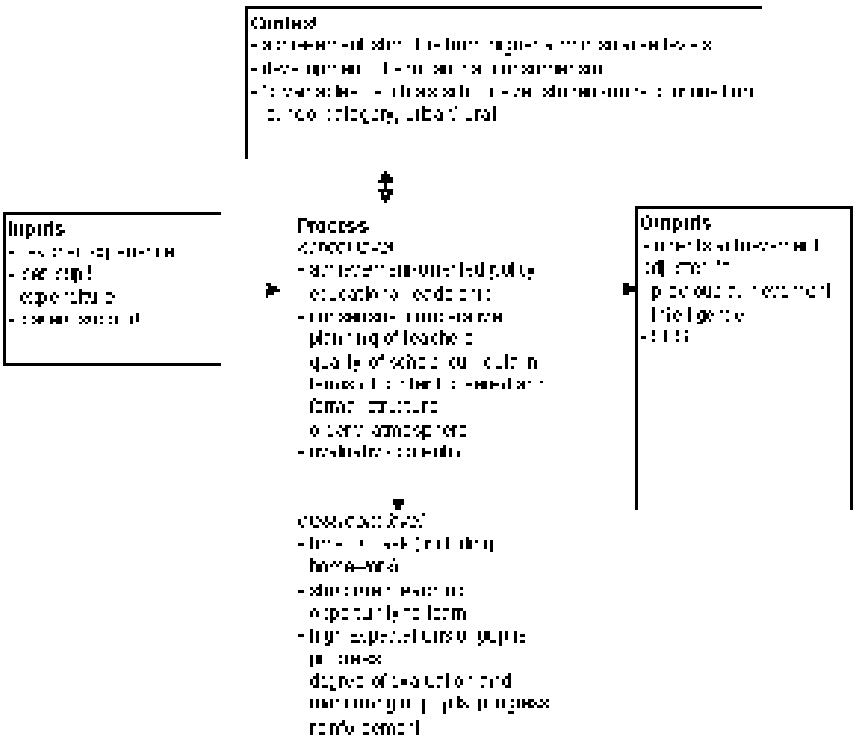
2.3.3 Een integraal schooleffectiviteitsmodel

In het voorgaande is geconcludeerd dat de eenvoud van de diverse modellen (vergelijk Edmonds, 1979) niet tegemoet komt en geen recht doet aan de complexiteit van de onderwijspraktijk. Als reactie daarop zijn diverse modellen voor schooleffectiviteit opgesteld, waarbij met deze complexiteit rekening wordt gehouden, (zie bijvoorbeeld: Scheerens, 1990; Creemers, 1994; Stringfield & Slavin, 1992). Hoewel deze modellen verschillen ten aanzien van het aantal variabelen en de nadruk op de te onderscheiden niveaus, geven ze alle wel het belang aan van de variabelen ‘time on task’, ‘Opportunity To Learn’ en de ‘instructional quality of schooling’. In zijn ‘comprehensive model of educational effectiveness’ laat Creemers (1994) zien dat de onderscheiden variabelen op de diverse niveaus elkaar beïnvloeden en dat variabelen op het niveau van de leerling afhangen van variabelen op klas- en schoolniveau. Door deze afhankelijkheid zichtbaar te maken, maakt hij ook hun onderlinge beïnvloeding en de uiteindelijke bijdrage aan de prestaties van leerlingen zichtbaar.

Als beschrijvingskader voor de ontwikkeling van een procedure voor schoolzelfevaluatie wordt in dit proefschrift uitgegaan van het integraal model voor schooleffectiviteit van Scheerens (1989). Zijn model is mede gebaseerd op onderzoeksresultaten naar schooleffectiviteit. Bovendien is van de in het model opgenomen variabelen vastgesteld dat deze positief correleren met de leerresultaten van leerlingen in kernvakken, zoals rekenen en taal. Het model staat weergegeven in figuur 2.2.

Het model vindt zijn oorsprong in het economisch georiënteerd onderzoek uit de zeventiger jaren. Deze oorsprong verklaart voor een deel de aard van de opgenomen variabelen die genoemd staan onder 'inputs'. In de economische benadering is de vraag van belang welke manipuleerbare 'inputs' de 'outputs' kunnen vergroten. Hoewel in het model van Scheerens uitgegaan wordt van theorievorming uit de economie, worden ook theorieën uit andere disciplines in het model geïncorporeerd, zoals onderwijspsychologie, cognitieve psychologie, contigentiebenadering en onderwijssociologie. Om deze reden wordt het model ook wel een integraal model genoemd. Voorbeelden uit deze andere disciplines zijn de kenmerken die ingaan op het gedrag van de leerkracht, instructieprocedures en organisatie van de lessen. Bovendien worden in het model verschillende organisatieniveaus onderscheiden. Van Petegem (1997) geeft het belang van het model als volgt weer:

- Het model onderscheidt input-, proces-, output- en contextvariabelen.
- Het model vermeldt indicatoren op verschillende niveaus: leerling, klas, school en omgevingrelaties, en maakt interrelaties tussen variabelen op verschillende niveaus mogelijk.
- Het model integreert bevindingen van verschillende soorten effectiviteitsonderzoek, namelijk het onderzoek naar ongelijke kansen in het onderwijs, de economische benadering van het onderwijs, het klassieke schooleffectiviteitsonderzoek, en het onderzoek naar effectieve instructie en effectieve leerkrachten.



Figuur 2.2
Geïntegreerd model voor schooleffectiviteit.
(bron: Scheerens, 1990)

Scheerens (1990) geeft aan dat het integrale schooleffectiviteitsmodel vooral gebruikt dient te worden als een rationale voor de wisselwerking tussen klas-, school- en contextkenmerken. Het is een model waarin organisatorische en curriculaire condities op schoolniveau in samenhang met instructiekenmerken op klasniveau kunnen worden geanalyseerd. Daarnaast kan het model volgens hem gezien worden als een ordeningskader voor de meer inhoudelijke bespreking van kenmerken die van belang zijn voor de effectiviteit van scholen.

2.3.4 Een beschrijvingskader voor het evalueren van scholen

Het uitgangspunt dat EVADOS zich richt op de zelfevaluatie door de school is van belang voor de keuze van variabelen die opgenomen moeten worden in een beschrijvingskader. Hoewel het model van Scheerens goede uitgangspunten voor deze zelfevaluatie biedt, is het wellicht nodig aan dit model extra variabelen toe te voegen. In dit kader kan bijvoorbeeld verwezen worden naar Hendriks (1997), die aangeeft dat recente overzichten van effectiviteitsbevorderende factoren laten zien dat 'de betrokkenheid van ouders' en 'het werkklimaat' factoren zijn die op schoolniveau hun invloed hebben.

Voor een school zijn die variabelen van belang die door haar te manipuleren zijn en een relatie hebben met de effectiviteit van het onderwijs. Informatie over deze vaardigheden en de mogelijkheid deze te beïnvloeden, stelt een school in staat een actief beleid te voeren in haar streven een zo hoog mogelijke kwaliteit van haar onderwijs te bereiken. Naast door scholen te manipuleren variabelen zijn ook variabelen te onderscheiden die weliswaar niet door een school te manipuleren zijn, maar wel van invloed (kunnen) zijn op het prestatieniveau van de school. Voorbeelden hiervan zijn samenstelling van de schoolbevolking, denominatie van de school, bestuursvorm en geografische setting van de school. Om aan te geven hoe een school het doet in vergelijking met andere scholen, zullen referentiegegevens samengesteld moeten worden, die zowel door scholen manipuleerbare als niet-manipuleerbare variabelen bevatten.

In het hiernavolgende zal aan de hand van figuur 2.3 nader ingegaan worden op de voor de effectiviteit van een school relevante proceskenmerken. Deze proceskenmerken worden daarbij onderverdeeld in school- en klaskenmerken. Figuur 2.3 is een kleine bewerking van het door Hendriks (1997) samengesteld overzicht van mogelijke relevante effectiviteitsbevordende input- en procesvariabelen op school- en klasniveau. Hendriks gaat bij haar overzicht uit van het model van Scheerens, aangevuld met variabelen uit eigen onderzoek, recente overzichten en rapporten.

Variabelen op school- en klasniveau

In figuur 2.3 worden op schoolniveau zes en op klasniveau vier variabelen onderscheiden, met per variabele één of meer nadere uitwerkingen. In tegenstelling tot het overzicht van Hendriks is in figuur 2.3 als variabele op klasniveau 'Opportunity To Learn' opgenomen, die in dit proefschrift uitgewerkt wordt als 'Toets Curriculum Overlap'. Op deze variabele, die voor de interpretatie van de toetsresultaten van leerlingen erg belangrijk is, zal in hoofdstuk vier nader worden ingegaan.

School	
Prestatiegerichtheid:	Centraal stellen van basisvaardigheden Hoge verwachtingen
Onderwijskundig leiderschap:	Aandacht schoolleiding voor leerlingen Aandacht schoolleiding voor individuele leerkrachten Aandacht schoolleiding voor team Professionalisering
Samenwerking en overleg:	Frequentie van samenwerking Onderwerpen/aspecten van samenwerking
Evaluatie:	Gebruik van evaluatie-instrumenten Evaluatie van leerlingresultaten
Klimaat:	Ordelijk klimaat School/klassenklimaat Relatie leerkracht-leerling Relatie leerkrachten onderling Werksfeer in de klas
Werkklimaat:	Samenwerking binnen het team Schoolleiding Werkbelasting
Klas	
Effectieve leertijd:	Onderwijstijd op groepsniveau Klassenmanagement
Gestructureerd onderwijs:	Introductie Instructie Verwerking Feedback
Opportunity To Learn:	Toets Curriculum Overlap
Adaptief onderwijs:	Differentiatie Mogelijkheden voor extra onderwijsbehoeften

Figuur 2.3

Overzicht relevante school- en klaskenmerken

(bron: Hendriks, 1997)

In haar publicatie geeft Hendriks aan, dat uit haar onderzoek naar procesvariabelen blijkt dat aan dezelfde variabelen niet altijd dezelfde betekenis wordt toegekend en dat deze in de diverse instrumenten voor zelfevaluatie zeer verschillend geoperationaliseerd worden.
In figuur 2.3 worden op schoolniveau de volgende variabelen onderscheiden: prestatiegerichtheid, onderwijskundig leiderschap, samenwerking en overleg, evaluatie, klimaat en werkklimaat, en op klasniveau: effectieve leertijd, gestructureerd onderwijs, Opportunity To Learn en adaptief onderwijs. Hendriks kent aan deze variabelen de volgende betekenis toe (voor een gedetailleerde beschrijving zie Hendriks, 1997):

Prestatiegerichtheid schoolbeleid
Bij de prestatiegerichtheid van het schoolbeleid is met name het centraal stellen van basisvaardigheden een belangrijk aspect. Hiermee wordt bedoeld dat leerkrachten en schoolleiding gericht zijn op het aanleren van de basisvaardigheden binnen de kernvakken taal en rekenen en op de beheersing daarvan door de leerlingen. Hoge verwachtingen heeft betrekking op het verwachtingsniveau dat leraren en schoolleiding van elkaar en vooral van de leerlingen hebben.

Onderwijskundig leiderschap
De mate waarin de schoolleiding onderwijskundige voorwaarden schept waaronder leerkrachten goed kunnen functioneren, de leerkrachten informeert en stimuleert en hen begeleidt en adviseert.

Samenwerking en overleg
De mate van samenwerking tussen schoolleiding en het team van leerkrachten in bijvoorbeeld teambesprekingen (formele samenwerking) en de mate van samenwerking tussen leerkrachten onderling (functionele samenwerking).

Evaluatie
De registratie van schoolvorderingen in een leerlingvolgsysteem, alsmede de afspraken en/of regels met betrekking tot toetsing en evaluatie.

Klimaat

Onder klimaat vallen aspecten als ordelijk klimaat, schoolklimaat, werkklimaat en klasseklimaat. Klimaatfactoren op schoolniveau betreffen die factoren waar leerlingen niet direct bij betrokken zijn. Ook de samenwerking binnen het team valt onder klimaat, als ook de mate waarin de schoolleiding vertrouwen heeft in de teamleden, bevordert dat met plezier op school gewerkt wordt, gemakkelijk bereikbaar is voor geledingen binnen de school en conflicten probeert uit te praten en/of oproept.

Effectieve leertijd

Hierbij wordt een onderscheid gemaakt naar formele onderwijstijd en daadwerkelijk toegekende lestijd. Formele onderwijstijd heeft betrekking op de hoeveelheid tijd die per klas (per leerjaar) per week op het rooster voor de leer- en vormingsgebieden lezen, taal en rekenen-wiskunde staat aangegeven. De daadwerkelijk toegekende tijd betreft de feitelijke lestijd met betrekking tot de voornoemde leer- en vormingsgebieden.

Gestructureerd onderwijs

Dit betreft de wijze waarop de in een les te onderscheiden vier componenten introductie, instructie, verwerking en feedback uitgewerkt worden. Van deze componenten neemt het onderdeel directe instructie (nastreven duidelijke leerdoelen, structureren van de inhoud, heldere presentatie van de lesstof, stellen van vragen aan leerlingen, onmiddellijk inoefenen van de leerstof na de instructie en evaluatie, terugkoppeling en correctief onderwijs in en na de les) een belangrijke plaats in.

Opportunity To Learn

Deze variabele staat niet in het overzicht van Hendriks. Vanwege het belang ervan is deze variabele in figuur 2.3 alsnog opgenomen. Het begrip 'Opportunity To Learn' (OTL) kent in de literatuur meerdere betekenissen. Eén daarvan betreft de mate waarin de te gebruiken toetsen aansluiten bij het gegeven onderwijs. De Haan, (1992) spreekt van Toets Curriculum Overlap (TCO). In dit proefschrift wordt OTL als TCO opgevat en gedefinieerd als de mate waarin er

overeenstemming is tussen het 'beoogd curriculum' zoals dat geoperationaliseerd is in de toetsing, en het 'geïmplementeerd curriculum' zoals dat vorm krijgt in de lessen.

Adaptief onderwijs
Het kunnen vaststellen van verschillen tussen leerlingen en het afstemmen van het onderwijs op die verschillen, waarbij de verschillen niet mogen leiden tot negatieve consequenties voor leerlingen.

2.3.5 Selectie variabelen per component

In het voorgaande is aangegeven dat EVADOS tot doel heeft scholen te informeren over de kwaliteit van hun onderwijs. Om zich te kunnen vergelijken met andere scholen is het gewenst te corrigeren voor variabelen die niet door scholen te beïnvloeden zijn en waarvan de effecten dan ook niet toe te schrijven zijn aan scholen. Wanneer na correctie voor deze niet-manipuleerbare variabelen nog steeds verschillen in prestaties aanwezig zijn, dan kunnen scholen met behulp van de in de vorige paragraaf besproken school- en klaskenmerken proberen de kwaliteit in positieve zin te beïnvloeden. Welke variabelen opgenomen worden in EVADOS komt in hoofdstuk vier aan bod. In dat hoofdstuk zullen de componenten input, proces en output nader uitgewerkt worden. In hoofdstuk drie wordt eerst nader ingegaan op de functionaliteit van EVADOS.

3 Functionaliteit EVADOS

EVADOS richt zich op de evaluatie van het onderwijs. De term evaluatie behoeft enige toelichting. Bosker, Houtveen en Meijnen (1998) maken een onderscheid tussen (effect-)evaluatie en monitoren (volgen). De functie van EVADOS ligt bij het onderdeel monitoren en niet bij (effect-)evaluatie. Niet EVADOS, maar de school zelf zal moet nagaan of genomen maatregelen effectief waren en correct zijn uitgevoerd. Ook zal EVADOS niet aangeven dat (een ander) beleid nodig is. Het is aan de school zelf te beoordelen of genomen maatregelen effect hebben gehad en of verdere bijstellingen nodig zijn. Als in het verdere verloop van dit proefschrift gesproken wordt over de bijdrage van EVADOS aan de evaluatie van het onderwijs, beperkt de functie van EVADOS zich tot het monitoren en niet tot het doen van uitspraken over het effect van genomen maatregelen. Op welke wijze EVADOS de school behulpzaam kan zijn bij het evalueren van de kwaliteit van haar onderwijs, komt in de volgende paragraaf aan bod.

3.1 Evaluatie van de kwaliteit van de school

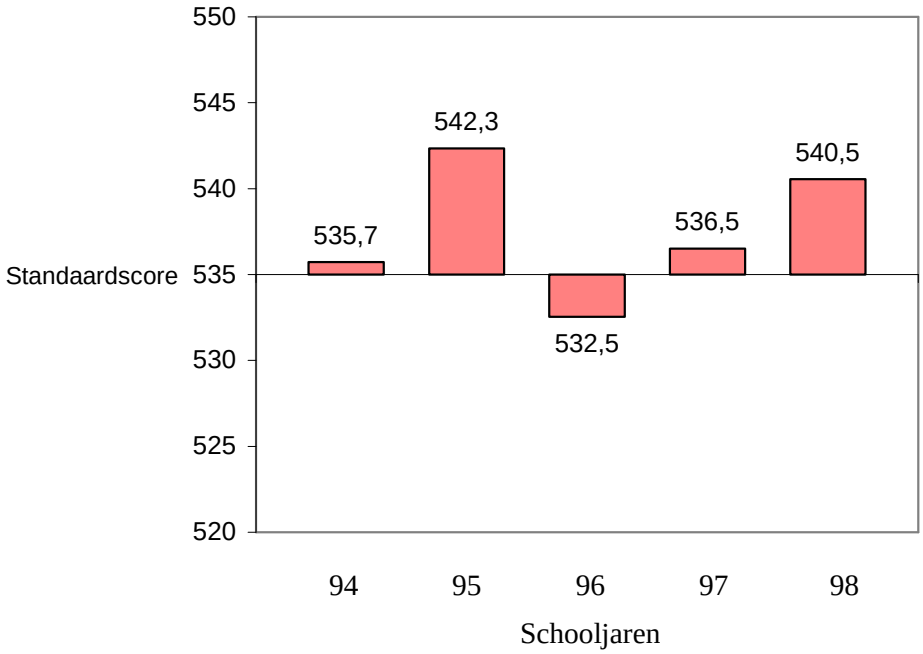
EVADOS dient scholen van informatie te voorzien over het verloop van de prestaties van leerlingen in de tijd op basis waarvan zij de kwaliteit van hun onderwijs kunnen beoordelen. Voorwaarde voor een dergelijke beoordeling is wel dat de prestaties op te gebruiken toetsen met elkaar in verband gebracht kunnen worden. Dit kan bijvoorbeeld door de toetsresultaten van leerlingen te transformeren naar standaardscores, zoals bij de Eindtoets basisonderwijs plaatsvindt, of door de behaalde resultaten te transformeren naar een met behulp van de itemresponstheorie geconstrueerde vaardigheidsschaal, zoals bij de

toetsen uit het Cito-leerlingvolgsysteem. Zowel de Eindtoets basisonderwijs als de toetsen uit het Cito-leerlingvolgsysteem lenen zich in principe voor een vergelijking van de resultaten in de tijd. Er is echter een groot verschil. De Eindtoets basisonderwijs richt zich op het einde van de basisschool (groep acht) en geeft geen informatie over de ontwikkeling van leerlingen tijdens hun schoolloopbaan. Het Cito-leerlingvolgsysteem daarentegen verschaft wel informatie over het verloop van het onderwijsproces van leerlingen tijdens de basisschoolperiode en biedt daardoor scholen de mogelijkheid voor deze leerlingen gepaste maatregelen te nemen. De Eindtoets basisonderwijs en het Cito-leerlingvolgsysteem hebben gemeenschappelijk dat beide niet verwijzen naar mogelijke verklarende factoren.

Bij het beoordelen van de kwaliteit van het door haar verzorgde onderwijs kan een school uitgaan van een intern referentiekader of van een extern referentiekader. In het eerste geval maakt de school gebruik van de door haar in het verleden behaalde resultaten en in het tweede geval van de resultaten behaald door andere scholen. Beide referentiemogelijkheden worden besproken. Bij de bespreking van het interne referentiekader zal nader ingegaan worden op een vergelijking van de resultaten van leerlingen in de tijd aan het einde van het basisonderwijs en een vergelijking in de tijd tijdens de schoolloopbaan van de leerlingen.

3.1.1 De eigen school als referentiekader

Een voorbeeld van een vergelijking op basis van resultaten van leerlingen op de Eindtoets basisonderwijs staat in figuur 3.1. In deze figuur zijn voor een bestaande school X over een periode van vijf jaren (1994 t/m 1998) de resultaten op deze toets weergegeven.

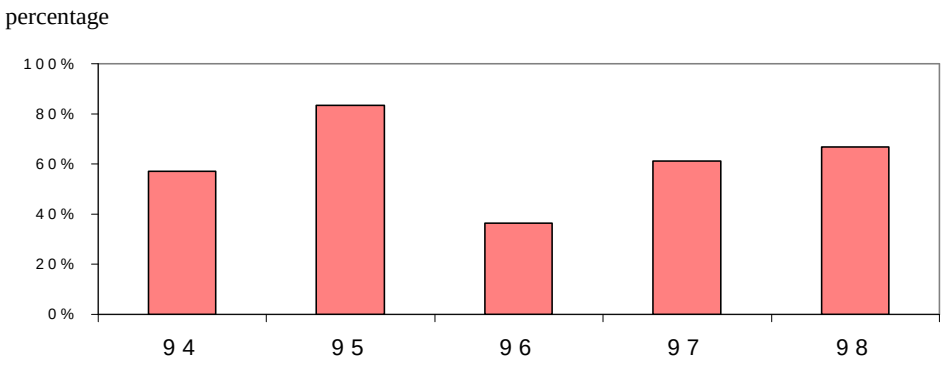


Figuur 3.1
Gemiddelde standaardscores school X
in de periode '94 – '98

De kolommen geven de in die jaren door de school behaalde gemiddelde standaardscores weer. Zo was het gemiddelde voor de school 535,7 (14 leerlingen) in 1994 en 542,3 (12 leerlingen) in 1995. De Eindtoets basisonderwijs hanteert een schaal van 501 tot 550, met als gemiddelde 535 en een standaarddeviatie van 10. De verticale as in figuur 3.1 geeft deze standaardscores weer. Uit figuur 3.1 is af te leiden dat de school in 1996 onder het landelijk gemiddelde scoorde.

De resultaten van een school op de Eindtoets basisonderwijs kunnen ook uitgedrukt worden in het percentage leerlingen met een standaardscore boven het landelijke gemiddelde (zie figuur 3.2). Uit deze figuur is af te lezen dat in 1994 57% van de leerlingen boven het landelijk gemiddelde scoorde. In 1995 was dit 83%. Bij de interpretatie van de figuren 3.1 en 3.2 dient men het geringe aantal leerlingen (range 11 tot 18) in ogenschouw te nemen. Een geringe wijziging in

samenstelling van deze leerlingpopulatie kan al gauw leiden tot een ander beeld. Bovendien is voor de interpretatie van de resultaten op de Eindtoets basisonderwijs het gegeven van belang dat niet steeds dezelfde leerlingen met elkaar vergeleken worden. Het zijn ieder jaar steeds andere leerlingen die deelnemen aan de Eindtoets basisonderwijs.

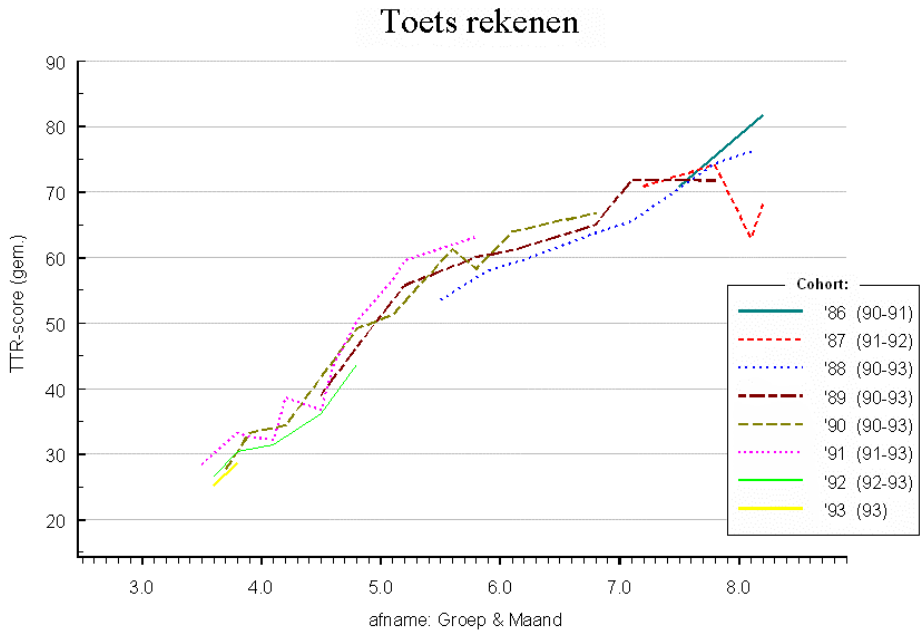


Figuur 3.2
Percentage leerlingen school X dat presteert
boven het landelijk gemiddelde

Hoewel een vergelijking als weergegeven in de figuren 3.1 en 3.2 een school informeert over de resultaten op de Eindtoets basisonderwijs in vergelijking met andere jaren, heeft deze toch een aantal beperkingen. Reeds genoemd is de onmogelijkheid om het onderwijs nog aan deze schoolverlaters aan te passen. Bovendien wordt bij een dergelijke vergelijking niet verwezen naar mogelijke verklarende factoren. EVADOS probeert aan die beperkingen tegemoet te komen. Door prestaties van leerlingen in de tijd te volgen, wordt een school geïnformeerd over de ontwikkeling van (groepen van) leerlingen en kan bij een afwijkend verloop (bijvoorbeeld in vergelijking met voorgaande jaren) maatregelen ter verbetering genomen worden. Deze weergave in de tijd en de mogelijkheid die scholen geboden wordt maatregelen te nemen, vormen de kern van EVADOS, en maken EVADOS in die zin uniek. Het principe van het kunnen volgen van de ontwikkeling van leerlingen in de tijd en de mogelijkheid

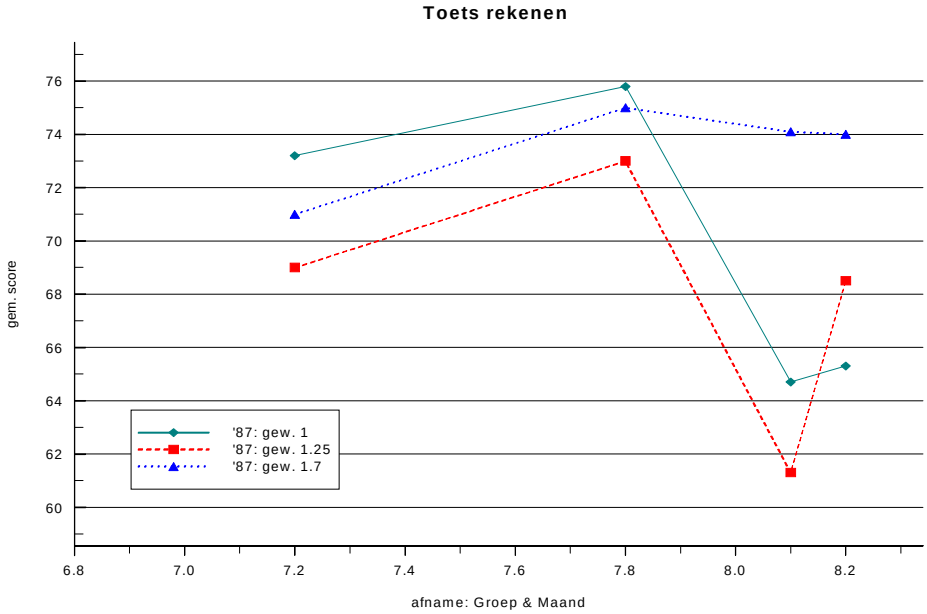
scholen te ondersteunen bij hun zoeken naar mogelijke verklarende factoren, zal aan de hand van de figuren 3.3 en 3.4 toegelicht worden.

Figuur 3.3 geeft een vergelijking voor de Tempo Test Rekenen (TTR) (Vos, 1992) van een bestaande basisschool weer. De afname van deze toets vindt een aantal malen per jaar plaats. In deze figuur zijn de ontwikkelingen van de leerlingen van diverse cohorten tijdens hun schoolloopbaan in kaart gebracht. Onder een cohort wordt hier een groep leerlingen verstaan die aan het begin van het schooljaar het onderwijs aanvangt in groep 3 en vervolgens zonder doubleren doorstroomt naar groep 8. De gemiddelde resultaten van de leerlingen op de diverse afnametijdstippen zijn door een lijn met elkaar verbonden.
Op de horizontale as van figuur 3.3 staat vermeld wanneer de afnamen plaatsvonden. De getallen 3 t/m 8 hebben betrekking op de te onderscheiden groepen van de basisschool. De verdeling tussen twee getallen geeft de schoolmaanden in het desbetreffende schooljaar weer, waarbij het schooljaar verdeeld is in tien (onderwijs)maanden. De gemiddelde score van het cohort is af te lezen op de verticale as. De hellingshoek van de afzonderlijke lijnen is een indicatie van de vorderingen van de desbetreffende cohorten: hoe steiler de lijn, des te groter is de gemiddelde toename in (reken)vaardigheid.



Figuur 3.3
Ontwikkelingspatroon rekenvaardigheid diverse cohorten gemeten met de TTR

Uit figuur 3.3 kan een school afleiden of de resultaten van een bepaald cohort in de tijd overeenkomen met het patroon dat voor de school gebruikelijk is (intern referentiekader). Bij opvallende resultaten - in dit geval de prestaties van cohort ‘87 in groep 8 - kan een school onderzoeken wat (een) mogelijke reden(en) daarvoor zou(den) kunnen zijn. Bijvoorbeeld door relevante leerlingkenmerken uit het schooladministratiepakket te koppelen aan de toetsresultaten van leerlingen. In figuur 3.4 is dit gedaan voor het leerlinggewicht.



Figuur 3.4
Resultaten cohort '87 in groep 7 en 8 op de TTR
uitgesplitst naar leerlinggewicht

In figuur 3.4 staan op de horizontale as de afnamemomenten en op de verticale as de behaalde gemiddelde scores op de TTR. Omdat figuur 3.4 betrekking heeft op de terugval in resultaten in groep acht, is slechts een beperkt deel van de schaal uit figuur 3.3 uitvergroot. Uit figuur 3.4 is af te lezen dat zowel in groep zeven als in groep acht de toets tweemaal afgenomen is, zij het dat de afnamen in groep acht snel achter elkaar plaatsvonden. Het blijkt dat het minder goede resultaat van de leerlingen uit groep acht vooral het gevolg is van de slechtere prestaties van de 1,0- en 1,25-leerlingen, en niet - zoals op voorhand wellicht verwacht zou worden - van de 1,7-leerlingen. Informatie op basis waarvan de school gerichter kan nagaan wat een mogelijke verklaring zou kunnen zijn. Door patronen op deze wijze nader uit te splitsen, verkrijgt een school informatie over groepen van leerlingen tijdens hun schoolloopbaan en heeft zij de mogelijkheid maatregelen te nemen. Of de genomen maatregelen tot het gewenste effect

hebben geleid, kan een school uit een volgend meetmoment of uit volgende meetmomenten afleiden.

Dat een school zichzelf als referentiekader neemt, is voor een school met name zinvol wanneer de populatie van de school afwijkt van andere scholen, en de school op basis daarvan structureel hogere of lagere scores verwacht. School-eigen standaarden kunnen afwijken van landelijke standaarden. Het is mogelijk dat een op basis van een intern referentiekader hoog scorende school, landelijk gezien (nog steeds) onder het gemiddelde scoort, hetgeen kan impliceren dat bepaalde doelstellingen niet gehaald worden. Het is voor een school dus van belang ook over een extern referentiekader te kunnen beschikken.

3.1.2 Extern referentiekader

Een extern referentiekader biedt een school de mogelijkheid de kwaliteit van haar onderwijs te vergelijken met andere scholen of met een vooraf vastgelegd criterium, zoals bijvoorbeeld de kerndoelen basisonderwijs. Bij een vergelijking met andere scholen is het gewenst voor bij aanvang aanwezige verschillen te corrigeren, omdat het voorbijgaan aan deze verschillen tot verkeerde conclusies zou kunnen leiden. Zo is het wellicht op voorhand al mogelijk aan te geven dat scholen met veel allochtone leerlingen minder goed zullen presteren dan scholen met alleen autochtone leerlingen met hoog opgeleide ouders. Een vergelijking tussen deze twee in populatie zo verschillende scholen doet geen uitspraak over de kwaliteit van het geboden onderwijs. Of anders geformuleerd: een dergelijke vergelijking laat niet zien wat de specifieke bijdrage van de school is aan de ontwikkeling van de leerlingen. Wel laat zo'n vergelijking zien dat leerlingen van de ene school gemiddeld hoger of lager presteren dan leerlingen van de andere school. Ook de financiële middelen waarover een school kan beschikken, kunnen de prestaties van leerlingen positief beïnvloeden. Te denken daarbij valt aan de mogelijkheid om extra leermiddelen aan te schaffen, extra aandacht te geven aan leerlingen waar dat gewenst is door groepen te splitsen, het aanstellen van een remedial teacher en/of een interne begeleider.

Een vergelijking op basis van uitsluitend output is zinvol als er sprake is van een criteriumgerichte vergelijking. Centraal staat dan de vraag of een school gemiddeld genomen boven of onder een van te voren gedefinieerd criterium presteert. Mocht een school boven of onder een van te voren gedefinieerd criterium presteren, dan impliceert dat overigens niet dat dit voor alle leerlingen binnen de school geldt (zie Scheerens & Bosker, 1997). Het hanteren van een criterium gerichte vergelijking geeft uitsluitsel over het (gemiddeld) eindniveau dat een school met haar leerlingen behaalt. Na correctie voor bij aanvang aanwezige verschillen kan blijken dat de output (de kwaliteit) van het onderwijs op een school hoog is in vergelijking met andere (vergelijkbare) scholen. Dat wil echter niet zeggen dat de school ook het gewenste eindniveau bereikt. Een aspect dat met name ook voor ouders belangrijk kan zijn, omdat een vergelijking met een extern criterium aangeeft of hun kinderen over het gewenste (start)niveau voor een vervolgstudie beschikken. Als voorbeeld van zo'n gewenst eindniveau kunnen de reeds eerder aangehaalde kerndoelen basisonderwijs gezien worden. EVADOS gaat niet in op de vergelijking met een extern criterium. EVADOS biedt scholen de mogelijkheid de ontwikkeling van leerlingen te vergelijken met de behaalde resultaten uit voorgaande jaren en met de prestaties van andere vergelijkbare scholen. Bij deze laatste vergelijking zullen relevante input- en procesfactoren betrokken worden. Wil een basisschool nagaan in hoeverre de voor het basisonderwijs geformuleerde kerndoelen behaald zijn, dan zal zij gebruik moeten maken van een andere procedure of instrument.

EVADOS informeert scholen over de ontwikkeling van vergelijkbare scholen. Dit impliceert dat in deze vergelijking input- en procesfactoren opgenomen zullen zijn. Voor een juiste beoordeling over de bijdrage van een school aan de ontwikkeling van haar leerlingen, dient voor deze factoren gecorrigeerd te worden. Het komt voor dat uitspraken over de 'kwaliteit' van het geboden onderwijs gedaan worden zonder (in voldoende mate) rekening te houden met de omstandigheden waaronder deze 'kwaliteit' bereikt is. Indien de vergelijking plaatsvindt op basis van ruwe, niet gecorrigeerde scores, spreekt men van 'gross achievement measures'. Men spreekt van 'net achievement measures' als in een

vergelijking wel input- en procesfactoren betrokken worden. In het hiernavolgende wordt op deze twee vergelijkingswijzen ingegaan.

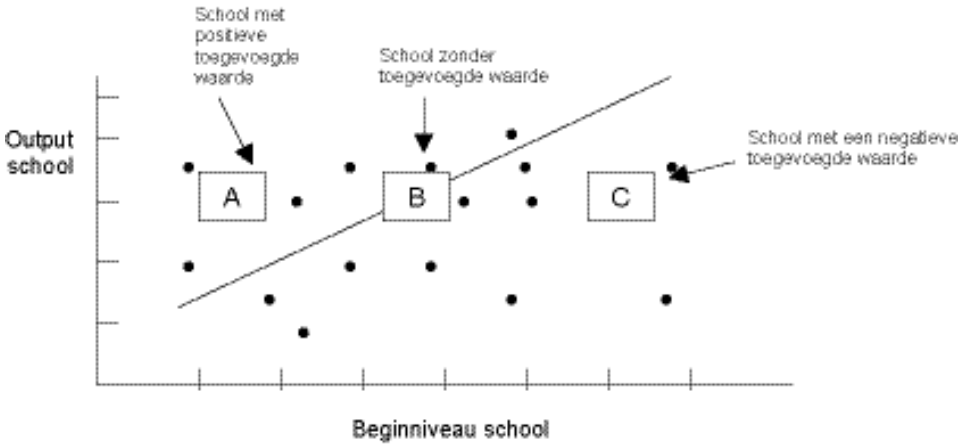
Gross achievement measures

Een voorbeeld van gross achievement zijn de in Engeland ingevoerde 'league tabels'. In deze vergelijking worden scholen gerangordend op basis van de prestaties van hun 11-jarige leerlingen op een landelijke toets. De 'league tabels' zijn openbaar en zelfs op het internet te raadplegen. Een dergelijk lijst suggereert dat het onderwijs op de scholen die bovenaan op de lijst staan van een betere kwaliteit is dan het onderwijs op de scholen die onderaan staan. Dit hoeft natuurlijk niet zo te zijn. In het voorgaande is al betoogd dat achtergrondvariabelen van leerlingen een belangrijke rol kunnen spelen in het bereiken van bepaalde doelstellingen. Kritiek op de 'league tables' is dan ook alom te horen. Ook in Nederland heeft een dergelijke discussie naar aanleiding van een publicatie van Dronkers in het dagblad Trouw (Agerbeek, Hageman & Lakmaker, 1997) plaatsgevonden. In het door velen gewraakte artikel rangorde Dronkers scholen uit het voortgezet onderwijs op basis van een door hem gegeven rapportcijfer dat gebaseerd was op een aantal criteria, zoals het examencijfer op een aantal vakken, het aantal geslaagden, het percentage zittenblijvers, het percentage allochtonen en de grootte van de uitstroom. De kritiek op Dronkers betrof met name het ontbreken van een aantal relevante variabelen in zijn analyses, zoals bijvoorbeeld het sociaal milieu en met name ook het instroomniveau van de leerling.

Net achievement measures

In het voorgaande is aangegeven dat achtergrondvariabelen van belang zijn voor het vaststellen van het aandeel van de school in de toename in leerprestaties. Deze factoren zijn mede van invloed op de door een school te behalen prestaties en dienen bij de interpretatie van deze prestaties meegenomen te worden. Nagegaan zal moeten worden onder welke omstandigheden het onderwijs heeft plaatsgevonden en wat de invloed daarvan geweest kan zijn. Een uitspraak over de kwaliteit van het onderwijs op basis van leerresultaten van leerlingen moet (zoveel mogelijk) gebaseerd zijn op de feitelijke bijdrage van een school aan het

behalen van deze resultaten en niet toegeschreven kunnen worden aan leerlingkenmerken zoals bijvoorbeeld het startniveau of taalvaardigheid van leerlingen aan het begin van een opleiding. Voor deze verschillen in leerlingkenmerken tussen scholen dient gecorrigeerd te worden als het erom gaat vast te stellen wat het aandeel van de school in de toename van leerprestaties is. In de literatuur spreekt men in dit kader van de toegevoegde waarde (value added). Toegevoegde waarde wordt daarbij omschreven als de waarde die een school toevoegt aan de ontwikkeling van leerlingen, nadat is gecontroleerd voor contextuele kenmerken en het beginniveau van leerlingen (zie Willms, 1992; Willms & Kerckhoff, 1995; Hill, 1995; Scheerens & Bosker, 1997; Fitz-Gibbon, 1997). Onderstaande figuur geeft het principe van toegevoegde waarde weer.



Figuur 3.5
Principe toegevoegde waarde
(bewerking van Fitz-Gibbon, 1997)

Figuur 3.5 is een grafische illustratie van de 'toegevoegde waarde' van een school. In deze figuur staat de output van een aantal scholen op een bepaalde toets weergegeven. De zwarte rondjes stellen de scholen voor. Drie scholen (A, B en C) zijn eruit gelicht. Op de horizontale as staat het beginniveau van de scholen. De regressielijn geeft de voorspelde relatie aan tussen het beginniveau en de output van een school. Zo is uit deze regressielijn af te lezen dat scholen

met een hoog beginniveau naar verwachting tot een hogere output komen dan scholen met een laag beginniveau. Dat dit niet altijd zo hoeft te zijn, laat figuur 3.5 zien. Hoewel de drie scholen A, B en C alle dezelfde output hebben, is hun relatieve progressie verschillend. School A blijkt beter te scoren dan op basis van het beginniveau verwacht mag worden. De werkelijke output van deze school minus de voorspelde output is positief. Bij school A kan gesproken worden van een positieve toegevoegde waarde. Bij school C blijkt het verschil tussen de werkelijke output en de voorspelde output negatief te zijn. Deze school kent daarom een negatieve toegevoegde waarde. Bij school B ten slotte is de werkelijke score gelijk aan de voorspelde, hetgeen impliceert dat de toegevoegde waarde van deze school nul is.

Bij het vaststellen van de toegevoegde waarde is het belangrijk vast te stellen op basis van welke variabelen de voorspelde waarde berekend gaat worden. Of anders gezegd: voor welke variabelen de outputresultaten gecorrigeerd gaan worden. Scheerens en Bosker (1997) en Hill (1995) noemen de volgende drie mogelijkheden:

- correctie voor alleen achtergrondvariabelen zoals bijvoorbeeld sociaal economische status, geslacht, leeftijd en etniciteit ('unpredicted student achievement');
- correctie voor het beginniveau van de leerlingen ('learning gain');
- correctie voor zowel het beginniveau van de leerlingen als voor achtergrondvariabelen ('net progress achievement' of 'unpredicted learning gain').

Welke keuze men maakt voor het corrigeren van de output hangt mede van de vraagstelling af. Willms (1992) illustreert dit aan de hand van wat hij noemt het type A- en het type-B-effect. Bij het type A-effect staat de vraag centraal hoe een leerling het op een specifieke school doet in vergelijking met andere scholen. Anders geformuleerd: zouden de prestaties van deze leerling anders zijn geweest als hij naar een andere school zou zijn gegaan? Deze vergelijking komt overeen met wat eerder ‘unpredicted student achievement’ genoemd is. Bij deze vergelijking wordt alleen gecorrigeerd voor leerlingkenmerken en niet voor schoolkenmerken.

Indien de onderlinge vergelijking van scholen ten aanzien van het door hen gevoerde schoolbeleid en hun schoolpraktijk centraal staat, is er sprake van het type B-effect. Bij dit effect vindt niet alleen een correctie plaats voor leerlingkenmerken, maar ook voor niet beïnvloedbare schoolkenmerken. Een dergelijke vergelijking maakt zichtbaar wat de specifieke bijdrage van de school is op de verkregen output. De te constateren verschillen tussen scholen zijn dan toe te schrijven aan het beleid en de organisatie van de scholen. Het type A-effect geeft informatie over de resultaten van leerlingen die een bepaalde school bezoeken waarbij gecorrigeerd is voor leerlingkenmerken. Het type B-effect geeft informatie over hoe een school presteert in vergelijking met andere scholen die een vergelijkbare samenstelling kennen en in een vergelijkbare sociale en economische context opereren. Voor scholen - en ook voor de overheid ingeval van het verantwoordelijk stellen van een school voor zijn resultaten - is met name deze laatste vergelijking belangrijk. Het type A-effect daarentegen komt meer tegemoet aan de informatiebehoefte van ouders. Zij zullen met name geïnteresseerd zijn in de vorderingen die zij van hun kinderen in een bepaalde school mogen verwachten. Voor hen zal het eindresultaat tellen en zij zullen minder geïnteresseerd zijn in allerlei redenen (verklarende variabelen) waarom het resultaat bij deze school anders is dan bij een andere school. Ouders zullen geïnteresseerd zijn in die school die de meeste bijdrage levert aan de ontwikkeling van hun kinderen, ongeacht de sociale en economische context waarin een school opereert of moet opereren. Ouders willen weten of hun kind beter naar school A dan naar school B kan gaan.

3.2 Toegevoegde waarde

In het voorgaande is 'toegevoegde waarde' omschreven als de waarde die een school toevoegt aan de ontwikkeling van leerlingen, nadat is gecontroleerd voor contextuele kenmerken en het beginniveau van leerlingen. Uit de bespreking van het type A- en het type B-effect is aangegeven dat het van belang is vast te

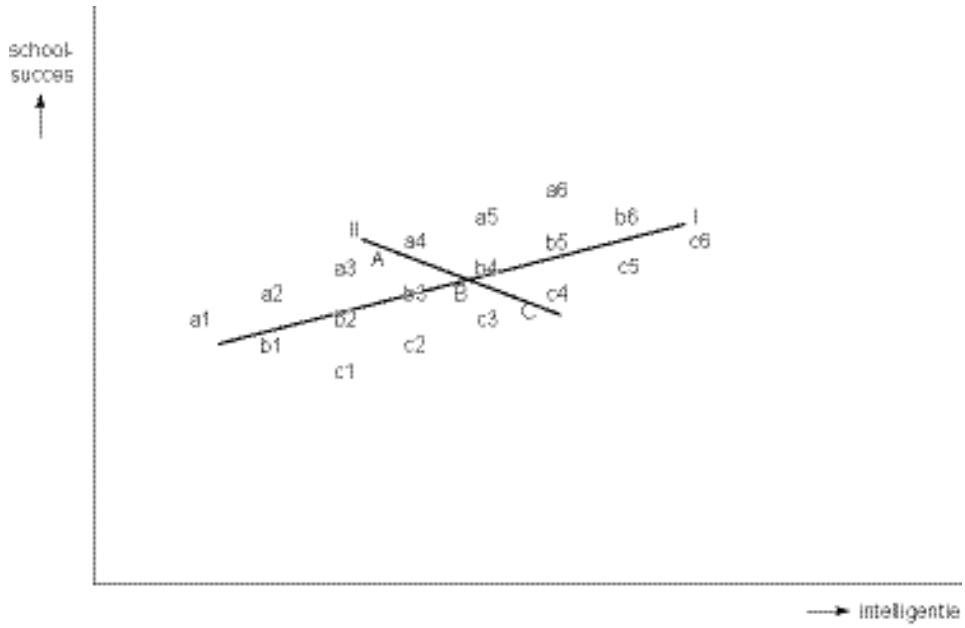
stellen waarover men een uitspraak wil doen, omdat op basis daarvan keuzes gemaakt moeten worden voor relevante variabelen die in de berekening van de toegevoegde waarde meegenomen moeten worden.

Aangezien EVADOS de school als doelgroep heeft, zal nagegaan moeten worden welke informatie scholen optimaal informeert over de kwaliteit van hun onderwijs. Voor een school is het belangrijk haar prestaties (kwaliteit) te kunnen vergelijken met voor haar vergelijkbare scholen (vergelijk het type B-effect). Een school zal moeten weten hoe andere scholen het doen (extern referentiekader) om op basis van deze informatie haar beleid te kunnen evalueren en zo mogelijk haar onderwijs te kunnen optimaliseren.

Berekenen toegevoegde waarde

Uitspraken over de toegevoegde waarde hebben betrekking op de school. Bij de berekening ervan moet nagegaan worden welke analysetechnieken zich daar het beste voor lenen. Dat er verschillende mogelijkheden zijn om een uitspraak op schoolniveau te doen, laten Dronkers (Agerbeek e.a., 1997) en Veenstra, Dijkstra, Peschar en Snijders (1998) zien. Dronkers heeft voor het eerder aangehaalde artikel in Trouw, waarin hij rapportcijfers vaststelt voor scholen uit het voortgezet onderwijs, gebruik gemaakt van op schoolniveau geaggregeerde gegevens. Zijn werkwijze onvervond veel kritiek. Zo gaven Veenstra e.a. (1998) aan dat Dronkers gebruik had moeten maken van multilevel analyses, omdat daarbij rekening gehouden wordt met de hiërarchische structuur van de gegevens - school en leerling - en variantiecomponenten per niveau geschat kunnen worden. Zij verwijzen daarbij naar Bryk en Raudenbush (1992) en Van den Eeden en Meijnen (1990). Bij multilevel analyses vindt een controle plaats voor de samenstelling van de leerlingenpopulatie en de nauwkeurigheid van de resultaten. Aan deze controle voor de samenstelling van de leerlingpopulatie gaat Dronkers voorbij. Door deze keuze is het onderzoek van Dronkers geen onderzoek naar de toegevoegde waarde van scholen, maar beantwoordt hij de vraag hoe scholen presteren ten opzichte van andere scholen. Door het aggregeren van leerlinggegevens op schoolniveau, neemt Dronkers aan dat leerlingen binnen eenzelfde school allen gelijk zijn. Bovendien blijft door een

dergelijke aggregatie de vraag onbeantwoord of eventueel gevonden verschillen moeten worden toegeschreven aan de ‘treatments’ (het onderwijs zoals dat op de school heeft vormgekregen) of aan vooraf bestaande verschillen tussen de groepen (Bosker, 1990). Bij het schatten van schooleffecten moet met instroomverschillen tussen scholen rekening gehouden worden. Bosker (1990) laat dit zien met een ideaaltypische grafiek zoals afgebeeld in figuur 3.6. Zowel de figuur als de beschrijving is ontleend aan Bosker.



Figuur 3.6
Mogelijk effect van correctie voor covariaten op leerling-niveau na aggregatie tot schoolniveau (Bosker, 1990, p. 39)

In figuur 3.6 geven de punten a1 tot en met a6 de scores van de leerlingen van school a weer. Op vergelijkbare wijze zijn de scores van de leerlingen van school b en school c afgebeeld. De punten A, B en C symboliseren de scores van de leerlingen na aggregatie. De regressie van het schoolgemiddelde onderwijs-succes op het schoolgemiddelde IQ is getekend in lijn II. De schoolgemiddelden liggen precies op de regressielijn, zodat de correctie voor de geaggregeerde IQ-

data perfect is: de scholen verschillen na controle niet in hun effectiviteit. Merk op dat lijn II aangeeft dat voor geaggregeerde scores geldt dat hoe hoger het IQ is des te lager het schoolsucces is. Het ligt in de rede het omgekeerde te verwachten.
Uitgaande van het gemiddelde van de metingen (a, b en c) van de leerlingen binnen een school, zal de regressielijn voorgesteld worden door lijn I. School c is dan ineffectief, aangezien alle leerlingen beneden verwachting presteren. Alle leerlingen van school c liggen immers onder de regressielijn. School b heeft evenzoveel leerlingen boven als onder de gepoolde regressielijn liggen, zodat deze school als gemiddeld effectief getypeerd kan worden. School a is zeer effectief: alle leerlingen liggen boven de regressielijn.

Een tweede reden voor schoolzelfevaluatie om niet uit te gaan van op schoolniveau geaggregeerde leerlinggegevens is het verwaarlozen of negeren van de binnengroepsvariantie. Aggregatie op schoolniveau ontneemt de mogelijkheid vast te stellen of scholen in hun effecten interacteren met leerlingkenmerken (Bosker, 1990). Zo is het mogelijk dat een school wel effectief is voor leerlingen uit de lagere sociale strata, maar niet voor leerlingen uit de hogere sociale strata.

Uit de aangehaalde voorbeelden van Dronkers en Bosker kan geconcludeerd worden dat verschillende statistische modellen tot geheel andere conclusies kunnen leiden. Ook Veenstra e.a. (1998) laten in hun replicatie van het Trouwonderzoek zien dat er verschillen tussen de resultaten op basis van geaggregeerde data en multilevel analyses bestaan. In hun onderzoek maken zij gebruik van dezelfde analysemethode die Dronkers heeft toegepast en van multilevel analyses. Omdat zij bij hun analyses ook de invloed van achtergrondkenmerken wilden meenemen, hebben zij gebruik gemaakt van een dataset uit het cohortonderzoek VOCL’89. Bij de door Dronkers gebruikte inspectiegegevens zijn geen achtergrondkenmerken beschikbaar
Per analysemethode hebben zij twee varianten in hun onderzoek meegenomen. Analoog aan de werkwijze van Dronkers hebben zij in eerste instantie alleen gecorrigeerd voor het percentage allochtone leerlingen van de school. In tweede instantie hebben zij extra correcties toegepast. Op leerlingniveau zijn dit etnici-

teit, opleiding vader en de aanvankelijke leerprestaties. Op schoolniveau het percentage allochtone leerlingen, het gemiddeld opleidingsniveau van de vader en de gemiddelde prestaties bij instroom. Zij concluderen dat het corrigeren voor extra variabelen in de methode Dronkers weinig uitmaakt voor de scores van scholen zoals Dronkers deze in het Trouw-onderzoek aan hen toekent. Wel concluderen zij dat de meer uitgebreide controles bij het gebruik van multilevel analyses tot aanzienlijke verschillen in scores van scholen leiden. In hun eindoordeel pleiten zij voor een multilevel analyse, omdat alleen in deze methode adequaat kan worden gecontroleerd voor de samenstelling van de leerlingpopulatie. Bovendien geven zij aan - daarbij verwijzend naar Bosker (1990) - dat de resultaten van een multilevel analyse nauwkeuriger zijn dan de uitkomsten van een regressieanalyse en ook een meer valide beeld geven van de effectiviteit van een school.

3.3 Keuze van de analyses

In hoofdstuk 2 is een beschrijvingskader besproken dat laat zien met welke variabelen op school- en klasniveau een school mogelijk de kwaliteit van haar onderwijs kan beïnvloeden. EVADOS heeft tot doel scholen te informeren over de wenselijkheid daarvan. Belangrijke indicatoren daarvoor zijn de resultaten van leerlingen in de tijd, referentiegegevens, zowel interne als externe, en de specifieke bijdrage (de toegevoegde waarde) van de school.

In de analyses zullen de resultaten van groepen van leerlingen in de tijd weergegeven worden. De weergave zal zich richten op reguliere cohorten, waaronder leerlingen verstaan worden die hun onderwijs in het begin van het schooljaar aanvangen in groep drie en zonder vertraging doorstromen naar groep acht. Leerlingen die behoren tot de niet-reguliere cohorten (waaronder bijvoorbeeld zittenblijvers en zij-instromers) worden in de analyses niet meegenomen. De reguliere en niet-reguliere cohorten vormen samen de populatie van de school.

De gedachte is dat een school haar kwaliteit het beste kan afleiden uit de resultaten van de reguliere cohorten, omdat de niet-reguliere cohorten deels een ander onderwijsaanbod hebben genoten. EVADOS zal scholen de mogelijkheid bieden hun resultaten uit te splitsen naar doelgroepen met eigen kenmerken, waarvan geslacht en leerlinggewicht twee voorbeelden zijn.

Een school zal haar resultaten (in een bepaald jaar) kunnen spiegelen aan referentiegegevens. Deze referentiegegevens zullen zowel een intern als een extern karakter hebben. Het interne referentiekader wordt gevormd door de door de school behaalde resultaten in voorgaande jaren. Het extern referentiekader wordt gevormd door andere scholen. Afhankelijk van de vergelijkingsbasis kunnen dat scholen uit een bepaalde regio of gemeente zijn. Ook een landelijk referentiekader behoort tot de mogelijkheden.

EVADOS zal bij het vaststellen van het aandeel van de school in de toename van de vaardigheid bij leerlingen uitgaan van het principe van ‘net achievement’. Bij de berekening van deze toename zal voor een aantal achtergrondvariabelen gecorrigeerd worden. De inputkenmerken zijn opgeslagen in de op scholen aanwezige schooladministratiepakketten en behoeven niet apart verzameld te worden. In het volgende hoofdstuk zal hier nader op ingegaan worden. Voor zover scholen procesvariabelen willen betrekken in de beoordeling van de kwaliteit van het onderwijs, zullen zij gebruik moeten maken van andere procedures of instrumenten. Het ontwikkelen van dergelijke instrumenten of procedures valt buiten het kader van dit proefschrift. Een uitzondering wordt gemaakt voor de variabele ‘Toets Curriculum Overlap’. In paragraaf 2.3.4 is bij de bespreking van procesvariabelen het belang van TCO voor de interpretatie van toetsresultaten van leerlingen aangegeven. Vanwege het directe belang van TCO voor de interpretatie van de toetsresultaten is aan deze variabele wel aandacht besteed. Op de ontwikkeling van een instrument TCO zal in hoofdstuk vier nader ingegaan worden.

4 Drie componenten van het CIPO-model nader uitgewerkt

In hoofdstuk 2 is op basis van resultaten van het schooleffectiviteitsonderzoek geconcludeerd dat diverse factoren de effectiviteit van een school bepalen. Deze factoren zijn gecategoriseerd in de componenten Context, Input, Proces, en Output van het CIPO-model. In dit hoofdstuk worden deze componenten nader uitgewerkt. Hoewel ook de Context waarin het onderwijs zich afspeelt een rol van betekenis kan spelen, wordt hierop niet verder ingegaan. Contextvariabelen zijn voor een school in de regel een gegeven en niet manipuleerbaar.

Informatie over de component Input is opgeslagen in schooladministratiepakketten. In dit hoofdstuk wordt in paragraaf 4.2 exemplarisch één schooladministratiepakket besproken. Bij deze bespreking komt de gegevensstructuur van dit pakket aan bod en wordt ingegaan op de gegevens die in een dergelijk pakket opgeslagen (kunnen) worden.
In hoofdstuk 2 werd de component Proces uiteengelegd in een aantal variabelen op school- en klasniveau. Eén van deze variabele was Opportunity to Learn. In paragraaf 4.3 wordt deze variabele nader uitgewerkt. Als outputvariabele is gekozen voor de resultaten op de toetsen van het Cito-LVS. In paragraaf 4.1 komt een bespreking van het Cito-LVS aan bod.

4.1 Output

De resultaten van leerlingen op toetsen uit het Cito-LVS worden hier opgevat als een indicatie voor de kwaliteit van het onderwijs. Deze toetsen stellen de gebruiker in staat de prestaties van leerlingen in de tijd te volgen en deze vervolgens te

vergelijken met een intern en extern referentiekader. Met deze eigenschap onderscheiden de toetsen uit het Cito-LVS zich niet alleen van vele andere voor het onderwijs ontwikkelde toetsen, maar zijn ze bovendien uitermate geschikt voor EVADOS. Voor zover andere toetsen, zoals bijvoorbeeld de Tempo Test Rekenen, leerkrachten ook informatie kunnen geven over de ontwikkeling van leerlingen in de tijd, kennen deze veelal beperkingen. Zo laten Goffree en Frowijn (1996) zien dat de Tempo Test Rekenen slechts een beperkt gedeelte van het vakgebied bestrijkt en zeker niet als een operationalisatie van de kerndoelen basisonderwijs gezien kan worden. Bovendien gaat de leerstof waarop de toets betrekking heeft niet verder dan groep vijf van de basisschool. Een andere tekortkoming van de Tempo Test Rekenen is dat steeds dezelfde toets moet worden afgenomen om leerlingen in de tijd te kunnen volgen.

Aan de voornoemde drie bezwaren komt het Cito-LVS tegemoet. Goffree en Frowijn (1996) laten aan de hand van de toetsen Rekenen-Wiskunde zien dat de Cito-LVS-toetsen een goede operationalisatie zijn van de kerndoelen basisonderwijs. Bovendien bestrijken de LVS-toetsen alle groepen van de basisschool. Ten slotte behoeft bij de toetsen van het Cito-LVS niet steeds dezelfde toets te worden afgenomen om zicht te krijgen op de vorderingen van de leerlingen in de tijd. De toetsen uit het Cito-LVS zijn zo ontwikkeld dat de resultaten op verschillende toetsen uit dit systeem op eenzelfde meetschaal met elkaar vergeleken kunnen worden. In paragraaf 4.1.2 wordt dit nader toegelicht. Bovendien sluiten de toetsen uit het Cito-LVS aan bij het vaardigheidsniveau van de leerlingen.

Het Cito-LVS maakt het mogelijk de vorderingen van leerlingen op een aantal relevante vakgebieden in de tijd te volgen. In tabel 4.1 staat een overzicht van de toetspakketten van het Cito-LVS. Uit dit overzicht is ook af te leiden voor welke leerjaren de pakketten ontwikkeld zijn.

Tabel 4.1

Overzicht Toetspakketten Cito-LVS

Doelgroep/vakgebied	Groepen							
	1	2	3	4	5	6	7	8
Kleuters	Ordenen							
	Taal							
	Ruimte en tijd							
	Tweetaligheid							
Technisch lezen			Drie-Minuten-Toets					
			Leestechniek Leestempo					
Begrijpend lezen			Lezen met begrip		Toets Begrijpend Lezen			
Luisteren			Luisteren 1		Luisteren 2		Luisteren 3	
Schrijfvaardigheid					Schrijfvaardigheid			
Spelling			SVS-1		SVS-2		SVS-3	
Taal				Taalschaal 1			Taalschaal 2	
Woordenschat			WST		Leeswoordenschat			
Rekenen			Rekenen-W2002					
Wereldoriëntatie					Wereldoriëntatie			

In de volgende paragraaf wordt ingegaan op het principe van het Cito-LVS. Vervolgens komt de itemresponstheorie (IRT) aan bod, welke het Cito-LVS in staat stelt de vorderingen van leerlingen in de tijd te volgen. Bij de bespreking van de IRT komen twee modellen aan de orde: het Raschmodel en het één-parameter logistisch model (OPLM).

4.1.1 Het Cito-leerlingvolgsysteem

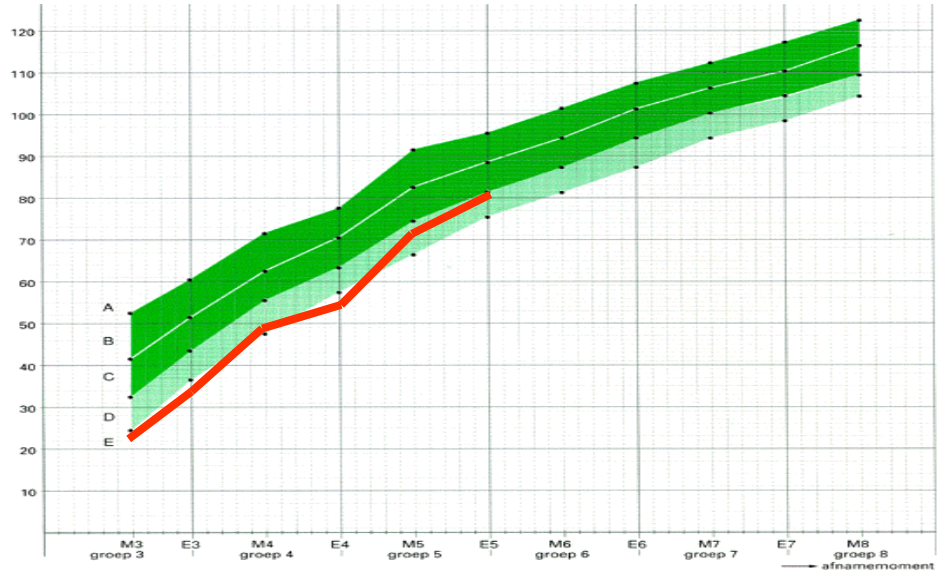
Het Cito-LVS bestaat uit een verzameling opgaven waaruit toetsen zijn samengesteld, die leerkrachten in staat stellen de ontwikkeling in de tijd in een bepaald vak of vakgebied bij leerlingen vast te stellen. De toetsafname vindt doorgaans tweemaal per jaar plaats, halverwege en aan het einde van het leerjaar. De inhoud van de toetsen zijn in globale zin inhoudelijk dekkend voor de te toetsen leerstofonderdelen.
Bij een LVS zijn in principe de antwoorden op twee vragen van belang. Gaat de leerling in vaardigheid vooruit? Is die vooruitgang bevredigend?

Het Cito-LVS biedt leerkrachten de mogelijkheid de vorderingen van de leerling periodiek, tijdens de gehele schoolperiode, te volgen. Bovendien zijn in het systeem per afnamemoment referentiegegevens opgenomen die de leerkrachten een referentiekader bieden om te bepalen of zij de vooruitgang bevredigend vinden. Om vast te stellen of geconstateerde vooruitgang bevredigend is, heb je een maatstaf nodig. Het Cito-LVS biedt in principe drie mogelijkheden daartoe:

- zelfgerichte: is de vooruitgang conform verwachting op basis van het verleden?
- normgerichte: is de vooruitgang conform verwachting in vergelijking met anderen?
- domeingerichte: is de vooruitgang conform verwachting in vergelijking met een domeingericht criterium?

In het Cito-LVS vormen de schaalscores de basis voor het bepalen van de voortgang in de tijd. Deze schaalscores kunnen in een grafiek worden weergegeven,

waardoor een leerlingrapport ontstaat. In dit rapport is de ontwikkeling van de leerling in de tijd af te lezen. Figuur 4.1 is een voorbeeld van zo'n leerling-rapport.



Figuur 4.1
Leerlingrapport Cito-LVS

Op de horizontale as staan de afnamenmomenten. De met een M en E gecodeerde momenten geven aan dat de toetsen halverwege (M) en aan het einde (E) van het schooljaar afgenomen worden. Op de verticale as staan schaal- of vaardigheidsscores.

De leerkracht geeft de door de leerling behaalde resultaten weer in de grafiek. Door dit voor een aantal meetmomenten te doen, ontstaat een beeld van de ontwikkeling van de leerling in de tijd. Het gearceerde gedeelte in figuur 4.1 zijn referentiegegevens, die aangeven hoe vergelijkbare leerlingen op de toetsen gescoord hebben. Deze gegevens zijn verkregen door items in een longitudinaal onderzoek over een reeks van jaren aan een landelijk cohort van leerlingen voor

te leggen. De referentiegegevens zijn verdeeld in vijf niveaugroepen (zie figuur 4.1) met de volgende betekenissen:

Niveau A: Goed tot zeer goed (25 procent hoogst scorende leerlingen).

Niveau B: Ruim voldoende tot goed (25 procent net boven het landelijk gemiddelde).

Niveau C: Matig tot voldoende C (25 procent net onder het landelijk gemiddelde).

Niveau D: Zwak tot matig (15 procent ruim onder het landelijk gemiddelde).

Niveau E: Zwak tot zeer zwak (10 procent laagst scorende leerlingen).

De witte lijn in figuur 4.1 correspondeert met het landelijk gemiddelde. Komt de score van de leerling boven de witte lijn, dan doet de leerling het goed in vergelijking met de landelijke populatie. Indien de resultaten van een leerling overeenkomen met de resultaten van de niveaugroepen D en E dan is dat een teken voor de school de leerling (extra) in de gaten te houden of actie te ondernemen. Het systeem ondersteunt de leerkracht daarbij met toetsen en didactische hulpmaterialen die zij kunnen gebruiken om hiaten vast te stellen en vervolgens gericht te remediëren.

4.1.2 Toepassing van de itemresponstheorie

Het LVS geeft informatie over de vorderingen van leerlingen over een aantal leerjaren, waarbij het gebruik maakt van (inhoudelijk) verschillende toetsen. Om te kunnen vaststellen hoeveel een leerling vooruitgaat, dienen zijn resultaten op deze toetsen in de tijd vergelijkbaar te zijn. Bovendien is het voor een dergelijke vaststelling van belang er zeker van te zijn dat de items in de toetsen dezelfde onderliggende vaardigheid meten: op dezelfde schaal liggen.

Voor het ontwikkelen van zo'n schaal maakt het Cito-LVS gebruik van de itemresponstheorie (zie Verhelst, 1992, 1993). De itemresponstheorie (IRT) definieert in een wiskundig model het verband tussen een niet observeerbare (latente) vaardigheid en de beantwoording van een opgave/item waarvoor die

vaardigheid vereist is. Aan elke persoon kan een getal toegekend worden dat een uitdrukking is van de mate waarin die persoon over de vaardigheid beschikt. IRT veronderstelt dat een correct antwoord op een bepaalde opgave op een grotere vaardigheid duidt dan een foutief antwoord. Bovendien gaat men ervan uit dat het antwoord op een item nooit volledig vastligt, ongeacht hoe groot of klein de vaardigheid van de persoon is.

Voor EVADOS heeft de IRT in vergelijking met de klassieke testtheorie (KTT) het voordeel van populatie-(on)afhankelijkheid en toetsspecificiteit. Bij toepassing van de KTT is het niet duidelijk of een verschil in de gemiddelde vaardigheid van een steekproef van leerlingen op toets E3 ten opzichte van toets M3 te wijten is aan een toename van de gemiddelde vaardigheid van de leerlingen (populatie-afhankelijkheid) of aan het feit dat toets E3 gemakkelijker was dan toets M3 (toetsspecificiteit), of aan beide. IRT komt aan deze beide problemen tegemoet.

Er kunnen diverse IRT-modellen onderscheiden worden. Eén van deze modellen is het Raschmodel, dat met de volgende formule gepresenteerd kan worden:

$$f_i(\vartheta) = \frac{e^{\vartheta - \sigma_i}}{1 + e^{\vartheta - \sigma_i}} \qquad (4.1)$$

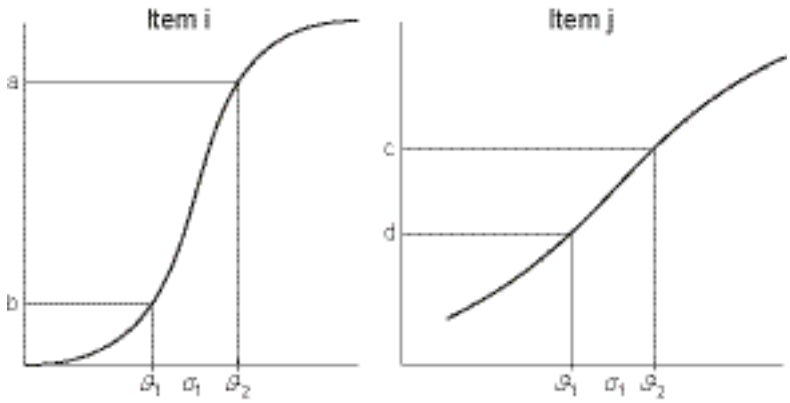
De formule geeft aan hoe groot de kans is dat het item juist wordt beantwoord als functie van de vaardigheid. In het Raschmodel is de itemresponsfunctie een logistische functie, met als argument het verschil tussen de latente vaardigheid van een persoon (ϑ) en een kengetal dat item i karakteriseert (σ_i). Uit formule (4.1) is af te leiden dat bij een item met een moeilijkheidsgraad (σ_i) die overeenkomt met de vaardigheid ϑ van een persoon, deze persoon 50% kans heeft het item goed op lossen. Omgekeerd kan σ_i gezien worden als de hoeveelheid vaardigheid die nodig is om een kans van 50% te hebben op een juist antwoord op dat item. Merk op dat formule (4.1) veralgemeniseerd kan worden door zowel in de teller als in de noemer het verschil tussen de vaardigheid van de persoon (ϑ) en het kengetal (σ_i) dat het item karakteriseert te vermenigvuldigen met de discriminatieparameter *a*. Onder het Raschmodel is *a* een constante en heeft voor alle items dezelfde waarde (in de regel de waarde 1), waardoor deze niet in

formule (4.1) opgenomen staat. Opname van de discriminatieparameter *a* in formule (4.1) leidt tot (4.2)

$$f_i(\vartheta) = \frac{e^{a(\vartheta-\sigma_i)}}{1+e^{a(\vartheta-\sigma_i)}} \qquad (4.2)$$

Eénparameter logistisch model

Het Raschmodel gaat uit van identieke discriminatieparameters. Een model dat niet uitgaat van identieke discriminatieparameters is het éénparameter logistisch model (OPLM). Een voorbeeld van twee itemresponscurves met een verschillend discriminerend vermogen staat in figuur 4.2. Op de verticale as staat de kans weergegeven dat een item goed beantwoord wordt en op de horizontale as de (latente) vaardigheid.



Figuur 4.2
Twee items die verschilend discrimineren
(Bron: Verhelst, 1993, p. 122)

In figuur 4.2 is duidelijk te zien dat de itemresponscurve van item i afwijkt van die van item j. De snelheid waarmee de functies in de buurt van de moeilijkheidsgraad σ_i en σ_j verandert, is verschillend. Uit figuur 4.2 is af te leiden dat voor de items i en j geldt dat een zelfde toename in vaardigheid (weergegeven door $\vartheta_2 - \vartheta_1$) leidt tot een grotere kans op het goed beantwoorden van item i in

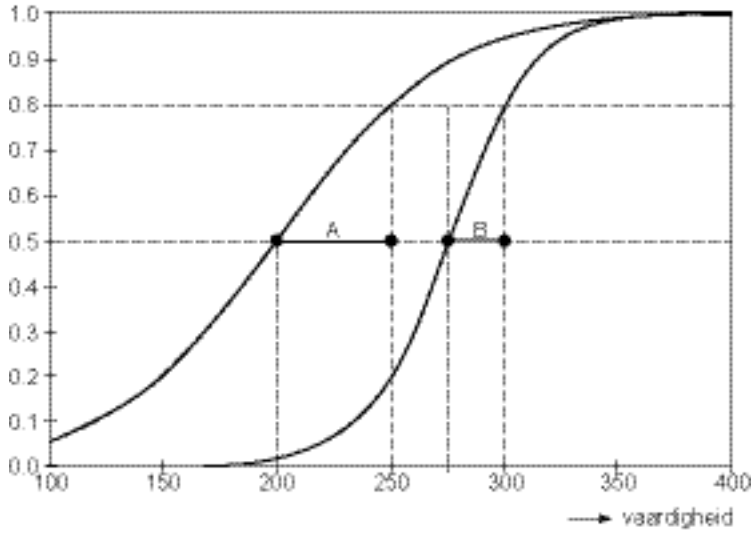
vergelijking met item j. Deze toename in kans wordt weergegeven door de lijnstukken (a-b) voor item i en (c-d) voor item j. Of met andere woorden: het onderscheidend (discriminerend) vermogen van item i is groter dan dat van item j. Dit verschil in discriminerend vermogen wordt tot uitdrukking gebracht door de constante *a* in formule (4.2) te voorzien van een index, waardoor het tweeparameter logistisch model ontstaat (zie formule 4.3)

$$\text{Prob}\,(X_i = 1 \mid \vartheta) = f_i(\vartheta) = \frac{e^{a_i(\vartheta - \sigma_i)}}{1 + e^{a_i(\vartheta - \sigma_i)}} \quad (a_i > 0) \tag{4.3}$$

Door nu de grootheden a_i niet langer te beschouwen als onbekende parameters, maar als gegeven constanten die uit de data berekend kunnen worden, verliest a_i zijn status als parameter. Om het onderscheid in de terminologie goed aan te geven, wordt a_i discriminatie-index genoemd en spreken Verhelst en Eggen (1989) van het éénparameter logistisch model (OPLM).

Toepassing van OPLM

OPLM biedt de mogelijkheid om op een vaardigheidsschaal ook de moeilijkheidsgraad van opgaven te typeren. Hierdoor ontstaat de mogelijkheid een uitspraak te doen in hoeverre een leerling bepaalde type opgaven beheerst, hetgeen vervolgens een leerkracht in staat stelt gerichte acties te ondernemen. Aan de hand van onderstaand voorbeeld dat ontleend is aan de PPON-publicatie Balans van het rekenonderwijs halverwege de basisschool 2 van Bokhove, Van der Schoot en Eggen (1996), wordt dit nader toegelicht.



Figuur 4.3
Relatie tussen vaardigheid en de kans op het goed maken van twee opgaven (bron: Bokhove e.a., 1996, p. 9)

In figuur 4.3 is voor twee opgaven de relatie weergegeven tussen de vaardigheid en de kans op een goed antwoord. Voor beide opgaven geldt dat de kans op een goed antwoord stijgt met toename in vaardigheid. Opgave B blijkt aanmerkelijk moeilijker te zijn dan opgave A. De vaardigheid die nodig is om met een bepaalde kans opgave B goed te maken, is aanmerkelijk hoger dan de vaardigheid voor diezelfde kans bij opgave A. Tevens is te zien dat het discriminerend vermogen, de snelheid waarmee die kans oploopt met toenemende vaardigheid, tussen beide opgaven varieert. Zo is bij opgave B de helft (300-275) minder vaardigheid nodig dan bij A (250-200) om van 50% goed naar 80% goed te komen.

In het LVS onderscheidt men drie niveaus van beheersing. Wanneer de kans op een goed antwoord hoger is dan 80% spreekt men van een goede beheersing. Wanneer de kans op een goed antwoord ligt tussen 50% en 80%, is er sprake van een matige beheersing van de opgave. Van een onvoldoende beheersing is sprake als de kans op een goed antwoord kleiner is dan 50%. In het voorbeeld

van figuur 4.3 beheersen leerlingen met een vaardigheid 270 opgave A dus goed, maar opgave B onvoldoende.
Afhankelijk van de vaardigheidsscore van een leerling ten opzichte van de kans die een leerling heeft een item goed op te lossen, kan aangegeven worden of een leerling het type opgaven waartoe het item behoort onvoldoende, matig of goed beheerst. Om de mate van beheersing in relatie met het type opgave voor leerkrachten inzichtelijk te maken, is in het LVS per toets een vaardigheidsprofiel ontwikkeld. Een voorbeeld van een dergelijk profiel voor de toets Rekenen-Wiskunde M3 staat in figuur 4.4.

Aan de rechterkant van figuur 4.4 staan de bij het LVS onderscheiden vijf vaardigheidsniveaus (A tot en met E) weergegeven. De horizontale lijntjes geven een indicatie van de moeilijkheidsgraad van de opgaven. De linkerkant van elk lijntje komt overeen met een vaardigheidsniveau waarmee een leerling 50% kans heeft de desbetreffende opgave goed te maken. De rechterkant van het lijntje komt overeen met een vaardigheidsniveau met een kanspercentage van 80%. Aan de hand van de behaalde vaardigheidsscore op de toets (in dit geval Rekenen-Wiskunde E3) krijgt de leerkracht een indruk van het type opgaven dat een leerling onvoldoende, matig of voldoende beheerst. Met behulp van de niveau-indicaties A tot en met E kan de leerkracht bovendien zien hoe dit beheersingsniveau zich verhoudt tot het niveau van de landelijke referentiegroep.

Zo zijn in figuur 4.4 voor het onderdeel ‘structuur van de telrij’ drie opgaven opgenomen die staan voor een groep opgaven met een vergelijkbare moeilijkheidsgraad. De eerste opgave (Welk getal komt vòòr 7?) is de gemakkelijkste van de drie. In het vaardigheidsprofiel aan de rechterkant komt deze opgave overeen met het linkse (het bovenste) lijntje. De tweede opgave komt overeen met het middelste lijntje en de derde opgave met het rechtse lijntje.

Figuur 4.4

Vaardigheidsprofiel Rekenen-Wiskunde toets M3

Uit figuur 4.4 is af te leiden dat een persoon met een vaardigheidsniveau van bijvoorbeeld 40, vergelijkbare opgaven als opgave 1 waarschijnlijk beheerst. Immers de vaardigheid van de leerling is groter dan de benodigde vaardigheid om 80% kans te hebben opgaven van het type 1 goed te beantwoorden (de rechterkant van het linkse lijntje). Opgaven van het type 2 zal de leerling matig beheersen. Zijn vaardigheidsscore valt binnen de 50% - 80% kans om opgaven van dit type goed te beantwoorden. Het derde type opgave is te moeilijk voor deze leerling. Zijn vaardigheidsscore komt overeen met een kanspercentage van minder dan 50% om vergelijkbare opgaven goed te beantwoorden. In het algemeen zal deze persoon dit type opgaven niet goed beantwoorden. Door de vaardigheidsscore van een persoon te vergelijken met de kans bepaalde typen opgaven goed te beantwoorden, krijgt een leerkracht meer inzicht in het type vragen dat de persoon wel of niet aan kan en kan hij zijn onderwijs daarop afstemmen.

Aan de linkerkant van figuur 4.4 staat een inhoudsbeschrijving van de toets, uitgesplitst in domeinen en subdomeinen. Door deze uitsplitsing voorziet het LVS een leerkracht niet alleen van normgerichte informatie, maar ook van criteriumgerichte. Een leerkracht wordt niet alleen geïnformeerd over hoe zijn leerlingen het doen ten opzichte van anderen, maar ook ten opzichte van de leerstof (vergelijk het zojuist geschetste voorbeeld). Omdat het Cito-LVS zoveel mogelijk gebruik maakt van één vaardigheidsschaal voor alle groepen van de basisschool ontstaat de mogelijkheid het ontwikkelingsniveau van een leerling in een bepaald vakgebied over leerjaren (ook vakinhoudelijk) te interpreteren en het onderwijs daar zoveel mogelijk op af te stemmen.

4.2 Input

Het CIPO-model in hoofdstuk twee laat zien dat de component input van invloed is op de output van het onderwijs. Deze inputcomponent bestaat uit variabelen die betrekking hebben op kenmerken van:

- leerlingen, zoals sociaal economische status, leerlinggewicht, geslacht, geboortedatum, nationaliteit en geloofsovertuiging;
- leerkrachten, zoals onderwijservaring, leeftijd, geslacht, betrekking en aantal uren na- of bijscholing;
- materiële en financiële voorzieningen, zoals de beschikbare hoeveelheid formatie, teamsamenstelling, aantal klassen, klassengrootte, huisvesting, en de aanwezige voorzieningen, hulpmiddelen en materialen.

Vele van deze kenmerken zijn opgeslagen in de door scholen gebruikte schooladministratiepakketten. In het kader van dit proefschrift zijn twee pakketten nader bekeken. Uit de analyse blijkt dat beide pakketten niet alleen van elkaar verschillen ten aanzien van de inhoud, maar ook ten aanzien van de wijze waarop de informatie (electronisch) is opgeslagen. Omdat één van deze pakketten, het pakket ESIS (Onderwijs Automatiseringsbureau, 1994), veruit de grootste verspreiding kent, en omdat de aan het onderzoek deelnemende scholen met name dit pakket gebruikten, is besloten het pakket ESIS nader te onderzoeken op voor schoolzelfevaluatie relevante inputfactoren.

Het pakket ESIS bestaat uit de modulen ESIS-A en ESIS-B. Module ESIS-A is ontwikkeld voor de schooladministratie, terwijl module ESIS-B zich richt op administratie en planning van het onderwijsleerproces van (groepen van) leerlingen en in een beperkte mate ook op de evaluatie van dit proces. Als opslagmedium voor inputkenmerken is module ESIS-A van belang en zal om die reden nader toegelicht worden. Het pakket ESIS-B zal niet verder besproken worden.

Het pakket ESIS-A richt zich op algemene administratieve handelingen die voor een school voor primair onderwijs van belang zijn. In de handleiding van het

pakket staat aangegeven dat het doel van de module is de administratieve werkzaamheden van een school op de volgende terreinen te vergemakkelijken:

- de leerlingadministratie;
- de absentenadministratie;
- de inventarisadministratie;
- de financiële administratie;
- het plannen van alle schoolse activiteiten op een termijnkalender;
- het vervaardigen van een volledig activiteitenplan, inclusief alle lesroosters;
- het vastleggen en bewaken van besluiten in een besluitenlijst;
- het administreren van alle voor school benodigde personeelsgegevens;
- het vervaardigen van roosters ten behoeve van ouderavonden;
- het raadplegen van gegevens van oud-leerlingen;
- het berekenen van het formatiebudget.

De gegevens in ESIS-A zijn ondergebracht in acht categorieën die weer onderverdeeld zijn in een aantal subcategorieën. Elke subcategorie heeft een eigen tabel waarin de gegevens opgeslagen zijn. Alle in ESIS aanwezige subcategorieën zijn niet van belang voor schoolzelfevaluatie. Voor een deel komt dit omdat de opgeslagen informatie te schoolspecifiek is en met name administratief van aard. Als voorbeeld kan de subcategorie met als codering TABSB genoemd worden waarin de door de school genomen besluiten vastgelegd worden. Ook tabellen waarin het financieel betalingsverkeer vastgelegd wordt, bevatten gegevens die niet interessant zijn voor schoolzelfevaluatie. Voor schoolzelfevaluatie zijn die tabellen van belang die informatie geven over relevante input- en procesfactoren. Een aantal van deze tabellen staat in tabel 4.2 opgenomen, waarbij een onderscheid is aangebracht tussen datatabellen en coderingstabellen. Datatabellen bevatten gegevens over de leerlingen en het personeel van de school en de wijze waarop in de school de organisatie (bijvoorbeeld groepssamenstelling) vorm heeft gekregen. In de coderingstabellen staan omschrijvingen van afkortingen die het systeem hanteert, bijvoorbeeld coderingen van landen, meubilair, vormen van onderwijs en vormingsgebieden. Coderingstabellen zijn standaard in het pakket aanwezig. Voor een deel kunnen deze tabellen ook schoolspecifiek zijn. Het is aan de school te beslissen of zij van de

afkorting en omschrijving gebruik wil maken. Als voorbeeld hiervan kunnen functies van medewerkers genoemd worden. Voor vele functies zijn afkortingen en omschrijvingen opgenomen. Niet alle functies echter zullen op elke school vervuld worden, bijvoorbeeld de functie van conciërge. In tabel 4.2 zijn deze 'schoolspecifieke' tabellen aangeduid met een asterisk.

Belangrijk voor de te ontwikkelen procedure voor schoolzelfevalutie is de kwantificeerbaarheid van de op te nemen factoren. Bij de keuze van deze factoren zal dit steeds een belangrijk aandachtspunt dienen te zijn. Uit onderzoek (Scheerens, 1989; Brandsma, 1993) blijkt dat het niet altijd eenduidig is welke factoren een positieve correlatie hebben met leerresultaten. Onderzoeksresultaten leiden niet altijd tot dezelfde conclusies. Wel wordt algemeen aanvaard dat factoren als intelligentie, sociaal economische status en opleidingsniveau van ouders belangrijk zijn. Aangezien de module ESIS-A de mogelijkheid biedt om dergelijke informatie op te slaan, is ESIS-A als informatiebron voor (met name) inputfactoren prima bruikbaar voor schoolzelfevalutie. Dit ondanks het feit dat scholen in de praktijk verschillend met het pakket omgaan.

Tabel 4.2

Overzicht van een aantal voor schoolzelfevaluatie belangrijke tabellen uit ESIS-A

Aard tabel	Omschrijving
Datatabellen	Overzicht absentie leerlingen
	Informatie over leerlingen, w.o. gezinssamenstelling
	Schoolloopbaan leerlingen
	Historische tabel leerlingen
	Leerlingen tabel
	Verzuimgegevens leerlingen
	Code leerkracht, taakomschrijving, bevoegdheid en aanstelling
	Verdeling groepen over leerkrachten en schooljaar
	Diverse opmerkingen over leerkrachten
	Diverse gegevens over leerkrachten
	Rooster groepen, w.o. vak en leerkracht
Coderingstabellen	Afkorting landen en nationaliteit
	Diverse sleutels en omschrijvingen
	Code en omschrijving meubilair en materialen leslokalen
	Code en omschrijvingen werkruimtes*
	Code en omschrijving vormingsgebieden*
	Code en omschrijving functies*
	Code en omschrijving schoolactiviteiten*

4.3 Proces

In figuur 2.3 is een overzicht van relevante school- en klaskenmerken gegeven die mogelijk van invloed zijn op de output. Welke variabelen meegenomen moeten worden bij de interpretatie van de output hangt af van het gewenste referentiekader. Indien een school haar eigen referentiekader is, dan zijn de variabelen die in de tijd (min of meer) stabiel zijn niet zo relevant. De aanname is dan dat deze variabelen geïntegreerd zijn in het onderwijsproces en daar structureel deel van uitmaken. Indien een school haar resultaten vergelijkt met andere scholen, dan zijn, om voor de school vergelijkbare groepen van scholen samen te stellen, ook de in de tijd stabiele variabelen van belang.

In een voor een school stabiele situatie zullen de mogelijke (jaarlijkse) veranderingen in de procescomponent met name terug te vinden zijn in de variabelen 'time on task', instructional quality of schooling' en 'Opportunity To Learn'. De eerste twee genoemde variabelen zullen in het kader van dit proefschrift niet verder uitgewerkt worden. Uit literatuur (Schaffer & Nesselrodt, 1992; Van de Tuin en Van der Werf, 1996; Van der Tuin, 1997) blijkt dat voor het verkrijgen van informatie over deze beide variabelen gebruik gemaakt wordt van logboeken en observaties in de klas, wat twee relatief arbeidsintensieve methoden zijn. Met name vanwege de arbeidsintensiviteit van beide methoden zullen deze niet snel door een school onderzocht worden. In hoofdstuk twee is aangegeven dat de variabele 'Opportunity To Learn' in dit proefschrift uitgewerkt wordt als 'Toets Curriculum Overlap'. Voor interpretatie van de toetsresultaten van leerlingen is deze variabele erg belangrijk. Vanwege dit belang komt een bespreking van de variabele 'Opportunity To Learn' en in het verlengde daarvan de variabele 'Toets Curriculum Overlap' wel aan bod[1].

[1] De bespreking van de variabele Toets Curriculum Overlap (TCO) en de ontwikkeling van een instrument TCO zoals weergegeven in dit proefschrift is een kleine bewerking van het artikel dat gepubliceerd is in het tijdschrift Pedagogische Studiën (jaargang 81, nummer 3, 2004).

EVADOS maakt het mogelijk uitspraken te doen over de kwaliteit van het onderwijs door gebruik te maken van de resultaten van leerlingen op de aan hen voorgelegde toetsen. Twee belangrijke vragen in dit kader die verband houden met het proces zijn:

- In hoeverre is het geboden onderwijs afgestemd op de kerndoelen zoals deze gelden voor het basisonderwijs?
- Sluiten de opgaven van het gebruikte instrumentarium in voldoende mate aan op het geboden onderwijs?

Beide vragen zullen in het hiernavolgende behandeld worden, waarbij met name de tweede vraag voor EVADOS erg belangrijk is.

4.3.1 Relatie Cito-LVS-toetsen met kerndoelen basisonderwijs

Voor een (maatschappelijke) waardering van het geboden onderwijs is een extern referentiekader nodig. Voor het basisonderwijs zijn kerndoelen (Ministerie van Onderwijs Cultuur en Wetenschappen, 1998) geformuleerd die als zodanig kunnen gelden. Om uitspraken te kunnen doen over de kwaliteit van het onderwijs is het van belang te weten in hoeverre het onderwijs is afgestemd op deze kerndoelen. Of anders geformuleerd: in hoeverre representeren de toetsen uit het Cito-LVS de kerndoelen.

De kerndoelen basisonderwijs zijn geformuleerd op macro-niveau. Zij geven de beoogde (‘intended’) doelstellingen weer. Of deze doelstellingen inderdaad bereikt worden, is van een groot aantal factoren afhankelijk. Schmidt en McKnight (1995, p. 348) geven dit als volgt aan: ‘transition from intended to implemented curricula is reflected by movement from the halls of government agencies to the halls of schools - to classrooms and teachers’. Zij geven daarmee aan dat de uiteindelijke implementatie van een curriculum bepaald wordt door een groot aantal factoren zoals nationale doelstellingen, politieke opvattingen, gebruikte methoden, visie van school en leerkrachten en de wijze waarop het onderwijs in de klas concreet vorm krijgt. Naar hun opvatting is het niet realistisch prestaties als enige indicatie te zien, maar dienen deze steeds in relatie met

andere factoren - zoals het beoogde en geïmplementeerde curriculum - beschouwd te worden. Daarnaast merken zij op dat eigenlijk niet de leerkracht de laatste schakel in het traject van het beoogde naar het geïmplementeerde curriculum is, maar dat dat de leerlingen zijn. Zij zijn degenen die uiteindelijk 'bepalen' wat het gerealiseerd ('attained') curriculum is.

Ten behoeve van EVADOS zijn Goffree en Frowijn (1996) nagegaan in hoeverre de items uit de toets Rekenen-Wiskunde van het Cito-LVS een goede operationalisatie zijn van de op macro-niveau geformuleerde kerndoelen basisonderwijs. In hun analyse laten ze zien dat een directe vergelijking tussen concrete items en kerndoelen niet eenvoudig, zo niet onmogelijk is, omdat de kerndoelen daar te abstract voor geformuleerd zijn en daardoor allerlei interpretaties toelaten. Wel concluderen zij op basis van hun analyses - weliswaar exemplarisch uitgevoerd voor de leergang vermenigvuldigen - dat Cito-LVS-toetsen de kerndoelen (in grote mate) representeren, waardoor de toetsen van het Cito-LVS uitgangspunt kunnen zijn voor een uitspraak over de kwaliteit van het onderwijs.

4.3.2 Aansluiting Cito-LVS-toetsen bij het geboden onderwijs

Voor een oordeel over de kwaliteit van het geboden onderwijs op basis van de resultaten van leerlingen op toetsen, is het belangrijk dat deze toetsen aansluiten bij het geboden onderwijs. Indien de toetsen niet (voldoende) aansluiten, kan dat leiden tot een verkeerd beeld van de geboden kwaliteit. Ten eerste omdat door de hele toets af te nemen, de suggestie gewekt wordt dat alle onderwerpen die in de toets aan bod komen ook behandeld zijn. Ten tweede omdat de vaardigheid van leerlingen mogelijk onderschat wordt. Centraal bij de aansluiting tussen de toetsen en het geboden onderwijs staat de vraag of de leerstof waarop de toetsing betrekking heeft tijdens de lessen behandeld is. Of met andere woorden: sluiten de toetsen aan bij het geïmplementeerd curriculum. Husén en Tuijnman (1994, p. 2) formuleren het als volgt: 'Before performance can be fairly assessed, it is

necessary to determine whether all the students have had the opportunity to learn the prescribed content'.

Methoden om OTL vast te stellen

'Opportunity To Learn' (OTL) wordt wel gedefinieerd als de mate waarin leerlingen in de gelegenheid zijn gesteld zich de vereiste lesstof eigen te maken. Deze definitie van OTL is ruim, omdat het ook de wijze waarop de instructie heeft plaatsgevonden en de tijd die aan het leren is besteed, kan betreffen. In hoofdstuk twee is OTL omschreven als de mate waarin de toetsen aansluiten bij het gegeven onderwijs. Of met andere woorden: de mate waarin er overeenstemming is tussen het 'beoogde curriculum' en het 'gerealiseerde curriculum', zoals gemeten door de toets. Ook Pelgrum, Voogt en Plomp (1995, p. 90) geven het belang van deze overeenstemming aan. Zij zien OTL als 'a measure for the implemented curriculum' en, zo vervolgen zij, 'it is often used in determining the curricular validity of student achievement tests'.

Internationaal zijn vele studies naar OTL verricht, waarbij aan leerkrachten zowel op item- als op leerstofniveau gevraagd is naar de aansluiting met het geboden onderwijs. Op beide methoden zal kort worden ingegaan.

Pelgrum (1989) heeft in zijn studie naar peilingsonderzoek in het onderwijs onderzocht hoe valide en betrouwbaar een op toetsitems en een op basis van leerstofcategorieën gebaseerde maat van het feitelijk uitgevoerde leerplan is. Hij concludeert dat het gebruik van de itemmethode de voorkeur verdient.
Pelgrum, Voogt en Plomp (1995) maken melding van acht studies waarbij gebruik gemaakt is van een 'item-based approach'. Bij deze methode dienen leerkrachten aan te geven of de items uit een toets aansluiten bij het geboden onderwijs (geïmplementeerd curriculum). Deze benaderingswijze ondervindt zowel bijval als kritiek. Pelgrum e.a. (1995, p. 19) verwijzend naar Oakes (1989) en McKnight en Curtis (1987) stellen: 'Measures using an item-based approach to curriculum content appear to be particularly promising, because of their direct focus on the curriculum content of the implemented curriculum and not on indirect measures such as curriculur emphasis (Oakes, 1989) or curricular

intensity (McKnight & Curtis, 1987) which only refer to time allocated to (parts) of subjects'.

Schmidt en McKnight (1995) daarentegen wijzen op het gevaar dat bij een 'item-based approach' de aandacht van de leerkrachten meer gericht zou kunnen zijn op de itemvorm dan op de inhouden waarop de items betrekking hebben. Het resultaat zou dan veel meer een voorkeur van leerkrachten voor bepaalde itemvormen weergeven, dan een antwoord op de vraag in hoeverre de items aansluiten bij het geboden onderwijs. Wiley en Yoon (1995, p. 357) geven aan dat er opvattingen zijn die niet uitgaan van een 'item-based approach'. Zij verwoorden dit als volgt 'newer thinking about OTL focusses on learning goals and the instructional activities bearing on them rather than on the specific items or tasks used in the tests'. Bij deze methode staan niet de items, maar leerstofcategorieën waarop de items betrekking hebben centraal. Aan leerkrachten wordt gevraagd aan te geven of bepaalde leerstofcategorieën behandeld zijn. Op basis van deze informatie wordt aangenomen dat de items die daarop betrekking hebben, aansluiten bij het geïmplementeerd curriculum.
OTL kan omschreven worden als de mate waarin leerlingen in de gelegenheid zijn geweest zich de vereiste leerstof eigen te maken. De omschrijving kan ruim geïnterpreteerd worden, door ook de instructiewijze en de hoeveelheid bestede tijd erbij te betrekken. In het voorgaande is OTL in beperkte zin gebruikt, door alleen te vragen naar de mate waarin de toetsen aansluiten bij het gegeven onderwijs. Ook bij internationale studies naar OTL wordt deze beperkte omschrijving van OTL gehanteerd. De Haan (1992) spreekt dan niet meer van OTL, waarvan, zoals in het voorgaande reeds is aangegeven, ook aspecten als leertijd (time on task) en de instructiewijze van leerkrachten deel uit (kunnen) maken, maar van Toets Curriculum Overlap. Dit is ook de term die in dit proefschrift gehanteerd wordt

4.3.3 Toets Curriculum Overlap (TCO)

Aan de ontwikkeling van een instrument TCO voor EVADOS lagen de volgende vier onderzoeksvragen ten grondslag:

1 Welke meetmethode om TCO te meten is het meest adequaat?
2 Is deze meetmethode valide en betrouwbaar?
3 Hoe kan het TCO-instrument het beste ingezet worden, rekening houdend met psychometrische en praktische overwegingen?
4 Wat betekent het instrument voor de leerkracht in de klas en hoe dient deze het instrument te gebruiken?

De ontwikkeling van het instrument heeft plaatsgevonden aan de hand van de toets Rekenen-Wiskunde E3 van het Cito-leerlingvolgsysteem. Deze toets is bestemd voor de leerlingen van (eind) groep 3 van het basisonderwijs. De toets is methode-onafhankelijk en bestaat in totaal uit 53 items. Voor de ontwikkeling van het instrument TCO is aan leerkrachten een vragenlijst voorgelegd. Daartoe is gebruik gemaakt van een naar postcode-gebied gestratificeerde steekproef van 450 scholen uit een bestand van 1490 gebruikers van de toets Rekenen-Wiskunde E3. Voor deelname aan het onderzoek zijn een drietal voorwaarden gesteld:

1 De deelnemers moeten de toets conform de handleiding aan het einde van groep 3 afnemen. Hiervoor is gekozen om tijd van afname als variabele constant te houden.
2 Slechts één leerkracht per school mag participeren in het onderzoek. De kans is groot dat twee leerkrachten van dezelfde school hetzelfde TCO-profiel opleveren, terwijl het in het kader van het onderzoek wenselijk is verschillende TCO-profielen te meten.
3 De leerkracht dient het hele leerjaar voor de klas te hebben gestaan, omdat de leerkracht dan een goed overzicht heeft op wat wel en wat niet behandeld is.

Van de 450 aangeschreven scholen hebben 170 leerkrachten (38%) positief gereageerd. Een aantal leerkrachten gaf aan dat het tijdstip (einde schooljaar) waarop de vragenlijst naar de school gestuurd werd ongunstig was. Mogelijk verklaart dit de relatief lage respons. Bij twee scholen bleken de vragenlijsten

niet volledig te zijn ingevuld. Deze twee scholen zijn uit het bestand verwijderd. Aan de leerkrachten is ook gevraagd de resultaten van hun leerlingen op de toets E3 mee te sturen. In totaal zijn de resultaten van 3265 leerlingen verzameld en in het onderzoek betrokken.

Bij de ontwikkeling van een instrument TCO zijn de volgende drie meet-methoden onderzocht:

- het voorleggen van items aan leerkrachten (itemmethode);
- het voorleggen van leerstofcategorieën (categoriemethode);
- het vragen naar de gebruikte rekenmethode (lesmethode).

De 53 items waaruit de toets Rekenen-Wiskunde E3 bestaat zijn verdeeld over twee boekjes. De leerstof waar de toetsen betrekking op heeft, is toegewezen aan de volgende vijf hoofdcategorieën met de daarbij onderscheiden 14 sub-categorieën:

Tellen en ordenen

1 Structuur van de telrij
2 Resultatief en structurerend tellen
3 Vergelijkingen en ordenen

Structureren

4 Splitsen
5 Samenstellen
6 Aanvullen

Bewerkingen

7 Optellen
8 Aftrekken
9 Diversen

Rekendictee
10 Optellen
11 Aftrekken
12 Splitsen

Meten en Tijd
13 Meten
14 Tijd

De voor het onderzoek ontwikkelde vragenlijst is voorgelegd aan de bij het onderzoek betrokken leerkrachten van groep 3 van het basisonderwijs. De vragenlijst is mede gebaseerd op het resultaat van het onderzoek van De Haan (1992) om TCO te meten. In haar onderzoek vergelijkt De Haan twee meetmethoden: een gedetailleerde TCO-vragenlijst en een holistische. Bij deze laatste vragenlijst wordt aan leerkrachten gevraagd aan te geven of een item 'maakbaar'of 'niet-maakbaar' (De Haan spreekt van 'taught') is. Zij komt tot de conclusie dat om praktische overwegingen de holistische vragenlijst een goed alternatief is.
In de ontwikkelde vragenlijst werden de leerkrachten naar de door hen gebruikte rekenmethode gevraagd en de wijze waarop ze de rekenmethode gebruikten. Mocht het zo zijn dat een bepaalde rekenmethode automatisch leidt tot een voldoende hoge maakbaarheidsscore op de toets E3, dan zou volstaan kunnen worden met het vragen naar de gebruikte rekenmethode. Na het aangeven van de rekenmethode dienden de leerkrachten per onderscheiden subcategorie aan te geven of naar hun oordeel de leerlingen de bij de subcategorie behorende leerstof zich eigen hebben kunnen maken (of de subcategorie maakbaar is, hetgeen betekent dat de leerstof behandeld is en dat de leerlingen er mee hebben kunnen oefenen). Om leerkrachten te informeren waaruit de leerstof van de onderscheiden subcategorieën bestaat, is gebruik gemaakt van de in de handleiding van de toets E3 gebruikte omschrijvingen. Tot slot werd aan hen gevraagd ook per item aan te geven of deze gegeven het door hen verzorgde onderwijs maakbaar is. Merk op dat het begrip 'maakbaar' zich onderscheidt van het begrip 'moeilijkheid'. 'Maakbaar' verwijst naar het wel of niet behandeld zijn van leer-

stof ongeacht de moeilijkheidsgraad van een bepaald item. Als er gesproken wordt over de ‘moeilijkheidsgraad’ van een item dan wordt daarmee aangegeven of het een ‘makkelijk’ of ‘moeilijk’ item voor de leerling is, waarbij je er impliciet van uitgaat dat de leerling over de vereiste kennis en vaardigheden voor het oplossen van het item beschikt. Of anders gezegd: de leerstof is onderwezen en het item is maakbaar.

Bij de bespreking van het onderzoek is voor een deel gebruik gemaakt van een door Van Abswoude (1999) in het kader van de ontwikkeling van EVADOS uitgevoerd onderzoek naar TCO. Daarnaast zijn in het kader van dit proefschrift op de verzamelde data secundaire analyses uitgevoerd om te komen tot een instrument TCO.
In de bespreking van de ontwikkeling van het instrument TCO zal in eerste instantie ingegaan worden op een vergelijking van de ‘itemmethode’ met de ‘categoriemethode’. Daarna volgt een bespreking van de ‘lesmethode’, daar deze meetmethode zich onderscheidt van de twee andere onderzochte meetmethoden.

Itemmethode en categoriemethode

Tabel 4.3 geeft een overzicht van het oordeel van de leerkrachten over de maakbaarheid van toets E3 wat betreft de subcategorieën en de items. In de kolom 'Categorie' staat aangegeven hoeveel procent van de leerkrachten vindt dat - uitgaande van de in de vragenlijst opgenomen omschrijving - items over deze subcategorieën aan de leerlingen voorgelegd mogen worden. In de kolom ‘Item’ staan de gemiddelde maakbaarheidsscores op de items, hier geclusterd per subcategorie. Tussen haakjes staat het aantal items dat tot de desbetreffende subcategorie behoort. De laatste kolom geeft het percentage leerlingen weer dat de items correct heeft beantwoord.

Tabel 4.3

Overzicht van het oordeel van leerkrachten over de maakbaarheid van toets E3

Categorieën	Opvatting leerkrachten		Leerlingresultaten (% goed beantwoord)
	Categorie (% maakbaar)	Item (% maakbaar)	
Tellen en ordenen			
Structuur van de telrij	98	98 (5 items)	90
Resultatief en structurerend tellen	97	95 (2 items)	80
Vergelijken en ordenen	95	76 (3 items)	77
Structureren			
Splitsen	79	85 (4 items)	79
Samenstellen	89	92 (1 item)	91
Aanvullen	81	83 (5 items)	84
Bewerkingen			
Optellen	96	89 (5 items)	87
Aftrekken	92	76 (7 items)	71
Diversen	81	80 (3 items)	76
Rekendictee			
Optellen	97	99 (3 items)	74
Aftrekken	96	98 (4 items)	90
Splitsen	83	80 (3 items)	71
Meten en tijd			
Meten	88	83 (5 items)	80
Tijd	59	63 (3 items)	68

Uit tabel 4.3 blijkt dat de leerkrachten van mening zijn dat de items uit de toets E3 in het algemeen goed aansluiten bij het onderwijs. Dit geldt voor zowel de subcategorieën als voor de items. Ook is er sprake van een positieve correlatie tussen de opvattingen van de leerkrachten op de categorieën met de leerling-scores (.506) en hun opvattingen op basis van de items met de leerlingscores

(.692). Deze correlatie behoeft niet perfect te zijn, daar het maakbaar zijn van een item niet per se betekent dat het item ook door de leerlingen goed beantwoord zal worden, omdat moeilijkheidsgraad en maakbaarheid immers als twee verschillende concepten beschouwd worden.
Tabel 4.3 geeft geen informatie over de verdeling van de maakbaarheidsscores van de leerkrachten over de subcategorieën en de items. Het is dus niet bekend of het steeds dezelfde leerkrachten zijn die bepaalde subcategorieën of items als niet-maakbaar zien of dat per subcategorie of per item steeds andere leerkrachten aangeven de subcategorie of het item als niet-maakbaar te zien. Indien bepaalde leerkrachten een relatief groot aantal subcategorieën of items als niet-maakbaar zien, kan dat betekenen dat gegeven het geboden onderwijs de toets E3 onterecht aan hun leerlingen is voorgelegd. Een onderschatting van de vaardigheid van deze leerlingen kan hiervan het gevolg zijn. Bij het gebruik van het TCO-instrument zal de maakbaarheidsscore van de leerkracht een belangrijk element zijn. In de verdere bespreking van de ontwikkeling van het TCO-instrument zal hierop teruggekomen worden.

Rekenmethode

In tabel 4.4 staat aangegeven welke rekenmethoden door de aan het TCO-onderzoek deelnemende leerkrachten gebruikt worden. Als een leerkracht een andere methode gebruikt, staat dit aangegeven onder ´andere rekenmethode´. Leerkrachten waarvan niet bekend is welke rekenmethode zij hanteren, vallen onder de categorie ´onbekend´.

Tabel 4.4

Overzicht van het aantal gebruikte rekenmethoden

Rekenmethode	Aantal leerkrachten	Percentage
Wereld in getallen (oud)	19	11
Wereld in getallen (nieuw)	47	28
Operatoir rekenen (oud)	10	6
Operatoir rekenen (nieuw)	7	4
Rekenen en wiskunde	46	27
Pluspunt	34	20
Andere lesmethode	5	3
Onbekend	2	1

Uit tabel 4.4 blijkt dat de meeste aan het onderzoek deelnemende leerkrachten gebruik maken van 'Rekenen en wiskunde' en de nieuwe versie van ' Wereld in getallen'. Leerkrachten die de nieuwe versie van 'Operatoir rekenen' gebruiken zijn het minst vertegenwoordigd in de steekproef. De vraag is nu of er een relatie te leggen is tussen het gebruik van een bepaalde rekenmethode en de maakbaarheid van de toets E3? Bovendien is het van belang te weten of bij het gebruik van een bepaalde rekenmethode systematisch bepaalde leerstofonderdelen onvoldoende aan bod komen, waardoor deze voor leerlingen niet maakbaar zouden zijn. Indien een duidelijke positieve relatie tussen rekenmethode en maakbaarheid aanwezig is, kan wellicht volstaan worden met het vragen naar de rekenmethode. De rekenmethode bepaalt vervolgens de in een toets op te nemen items.

In tabel 4.5 staat een overzicht van de gemiddelde maakbaarheidsscores per rekenmethode. De standaarddeviatie geeft de spreiding aan van de maakbaarheidsscores van de leerkrachten die dezelfde methode hanteren.

Tabel 4.5

Gemiddelde maakbaarheidsscore per rekenmethode

Rekenmethode	Gemiddelde maakbaar-heidsscore over items	Standaarddeviatie
Wereld in getallen (oud)	43,3	6,3
Wereld in getallen (nieuw)	44,5	6,6
Operatoir rekenen (oud)	47,8	4,2
Operatoir rekenen (nieuw)	47,9	4,9
Rekenen en wiskunde	47,0	5,5
Pluspunt	42,1	6,9

Uit tabel 4.5 blijkt dat er niet alleen verschillen zijn in maakbaarheidsscores tussen rekenmethoden (F = 3,429, p = 0,006), maar ook binnen rekenmethoden. Blijkbaar is de rekenmethode alleen geen goede indicatie voor de mate van maakbaarheid van een toets. Daar komt bij dat uit de vragenlijst naar voren is gekomen dat leerkrachten zich niet altijd alleen beperken tot de rekenmethode, maar dat zij ook gebruik maken van extra materialen of juist onderwerpen uit de rekenmethode niet behandelen. Slechts 61% van de leerkrachten gaf aan strikt volgens de rekenmethode te werken.

De in tabel 4.5 weergegeven resultaten zijn gemiddelde oordelen van leerkrachten per rekenmethode over de 53 items. Deze oordelen geven niet aan in welke mate binnen een rekenmethode aandacht besteed wordt aan een bepaalde subcategorie. Dat deze aandacht verschilt, laat tabel 4.6 zien.

Tabel: 4.6

Overzicht maakbaarheid categorieën uitgedrukt in procenten per rekenmethode

Rekenmethode	% maakbare categorieën													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wereld in getallen (oud)	100	89	89	74	84	68	95	89	84	95	95	100	100	53
Wereld in getallen (nieuw)	98	98	98	64	83	66	87	85	68	98	98	91	91	70
Operatoir rekenen (oud)	100	90	90	80	80	90	100	100	90	100	100	80	80	100
Operatoir rekenen (nieuw)	100	100	100	100	86	100	100	100	100	100	100	71	71	100
Rekenen en wiskunde	91	88	74	74	85	82	100	88	79	91	97	68	68	21
Pluspunt	96	96	89	89	100	91	98	96	85	93	93	91	91	57

De kolommen 1 tot en met 14 in tabel 4.6 corresponderen met de in toets E3 onderscheiden 14 subcategorieën. In tabel 4.6 is per categorie het percentage docenten aangegeven dat deze categorie als maakbaar beschouwt. Zo beschouwt 89% van de docenten die de methode ‘Wereld in getallen’ (oude uitgave) gebruikt categorie 2 (Resultatief en structurerend tellen) als maakbaar en mogen volgens het door hen gegeven onderwijs over deze categorie items aan leerlingen voorgelegd worden.

Tabel 4.6 laat zien dat de 14 subcategorieën niet allemaal evenveel aandacht krijgen in de methoden. Opvallend is subcategorie ‘Tijd’ (kolom 14). Met name de leerkrachten die de rekenmethode ‘Pluspunt’ hanteren, maar ook de leerkrachten die gebruik maken van ‘Rekenen en wiskunde’ en ‘Wereld in getallen’ (oude uitgave) geven aan deze categorie niet zo maakbaar te vinden. Daarentegen vinden de leerkrachten die ‘Operatoir rekenen’ gebruiken deze categorie maakbaar. Dit geldt zowel voor de oude als voor de nieuwe uitgave van deze rekenmethode. Ook bij de andere onderscheiden categorieën zijn er verschillen in maakbaarheid tussen de verschillende rekenmethoden.

4.3.4 Het maken van een keuze voor een meetmethode TCO

In het voorgaande zijn drie mogelijke meetmethoden voor het meten van TCO besproken. Welke meetmethode het meest geschikt is, hangt van een aantal factoren af.
Een belangrijke factor is de tijdsinvestering die een meetmethode vraagt van de gebruiker. Alle onderzochte meetmethoden vragen weinig tijd van de gebruiker. Hoewel de 'itemmethode' het meest arbeidsintensief is, blijkt deze methode minder dan tien minuten aan tijd te vragen van de leerkrachten. Daar komt bij dat de afname van de toetsen slechts tweemaal per jaar plaatsvindt. Op grond van deze resultaten is het verschil in tijdsinvestering tussen de drie meetmethoden geen reden om aan één van deze drie methoden de voorkeur te geven.

Uit het onderzoek blijkt dat leerkrachten zich bij hun onderwijs niet altijd beperken tot dat wat de rekenmethode hen aanreikt. Bovendien blijken de maakbaarheidsscores van leerkrachten die dezelfde rekenmethode hanteren, te verschillen. Op basis van deze twee constateringen wordt geconcludeerd dat het vragen naar de 'lesmethode' geen goede meetmethode is voor het vaststellen van TCO.

Uit het onderzoek blijkt ook dat als een leerkracht aangeeft een(sub)categorie als maakbaar te zien, dat niet altijd geldt voor alle items binnen deze (sub)categorie. Zo blijkt dat 92% van de leerkrachten categorie 8 'Aftrekken' als maakbaar te beschouwen. Deze categorie bevat 7 items, met een gemiddelde maakbaarheidspercentage van 76%. De oordelen per item lopen uiteen van 58% tot 94%. De 'itemmethode' geeft dus concretere informatie dan de 'categoriemethode' en sluit beter aan bij de onderwijspraktijk van de individuele leerkracht.

Op grond van deze bevindingen wordt geconcludeerd dat op basis van onderwijsinhoudelijke redenen de 'itemmethode' als meetmethode voor TCO het beste aansluit bij de onderwijspraktijk. Zowel met verschillen in keuze voor en gebruik van een rekenmethode, als met discrepanties tussen de opvatting over de maakbaarheid van een categorie met de bij deze categorie behorende items,

wordt bij deze meetmethode rekening gehouden. Ook Pelgrum (1989) en De Haan (1992) geven de voorkeur aan de ‘itemmethode’.
De keuze voor een adequate meetmethode wordt echter niet alleen bepaald door onderwijsinhoudelijke en praktische redenen, maar ook door psychometrische. Duidelijk zal moeten zijn dat de ontwikkelde vragenlijst waarmee TCO vastgesteld gaat worden, inderdaad het concept 'maakbaarheid' op een valide en betrouwbare manier meet.

Validiteit en betrouwbaarheid van de 'itemmethode'

Met behulp van de itemresponstheorie is nagegaan of de 'itemmethode' schaalbaar is, dat wil zeggen dat de vragen uit de vragenlijst alle hetzelfde concept 'maakbaarheid' meten. De resultaten van het onderzoek naar schaalbaarheid worden hier kort samengevat weergegeven. Voor een uitvoerige beschrijving wordt verwezen naar Van Abswoude (1999).
Allereerst is de vragenlijst met het Raschmodel onderzocht wat een slechte modelfit opleverde. Ook het verwijderen van slecht fittende items of items met een extreme hoge p-waarde, leverde geen betere modelpassing op. Toepassing van het model OPLM (Verhelst & Eggen, 1989) leverde een goede passing van de 53 vragen uit de vragenlijst op. Het verwijderen van vragen met een extreem hoge p-waarde leverde geen betere passing op. Gegeven het voorgaande werd geconcludeerd dat het mogelijk is een eendimensionele schaal te ontwikkelen met als latente trek 'maakbaarheid'. Dat met de itemmethode ook betrouwbaar gemeten kan worden, blijkt uit een Cronbach’s alpha van .88.
Het door Van Abswoude uitgevoerde onderzoek laat zien dat de 'itemmethode' een valide en betrouwbare manier is om het concept 'maakbaarheid' te meten. In de volgende paragraaf komt de implementatie van het TCO-instrument aan bod.

4.3.5 Implementatie van het TCO-instrument in EVADOS

Bij de implementatie van het instrument TCO spelen praktische en psychometrische overwegingen een rol. In deze paragraaf komen 3 mogelijkheden om het TCO-instrument in te zetten aan bod. Ook wordt ingegaan op de praktische en psychometrische overwegingen en hoe deze van invloed kunnen zijn op de te maken keuze.

Drie mogelijkheden om het instrument TCO in te zetten zijn:

1 *Het TCO-instrument als 'entreemeting'.*
Bij deze toepassing gaat de leerkracht op basis van de toets E3, of een parallelle vorm daarvan, na in hoeverre er sprake is van TCO. Indien blijkt dat er een (groot) verschil is tussen de inhoud van de toets en het geboden onderwijs, heeft de leerkracht de mogelijkheid om voor de afname van de toets de nog niet onderwezen leerstof alsnog te onderwijzen. Is de leerkracht daartoe in staat, dan kan de toets daarna zonder probleem worden afgenomen en behoeft er geen correctie voor de vaardigheidsschattingen van leerlingen plaats te vinden. De schatting van de vaardigheid van de leerlingen vindt dan plaats op basis van alle items. Indien blijkt dat de discrepantie tussen toets en het geboden onderwijs te groot blijft, dient de toets niet te worden afgenomen. Later zal nader worden ingegaan op de beslissingsregel om een toets wel of niet af te nemen.

2 *Voor afname van de toets met het TCO-instrument vaststellen welke items maakbaar zijn en de vaardigheid van de leerlingen schatten op basis van hun score op deze voor hen maakbare items.*
Bij deze toepassing krijgen de leerlingen alleen die items voorgelegd die de leerkracht als maakbaar beschouwt. Niet maakbare-items worden uit de toets verwijderd. Deze toepassing wordt in het vervolg aangeduid als 'correctie vooraf'.

3 *Na afname van de toets met het TCO-instrument vaststellen welke items maakbaar zijn en de vaardigheid van de leerlingen schatten op basis van hun score op deze items.*

Bij deze toepassing krijgen de leerlingen alle items voorgelegd, ongeacht of een item voor hen wel of niet maakbaar is. Vervolgens gaat de leerkracht met het TCO-instrument na welke items niet maakbaar zijn. Na afname van de toets vindt een correctie plaats voor het aantal niet-maakbare items. Deze toepassing wordt in het vervolg aangeduid als 'correctie achteraf'.

Hoewel het onderzoek naar schaalbaarheid van de vragenlijst heeft aangetoond dat het mogelijk is de opvatting van leerkrachten over de maakbaarheid van items te kwantificeren, valt het toepassen van een 'correctie achteraf' om psychometrische redenen af. Deze toepassing vereist namelijk een hoge latente correlatie tussen de opvatting van leerkrachten over de maakbaarheid van toets E3 en de feitelijke leerlingresultaten. Deze correlatie blijkt echter slechts .17 te zijn. Merk op dat leerlingen genest zijn binnen leerkrachten. Als een leerkracht een oordeel geeft over de maakbaarheid van een item, geldt zijn oordeel voor alle leerlingen uit zijn klas. Indien de opvattingen van de leerkrachten gecorreleerd worden met de gemiddelde leerresultaten van hun leerlingen, is de correlatie .34. Uit deze lage correlaties wordt geconcludeerd dat 'correctie achteraf' geen goede optie is. Ook om praktische redenen is 'het corrigeren achteraf' niet aan te bevelen. Om te kunnen corrigeren zal bij implementatie van het instrument in de praktijk een invoerscherm op de computer of een formulier ontwikkeld moeten worden waarmee leerkrachten kunnen aangeven welke items maakbaar zijn.

Ook bij de toepassing 'correctie vooraf' dient een invoerscherm of een formulier ontwikkeld te worden waarmee leerkrachten kunnen aangeven welke items maakbaar zijn. Bovendien vraagt deze methode mogelijk om organisatorische aanpassingen. Leerkrachten zullen aan leerlingen op de een of andere manier duidelijk moeten maken dat zij niet alle opgaven uit de toetsen hoeven te maken, maar dat zij er een aantal mogen overslaan. En bij deze methode geldt dat nieuwe omzettingstabellen voor de transformatie van ruwe scores naar schaalscores geconstrueerd moeten worden.

Zowel 'correctie vooraf' als 'correctie achteraf' hebben nog een belangrijk nadeel. De ontwikkelde pakketten die scholen gebruiken voor de opslag van toetsresultaten gaan uit van de hele toets. Voor scholen betekent dit dat zij de resultaten niet kunnen invoeren in de pakketten en derhalve ook geen gebruik kunnen maken van de faciliteiten die deze pakketten bieden, zoals bijvoorbeeld het gebruiken van normeringsgegevens bij de interpretatie van resultaten van leerlingen.

TCO als 'entreemeting' geniet de voorkeur. Bij deze methode wordt rekening gehouden met TCO en kan de hele toets, zonder dat er correctie hoeft plaats te vinden of extra gegevens verzameld dienen te worden, afgenomen worden. Deze methode heeft wel als uitgangspunt dat leerkrachten in staat zijn extra aandacht te besteden aan die onderwerpen die - gegeven de toets - nog onvoldoende in zijn onderwijs aan bod zijn geweest. Als blijkt dat TCO op het moment van de entreemeting te gering is, kan dat betekenen dat de leerkracht niet meer in staat is in voldoende mate (extra) aandacht te besteden aan bepaalde onderwerpen. In een dergelijke situatie zou besloten moeten worden de toets (op dat moment) niet af te nemen.

Wanneer dient een leerkracht geadviseerd te worden de toets niet meer af te nemen? De keuze is gelegd bij 20% van het totaal aantal items in de toets, waarbij aangesloten wordt bij het Cito-LVS dat (weliswaar op itemniveau en niet op toetsniveau) het 80%-niveau als beheersniveau hanteert. Indien slechts enkele items niet maakbaar zijn, is de aanname dat deze een verwaarloosbaar effect hebben op de schatting van de vaardigheid wanneer toch de hele toets wordt voorgelegd. Voor hoeveel procent van het aantal items dat geldt, is niet bekend. Arbitrair is gesteld dat meer dan 90% van het aantal items maakbaar moet zijn. Aan de hand van de resultaten op de vragenlijst en de resultaten van de leerlingen van deze leerkrachten op de toets E3, is nagegaan in hoeverre er empirische evidentie aanwezig is voor de gemaakte keuzen.

4.3.6 Effect beslisregel op niveau-indicatie Cito-LVS

In paragraaf 4.1.1 is aangegeven dat het Cito-LVS bij haar rapportage gebruik maakt van een vijftal niveaus. Welk niveau aan een leerling wordt toegekend, is afhankelijk van zijn vaardigheidsscore die afhangt van het aantal goed beantwoorde items. Bij het indelen in niveaus hanteert het Cito-LVS bij de toets Rekenen-Wiskunde E3 de volgende indeling:
A-niveau: 46 of meer items goed beantwoord;
B-niveau: 40 tot en met 45 items goed beantwoord;
C-niveau: 32 tot 40 items goed beantwoord;
D-niveau: 24 tot 32 items goed beantwoord;
E-niveau: minder dan 24 items goed beantwoord.

Het is evident dat de schatting van de vaardigheid bepaald wordt door het aantal correct beantwoorde (maakbare) items. Een verschil in geschatte vaardigheid op basis van de toets en op basis van alleen de maakbare items, behoeft echter niet per se te leiden tot een verschil in niveau-indicatie zoals het Cito-LVS dat hanteert. En in hun praktijk gaan leerkrachten uit van deze niveau-indicaties. Om het effect van maakbaarheid vast te stellen, zijn de leerkrachten op basis van hun maakbaarheidsscores ingedeeld in de volgende drie groepen[2]:

- leerkrachten met een maakbaarheidsscore van 42 (of minder), hetgeen overeenkomt met (ongeveer) 80% van het totaal aantal van 53 items;
- leerkrachten met een maakbaarheidsscore van 43 tot en met 48;
- leerkrachten met een maakbaarheidsscore van 49 (of meer), hetgeen overeenkomt met (ongeveer) 90% van het totaal aantal van 53 items.

Per groep zijn in een kruistabel de niveau-indicaties op basis van de gehele toets (53 items) en op basis van alleen de maakbare items met elkaar vergeleken (zie

[2] Merk op dat de toets Rekenen-Wiskunde E3 in totaal uit 53 items bestaat. Gesteld is dat als minder dan 80% van de items door de leerkracht als maakbaar beschouwd wordt, de toets niet meer afgenomen dient te worden. 80% van het aantal items komt overeen met 42. Ingeval minder dan 10% van het aantal items niet maakbaar is, kan de toets in zijn geheel afgenomen worden. 10% van het aantal items is gelijk gesteld aan 5 items.

tabel 4.7). In de bespreking van tabel 4.7 worden de drie groepen aangeduid als 'groep < 43', 'groep 43-48' en 'groep > 48'.

Tabel 4.7

Indeling in Cito-LVS schaalscores (SS) van leerlingen verdeeld over drie groepen van maakbaarheidsscores in absolute aantallen

		SS-LVS Alle items (53)					
SS-LVS		A	B	C	D	E	Tot.
Maakbaar-heidsscore < 43	A	365	97	3			465
	B	35	228	73			336
	C		46	151	39		236
	D			13	60	5	78
	E				6	17	23
	Tot.	400	371	240	105	22	1138
Maakbaar-heidsscore 43 – 48	A	391	28				419
	B	28	201	17			246
	C		19	142	6		167
	D			7	52	2	61
	E					26	26
	Tot.	419	248	166	58	28	919
Maakbaar-heidsscore > 48	A	586	20				606
	B	20	334	8			362
	C		9	162			174
	D			3	434	2	48
	E				2	16	18
	Tot.	606	363	173	48	18	1208

Voor de bespreking van tabel 4.7 zijn drie opmerkingen van belang.

1 Uit het TCO-onderzoek blijkt dat in totaal 133 verschillende antwoordpatronen van leerkrachten over de maakbaarheid van items te onderscheiden zijn. Met een antwoordpatroon wordt de combinatie van (uit de in totaal 53) items bedoeld die door leerkrachten als maakbaar worden beschouwd. Bij de indeling in drie groepen en het berekenen van de daarmee corresponderende LVS-niveaus is geen rekening gehouden met deze antwoordpatronen. Ook aan de bijdrage van afzonderlijke items aan de geschatte vaardigheid is voorbijgegaan. In principe is het mogelijk dat toetsen met hetzelfde aantal maakbare doch verschillende items tot een andere niveau-indeling leiden.
2 Het vaststellen van het vaardigheidsniveau van de leerlingen gaat gepaard met schattingsfouten. Op basis van deze schattingsfouten is het mogelijk dat niveau A-leerlingen ook geplaatst zouden kunnen worden in niveau B en omgekeerd. Hetzelfde geldt voor de andere niveaus. Bij de indeling van de leerlingen in de diverse niveaus is geen rekening gehouden met de schattingsfout die kan optreden. Uit een onderzoek naar de verschillen tussen de geschatte vaardigheden op basis van alle items en op basis van alleen de maakbare items, bleek het aantal significante verschillen zeer beperkt te zijn. In de niveau-toekenning van de leerlingen zijn de mogelijke misclassificaties ten gevolge van schattingsfouten dan ook buiten beschouwing gelaten.
3 De vaardigheid van de leerlingen is in de tijd toegenomen. Verhoudingsgewijs hebben meer leerlingen een hogere niveau-indicatie dan een aantal jaren geleden. Zo blijkt voor de 'groep > 48' dat op basis van de huidige resultaten 48,5% van de leerlingen het hoogste niveau (A) krijgt, terwijl de in 1990 opgestelde normeringsgegevens ervan uitgaan dat dit percentage 25% is. Ook bij de andere niveaus zien we een dergelijke verschuiving. Slechts 1,3% van deze leerlingen bevindt zich op het laagste niveau (E), in tegenstelling tot de 10% volgens de normeringsgegevens.

Uit tabel 4.7 blijkt dat voor de 'groep < 43' geldt dat bij 72,1% van de leerlingen de niveau-indicatie hetzelfde blijft, ongeacht of deze gebaseerd is op de resultaten van de hele toets of alleen op de resultaten op de maakbare items. Bij

de 'groep 43-48' is dit 88,4% en bij de 'groep > 48' geldt dit voor 94,4% van de leerlingen.

Ingeval er sprake is van een verschil in niveau-indicatie, dan is dit verschil met name terug te vinden bij de hogere niveaus. Voor alle niveaus geldt dat als uitgegaan wordt van de maakbare items, meer leerlingen een hogere indicatie zouden krijgen dan wanneer uitgegaan zou worden van alle items. Het aantal leerlingen waarvoor dit geldt, neemt (verhoudingsgewijs) af met de toename van het aantal maakbare items.

Ter illustratie:

- Voor de 'groep < 43' (totaal 1138 leerlingen) geldt dat 72,1% van de leerlingen dezelfde niveau-indicatie zou krijgen als uitgegaan wordt van alle items of alleen van de maakbare items. Voor 217 leerlingen (19%) geldt dat hun niveau op alleen de maakbare items hoger is. Van deze leerlingen behoren 173 (15,2%) tot de categorieën A t/m C en 44 leerlingen (3,9%) tot de categorieën D en E. Voor 100 leerlingen (8,8%) geldt dat hun niveau op alleen de maakbare items lager is. Van de leerlingen behoren 94 (8,3%) tot de categorieën A t/m C en 6 leerlingen (0,5%) tot de categorieën D en E.
- Voor de 'groep 43-48' (totaal 919 leerlingen) geldt dat 88,4% van de leerlingen dezelfde niveau-indicatie zou krijgen als uitgegaan wordt van alle items of alleen de maakbare items. Voor 53 leerlingen (5,8%) geldt dat hun niveau op alleen de maakbare items hoger is. Van deze leerlingen behoren er 45 (4,9%) tot de categorieën A t/m C en 8 leerlingen (0,9%) tot de categorieën D en E. Voor 54 leerlingen (5,9%) geldt dat hun niveau op alleen de maakbare items lager is. Deze leerlingen behoren allen tot de categorieën A t/m C.
- Voor de 'groep > 48' (totaal 1208 leerlingen) geldt dat 94,4% van de leerlingen dezelfde niveau-indicatie zou krijgen als uitgegaan wordt van alle items of alleen de maakbare items. Voor 33 leerlingen (2,7%) geldt dat hun niveau op alleen de maakbare items hoger is; 28 van deze leerlingen (2,3%) behoren tot de categorieën A t/m C en 5 leerlingen (0,4%) tot de categorieën D en E. Voor 34 leerlingen (2,8%) geldt dat hun niveau op alleen de

maakbare items lager is; 32 van deze leerlingen (2,6%) behoren tot de categorieën A t/m C en 2 leerlingen (0,2%) tot de categorieën D en E.

Voor klassen met veel D- of E-leerlingen maakt het gemiddeld genomen minder uit of hun niveau-indeling plaatsvindt op basis van alle items of op basis van alleen de maakbare items. Voor klassen met vooral A- en B-leerlingen maakt het mogelijk wel enig verschil, zij het dat de mate waarin bepaald wordt door de maakbaarheidsscore van de leerkracht. Hoe hoger de maakbaarheidsscore des te geringer is het effect.

Conclusie

Het onderzoek naar de beslisregel laat zien dat het verantwoord lijkt te komen tot de volgende tweedeling:

- Indien het aantal maakbare items gelijk is aan 42 of minder is het advies de toets niet af te nemen. Het aantal misclassificaties om te komen tot een niveau-indeling op basis van alle 53 items en op basis van alleen de maakbare items is te groot (27,9%).
- Indien het aantal maakbare items groter is dan 42 van de in totaal 53 items, kan de toets in zijn geheel afgenomen worden. Het aantal misclassificaties is bij deze groep erg klein en voor zover er sprake is van misclassificaties komen deze met name voor bij de hogere niveaus. Merk op dat bij een zelfgerichte vergelijking (in de tijd) een verschil in maakbaarheid tussen opeenvolgende afnames mogelijk wel een effect heeft.

Hoewel uit de kruistabellen blijkt dat het verantwoord is een tweedeling te maken, is het toch zinvol bij de implementatie van het TCO-instrument ook stil te staan bij de 'groep 43-48'. Ten eerste om onderwijsinhoudelijke redenen. Hoe lager de maakbaarheidsscores des te minder sluit de toets aan bij het geboden onderwijs. Ten tweede omdat uit de analyse blijkt dat circa 43% van de leerkrachten in het TCO-onderzoek een maakbaarheidsscore van 47 of 48 heeft. Als het voor deze leerkrachten mogelijk is de ontbrekende lesstof te behandelen, behoren zij ook tot de 'groep > 48'. Van deze groep is vastgesteld dat het aantal misclassificaties zeer beperkt is.

In de vorige paragrafen is vastgesteld hoe het TCO-instrument vormgegeven kan worden en welke beslisregel gehanteerd kan worden. Op basis van de resultaten van de analyses wordt voorgesteld dat de leerkracht twee à drie maanden voor de geplande afname van de toets E3 de items van deze toets of een inhoudelijk vergelijkbare toets doorneemt met als doel na te gaan of de vereiste lesstof voor het kunnen beantwoorden van deze items in voldoende mate aan bod is geweest (maakbaar is). Dat wil zeggen dat de leerkracht de lesstof behandeld heeft en de leerlingen voldoende met de lesstof hebben kunnen oefenen. Bij het beoordelen van de mate van maakbaarheid van de items houdt de leerkracht rekening met het nog resterende onderwijsprogramma.

Indien blijkt dat op het geplande moment van afname van toets E3 meer dan 48 items maakbaar zijn, neemt de leerkracht de volledige toets af. De indeling van de leerlingen in de onderscheiden LVS-niveaus vindt plaats op basis van de volledige toets.

Indien blijkt dat op het geplande moment van afname van toets E3 42 of minder items maakbaar zullen zijn, neemt de leerkracht de toets niet af.

Indien blijkt dat op het geplande moment van afname van toets E3 meer dan 42, maar minder dan 49 items maakbaar zijn, neemt de leerkracht de volledige toets af. Ook hier geldt dat de leerkracht bij het vaststellen van het aantal maakbare items op het moment van afname van de toets rekening houdt met het nog resterende onderwijsprogramma. Leerkrachten die tot deze groep behoren, proberen voor zover het programma het toelaat, door extra lesstof te behandelen deel uit te gaan maken van de 'groep > 48'.

Leerkrachten die op het moment van afname deel uitmaken van de 'groep 43-48', dienen er rekening mee te houden dat zich in zijn klas, met name bij de hogere niveaus, kleine afwijkingen in de toekenning van het LVS-niveau kunnen voordoen.

4.3.7 Toelichting op de implementatie van TCO

Bij een maakbaarheidsscore van minder dan 43, heeft slechts 72,1% van de leerlingen een niveau-indicatie die vergelijkbaar is met de indicatie die zij zouden krijgen op basis van de hele toets. 27,9% van de leerlingen krijgt een andere indeling, hetgeen als te veel gezien wordt. Uit het TCO-onderzoek blijkt dat bepaalde leerstofonderdelen bij bepaalde methoden duidelijk minder aan bod komen. Dit betreft dan met name de categorie ‘Meten en Tijd’. Hoewel slechts 168 leerkrachten bij het onderzoek betrokken zijn, blijkt uit een analyse dat bij die leerkrachten die aangeven minder dan 43 items als maakbaar te beschouwen, meer leerstofonderdelen als niet behandeld naar voren komen. Een extra (onderwijskundige) reden om de toets niet af te nemen.

Bij een maakbaarheidsscore van 49 of meer blijkt dat 94,4% van de leerlingen eenzelfde niveau-score krijgt als bij afname van de volledige toets. Het aantal leerlingen dat een ander niveau krijgt, is zeer beperkt. In een klas van 30 leerlingen betreft het gemiddeld 1 tot 2 leerlingen, die dan met name tot de hogere LVS-niveaus behoren. Voor de lagere niveaus (D en E) maakt het niets uit (0,2% van de leerlingen).

De meeste aandacht vraagt de groep met een maakbaarheidsscore van 43 tot en met 48. Uit de analyses blijkt dat als leerkrachten in staat zijn de lesstof van twee items te behandelen, ongeveer 45% van deze leerkrachten gaan behoren tot de 'groep > 48'. Van deze groep is vastgesteld dat correctie voor de maakbaarheid van items niet behoeft plaats te vinden. Voor zover er in deze groep sprake is van 'misclassificatie' betreft dat ook hier met name leerlingen die behoren tot de hogere niveaus. Voor de lagere niveaus (D en E) is het gemiddeld minder dan één leerling per klas.

Voor de keuze om alleen te kiezen voor de genoemde tweedeling zijn ook praktische argumenten op te voeren. Ten eerste heeft deze keuze tot gevolg dat er geen voorziening ontwikkeld behoeft te worden, waarmee de leerkracht aangeeft hoeveel items maakbaar zijn. Immers de toets wordt wel of de toets

wordt niet in zijn geheel afgenomen. Ook vraagt deze tweedeling niet om een aanpassing van het huidige LVS-systeem waar de vaardigheid geschat wordt op basis van alle items. Zowel aan de kant van de leerkracht als aan de kant van de instrumentontwikkelaars levert deze keuze een voordeel op. Ten tweede sluit het afnemen van de volledige toets en niet een gedeelte daarvan goed aan op de huidige systematiek zoals deze momenteel toegepast wordt in het computer-programma van het Cito-LVS en andere toetsregistratiepakketten, zoals bijvoorbeeld ESIS-B. Deze beide systemen gaan uit van de resultaten op de totale toets. Door te kiezen voor het afnemen van de volledige toets, kunnen de resultaten van de leerlingen ook direct in beide systemen ingevoerd worden en behoeft er geen extra inspanning van de leerkrachten gevraagd te worden.

Naast praktische argumenten zijn er ook toekomstgerichte argumenten. Het onderzoek naar de ontwikkeling van EVADOS is gestart in de context van 1995. Sinds die tijd hebben zich vele ontwikkelingen voorgedaan. Zo vindt in het onderwijs een verschuiving plaats van ‘paper & pencil tests’ naar ‘computer based tests’ (CBT’s). In één van de te onderscheiden vormen van CBT’s, de adaptieve toetsen, speelt TCO een belangrijke rol. Een kenmerk van adaptieve toetsen is dat deze tijdens de afname samengesteld worden. Afhankelijk van het antwoord (goed of fout) op een item wordt een leerling een ander item aangeboden. In principe krijgt dus elke leerling een andere toets, met als gevolg dat niet van alle leerlingen dezelfde gegevens verzameld worden (onvolledig design). Ook leerkrachten weten niet welke items uit de beschikbare itembank aan de leerlingen worden voorgelegd. Onderzoek naar de relatie tussen TCO en deze vorm van toetssamenstelling en toetsafname lijkt erg zinvol. Voor zover een leerkracht niet de mogelijkheid heeft kennis te nemen van alle mogelijke items uit de itembank waaruit geselecteerd kan worden, is hij of zij niet in staat een uitspraak te doen over de mate van TCO. Bovendien kan deze, gegeven de aard van de afname, per afname verschillend zijn. Het lijkt het meest praktisch om ook gegeven deze ontwikkelingen het beperkt aantal misclassificaties bij de geschetste inzet van het TCO-instrument te accepteren. Wat de relatie is tussen TCO en CBT’s is een onderwerp dat om nader onderzoek vraagt.

5 Multilevel analyses

Het in hoofdstuk 2 besproken geïntegreerd model voor schooleffectiviteit laat zien welke variabelen van invloed kunnen zijn op de door leerlingen behaalde resultaten. In het model wordt een onderscheid gemaakt tussen variabelen op schoolniveau en variabelen op klas- en leerlingniveau. Gegevens van een hogere orde (school ten opzichte van klas en klas ten opzichte van leerling) kunnen de gegevens van een lagere orde beïnvloeden. In de multilevel literatuur spreekt men van clustering (nesting) wanneer de resultaten van bijvoorbeeld leerlingen beïnvloed worden door de klas waarvan zij deel uitmaken: de gegevens zijn zogenaamd hiërarchisch gestructureerd. Vanwege het hiërarchisch karakter van de gegevens uit de schooladministratiepakketten en de toetsadministratiesystemen zijn multilevel methoden nodig om effecten van variabelen die de resultaten van leerlingen zouden kunnen beïnvloeden te kunnen schatten. Ook Goldstein (1999), Snijders en Bosker (1999), Kreft en De Leeuw (1998) en Bosker en Scheerens (1995) onderschrijven het belang van het gebruik van multilevel modellen wanneer de te analyseren gegevens hiërarchisch gestructureerd zijn.

Goldstein (1999, p 2-3) noemt de volgende voordelen van multilevel analyses:

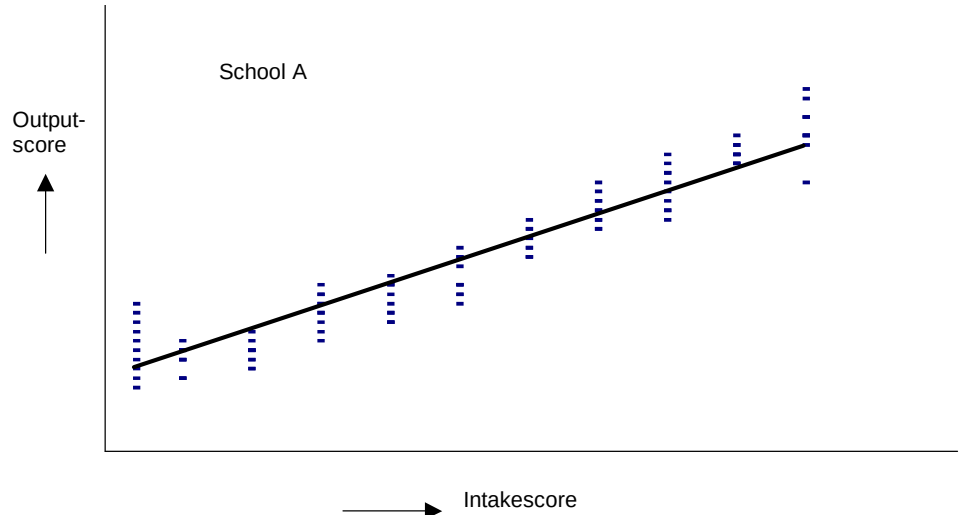
- verkrijgen van statistisch efficiënte schattingen van de regressiecoëfficiënten;
- correct schatten van standaardfouten, betrouwbaarheidsintervallen en significantietests;
- kunnen meten van covariaten op elk niveau, waardoor de invloed van bijvoorbeeld achtergrondkenmerken op de afhankelijke variabele (zoals de resultaten op toetsen) vastgesteld kan worden;
- vaststellen van schooleffecten op verschillende typen leerlingen (bijvoorbeeld uitgesplitst naar sociaal economische status);

- het plaatsen van scholen in een bepaalde rangorde op basis van de prestaties van leerlingen op toetsen, na controle voor beginniveau en achtergrondkenmerken.

In dit hoofdstuk worden eerst de principes van multilevel analyses uiteengezet. Aansluitend worden vijf multilevel modellen om verschillen in eindniveau en leerwinst te analyseren, drie univariate modellen en twee multivariate modellen, besproken. Tot slot wordt kort ingegaan op het vaststellen van schooleffecten

5.1 Het principe van multilevel analyses

Figuur 5.1 geeft voor een willekeurige school A de relatie weer tussen de resultaten op een intaketoets en de resultaten op een toets na een bepaalde onderwijsperiode (outputtoets). De puntjes in de figuur zijn de (fictieve) geobserveerde scores van de leerlingen.



Figuur 5.1
Relatie intakescore met outputscore

Uit figuur 5.1 blijkt dat er voor deze school een positieve relatie is tussen intakescore en outputscore. De in figuur 5.1 weergegeven regressielijn is als volgt weer te geven:

$$Y_i = \alpha + \beta x_i + e_i \qquad (5.1)$$

In (5.1) stelt Y_i de outputscore van leerling *i* in de school voor en x_i zijn intakescore. Het intercept - het snijpunt van de regressielijn met de y-as - wordt weergegeven door α en de helling van de regressielijn door β. De afwijking (residu) van de outputscore van leerling *i* ten opzichte van de voorspelde score op basis van de regressielijn voor de school is e_i.

Vergelijking (5.1) beschrijft een één-niveau relatie, die door het plaatsen van het subscript *j* zoals weergegeven in vergelijking (5.2) uitgebreid kan worden voor meerdere scholen.

$$Y_{ij} = \alpha_j + \beta_j x_{ij} + e_{ij} \qquad (5.2)$$

In (5.2) stelt Y_{ij} nu de outputscore van leerling *i* in school *j* voor, en x_{ij} zijn intakescore. Het subscript *j* staat voor de scholen ($j = 1 \ldots J$) en het subscript *i* staat voor individuele leerlingen ($i = 1 \ldots n_j$). De toevoeging van het subscript *j* aan het intercept α_j en de helling β_j in vergelijking (5.2) geeft aan dat elke school zijn eigen intercept en helling heeft. Met β_j (scholen hebben een eigen hellingshoek) geven we aan dat de relatie tussen de verklarende variabele (bijvoorbeeld intakescore) en de afhankelijke variabele (outputscore) voor scholen verschillend is. Een hoge β duidt op een grote invloed en een lage β duidt op een geringe invloed van de intakescore op de outputscore. Als we in een model veronderstellen dat alle scholen dezelfde helling hebben, geven we dit aan door de hellingscoëfficiënt β niet te voorzien van een subscript *j*. Wanneer we te maken hebben met meerdere scholen die elk hun eigen intercept hebben zoals weergegeven door α_j, maar daarentegen dezelfde helling (β), dan resulteert

dat in een grafiek met evenwijdige lijnen, waarbij de hellingshoek van de lijnen bepaald wordt door de grootte van β. Het snijpunt van de lijnen met de Y-as wordt bepaald door de grootte van het intercept α van elke school. De afwijking van de outputscore van leerling *i* ten opzichte van de voorspelde score op basis van de regressielijn voor school *j* is e_{ij}. Het gemiddelde over de leerlingen van de error-term e_{ij} binnen een school kan geïnterpreteerd worden als een schooleigenschap.

In sommige situaties waarbij er sprake is van slechts een aantal scholen, kan van vergelijking (5.2) gebruik gemaakt worden door alle 2n + 1 parameters te schatten, te weten: α_j, β_j, voor j=1,…,n en σ_e^2, waarbij ervan uitgegaan wordt dat de 'binnenschoolse' residuele variantie voor alle scholen gelijk is en elke school een afzonderlijke regressielijn heeft.
Een bijzondere situatie ontstaat wanneer we met name geïnteresseerd zijn in individuele scholen, maar waar we bovendien een groot aantal scholen hebben, zodat het toepassen van vergelijking (5.2) leidt tot de schatting van een groot aantal parameters. Bovendien kunnen sommige scholen slechts een klein aantal leerlingen hebben, waardoor toepassing van vergelijking (5.2) leidt tot onnauwkeurige schattingen voor die scholen. Door nu in dergelijke gevallen de scholen te zien als een deel van een grotere populatie scholen en door gebruik te maken van de populatie-schattingen van het gemiddelde en de tussen-school variantie, kunnen we op basis van deze informatie nauwkeurigere schattingen krijgen voor elke individuele school.
Om het voorgaande toe te kunnen passen, dient vergelijking (5.2) gewijzigd te worden in een twee-niveau model, waarbij we het intercept α en de helling β als random variabelen beschouwen. Praktisch betekent dit dat α en β per school kunnen verschillen. In navolging van Goldstein (1999) wordt α_j vervangen door β_{0j} en β_j door β_{1j}, en wordt aangenomen dat $\beta_{0j} = \beta_0 + \mu_{0j}$ en dat $\beta_{1j} = \beta_1 + \mu_{1j}$.
In het bovenstaande is β_{0j} het intercept van school *j*. β_0 is het algemeen schoolgemiddelde en μ_{0j} is de afwijking van school *j* ten opzichte van het algemeen schoolgemiddelde. β_{1j} is de helling van de regressielijn van school *j*. β_1 is de gemiddelde helling van alle scholen, en μ_{1j} is de afwijking van de helling

van de regressielijn van school *j* ten opzichte van de gemiddelde helling β_1. Hierbij zijn μ_{0j} en μ_{1j} random variabelen met parameters:
$\mathrm{E}(\mu_{0j}) = \mathrm{E}(\mu_{1j}) = 0$ en
$\mathrm{var}(\mu_{0j}) = \sigma^2{}_{\mu 0}$, $\mathrm{var}(\mu_{1j}) = \sigma^2{}_{\mu 1}$ en $\mathrm{cov}(\mu_{0j}, \mu_{1j}) = \sigma_{\mu 01}$

Gegeven voornoemde aannames kan vergelijking (5.2) geschreven worden als:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \left(\mu_{0j} + \mu_{1j} x_{ij} + e_{ij}\right) \tag{5.3}$$

In vergelijking (5.3) kan de afhankelijke variabele Y_{ij} gezien worden als de som van een 'fixed' deel en een 'random' deel. De variabelen uit het random deel staan in (5.3) tussen haakjes. De variantie μ_0 en μ_1 zijn variantiecomponenten op schoolniveau (niveau 2). De variantie e_{ij} is de variantie op leerlingniveau (niveau 1). In (5.3) is een verklarende variabele opgenomen die random is op niveau 2. β_1 geeft de algemene relatie weet tussen de verklarende variabele en de afhankelijke variabele. Dit is het 'fixed' deel. μ_{1j} is het random deel, dat voor een school *j* de afwijking weergeeft van de algemene relatie die door het fixed deel beschreven wordt.
Vergelijking (5.3) onderscheidt zich van een standaard lineair model (vergelijking 5.2) door de aanwezigheid van meer dan één residu, namelijk een residu op het eerste niveau, e_{ij} (leerlingen), en twee residuen op het tweede niveau: μ_{0j} en μ_{1j} (scholen).

Het basismodel zoals weergegeven door vergelijking (5.3) kan verder worden uitgebreid door aan de te onderscheiden niveaus extra verklarende variabelen toe te voegen. Bij het eerste niveau kan gedacht worden aan leerlingkenmerken, zoals sekse, en bij het tweede niveau aan schoolkenmerken, zoals bijvoorbeeld de denominatie van de school. Stel dat de verklarende variabelen op het eerste niveau (dat wil zeggen de leerling) weergegeven worden door $X_{1,\ldots,}X_p$ en die op het tweede niveau (dat wil zeggen de school) door $Z_{1,\ldots,}Z_q$, dan leidt dat tot[3]:

[3] In vergelijking (5.4) variëren de regressiecoëfficiënten $\beta_{10}...\beta_{p0}$ en $\beta_{01}...\beta_{0q}$ niet tussen scholen. Om die reden zijn ze niet voorzien van een subscript *j*.

$$Y_{ij} = \beta_0 + \beta_{10} x_{1ij} + ... + \beta_{p0} x_{pij} + \beta_{01} z_{1j} + ... + \beta_{0q} z_{qj} + \left(\mu_{0j} + e_{ij} \right) \qquad (5.4)$$

Hierbij is het eerste deel van de rechterkant van de vergelijking,

$$\beta_0 + \beta_{10} x_{1ij} + ... + \beta_{p0} x_{pij} + \beta_{01} z_{1j} + ... + \beta_{0q} z_{qj},$$

het 'fixed' deel van het model, daar de coëfficiënten 'fixed' zijn (niet stochastisch). Het deel tussen haakjes, $\mu_{0j} + e_{ij}$, is het 'random' gedeelte van het model. Ook nu wordt verondersteld dat de residuen μ_{0j} en e_{ij} onafhankelijk van elkaar zijn en dat
$\mu_{0j} \sim N\left(0, \tau_0^2\right)$ en $e_{ij} \sim N\left(0, \sigma^2\right)$.
Merk op dat de in het model op te nemen variabelen niet per definitie lineaire variabelen behoeven te zijn, maar dat het ook mogelijk is dat deze niet-lineair (bijvoorbeeld kwadratisch) zijn. Bovendien geldt dat er ook sprake kan zijn van een interactie tussen variabelen (cross-level interactie). Voor meer informatie over cross-level interactie effecten wordt verwezen naar Snijders en Bosker (1999, p. 73 e.v.).

Het nulmodel

De meeste multilevel analyses starten met het passen van een nulmodel. Dit model bevat slechts één responsvariabele (afhankelijke variabele) en geen verklarende variabelen anders dan het intercept. Door de resultaten van het nulmodel te vergelijken met de resultaten van een uitgebreider model met onafhankelijke variabelen kan vastgesteld worden hoeveel extra variantie verklaard wordt door de in het uitgebreider model opgenomen onafhankelijke variabelen.

Ter illustratie van de passing van het nulmodel is gebruik gemaakt van een voorbeeld dat ontleend is aan Snijders en Bosker (1999) die data analyseerden van 2287 leerlingen van groep 8 van in totaal 131 basisscholen. In hun voorbeeld is de afhankelijke variabele de score van de leerlingen op een taaltoets. In hun onderzoek gaan ze na in welke mate de intelligentie van een leerling en zijn

of haar sociaal economische status van invloed is op de in het onderzoek gebruikte taaltoets.

Het nulmodel laat zich als volgt weergeven:

$$Y_{ij} = \beta_0 + \mu_{0j} + e_{ij} \tag{5.5}$$

waarbij verondersteld wordt dat de residuen μ_{0j} en e_{ij} onafhankelijk van elkaar zijn en dat
$\mu_{0j} \sim N\left(0, \tau_0^2\right)$ en $e_{ij} \sim N\left(0, \sigma^2\right)$.

In vergelijking (5.5) is Y_{ij} de score van leerling i in school j, β_0 het algemeen gemiddelde, μ_{0j} de afwijking van school j ten opzichte van het algemeen gemiddelde en e_{ij} is de afwijking van leerling i ten opzichte van het gemiddelde van zijn school j.

De resultaten van het passen van het nulmodel staan weergegeven in tabel 5.1.

Het voorbeeld uit tabel 5.1 laat zien dat het algemeen gemiddelde 40,36 is met een standaarddeviatie van $\sqrt{19,42 + 64,57} = 9,16$. Het gemiddelde dient gezien te worden als de verwachte score op de taaltoets van een willekeurige leerling.

In het nulmodel zoals weergegeven door (5.5) is de afhankelijke variabele Y_{ij} (in dit geval de score op een taaltoets) gelijk aan de som van het algemeen gemiddelde (β_0) en een random-effect op schoolniveau (μ_{0j}) en een random-effect op individueel niveau (e_{ij}).

Tabel 5.1

Voorbeeld nulmodel (ontleend aan Snijders & Bosker, 1999)

Fixed effects	Coëfficiënt	S.E.
β_0 = intercept	40,36	0,43
Random Effect	**Variantiecomponent**	**S.E.**
Niveau 1 variantie: $\sigma^2 = \text{var}(e_{ij})$	64,57	1,97
Niveau 2 variantie: $\tau_0^2 = \text{var}(\mu_{0j})$	19,42	2,92
Deviantie	16253,2	

Uit tabel 5.1 kan afgeleid worden dat het algemeen gemiddelde op de taaltoets gelijk is aan 40,36 met een standaardfout van 0,43. De standaardfout is een indicatie voor de onnauwkeurigheid van de schatting. De deviantie is gelijk aan 16253,2. De deviantie, die niet direct interpreteerbaar is, is een maat die gebruikt wordt om na te gaan of het ene model beter past ten opzichte van een ander model, gegeven dezelfde data. Zowel de standaardfout als de deviantie zullen in het volgende onderdeel nader toegelicht worden bij de bespreking van de resultaten wanneer in het nulmodel een verklarende variabele opgenomen wordt (zie tabel 5.2).

Gegeven model (5.5) kan de totale variantie van Y_{ij} uiteengelegd worden in de som van de variantie op niveau twee en niveau één,

$$\text{var}(Y_{ij}) = \text{var}(u_{0j}) + \text{var}(e_{ij}) = \tau_0^2 + \sigma^2$$

en stelt het ons in staat de intraklassecorrelatiecoëfficiënt ρ uit te rekenen die gedefinieerd wordt als de proportie variantie op niveau twee (tussen groepsvariantie) ten opzichte van de totale variantie. In formulevorm:

$$\rho = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$$

Deze intraklassecorrelatiecoëfficiënt is een maat voor de (on)afhankelijkheid van individuen. Hoe sterker individuen binnen een groep op elkaar lijken des te hoger is de intraklassecorrelatiecoëfficiënt. De intraklassecorrelatiecoëfficiënt wordt bij traditionele regressiemethoden gebruikt om te controleren of de aanname van onafhankelijke observaties geschonden wordt. Een hoge intraklassecorrelatiecoëfficiënt is hier een indicatie voor. In multilevel modellen wordt met afhankelijkheid rekening gehouden in de meerniveau structuur. In de praktijk neemt men vaak intraklassecorrelaties waar tussen 0,05 en 0,20 (Veldhuijzen & Kleintjes, 1993).

Toevoegen van verklarende variabelen aan het model

De volgende stap na de passing van het nulmodel bestaat uit het opnemen in het model van verklarende variabelen om een deel van de variantie van Y_{ij} te verklaren. Als één verklarende variabele opgenomen wordt, leidt dat tot model (5.6):

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{ij} . \qquad (5.6)$$

Ook hier wordt verondersteld dat de residuen μ_{0j} en e_{ij} onafhankelijk van elkaar zijn en dat $\mu_{0j} \sim N\left(0, \tau_0^2\right)$ en $e_{ij} \sim N\left(0, \sigma^2\right)$.

In model (5.6) is één extra variabele opgenomen (x_{1ij}) met een eigen regressiecoëfficiënt $\left(\beta_1\right)$[4]. Nagegaan wordt of deze variabele een verklaring kan zijn voor de geconstateerde variantie in Y_{ij} . Ter illustratie is hieronder tabel 5.2 opgenomen. Als mogelijke verklarende variabele op het eerste niveau is de verbale

[4] Merk op dat de toegevoegde variabele als 'fixed' wordt beschouwd. De regressiecoëfficiënt in (5.6) is onafhankelijk van de klasse (bijvoorbeeld school) waartoe de leerling behoort. Om die reden bevat de variabele geen subscript *j*.

intelligentie (IQ) van de leerling op genomen. Om interpretatie van de resultaten te vereenvoudigen is de IQ-score gecentreerd.

Tabel 5.2
Schattingen na modellering van IQ
(ontleend aan Snijders & Bosker, 1999)

Fixed effects	Coëfficiënt	S.E.
β_0 = intercept	40,61	0,31
β_1 = coëfficiënt voor IQ	2,49	0,070
Random Effect	**Variantiecomponent**	**S.E.**
Niveau 1 variantie: $\sigma^2 = \text{var}(e_{ij})$	42,23	1,29
Niveau 2 variantie: $\tau_0^2 = \text{var}(\mu_{0j})$	9,50	1,52
Deviantie	15251,8	

Uit tabel 5.2 blijkt dat door het opnemen van IQ het algemeen gemiddelde β_0 iets in grootte toeneemt. Uit tabel 5.2 is ook de grootte van de regressie-coëfficiënt voor IQ af te lezen. Bij de interpretatie van de coëfficiënt wordt na-gegaan of deze significant ($\alpha = .05$) afwijkt van nul[5] . Indien er geen sprake is van een significante afwijking dan wordt dit geïnterpreteerd als dat de variantie van Y_{ij} niet verklaard wordt door deze variabele. De significantie kan worden getoetst met behulp van een t-toets. Daartoe wordt de gevonden waarde van de coëfficiënt gedeeld door zijn S.E. Voor verdere toelichting zie Snijders en

[5] De ratio van de absolute waarde van de parameterschatting en de bijbehorende standaardfout is groter dan 1.96: $|\beta| / \text{se}(\beta) > 1.96$. Voor random parameters geldt: $\sigma^2 / \text{se}(\sigma^2) > 1.96$.

Bosker (1999, pp. 86-87). De regressiecoëfficiënt in tabel 5.2 blijkt significant te zijn, wat inhoudt dat IQ variantie in Y_{ij} verklaart.

De variantie op de niveaus 1 en 2 wordt gebruikt om vast te stellen hoeveel variantie verklaard wordt (met hoeveel de variantie afneemt) door het opnemen van (extra) verklarende variabelen in het model. De mate waarin een model passend is voor de data wordt weergegeven door de afname in de deviantie. De parameters (regressiecoëfficiënten) worden bepaald volgens de grootste aannemelijkheidsmethode (Engels: maximum likelihood). Deze methode levert ook een waarde voor de likelihood die vertaald wordt naar een deviantie die gelijk is aan minus 2 maal de logaritme van de likelihood (-2log likelihood). Deze deviantie kan gezien worden als een maat voor de misfit tussen model en data. De deviantie op zich levert geen interpreteerbaar resultaat op, wel de verschillen in deviantie tussen twee geneste modellen die gepast worden op dezelfde databestand. Verschillen in deviantie in geneste modellen zijn χ^2 – verdeeld en geven aan of een ruimer model leidt tot een significante verbetering ten opzichte van een ander meer restrictief model. Het aantal vrijheidsgraden voor de toetsing is gelijk aan het verschil in het aantal parameters tussen de twee modellen. Als vuistregel geldt dat een model significant beter is dan een ander model als het verschil in deviantiescore van de twee modellen minimaal tweemaal groter is dan het verschil in het aantal te schatten parameters van beide modellen (Kreft & De Leeuw, 1998). In de gebruikte voorbeelden is het verschil in deviantie groot genoeg gegeven de toename van het aantal parameters (in het weergegeven voorbeeld één, namelijk de variabele IQ). Dit wil zeggen dat model (5.6) beter passend is voor de data dan model (5.5) en dat door opname van de variabele IQ een groter deel van de variantie in Y_{ij} verklaard wordt.

5.2 Multilevel modellen nader bekeken

De onderzoeker staan diverse multilevel modellen ter beschikking om data met een hiërarchische structuur te analyseren. Zo beschrijven Van den Bergh en Kuhlemeier (1997) vijf multilevel modellen voor het analyseren van verschillen in eindniveau en leerwinst. Voor de onderzoeker en ook voor het gebruik in EVADOS is het van belang goed na te gaan welke hypothese getoetst moet worden en welk model daar het beste bij past. De vijf door Van den Bergh en Kuhlemeier genoemde modellen worden in paragraaf 5.2.1 kort besproken. Voor een uitgebreidere bespreking van de modellen aan de hand van een uitgewerkt empirisch voorbeeld wordt verwezen naar Van den Bergh en Kuhlemeier (1997).

Bosker en Scheerens (1995) bespreken acht verschillende manieren voor het vaststellen van schooleffecten, waarbij ook gebruik gemaakt wordt van multilevel modellen. Op een aantal van deze manieren zal in paragraaf 5.2.2 kort ingegaan worden.

5.2.1 Vijf multilevel modellen voor het analyseren van verschillen in eindniveau en leerwinst

Van den Bergh en Kuhlemeier onderscheiden de volgende vijf modellen:
- het univariate variantie-analytisch model;
- het univariate covariantie-analytisch model;
- het univariate variantie-analytisch leerwinstmodel;
- het multivariaat variantie-analytisch model;
- het multivariaat variantie-analytisch leerwinstmodel.

1 Het univariate variantie-analytisch model (UNVA)

In dit model is het bereikte eindniveau na een bepaalde onderwijsperiode, bijvoorbeeld aan het einde van het vijfde leerjaar, de afhankelijke variabele. Het model wordt vaak toegepast als er slechts informatie voorhanden is over de

prestaties van de leerlingen op één moment. Zijn de prestaties van de leerlingen op verschillende momenten gemeten dan komen de vier andere door hen genoemde modellen in aanmerking.

In de meest eenvoudige vorm ziet het model er als volgt uit:

$$Y_{ijk} = \beta_0 + \left(e_{ijk} + u_{0jk} + v_{00k}\right) \tag{5.7}$$

In (5.7) stelt Y_{ijk} de score van leerling *i* in klas *j* van school *k* voor. Vergelijking (5.7) bestaat uit twee delen: het fixed gedeelte en het random gedeelte (tussen ronde haken). In het fixed gedeelte wordt slechts één regressiecoëfficiënt (β_0) geschat, die in het model zonder verdere 'verklarende' variabelen te interpreteren is als een schatting van het populatiegemiddelde. Met de standaardfout kan getoetst worden of dit gemiddelde significant afwijkt van nul.
In het random deel worden drie residuele scores onderscheiden die de afwijking van de leerling *i* van het gemiddelde van zijn klas (e_{ijk}), de afwijking van klas *j* van het gemiddelde van school *k* (u_{0jk}) en de afwijking van school *k* van het geschatte populatiegemiddelde (v_{00k}) representeren. Verondersteld wordt dat de residuen onafhankelijk van elkaar zijn en dat $v_{00k} \sim N\left(0, \tau_v^2\right)$, $\mu_{0jk} \sim N\left(0, \tau_u^2\right)$ en $e_{ijk} \sim N\left(0, \sigma^2\right)$. Om te toetsen of de klas waarin of de school waarop een leerling zich bevindt er toe doet, dient getoetst te worden of de variantie tussen klassen (τ_u^2), respectievelijk de variantie tussen scholen (τ_v^2) afwijkt van nul.
Het UNVA-model kan eenvoudig uitgebreid worden met leerling-, klas- en/of schoolvariabelen.

Het UNVA-model geeft informatie over de relatieve grootte van de geobserveerde prestatieverschillen tussen scholen, klassen en leerlingen op een bepaald moment in het onderwijs. Het geeft echter géén eenduidig antwoord op de vraag in hoeverre verschillen tussen scholen en klassen verklaard kunnen worden door de effectiviteit van het geboden onderwijs. Hooguit kan geconcludeerd worden dat de gemiddelde prestaties van leerlingen aan bijvoorbeeld het einde van het vijfde leerjaar van school tot school en van klas tot klas verschillen. Door het

toevoegen van verklarende variabelen om de verschillen tussen scholen, klassen en leerlingen te verklaren, kunnen dus alleen verschillen in bereikt eindniveau en geen verschillen in leerwinst verklaard worden.

2 Het univariate covariantie-analytisch model (UNCO)

Kenmerkend voor dit model is dat het eindniveau gecorrigeerd wordt voor bij aanvang aanwezige verschillen. Het belang van deze correctie wordt door Van den Bergh en Kuhlemeier onderstreept door te verwijzen naar Hill en Rowe (1996, p.9) die stellen: 'By controlling for prior achievement, one is able to obtain estimates of learning gain over the intervening period. (…) Effectiveness is thus defined in terms of that part of the achievement not predicted by prior achievement.' In hun modellen laten Van den Bergh en Kuhlemeier de regressiecoëfficiënt en het intercept variëren van klas tot klas en van school tot school. Deze vorm van modelleren maakt het mogelijk dat zowel het voorspelde algemeen niveau als het effect van de beginmeting kunnen verschillen van klas tot klas en van school tot school. Met dergelijke modellen kan beschreven worden dat leerlingen met dezelfde beginscore in de ene klas meer leren dan in de andere, terwijl leerlingen met een andere beginscore in de andere klas meer leren dan in de ene.

Het UNCO-model kan als volgt geschreven worden:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 * \left(BEG_{ijk} - BEG_{000} \right) + \\ &[e_{ijk} + u_{00\,jk} + u_{10\,jk} * \left(BEG_{ijk} - BEG_{000} \right) + \\ &\nu_{000k} + \nu_{100k} * \left(BEG_{ijk} - BEG_{000} \right)] \end{aligned} \tag{5.8}$$

($i = 1, 2, \ldots, I_{jk}$; $j = 1, 2, \ldots, J_k$; $k = 1, 2, \ldots, K$).

Merk op dat de covariaat gecentreerd is rond het algemeen gemiddelde $\left(BEG_{ijk} - BEG_{000}\right)$. In het fixed deel van het model worden twee parameters onderscheiden: het intercept $\left(\beta_0\right)$ en een algemene regressiecoëfficiënt $\left(\beta_1\right)$ voor de relatie tussen de begin- en eindmeting. Ook in dit model kan aan de

hand van de geschatte standaardfouten bepaald worden of (β_0) en (β_1) afwijken van nul. In het random gedeelte (tussen vierkante haken) worden vijf residuele scores onderscheiden: één op leerlingniveau (e_{ijk}), twee op klasniveau $(u_{00\,jk}$ en $u_{10\,jk})$ en twee op schoolniveau $(v_{000k}$ en $v_{100k})$. Met toepassing van dit model kan worden getoetst of de interceptvariantie op klas- en schoolniveau afwijkt van 0 en of de variantie van het regressiegewicht afwijkt van 0, dat wil zeggen of de relatie tussen begin- en eindmeting invariant is over klassen respectievelijk scholen.
Ook dit basismodel kan verder uitgebreid worden met diverse (andere) verklarende variabelen op school-, klas- en leerlingniveau.

3 Het univariate leerwinstmodel (UNLW)

In dit model wordt de leerwinst gemodelleerd door de afhankelijk variabele te definiëren als het verschil tussen eind- en beginniveau. Het multilevel model luidt:

$$\left(EIND_{ijk} - BEG_{ijk}\right) = \beta_0 + \left(e_{ijk} + u_{0\,jk} + v_{00k}\right), \qquad (5.9)$$

(i = 1, 2, …, I_{jk}; j = 1, 2, …, J_k; k = 1, 2, …, K).

Hierbij is BEG_{ijk} de score van leerling i in klas j van school k op de beginmeting en $EIND_{ijk}$ is de score van deze leerling op de eindmeting. Het verschil tussen deze twee metingen ($EIND_{ijk} - BEG_{ijk}$) in (5.9) is dan gelijk aan de leerwinst. Merk op dat het verschil tussen dit model en het UNVA-model gelegen is in de afhankelijke variabele: bij het UNVA-model is deze gelijk aan de score op de eindmeting en bij het UNLW-model is deze gelijk aan het verschil tussen de eind- en de beginmeting. Ook in dit model is analoog aan de andere besproken modellen een fixed deel en een random deel te onderscheiden. Ook in het UNLW-model dient getoetst te worden of de regressiecoëfficiënten en de varianties van de drie residuele scores afwijken van nul, waarmee onder andere de vraag beantwoord kan worden of de leerwinst verschilt van klas tot klas en van school tot school.

4 Het multivariate variantie-analytisch model (MUVA)

Het kenmerkende van het multivariate variantie-analytisch model is het opnemen van meer afhankelijke variabelen, dit in tegenstelling tot het univariate model waar slechts één afhankelijke variabele onderscheiden wordt. Van den Bergh en Kuhlemeier onderscheiden in hun bespreking van het MUVA-model twee afhankelijke variabelen: de score op de beginmeting en de score op de eindmeting. Het te analyseren model bestaat in hun voorbeeld uit twee delen, voor elke afhankelijke variabele één. Stel Y_{1ijk} is de score van leerling *i* in klas *j* van school *k* op een beginmeting en Y_{2ijk} is de score van deze leerling op een eindmeting. Het te analyseren model kan dan geschreven worden als:

$$\begin{aligned} Y_{1ijk} &= \beta_1 + \left(e_{1ijk} + u_{10\,jk} + v_{100k}\right) \\ Y_{2ijk} &= \beta_2 + \left(e_{2ijk} + u_{20\,jk} + v_{200k}\right) \end{aligned} \tag{5.10}$$

($i = 1, 2, \ldots, I_{jk}$; j = 1, 2, …, J_k; $k = 1, 2, \ldots, K$).

De parameters in beide vergelijkingen hebben dezelfde betekenis als die in het eerder besproken univariate model. Ook met betrekking tot de residuen worden in dit model dezelfde assumpties gemaakt als in de voorgaande besproken modellen.

$$\begin{aligned} &Y_{tijk} = \beta_1 * Dbeg_{tijk} + \beta_2 * Deind_{tijk} + \\ &\left[Dbeg_{tijk} * \left(e_{1ijk} + u_{10\,jk} + v_{100k}\right) + Deind_{tijk} * \left(e_{2ijk} + u_{20\,jk} + v_{200k}\right) \right] \end{aligned} \tag{5.11}$$

($t = 1, 2$; $i = 1, 2, \ldots, I_{jk}$; $j = 1, 2, \ldots, J_k$; $k = 1, 2, \ldots, K$).

In vergelijking (5.11) zijn $Dbeg_{tijk}$ en $Deind_{tijk}$ dummy-variabelen die 'aanstaan' als een score de beginmeting ($Dbeg_{tijk} = 1$; $Deind_{tijk} = 0$) respectievelijk de eindmeting betreft ($Dbeg_{tijk} = 0$; $Deind_{tijk} = 1$). De parameters β_1 en β_2 geven de gemiddelden (intercepten) op de begin- en eindmeting aan. In het random gedeelte (tussen vierkante haken) worden zes residuele scores onder-

scheiden: op elk niveau twee, één voor de beginmeting en één voor de eindmeting. Bij de residuen worden dezelfde assumpties gemaakt als in de voorgaande besproken modellen. Voor alle parameters en residuen kan nagegaan worden of deze afwijken van nul. Van den Bergh en Kuhlemeier (1997) merken op dat bij het MUVA-model geen restricties opgelegd worden aan de aan- of afwezigheid op één der beide meetmomenten. Dit betekent dat in het MUVA-model ook leerlingen die op één van de beide metingen afwezig waren, in de analyse betrokken kunnen worden, dit in tegenstelling tot het UNCO- en het UNLW-model. Merk op dat hoewel door gebruik te maken van dummy-variabelen de beide vergelijkingen in (5.10) weliswaar omgezet kunnen worden naar één vergelijking zoals weergegeven in (5.11), we in principe nog steeds te maken hebben met 2 onafhankelijke regressievergelijkingen. In vergelijking (5.11) is er geen sprake van een groeimodel omdat er geen afhankelijkheid over tijdstippen gemodelleerd wordt.

5 Het multivariate leerwinstmodel (MULW)

Ook het MULW-model kent net als het MUVA-model twee afhankelijke variabelen (het begin- en het eindniveau). Door de introductie van een dummy-variabele ($Deind_{hijk}$), die alleen 'aanstaat' als het een score op de eindmeting betreft (h = 1), kan de score op de eindmeting geschat worden als afwijking van de score op de beginmeting.

Het te analyseren model kan als volgt geschreven worden:

$$\begin{aligned} Y_{hijk} &= \beta_0 + \beta_1 * Deind_{hijk} + \\ &\left(e_{0ijk} + e_{1ijk} * Deind_{hijk} + u_{00\,jk} + u_{10\,jk} * Deind_{hijk} + \right. \\ &\left. \nu_{000k} + \nu_{100k} * Deind_{hijk} \right) \end{aligned} \tag{5.12}$$

$h = 0, 1; i = 1, 2, \ldots, I_{jk}; j = 1, 2, \ldots, J_k; k = 1, 2, \ldots, K).$

Onder aanname van een normale verdeling voor de residuele scores en het niet-gecorreleerd zijn van de residuele scores op de verschillende niveaus hebben de modelparameters de volgende betekenissen: het intercept β_0 is de gemiddelde

score op de beginmeting, terwijl β_1 staat voor het gemiddelde verschil tussen begin- en eindmeting (de gemiddelde leerwinst van de gemiddelde leerling). In het random gedeelte staan zes residuele scores: drie voor verschillen op de beginmeting en drie voor verschillen in leerwinst. Ook nu geldt dat onder de standaardaannames (zie vorige modellen) nagegaan kan worden of de schattingen van de diverse parameters afwijken van nul.

Van den Bergh en Kuhlemeier (1997, p. 72) concluderen aan de hand van een door hen uitgewerkt empirisch voorbeeld dat de vijf besproken modellen deels andere hypothesen toetsen. Hoewel voor de meeste gevallen de parameterschattingen van de vier modellen voor de analyse van gemiddelden (UNVA, UNLW, MUVA en MULW) tot elkaar te herleiden zijn, dient bij de interpretatie rekening gehouden te worden met de specifiek getoetste hypothesen. De beide multivariate modellen bieden meer mogelijkheden voor interpretatie dan de beide univariate modellen. Het UNCO-model neemt een uitzonderingspositie in. Van den Bergh en Kuhlemeier (1997) concluderen dat, hoewel het UNCO-model meer mogelijkheden biedt voor nuanceringen van de resultaten dan modellen voor gemiddelden, de parameterschattingen niet zonder meer eenduidig te interpreteren zijn. In het UNCO-model wordt als enige het eindniveau als functie van de beginmeting uitgedrukt, waarbij de relatie tussen beginmeting en eindmeting geschat wordt. Essentieel hierbij is dat de schaal van de beginmeting invloed heeft op de schatting van de bijbehorende regressiecoefficiënt. Van belang is dus of (en hoe) de beginmeting gecentreerd wordt. Vooral conclusies over de vraag of een regressiecoefficiënt al dan niet significant afwijkt van 0 kan afhangen van de gekozen schaal (en wordt dus beïnvloed door de centrering).Van den Bergh en Kuhlemeier adviseren om bij een keuze voor het UNCO-model ook één van beide multivariate modellen toe te passen.

5.2.2 Vaststellen van schooleffecten

Bosker en Scheerens (1995) laten aan de hand van een uitgewerkt voorbeeld acht manieren zien hoe een μ_{0j} schooleffect bepaald kan worden. De door hen

besproken manieren lopen uiteen van het bepalen van de gemiddelde score van alle leerlingen van een school op een bepaald tijdstip tot aan het vaststellen van de gemiddelde groei met gebruikmaking van 'empirical Bayes' methoden en het toepassen van een correctie voor achtergrondkenmerken.

De Bayesiaanse methode kenmerkt zich door de waarde van μ_{0j}, het groepseffect (zie vergelijking 5.5 op bladzijde 99), te berekenen door beschikbare gegevens van de groep te combineren met informatie over de totale populatie. Men gaat er daarbij van uit dat μ_{0j} een random variabele is vergelijkbaar met random groepseffecten en ook een normale verdeling kent met een gemiddelde van nul en een variantie van τ_0^2.

Het principe van de Bayesiaanse methode is als volgt. In vergelijking (5.5) is het nulmodel weergeven als:

$$Y_{ij} = \beta_0 + \mu_{0j} + e_{ij}$$

waarbij verondersteld wordt dat de residuen e_{ij} en μ_{0j} onafhankelijk van elkaar zijn en dat $e_{0ij} \sim N\left(0, \sigma^2\right)$ en $\mu_{0j} \sim N\left(0, \tau_0^2\right)$.

In deze vergelijking stelt β_0 het algemeen (populatie)gemiddelde voor en μ_{0j} de afwijking van school *j* ten opzichte van dit gemiddelde. Voor een specifieke school *j* geldt dat $\beta_{0j} = \beta_0 + \mu_{0j}$.
Bij de berekening van het schooleffect wordt ervan uitgegaan dat β_0, het algemeen gemiddelde, bekend is. Door nu β_{0j} te bepalen, is het schooleffect μ_{0j} (ook) bekend. Volgens de Bayesiaanse methode (zie Snijders & Bosker, 1999) wordt het intercept β_{0j} als volgt bepaald:

$$\beta_{0j}^{EB} = \lambda_j \hat{\beta}_{oj} + \left(1 - \lambda_j\right) \hat{\beta}_o \qquad (5.13)$$

waarbij het schoolgemiddelde $\hat{\beta}_{0j} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j}$ en het algemeen gemiddelde $\hat{\beta}_0 = \sum_{j=1}^{K} \frac{n_j}{M} \hat{\beta}_{0j}$. *K* staat voor het aantal scholen en het totaal aantal leerlingen (*M*) is gelijk aan $M = \sum_j n_j$.
Het superscript 'EB' staat voor 'empirical Bayes' en λ_j is een wegingsfactor die overeenkomt met de betrouwbaarheid van het gemiddelde van groep *j*. In formulevorm:

$$\lambda_j = \frac{\tau_0^2}{\tau_0^2 + \frac{\sigma^2}{n_j}} \qquad (5.14)$$

Uit (5.14) valt af te leiden dat de grootte van de factor σ^2/n_j afhankelijk is van het aantal observaties n van school *j*. Hoe groter het aantal observaties des te meer zal λ_j naderen tot één en wordt β_{0j}^{EB} meer bepaald door de gegevens van de desbetreffende groep *j* en minder door het algemeen populatiegemiddelde. Bij een kleiner aantal observaties neemt λ_j verhoudingsgewijs af en wordt β_{0j}^{EB} meer bepaald door het populatiegemiddelde. Voor een nadere toelichting op deze berekeningsmethode wordt verwezen naar Snijders & Bosker (1999, p. 59 e.v.).

Op de vraag welk model te gebruiken, geven (ook) Bosker en Scheerens aan dat de keuze bepaald wordt door de vraagstelling. In het algemeen geven zij aan dat waar mogelijk uitgegaan moet worden van ware scores van leerlingen op toetsen in plaats van ruwe scores. Bovendien stellen zij dat de Bayesiaanse methode statistisch gezien de voorkeur geniet. Als - zo stellen Bosker en Scheerens - alleen gebruik gemaakt wordt van de resultaten van leerlingen op een bepaald tijdstip, zonder rekening te houden met de resultaten op een eerder tijdstip, dan zit in de verkregen informatie over de output ook informatie over de input verweven omdat er dan geen rekening wordt gehouden met bij aanvang reeds aanwezige verschillen. Het vaststellen van leerwinst vraagt om een correctie voor reeds bij aanvang aanwezige verschillen, zo mogelijk door het opnemen van achtergrondvariabelen in het model.

De opvatting dat voor het vaststellen van leerwinst rekening gehouden moet worden met al bij de aanvang aanwezige verschillen betekent niet dat modellen waarin dat niet gebeurt niet zinvol zouden zijn. Als de vraagstelling niet gericht is op het bepalen van de leerwinst, maar op de vraag of scholen bepaalde standaarden bereiken, dan zijn modellen waarbij niet gecorrigeerd wordt voor de beginsituatie van leerlingen toepasbaar. Vergelijk (in dit kader) ook het door Van den Bergh en Kuhlemeier besproken UNVA-model.

In hoofdstuk 6 laten we aan de hand van een aantal van de in dit hoofdstuk besproken modellen zien hoe de bijdrage van een school bepaald kan worden. We maken daarvoor gebruik van het programma MLWin (Rasbash e.a., 2000). Bij het vaststellen van de bijdrage van een school gebruiken we in eerste instantie de data van de vier projectscholen. Zoals uit hoofdstuk 6 zal blijken, is deze dataset te beperkt om zowel de univariate als de multivariate modellen te passen. Om die reden is in tweede instantie ook gebruik gemaakt van data die voor een ander onderzoeksdoel verzameld zijn. Aan de hand van de resultaten op deze tweede dataset worden de univariate en multivariate modellen met elkaar vergeleken.

6 Rapportage aan scholen

In dit hoofdstuk komt de rapportage aan scholen aan bod. Bij de rapportage aan scholen gaat het erom scholen van informatie te voorzien waarmee zij geïnformeerd worden over de kwaliteit van het door hen geboden onderwijs of geïnformeerd worden over factoren die van invloed kunnen zijn op de kwaliteit van het onderwijs. Drie vormen van voor scholen relevante informatie komen aan bod. In eerste instantie wordt getoond dat het mogelijk is om op basis van informatie uit schooladministratie- en toetsregistratiepakketten aan de hand van een aantal beschrijvende statistieken informatie te bieden over de samenstelling van de schoolpopulatie. Met deze gegevens kan een school nagaan in hoeverre haar populatie verschilt van andere scholen. Ook de mate waarin leerlingen vroegtijdig de school verlaten of instromen, kan met deze gegevens in beeld worden gebracht. Daarna wordt getoond dat met de gegevens uit schooladministratiepakketten resultaten van leerlingen zoals opgenomen in de toetsregistratiepakketten uitgesplitst kunnen worden naar leerlingkenmerken om op deze manier een beeld te krijgen van de vorderingen van bepaalde groepen binnen een school. Door gegevens uit de schooladministratiepakketten in de analyses te betrekken kunnen scholen bovendien vergeleken worden met voor hen vergelijkbare scholen of kan gecontroleerd worden voor verschillen in achtergrondvariabelen van leerlingen om een betere vergelijking tussen scholen mogelijk te maken. Voorzichtigheid hierbij is echter gewenst (vergelijk Verhelst e.a., 2001). Door scholen te informeren over de ontwikkeling van groepen van leerlingen in de tijd, kunnen zij nagaan of bepaalde groepen van leerlingen binnen de school zich ontwikkelen zoals gewenst of verwacht, wat de school de mogelijkheid geeft ‘tijdig’ acties te ondernemen. Deze vorm van rapportage biedt een school de mogelijkheid ook de bij de analyse betrokken leerlingen van eventuele acties te laten profiteren. De derde vorm gaat over de toegevoegde waarde. Bij de berekening van de toegevoegde waarde wordt de vraag beantwoord wat de bijdrage van de school is aan de ontwikkeling van leerlingen.

In dit proefschrift wordt getoond hoe de toegevoegde waarde berekend kan worden, waarbij gebruik gemaakt wordt van de in hoofdstuk 5 beschreven modellen.

Alvorens in te gaan op de uitgevoerde analyses wordt eerst kort ingegaan op het rapporteren over en aan scholen. Bij deze bespreking worden opmerkingen van Verhelst e.a. (2001) over het rapporteren over scholen meegenomen. Aansluitend wordt ingegaan op de data die gewenst zijn om te komen tot voor scholen zo zinvol mogelijke analyses. Zowel na de analyses over de ontwikkeling van groepen van leerlingen in de tijd als na de berekening van de toegevoegde waarde, zal aangegeven worden wat de betekenis van de resultaten voor de scholen zijn of kunnen zijn.

6.1 Het rapporteren over de kwaliteit van onderwijs van scholen

Rapportages over schoolprestaties zijn inmiddels sinds 1997 in Nederland gemeengoed. In oktober van dat jaar startte het dagblad Trouw met de jaarlijkse presentatie van scholen geordend naar prestatie. Voor het bepalen van de rangorde werden zes criteria gehanteerd: het percentage onvertraagd geslaagden, het percentage uitvallers en zittenblijvers, het gemiddeld examencijfer voor Nederlands, Engels en Wiskunde A en een door J. Dronkers toegevoegd samenvattend rapportcijfer. Verhelst e.a. (2001) geven aan dat de suggestie die van zo een rangordening uitgaat dubbel is. Aan de ene kant zou verwacht kunnen worden dat scholen die bovenaan de rangordening staan van een betere kwaliteit zijn dan scholen die onderaan staan. En aan de andere kant zou de suggestie gewekt kunnen worden dat als ouders hun kinderen naar een school van een betere kwaliteit sturen (hoger op de lijst staan) de prestaties van deze kinderen ook hoger zullen zijn. Oosterbeek en Webbink (2001) geven aan dat de laatst genoemde suggestie eigenlijk verwijst naar het vaststellen van causale effecten.

Zij verwijzen naar de onderwijseconomie waar men geconstateerd heeft dat veel klas- en schoolkenmerken die van invloed zijn op de prestaties van leerlingen, niet toevallig over leerlingen zijn gespreid maar systematisch samenhangen met niet-waargenomen kenmerken van de leerlingen die op hun beurt weer systematisch samenhangen met (verwachte) leerprestaties. Als met bedoelde kenmerken in de analyse geen rekening wordt gehouden, kunnen veranderingen in de leerprestaties ten onrechte worden toegeschreven aan schoolkenmerken, terwijl ze in werkelijkheid veroorzaakt worden door andere factoren. Oosterbeek en Webbink illustreren dit door te verwijzen naar mogelijke effecten van klassenverkleining. In hun voorbeeld nemen ze aan dat de betrokkenheid van ouders een onafhankelijk effect heeft op de prestaties van hun kinderen en dat de kinderen van deze (betrokken) ouders in relatief kleine klassen zitten. De betrokkenheid van de ouders is niet direct waarneembaar. Het gevaar bestaat dat het effect op de leerprestaties wordt toegeschreven aan de groepsgrootte, terwijl dit in werkelijkheid veroorzaakt zou kunnen worden door de (niet waargenomen) betrokkenheid van de ouders. Het geschatte effect van de groepsgrootte op de leerlingprestaties wordt vertekend als de groepsgrootte met een niet waargenomen variabele (in het voorbeeld van Oosterbeek en Webbink de betrokkenheid van de ouders), die tevens van invloed is op de prestaties van leerlingen, samenhangt. Verhelst e.a. (2001) pleiten in dit kader voor een goed uitgewerkte theorie over de determinanten van schoolse prestaties.

Verhelst e.a. (2001) constateren dat de rangordening van scholen op basis van hun prestaties gebaseerd is op de gemiddelde schoolprestaties die gecorrigeerd zijn voor verschillen in achtergrondvariabelen (covariaten). Bovendien constateren zij dat de betrouwbaarheid van de meetuitslag afhankelijk is van de intraklassecorrelatie ρ (zie hoofdstuk 5) en van de schoolgrootte. Beide bepalen de grootte van de standaardfout, die de variatie rond een voorspelde gemiddelde score uitdrukt. Verhelst e.a. tonen aan dat de standaardfout toeneemt als ρ en het aantal leerlingen binnen de school afnemen. Dit houdt in dat een voorspelling nauwkeuriger is (kleinere standaardfout) voor grote scholen dan voor kleine scholen (zie Verhelst e.a., 2001). Bovendien blijkt dat bij correctie voor verschillen in achtergrondvariabelen tussen scholen de ρ daalt, waardoor het

vergelijken van scholen onderling lastiger wordt, maar als er niet gecorrigeerd wordt voor deze verschillen dan zijn de verschillen tussen scholen voor een gedeelte toe te schrijven aan selectieverschillen.

Ranglijsten van scholen kunnen gezien worden als rapportages over scholen. Scholen worden afgezet tegen andere scholen, met het gevaar dat de positie van een school op de lijst geïnterpreteerd wordt als een absoluut gegeven over de kwaliteit van de school. Verhelst e.a. (2001) laten zien dat het toekennen van een betekenis aan de positie van een school op een dergelijke lijst met enige terughoudendheid bezien moet worden.
Visscher, Dijkstra, Karsten, en Veenstra (2001) hebben standaarden geformuleerd waaraan publicaties van schoolprestatie-indicatoren zouden moeten voldoen. In paragraaf 6.6 wordt nader op de geformuleerde standaarden ingegaan.

Belangrijk bij het rapporteren over scholen is de vraag voor wie de rapportages bedoeld zijn en of ze informatief zijn? Een andere vraag is of een school er iets mee kan in het kader van zelfevaluatie en schoolverbetering. Visscher e.a. (2001) noemen de volgende redenen om schoolprestaties te presenteren:

1 het ondersteunen van ouders en leerlingen bij hun schoolkeuze;
2 het afleggen van verantwoording over met welk succes gemeenschapsgelden gebruikt zijn;
3 het geven van informatie aan scholen over de kwaliteit van hun functioneren, met het oog op zelfevaluatie en schoolverbetering.

De ranglijsten van scholen zouden een bijdrage kunnen leveren aan alle drie genoemde redenen. Voorwaarde is wel dat de informatie toegankelijk is en correct wordt weergegeven en dat duidelijkheid bestaat over de achtergrondvariabelen waarvoor gecorrigeerd is. Voor scholen is het van belang te weten in hoeverre de ranglijsten voor hen informatief zijn om te komen tot schoolverbeteringen. Ranglijsten van scholen geven een school alleen informatie over de behaalde resultaten ten opzichte van andere scholen. De lijsten geven scholen geen concrete aangrijpingspunten hoe te komen tot schoolverbeteringen. Daartoe is een andere vorm van informatieverstrekking en rapportage gewenst. De publicatie van ranglijsten zoals in het voorgaande besproken, kan gezien worden als een

rapportage over scholen en niet als een rapportage aan (of voor) scholen. Hoe een rapportage aan scholen eruit zou kunnen zien, komt in de volgende paragraaf aan bod.

6.2 Het rapporteren over de kwaliteit van onderwijs aan scholen

Voor scholen is het van belang dat de informatie over de kwaliteit van hun onderwijs aangrijpingspunten biedt om te komen tot schoolverbeteringen. Een vergelijking met andere scholen is informatief voor zover relevante achtergrondvariabelen in de vergelijking betrokken zijn. De vraag die daarbij onmiddellijk gesteld kan worden is die naar relevante variabelen en daaraan gekoppeld de mate waarin deze op een betrouwbare en valide manier bepaald kunnen worden. Oosterbeek en Webbink (2001) geven onder andere als belangrijke noodzakelijke verbetering van de indicatoren van schoolprestaties aan dat het moet gaan om de toegevoegde waarde (zie ook hoofdstuk 3). Ook geven zij aan dat het van belang is de mobiliteit van leerlingen tussen scholen in kaart te brengen.

Bosker, Béguin en Rekers (2001) geven aan dat, om de toegevoegde waarde van een school te kunnen bepalen, gegevens over het instroomniveau van leerlingen en over het uitstroomniveau van diezelfde leerlingen nodig zijn. Onder toegevoegde waarde verstaan zij kwalificaties die leerlingen ontwikkelen dankzij de school. Omdat gegevens over instroom- en uitstroomniveau in de praktijk vaak andersoortig zijn, wordt bij het bepalen van de toegevoegde waarde gewerkt met (een vorm van) covariantie-analyse, waarbij:

1 met een statistische techniek (multilevel analyse) de vorm en de sterkte van de relatie tussen de instroomkenmerken enerzijds en het uitstoomniveau anderzijds voor de gehele populatie van uitstromende leerlingen geschat wordt;

2 op basis van de vergelijking uit stap 1 voor elke leerling zijn of haar uitstroomniveau voorspeld wordt (het verwachte uitstroomniveau);
3 het verschil tussen het werkelijke en het verwachte uitstroomniveau bepaalt in welke mate een leerling over- of juist onderpresteert;
4 de middeling van de gegevens uit stap 3 per school een schatting weergeeft van de toegevoegde waarde per school.

Door volgens de vier genoemde stappen de toegevoegde waarde van een school te bepalen, wordt voldaan aan de eerder door Visscher e.a. genoemde drie redenen om schoolprestaties te presenteren. Voor ouders is het van belang te weten dat de toegevoegde waarde kan verschillen voor leerlingen met een verschillend beginniveau (of meer algemener: een verschillende startkwalificatie). Afhankelijk van het beginniveau van leerlingen kan er sprake zijn van differentiële schooleffecten. Een school kan voor de ene leerling effectiever zijn dan voor de andere. Om te komen tot schoolverbetering dient een school inzicht te hebben in haar functioneren en de mate waarin voor verschillen in leerlingpopulatie is gecorrigeerd. Voor het afleggen van rekenschap is het ook van belang de beschikking te hebben over gegevens over instroom, bevordering en voortijdige uitstroom van leerlingen.

De rapportage over de kwaliteit van het onderwijs van een school op basis van de vorderingen van leerlingen in de tijd vindt in dit proefschrift plaats op basis van de resultaten van vier projectscholen. Alle vier de scholen maakten gebruik van het Cito-LVS, waarmee het mogelijk is resultaten van leerlingen op verschillende momenten in de tijd met elkaar te vergelijken. Aggregatie van leerlingresultaten maakt het mogelijk uitspraken te doen over de ontwikkeling van groepen van leerlingen in de tijd. In het kader van dit proefschrift wordt als rapportage-eenheid een cohort gezien, dat omschreven wordt als een groep leerlingen dat hetzelfde onderwijsprogramma doorloopt.

Door gebruik te maken van het Cito-LVS is het mogelijk de ontwikkeling van de cohorten in de tijd weer te geven (zie hoofdstuk 4). Voor zover achtergrondvariabelen van leerlingen (of de school) gebruikt worden, zijn deze mogelijk aanwezig in de door scholen gebruikte administratiepakketten. Deze achtergrondgegevens maken het mogelijk uitsplitsingen naar groepen van leerlingen,

bijvoorbeeld ingedeeld naar leerlinggewicht, te maken. De informatie uit de schooladministratiepakketten bieden ook de mogelijkheid de populatie van de scholen te beschrijven. Ook gewenste informatie over in- en uitstromers en over de mobiliteit van leerlingen (zie de eerder gemaakte opmerking van Oosterbeek en Webbink) kunnen in kaart gebracht worden.

Aan de hand van de verwerking van de data van de vier projectscholen zal nader ingegaan worden op mogelijke rapportages aan scholen. In hoofdstuk vier van dit proefschrift is aangegeven dat schooladministratiepakketten voldoende mogelijkheden bevatten voor het opslaan van relevante gegevens voor school-zelfevaluatie. Uit onderzoek (Moelands, Ouborg en Engelen, 1996) blijkt dat het mogelijk is een koppeling tot stand te brengen tussen diverse gegevens uit school administratiepakketten en pakketten waarin toetsresultaten elektronisch opgeslagen worden en dat deze beide pakketten het in principe mogelijk maken de resultaten van groepen van leerlingen in de tijd te volgen. Wel laat het voor-noemde onderzoek zien dat de diverse administratie- en toetsregistratiepakketten de gebruiker zoveel vrijheden toelaten dat a-priori er niet van uitgegaan mag worden dat de gegevens correct zijn. Steeds zullen de gegevens uit de pakketten op hun juistheid gecontroleerd - en zonodig aangepast - moeten worden.

Hierna worden eerst de extractie en validatie van de gegevens uit de school-administratie- en toetsregistratiepakketten besproken. De extractie van data uit deze pakketten brengt een aantal problemen met zich mee. Aangegeven zal worden welke dat zijn.

6.3 Extractie en validatie van gegevens uit schooladministratie- en toetsregistratiepakketten

Schooladministratie- en toetsregistratiepakketten bevatten allerlei gegevens over leerlingen die nodig zijn om groepen van leerlingen in de tijd te kunnen volgen.

Gegevens over de klassen of groepen waarin de leerlingen tijdens hun schoolloopbaan zaten, zijn nodig om de schoolloopbaan van leerlingen in beeld te brengen. Is het een leerling die is blijven zitten, betreft het een zij-instromer, is het een leerling die zonder doubleren de acht groepen van het basisonderwijs heeft doorlopen? Achtergrondgegevens van leerlingen bieden de mogelijkheid bepaalde groepen samen te stellen om zo de ontwikkeling van deze groepen te kunnen volgen.

Voor het volgen van de resultaten van groepen van leerlingen in de tijd zijn de volgende drie aandachtspunten van belang:

1 *Het volgen van dezelfde (groepen van) leerlingen in de tijd*
   Voor het volgen van de resultaten van groepen van leerlingen in de tijd is het van belang om duidelijkheid te hebben over de groepssamenstelling, waardoor ontwikkelingen in de tijd niet toegeschreven kunnen worden aan mogelijke wisselende samenstellingen van groepen.
2 *Het volgen van resultaten van leerlingen in de tijd*
   De resultaten van leerlingen op toetsen moeten aan elkaar gerelateerd kunnen worden door transformatie naar één onderliggende vaardigheidsschaal.
3 *Het samenstellen van cohorten en het relateren van de resultaten aan (achtergrond-) kenmerken*
   Om uitspraken te doen over de kwaliteit van onderwijs dienen leerlingen toegewezen te worden aan voor een school zinvolle eenheden die in de tijd te identificeren zijn.

De punten 1 en 3 worden in de twee navolgende paragrafen besproken. Wat punt 2 betreft is het van belang op te merken dat uitgegaan wordt van de toetsen uit het Cito-LVS. Aangezien in hoofdstuk 4 reeds besproken is hoe de resultaten op de toetsen uit dit systeem op één vaardigheidsschaal afgebeeld kunnen worden, wordt punt 2 hierna niet verder meer besproken.

6.3.1 Het volgen van dezelfde (groepen van) leerlingen in de tijd

Bij uitspraken over de kwaliteit van het onderwijs op basis van de resultaten van groepen van leerlingen, moet het (per definitie) gaan over dezelfde leerlingen of tenminste over vergelijkbare leerlingen. Voorkomen moet worden dat conclusies over het onderwijs toegeschreven kunnen worden aan wisselende samenstellingen van groepen en niet aan het onderwijs als zodanig. Om die reden neemt EVADOS cohorten als eenheid van rapportage. In hoofdstuk drie is een cohort omschreven als een groep leerlingen die aan het begin van het schooljaar hun onderwijs aanvangen in groep drie en vervolgens zonder doubleren doorstromen naar groep acht. Niet alleen zittenblijvers, maar ook in- en uitstromers maken geen deel uit van deze groep, daar zij wisselingen in groepssamenstellingen veroorzaken. Een cohort is dus een in de tijd te identificeren groep. Om een onderscheid aan te brengen tussen leerlingen die zonder vertraging en instroming de school doorlopen met leerlingen die op enigerlei wijze wel een vertraging oplopen of gedurende het schooljaar instromen, worden in dit proefschrift de begrippen ‘reguliere’ en ‘niet-reguliere’ cohorten gebruikt. Reguliere cohorten zijn gedefinieerd als leerlingen die, behorend tot een bepaald cohort, in groep drie hun onderwijs aanvangen en vervolgens hetzelfde groepsverloop kennen. Tot de niet-reguliere cohorten behoren de zij-instromers, de zittenblijvers en de vroegtijdig schoolverlaters. Uitspraken over de kwaliteit van het onderwijs van een school zullen gebaseerd zijn op de resultaten in de tijd van de leerlingen van de reguliere cohorten. De niet-reguliere cohorten worden niet in de rapportage betrokken.

6.3.2 Het samenstellen van cohorten

Schooladministratiepakketten geven scholen de mogelijkheid achtergrondgegevens van leerlingen op te slaan als ook de groepen (klassen) die leerlingen doorlopen hebben. Met deze beide gegevens uit de schooladministratiepakketten moet het mogelijk zijn de schoolloopbaan van leerlingen te bepalen, op basis waarvan leerlingen toegewezen kunnen worden aan cohorten. Ook toetsregistra-

tiepakketten bieden deze mogelijkheid. In deze pakketten kan een school per leerling opnemen welke toets hij heeft gemaakt, op welk tijdstip en in welke groep (klas) hij zich op dat moment bevond. Toch laat onderzoek (Moelands e.a., 1996) zien dat het niet gemakkelijk, zo niet onmogelijk is, om op basis van de twee genoemde pakketten de schoolloopbaan van leerlingen in kaart te brengen. De volgende problemen blijken zich namelijk in de pakketten voor te doen:

- De gegevens van de leerlingen die de school verlaten hebben, worden in het schooladministratiepakket weggeschreven naar een historisch bestand. Dit bestand bevat slechts een beperkte hoeveelheid gegevens van de leerlingen. Zo gaat informatie over de groepen waarin een leerling zat tijdens het wegschrijven verloren. Deze informatie is cruciaal voor het samenstellen van de schoolloopbaan van de leerlingen. Het blijkt dus niet mogelijk te zijn om op basis van het historisch bestand de schoolloopbaan van leerlingen vast te stellen.
- De pakketten kennen een unieke leerlingidentificatie, die aan de leerling wordt toegekend gedurende zijn schoolloopbaan op de desbetreffende school. Als een leerling uitgeschreven wordt, dan wordt het nummer niet aan een nieuwe leerling toegekend. Hoewel dit op zich voor het identificeren van leerlingen een goede zaak is, geldt dit echter ook voor die leerlingen die tussentijds de school verlaten, uitgeschreven worden en besluiten weer terug te keren. Deze groep leerlingen krijgt bij het opnieuw inschrijven een nieuw leerlingnummer dat niet gelijk is aan het eerder aan hen toegekende nummer. Voor het volgen van de leerling in de tijd is dat een probleem. Voor het samenstellen van de schoolloopbaan betekent dit dat naast de leerlingidentificatie ook andere informatie gebruikt moet worden om leerlingen te identificeren.
- De pakketten laten de gebruiker soms de keuze bepaalde informatie wel of niet op te nemen, wat tot gevolg kan hebben dat voor het volgen van de schoolloopbaan relevante informatie niet altijd aanwezig is. Ook behoeft de informatie niet altijd op gelijke wijze ingevoerd te worden. Zo kunnen scholen zelf namen aan toetsen toekennen, waardoor dezelfde toetsen onder verschillende namen in de pakketten opgeslagen staan. Bovendien staan de pakketten vergissingen en discrepanties toe bij het invoeren van gegevens,

zoals het maken van typefouten, het dubbel invoeren van leerlingen en/of toetsresultaten.

Geconcludeerd kan worden dat hoewel de pakketten in principe voldoende informatie bevatten voor het samenstellen van groepen van leerlingen en het volgen van de leerlingen in de tijd, bepaalde eigenschappen van de pakketten - zoals het niet vasthouden van gegevens over leerlingen wanneer zij de school verlaten - deze niet geschikt maken als opslagmedium voor EVADOS. In hoofdstuk zeven zal een alternatief besproken worden.

6.4 Beschrijving schoolpopulatie

Bij het verstrekken van informatie aan scholen over de kwaliteit van het door hen geboden onderwijs is het belangrijk rekening te houden met achtergrondkenmerken van de leerlingen, daar deze een verklaring kunnen zijn voor geconstateerde verschillen tussen scholen. Met de informatie uit de schooladministratiepakketten is het mogelijk de leerlingpopulatie van een school te beschrijven, inclusief het aantal in- en uitstromers. Door jaarlijks deze gegevens te bekijken, wordt een school (tijdig) geïnformeerd over wijzingen in de schoolpopulatie, wat een aanleiding kan zijn om daarop in te spelen.

Bij een beschrijving van de populatie van een school dient eerst vastgesteld te worden wat beschreven gaat worden. In het voorgaande is reeds aangegeven dat in dit proefschrift een onderscheid aangebracht wordt tussen reguliere en niet-reguliere cohorten. Wanneer in paragraaf 6.7 de analyse-resultaten van groepen van leerlingen in de tijd weergegeven gaan worden, zal dat alleen betrekking hebben op de reguliere cohorten. Zoals eerder vermeld worden de niet-reguliere cohorten niet in deze analyses meegenomen. Wanneer echter de populatie van een school beschreven gaat worden, inclusief in- en uitstromers, dan zullen ook de gegevens van leerlingen die niet tot de reguliere cohorten behoren in de rap-

portage meegenomen moeten worden. Voor het onderzoek betekende dit dat met verschillende databestanden gewerkt is. Bij de rapportage over de voortgang van groepen van leerlingen is alleen gebruik gemaakt van de gegevens van leerlingen uit de reguliere cohorten (zie voor een verdere toelichting op dit databestand paragraaf 6.5). Bij de beschrijving van de schoolpopulaties zijn ook de leerlingen uit de niet-reguliere cohorten betrokken. In het totale databestand van een school (leerlingen uit de reguliere en de niet-reguliere cohorten) bevinden zich gegevens over leerlingen die zijn blijven zitten of een klas hebben overgeslagen. Deze leerlingen kunnen ingedeeld worden bij meerdere (niet-)reguliere cohorten. Door een klas over te doen of een klas over te slaan, veranderen deze leerlingen immers van cohort. Dit betekent dat bij de beschrijving van de populatie van een school het aantal te gebruiken gegevens bepaald wordt door de eenheid waarover gerapporteerd wordt. Gaat het om een beschrijving van de school als geheel, dan dienen de leerlingen slechts éénmaal in het databestand opgenomen te zijn. Gaat het echter om een beschrijving van de cohorten binnen een school, dan is het mogelijk dat een leerling - bijvoorbeeld in het geval hij een keer niet bevorderd is - in meerdere cohorten voorkomt. Het databestand waarbij uitgegaan werd van het cohort als eenheid leidde in totaal tot 982 records. Voor de beschrijving van de populatie van de vier scholen is het van belang dat de achtergrondgegevens van deze leerlingen niet dubbel zijn ingevoerd. Opschonen van het bestand van 982 records op het dubbel aanwezig zijn van achtergrondgegevens van dezelfde leerlingen leverde een bestand op van 959 records. Het verschil is toe te schrijven aan leerlingen die zijn blijven zitten of een klas hebben overgeslagen en zodoende deel uitmaken van twee (of meer) cohorten.

Met de twee databestanden is het mogelijk de vier schoolpopulaties te beschrijven en de cohorten binnen de scholen. Met deze gegevens kan een school nagaan wat de gebruikelijke samenstelling van de populatie is en kunnen afwijkingen in een bepaald jaar gesignaleerd worden. Ook is het mogelijk het aantal in- en uitstromers in beeld te brengen, wat informatie geeft over de mobiliteit van leerlingen. Wijzigingen in ‘leerlingstromen’ kunnen voor een school redenen zijn om nader onderzoek te doen om na te gaan of (beleids)wijzigingen gewenst zijn. Opvallende veranderingen in de schoolpopulatie zouden voor een

school ook op voorhand een signaal kunnen zijn bepaalde groepen van leerlingen nadrukkelijker te volgen.

In deze paragraaf worden enkele voorbeelden gegeven van beschrijvende statistieken die scholen informeren in hoeverre hun populatie afwijkt van voorgaande schooljaren of van andere scholen.

Leerlinggewicht

Op basis van herkomst en opleiding van ouder(s)/verzorger(s) worden leerlingen op individuele basis toegerekend aan een categorie. Deze categorieën kunnen gezien worden als een indicatie voor (sociale) herkomst. De volgende vijf categorieën worden daarbij onderscheiden: 1,0 - 1,25 - 1,40 - 1,70 - 1,90. Elke categorie is gedefinieerd. Zo behoren leerlingen tot categorie 1,25 wanneer beide ouders of verzorgers een schoolopleiding hebben genoten tot of tot en met het niveau eindexamen voorbereidend beroepsonderwijs. Mocht het een leerling betreffen uit een 1-oudergezin, dan geldt deze opleidingseis ten aanzien van de desbetreffende ouder of verzorger. Categorie 1,90 betreft met name die leerlingen waarvan tenminste één ouder geen hogere opleiding heeft dan het voorbereidend beroepsonderwijs. Ook het land van herkomst kan mede bepalend zijn of een leerling toegerekend wordt aan deze categorie.

De schoolpopulatie van de vier scholen kent een relatief groot aandeel 1,25- en vooral 1,90-leerlingen. Tabel 6.2 geeft voor de vier scholen de verdeling van de leerlingen over de drie onderscheiden leerlinggewichten weer.

Tabel 6.1

Verdeling leerlinggewicht per school in procenten

		Leerlinggewicht		
		1	1,25	1,9
School	1 (36 lln)	2,7	-	97,3
	3 (240 lln)	32,1	30,4	37,5
	4 (166 lln)	-	7,8	92,2
	5 (517 lln)	5,6	3,3	91,1
Totaal		11,2	10,7	78,1

Uit tabel 6.1 blijkt dat het merendeel van de schoolpopulatie uit 1,90-leerlingen bestaat. School drie kent een andere verdeling. Bij deze school zijn de leerlingen min of meer gelijk verdeeld over de drie onderscheiden categorieën.

Opvallend is wellicht het verschil in aantallen leerlingen per school. Dit verschil wordt niet alleen veroorzaakt door het aantal leerlingen per school, maar ook door de datum van invoer van het Cito-LVS bij de deelnemende projectscholen. Het primaire doel van het onderzoek was het volgen van de resultaten van groepen van leerlingen in de tijd, waarbij als uitgangspunt is genomen de resultaten op toetsen Rekenen-Wiskunde uit het Cito-LVS. Niet alleen startten de scholen op verschillende tijdstippen met de invoering van het Cito-LVS, maar vond de invoering ook niet altijd in dezelfde groep plaats, met als consequentie dat het aantal beschikbare (bruikbare) databestanden per school verschilde. Om deze reden komt het voor dat de beschikbare data (zie tabel 6.2) van school 1 betrekking hebben op twee cohorten en die van school 5 op zeven.

Tabel 6.1 geeft informatie over de verdeling van de leerlingen naar leerlinggewicht over alle jaren, terwijl tabel 6.2 de verdeling van de leerlingen naar leerlinggewicht door de jaren heen laat zien. In deze tabel zijn de fluctuaties te zien die zich binnen een school over een aantal jaren voordoen. Een school kan hieruit afleiden in hoeverre de samenstelling van de schoolpopulatie in een bepaald jaar afwijkt van dat wat voor de school gebruikelijk is. Op basis van

overzichten uit voorgaande jaren, zowel ten aanzien van bijvoorbeeld de verdeling van leerlingen naar leerlinggewicht als die van de door leerlingen behaalde prestaties, kan een school vroegtijdig een signaal krijgen dat de prestaties van de leerlingen van een bepaald cohort mogelijk extra aandacht vragen. Zo is in tabel 6.2 de samenstelling van het cohort 95 voor school 3 opvallend in die zin dat het aantal leerlingen dat tot dat cohort behoort, afwijkt van de andere vermelde cohorten als ook de verdeling van de leerlingen naar leerlinggewicht. Dit cohort kent relatief weinig 1,0- en veel 1,25-leerlingen.

Tabel 6.2

Overzicht verdeling leerlingen naar leerlinggewicht per school per cohort

Cohort	Leerlinggewicht (in%)			Totaal
	1	1,25	1,9	(in aantallen)
School 1				
92	-	-	100	27
93	11,1	-	88,9	9
Totaal	2,8	-	97,2	36
School 3				
92	25,5	38,2	36,3	55
93	40,4	30,8	28,8	52
94	34,6	34,6	30,8	52
95	11,5	46,2	42,3	26
96	37,3	15,3	47,4	59
Totaal	32,0	31,1	36,9	244
School 4				
92	-	11,1	88,9	27
93	-	3,6	96,4	28
94	-	8,3	91,7	36
95	-	7,7	92,3	39
96	-	7,7	92,3	39
Totaal	-	7,7	92,3	169
School 5				
91	5,4	6,8	87,8	74
92	3,9	2,6	93,5	77
93	1,8	1,8	96,4	57
94	9,1	2,3	88,7	88
95	7,2	3,6	89,2	83
96	4,5	4,5	91,0	88
97	4,5	-	95,4	66
Totaal	5,4	3,2	91,4	533

Analoog aan de beschrijving van de samenstelling van de populatie uitgesplitst naar leerlinggewicht, is het mogelijk de populatie uit te splitsen naar andere leerlingkenmerken. Ter illustratie volgen zij-instromers en vroegtijdig schoolverlaters.

Zij-instromers en vroegtijdige schoolverlaters

Onder zij-instromers worden die leerlingen verstaan die in een cohort instromen. Of met andere woorden: leerlingen die niet vanaf de start van het onderwijs in groep drie deel uitmaken van het desbetreffende cohort. Deze leerlingen volgen dus niet hetzelfde onderwijsprogramma als de reguliere leerlingen uit dat cohort. Op basis van de groep waarin de leerlingen instromen wordt hun cohort bepaald. Een uitzondering vormen de leerlingen waarvan de resultaten op de Cito-LVS toetsen niet vanaf groep drie maar pas vanaf een hogere groep door de school verzameld werden, doordat de school op een later tijdstip gebruik is gaan maken van de toetsen uit het Cito-LVS. Deze leerlingen worden in dit onderzoek niet als zij-instromers beschouwd, maar opgenomen in het reguliere cohort. Zij worden daarbij toebedeeld aan het cohort dat correspondeert met het jaar dat zij gestart zijn in groep drie. Het aantal zij-instromers ten opzichte van de totale schoolpopulatie staat per school weergegeven in tabel 6.3.

Tabel 6.3

Overzicht aantal zij-instromers per school

	Aantal zij-instromers	
School	%	Aantal
1	27,0	10
3	10,0	24
4	7,8	13
5	16,1	83

Uit tabel 6.3 blijkt dat verhoudingsgewijs school 1 de meeste zij-instromers kent. Absoluut gezien kent school 5 de meeste zij-instromers. Tabel 6.4 bevat een overzicht van het aantal vroegtijdige schoolverlaters per school.

Tabel 6.4

Overzicht aantal vroegtijdige schoolverlaters

	Aantal schoolverlaters	
School	%	Aantal
1	37,8	14
3	27,9	67
4	0,6	1
5	14,5	75

Beschrijvende statistieken zoals in deze paragraaf weergegeven, geven een school informatie over de samenstelling van de populatie. Of een school hier altijd iets mee kan is nog de vraag. Zo leidt een wijziging in de verdeling van leerlingen naar leerlinggewicht niet per definitie tot andere resultaten op toetsen. Als een school constateert dat zij te maken heeft met een groot aantal zij-instromers of vroegtijdig schoolverlaters, dan kan dat een reden zijn voor nadere acties. Bij een groot aantal zij-instromers zou nagegaan kunnen worden wat de reden daarvoor is en of de instromers als groep en/of de groep waar de instromers geplaatst worden, extra aandacht vraagt. Als het aantal vroegtijdige schoolverlaters groot is, kan een school onderzoeken wat de redenen daarvoor zijn. Deze beschrijvende statistieken geven een school geen antwoord op de vraag hoe de school het doet. Wel biedt de informatie een school de gelegenheid op voorhand na te denken of bepaalde wijzigingen om acties vragen. Wil een school weten hoe zij het doet, dan zijn de resultaten van de leerlingen op toetsen van belang.

Alvorens echter ingegaan wordt op het volgen van de resultaten van de leerlingen in de tijd, wordt eerst het databestand met de leerresultaten van de leerlingen van de vier projectscholen op de toetsen Rekenen-Wiskunde van het Cito-LVS beschreven.

6.5 Beschrijving dataset

De extractie en validatie van de gegevens uit de schooladministratie - en toets-registratiepakketten van de projectscholen hebben uiteindelijk geleid tot één tabel met informatie over de schoolloopbaan van de leerlingen en de door hen behaalde vaardigheidsscores op de toetsen Rekenen-Wiskunde. De beschikbare gegevens betroffen de jaren 1992 tot en met 1997. Om privacy redenen worden de scholen in dit proefschrift gecodeerd als school 1 en school 3 tot en met 5. Het project startte aanvankelijk met vijf scholen. De school met codering 2 is niet in de analyses betrokken. Deze school maakte gebruik van een ander schooladministratiepakket dan de vier andere scholen. De codering van de vier overige scholen is gehandhaafd, vandaar dat in dit proefschrift gesproken wordt van school 1 en de scholen 3, 4 en 5. Zoals in paragraaf 6.4 op bladzijde 126 vermeld, bevatte het bestand in totaal 982 records. De verdeling van het aantal records per school is daarbij als volgt:
school 1: 36 records
school 3: 244 records
school 4: 169 records
school 5: 533 records
Het verschil in aantal records per school wordt bepaald door de grootte van de school en de datum waarop de scholen de pakketten Rekenen-Wiskunde van het Cito-LVS hebben ingevoerd.

Voordat in de volgende paragrafen nader ingegaan wordt op de resultaten van de leerlingen van de projectscholen, is de volgende opmerking van belang. De handleiding van het Cito-LVS adviseert de toetsen Rekenen-Wiskunde tweemaal per schooljaar af te nemen: tijdens het schooljaar in de maand januari en aan het einde van het schooljaar in de maand juni. Uit de toetsafnamegegevens blijkt dat scholen zich daar niet (altijd) aan houden. Soms nemen scholen een toets op een ander tijdstip af dan in de handleiding staat aangegeven. Het komt zelfs voor dat een school ervoor kiest dezelfde toets bij dezelfde leerlingen in een periode van één maand tweemaal af te nemen. Omdat scholen de LVS-toetsen vaker

afnemen en op wisselende tijdstippen, is het niet mogelijk en wenselijk een vaste rapportage te verzorgen die gebaseerd is op een afname van tweemaal per jaar op vaste tijdstippen, alleen al omdat dan veel informatie over de ontwikkeling van leerlingen niet meegenomen zou worden. Besloten is daarom te kiezen voor een weergave van de resultaten van leerlingen over een periode van telkens twee maanden. Voor zover in deze twee maandelijkse periode meerdere toetsafnames plaatsvonden, zijn de resultaten - uitgedrukt in vaardigheidsscores - gemiddeld.

Tabel 6.5 geeft een overzicht op welk moment een bepaalde toets door een bepaalde school is afgenomen. In de cellen staat het aantal leerlingen weergegeven dat op dat tijdstip een toets heeft gemaakt. Tabel 6.5 betreft zowel de reguliere als de niet-reguliere cohorten.

In tabel 6.5 staan de afkortingen ‘scl’, ‘coh’ en ‘nt’ voor respectievelijk de school, het cohort waarop de gegevens betrekking hebben en het aantal leerlingen dat in totaal betrokken was bij de toetsafnames van dat cohort. De kolommen 3.4 tot en met 6.6 geven aan wanneer de toetsafname plaatsvond. Het eerste getal correspondeert met de groep waarin de afname plaatsvond (groep 3, 4, 5 of 6) en het tweede getal met de periode in een bepaald schooljaar. Hierbij is de volgende codering gebruikt:

1 augustus/september
2 oktober/november
3 december/januari
4 februari/maart
5 april/mei
6 juni/juli

Een afname in de periode 4.5 moet dan als volgt gelezen worden: de toets is afgenomen in groep 4 in de periode april/mei. De getallen in de cellen van de kolommen 3.4 tot en met 6.6 geven het aantal leerlingen aan waarbij een toets is afgenomen in de desbetreffende perioden.

Tabel 6.5

Overzicht toetsafname per school, per cohort

scl	coh	nt	3. 4	3. 5	3. 6	4. 1	4. 2	4 .3	4. 4	4. 5	4. 6	5 .1	5. 2	5. 3	5. 4	5. 5	5. 6	6. 3	6 .4	6. 6
1	91	1														1				
1	92	68				19		23		26										
1	93	12	6																	6
3	92	197							50	49	47	51								
3	93	281	49	48	45	48				46			45							
3	94	172		51			43			42			36							
3	95	79		23		6	12	2		18	1		17							
3	96	107		56			50	1												
4	92	64																32		32
4	93	113												29			30		26	28
4	94	157						35			36				35		32	19		
4	95	162	41		41				42		38									
4	96	79	40		39															
5	91	139																	69	70
5	92	346							58		53				63		67		64	41
5	93	297	43		41			41	6		44			46		1	33		42	
5	94	333	59		58				55		56				51				54	
5	95	247	64		63				55						65					
5	96	134	79						55											
5	97	66	66																	

Uit tabel 6.5 blijkt dat op het afnamemoment 5.5 slechts bij twee kandidaten een toets is afgenomen. Gegeven dit geringe aantal is besloten dit moment niet in de analyses mee te nemen. Dit besluit heeft tot consequentie dat cohort 91 van school 1 uit het databestand verwijderd is. In de genoemde aantallen records is deze verwijdering steeds al geëffectueerd.

Om ontwikkelingen in de tijd niet toe te kunnen schrijven aan mogelijke wisselende samenstellingen van groepen is in paragraaf 6.3.1 aangegeven dat de uitspraken over de kwaliteit van het onderwijs van een school gebaseerd zullen zijn op de resultaten van de leerlingen van de reguliere cohorten. Verwijdering van alle niet-reguliere leerlingen uit het databestand leidde tot een tabel met in totaal 643 records.

In tabel 6.6 staan de beschikbare resultaten van de reguliere cohorten per school opgenomen. De eerste kolom geeft de school en het desbetreffende cohort van

deze school weer. De andere kolommen corresponderen met de afnametijdstippen van de toetsen. In de cellen staat het gemiddelde van de vaardigheidsscores van de leerlingen. Deze vaardigheidsscores zijn berekend op basis van het aantal goed beantwoorde vragen op de toetsen Rekenen-Wiskunde.

Tabel 6.6
Overzicht gemiddelde vaardigheidsscore leerlingen per school, per cohort

	3.4	**3.5**	**3.6**	**4.1**	**4.2**	**4.3**	**4.4**	**4.5**	**4.6**	**5.1**	**5.2**	**5.3**	**5.4**	**5.6**	**6.3**	**6.4**	**6.6**
1_92				37,2		51,3		61,9									
1_93	27,8																92,6
3_92							61,0	66,7	66,9	69,4							
3_93	33,2	46,1	52,8	53,4				67,5			73,5						
3_94		40,9			52,2			65,8			70,0						
3_95		46,4		49,1				63,7			69,4						
3_96		31,5			48,3						84,6						
4_92															92,1		102,2
4_93												75,3		84,3		88,7	94,7
4_94						60,3			68,3				75,9	83,2	88,6		
4_95	35,0		46,9				59,6		69,8								
4_96	35,8		45,0														
5_91																89,7	93,8
5_92							54,1		64,4				74,0	80,3		87,2	96,8
5_93	25,9		41,7			48,2	50,9		61,5			67,9		79,8		84,0	
5_94	31,8		45,6				52,7		64,3				72,4			85,5	
5_95	29,8		43,9				56,3						72,8				
5_96	31,2						53,6										
5_97	34,0																
Gem.	31,6	41,2	46,0	46,6	50,3	61,1	55,5	65,1	65,9	69,4	70,9	71,6	73,8	81,9	90,3	87,0	96,0

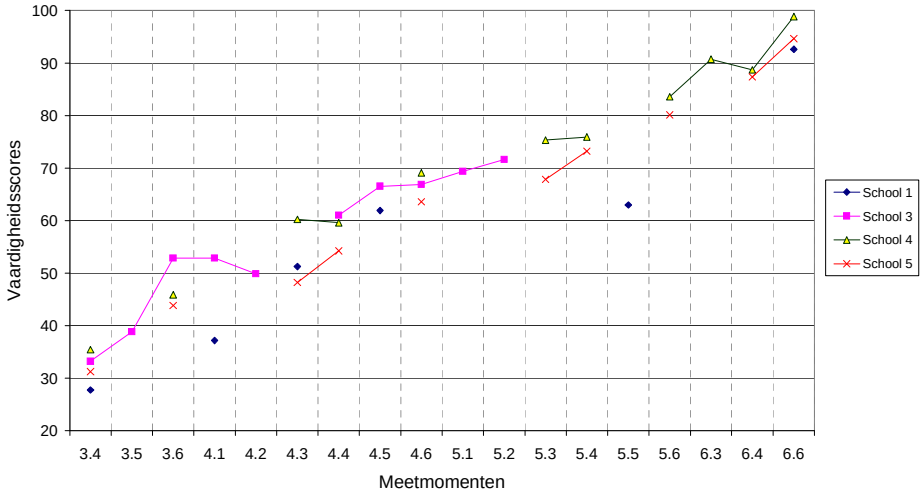
Uit tabel 6.6 is af te lezen dat de gemiddelde behaalde vaardigheidsscores van de leerlingen uit de diverse cohorten op dezelfde tijdstippen verschillen.

Aan de hand van bovenstaande dataset zal in de volgende paragrafen aangegeven worden hoe aan scholen gerapporteerd kan worden over de vorderingen van groepen leerlingen binnen de school. De data in tabel 6.6 zijn vaardigheidsscores. Deze scores representeren de rekenvaardigheid van de leerlingen op de diverse tijdstippen afgeleid uit hun scores op de toetsen Rekenen-Wiskunde uit het Cito-LVS. In het voorgaande is al aangegeven dat de afnametijdstippen en het aantal afnames van de toetsen verschillen tussen scholen. Ook is aangegeven

dat in de rapportage aan scholen ervoor gekozen is om als eenheid van rapportage een periode van twee maanden te nemen. Deze twee gegevens leiden ertoe dat per school niet op alle te onderscheiden tijdstippen afnamegegevens van leerlingen bekend zijn. In paragraaf 6.6. wordt nader ingegaan op het ontbreken van afnamegegevens van leerlingen op de onderscheiden tijdstippen en hoe daarmee omgegaan is.

6.6 Imputeren ontbrekende gegevens

In figuur 6.1 staan de resultaten van de vier projectscholen op de toetsen Rekenen-Wiskunde weergegeven.



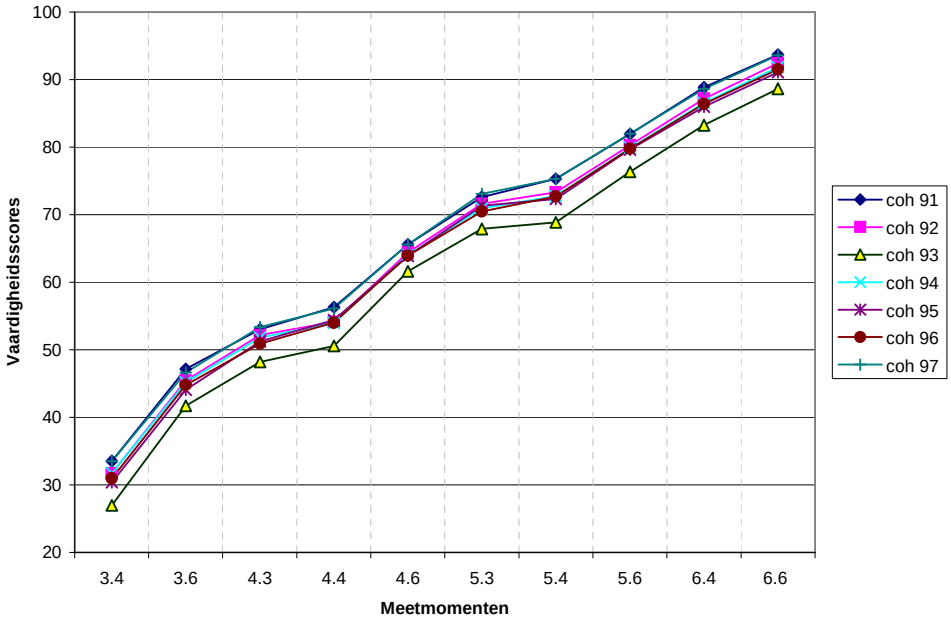
Figuur 6.1
Vaardigheidsscores per school over de periode 3.4 tot 6.6

Figuur 6.1 is gebaseerd op tabel 6.6. Horizontaal staan de onderscheiden meetmomenten, waarbij elk meetmoment een periode van twee maanden representeert. Verticaal staan vaardigheidsscores weergegeven. Deze vaardigheidsscores

zijn de gemiddelde scores per school. Figuur 6.1 is lastig te interpreteren vanwege het ontbreken van resultaten van leerlingen op de onderscheiden tijdstippen. Het beperkte aantal gegevens maakt het lastig om de ontwikkeling van de leerlingen van een school in de tijd te volgen. De problemen als gevolg van een onvolledig gegevensbestand zijn ondervangen door de ontbrekende resultaten (vergelijk tabel 6.6) te imputeren door gebruik te maken van het programma Multi (Frans Kamphuis, Cito). Dit programma schat de ontbrekende waarden op basis van de beschikbare waarden van de leerlingen op de diverse meetmomenten. Het programma werkt volgens een procedure die in Fahrmeier en Tutz (1994, hoofdstuk acht) beschreven staat (Kamphuis, 1998, persoonlijke mededeling). Het programma Multi biedt ook de mogelijkheid om bij het berekenen van de ontbrekende waarden rekening te houden met (effecten van) achtergrondvariabelen van leerlingen zoals bijvoorbeeld sekse, leerlinggewicht en cohort waartoe een leerling behoort. Tevens bestaat de mogelijkheid te corrigeren voor de (on)betrouwbaarheid van het meetinstrument (de toetsen Rekenen-Wiskunde) waarmee de scores van de leerlingen op de diverse meetmomenten zijn vastgesteld. Bij de toepassing van het programma Multi is gerekend met een betrouwbaarheid van de LVS-toetsen van .80. Het resultaat van het toepassen van het programma Multi levert een volledig gevulde data-matrix op. In deze matrix staan voor elke leerling(rijen) de schattingen van de vaardigheidsscores op alle tijdstippen(kolommen). De schatting van de vaardigheidsscore is gebaseerd op alle beschikbare waarnemingen van een specifieke leerling gecombineerd met de informatie die voorhanden is in het gespecificeerde structurele model. Door de schattingen van de vaardigheidsscores in de data-matrix te middelen over de cohorten per school kunnen we tabel 6.6 aanpassen. In de tabel zijn de lege cellen nu gevuld en bovendien zijn de gemiddelden nu gecorrigeerd voor meetfouten en effecten van de eerder besproken achtergrondvariabelen.

In figuur 6.2 zijn de uitkomsten van deze exercitie grafisch weergegeven voor 7 cohorten van een bepaalde school. In tegenstelling tot figuur 6.1 zien we in figuur 6.2 een duidelijk interpreteerbaar beeld: de ontwikkeling van de cohorten

verloopt vergelijkbaar in de tijd met dit verschil dat cohort 93 op alle tijdstippen iets lager scoort.



Figuur 6.2
Het ontwikkelingsverloop van 7 cohorten van een school na toepassing van het programma Multi

Besloten is om in de rapportages over de ontwikkelingen van de scholen geen gebruik te maken van de geïmputeerde gegevens, maar de rapportage te baseren op de gemiddelde vaardigheidsscores van de leerlingen van een school zoals berekend op basis van de behaalde resultaten op de LVS-toetsen. Als wel gebruik gemaakt zou worden van geïmputeerde gegevens zouden deze in de uit te voeren multilevel analyses (zie paragraaf 6.10.1) als inputdata genomen worden. Daarmee krijgen de geïmputeerde waarden de status van empirische geobserveerde waarden die ook nog eens perfect passen binnen het door Multi veronderstelde model, wat tot vertekeningen in de resultaten kan leiden.
Als referentie voor de ontwikkeling van de scholen en groepen binnen scholen is het gewogen gemiddelde over scholen op elk van de tijdstippen bepaald. Dit

gemiddelde kan omschreven worden als de ontwikkeling in de tijd van ‘de gemiddelde leerling’ over de (vier) projectscholen c.q. de groepen binnen een school. Voor het bepalen van het gewogen gemiddelde is wel gebruik gemaakt van alle verkregen data na toepassing van het programma Multi, dus inclusief alle geïmputeerde data. Het gewogen gemiddelde wordt in het vervolg van dit proefschrift aangeduid met de term ‘trendlijn’.

Vergelijking rapportages met de standaarden voor de publicatie van schoolprestaties

Het weergeven van de ontwikkeling van scholen of groepen binnen scholen vraagt om zorgvuldigheid. Vaak suggereren de overzichten (ten onrechte) een grote mate van zekerheid. Dat omzichtig met indicatoren en resultaten van scholen omgegaan moet worden, is reeds in paragraaf 6.1 aan de orde geweest. Op de in de komende paragrafen te presenteren ontwikkelingen van scholen of groepen binnen scholen, zijn niet alle door Visscher e.a. (2001) genoemde standaarden van toepassing.

Voordat we kort op de standaarden ingaan, dient opgemerkt te worden dat de te presenteren overzichten uiteen vallen in twee gedeeltes. Het eerste gedeelte betreft het grafisch weergeven van de ontwikkeling in de tijd van scholen en groepen van scholen. In het tweede gedeelte wordt ingegaan op de bijdrage van een school aan de ontwikkeling van leerlingen, waarbij gebruik gemaakt wordt van multilevel technieken. Gegeven de aard van beide presentaties zullen de door Visscher e.a. genoemde standaarden niet op beide presentaties (in gelijke mate) van toepassing zijn. Gesteld kan worden dat:

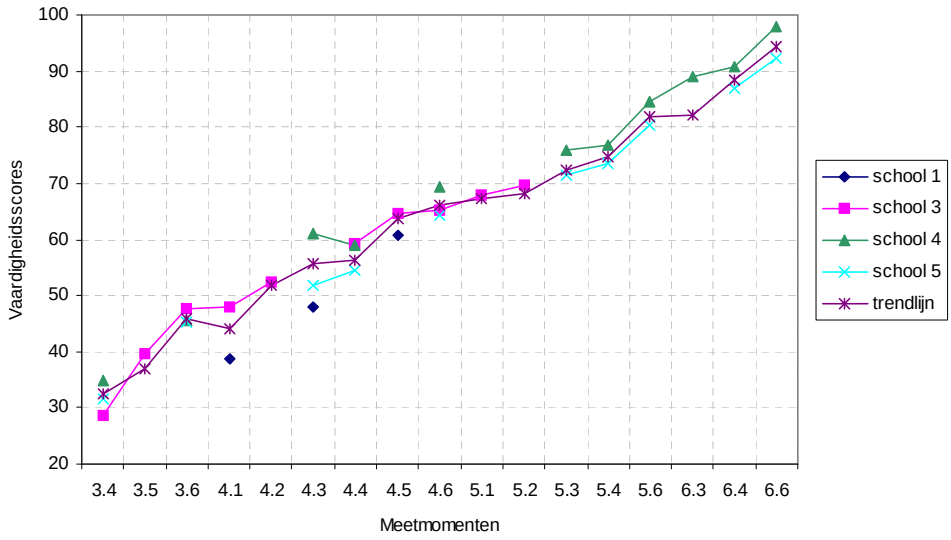
- de grafische overzichten geen inzicht geven in de toegevoegde waarde van scholen (standaard A1). Correctie voor verschillen die niet met de kwaliteit van de school te maken hebben, vindt niet plaats. In de later te presenteren resultaten naar aanleiding van multilevel analyses vindt deze correctie wel plaats.
- de berekening van de schooleffecten gebaseerd zijn op multilevel analyses (standaard A2). Bij de weergave van de grafische overzichten is dit niet van toepassing.

- de betrouwbaarheidsintervallen van de indicatoren (standaard A3) niet standaard vermeld worden. Exemplarisch komt dit bij één van de grafische presentaties wel aan bod.
- niet in alle gevallen voldaan is aan de eis dat het aantal leerlingen op grond waarvan een indicator wordt berekend minimaal 10 moet zijn (standaard A4). Door de flexibele afname van de toetsen door scholen komt het voor dat het aantal leerlingen waarop de grafiek is bepaald (in een enkel geval) minder dan 10 is. In het algemeen wordt aan deze standaard wel voldaan.
- in de presentaties geen expliciete rangordes van scholen gepubliceerd worden (standaard A5).
- de standaarden A6 (controleerbaar zijn van de berekeningen) en A7 (toepassen van heldere en zorgvuldige procedures) zoveel mogelijk als uitgangspunt bij het presenteren van de resultaten meegenomen worden.
- aan het geven van een adequate toelichting op de gepresenteerde gegevens (standaard B1) aandacht wordt besteed, wat ook geldt voor de functie van de gepresenteerde gegevens (B2).
- de standaarden B3 (bereiken ouders) en B4 (begrijpelijkheid informatie voor ouders) niet van toepassing zijn. De doelgroep is in eerste instantie de school.
- de rapportage plaats zal vinden zonder bevoorrechting van één der scholen of groepen binnen de scholen (standaard Z1).
- tot slot de deelnemende scholen reeds kennis genomen hebben van de ontwikkelingen zoals deze in dit proefschrift gepubliceerd worden.

Geconcludeerd kan worden dat de te presenteren gegevens in grote mate voldoen aan de standaarden zoals geformuleerd door Visscher e.a. (2001).

6.7 De ontwikkeling van scholen in de tijd

In figuur 6.3 staat de ontwikkeling van de vier projectscholen in de tijd weergegeven. Op de horizontale as staan de meetmomenten en op de verticale de vaardigheidsscores.



Figuur 6.3
Vaardigheidsscores per school over de periode 3.4 tot 6.6 met trendlijn

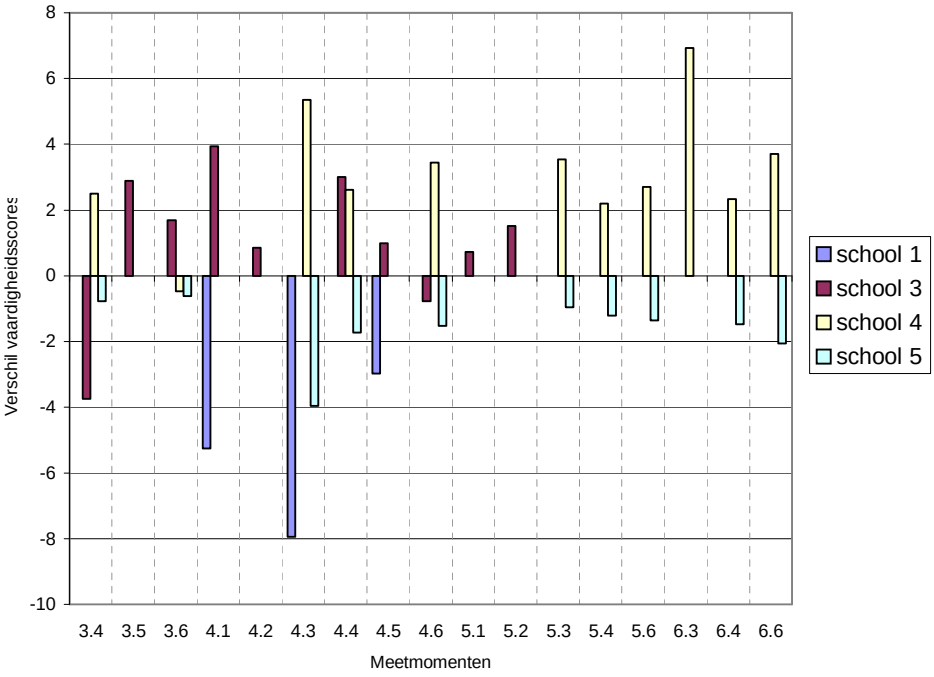
Figuur 6.3 is een nadere uitwerking van figuur 6.1. In figuur 6.3 is de trendlijn meegenomen als referentiekader voor de vier scholen. Hoewel deze trendlijn slechts gebaseerd is op deze vier scholen en hoewel deze trendlijn niet meer dan een exemplarische betekenis kan hebben, is deze toch opgenomen om te illustreren hoe op basis van het aantal deelnemende scholen een extern referentiepunt voor de afzonderlijke scholen tot stand kan worden gebracht. De scholen kunnen dan voor zichzelf bepalen hoe zij presteren ten opzichte van een (door het aantal deelnemende scholen bepaald) gemiddelde. In het onderhavige voorbeeld uit figuur 6.3 levert de school met het meeste aantal leerlingen de

grootste bijdrage aan de trendlijn. In die zin geeft de trendlijn zeker een vertekend beeld, en kan niet gezien worden als een ‘algemeen’ gemiddelde.

Het is van belang er nogmaals op te wijzen dat de grafiek bedoeld is om de ontwikkeling van de scholen in de tijd weer te geven en niet om aan te geven dat de ene school beter presteert dan de andere. Het ontbreken van betrouwbaarheidsintervallen maakt het doen van dergelijke uitspraken trouwens niet mogelijk. Het weergeven van de betrouwbaarheidsintervallen bij de diverse meetmomenten zou tot een niet leesbare grafiek leiden. Daarom wordt het weergeven van de betrouwbaarheidsintervallen besproken in paragraaf 6.8 bij het onderdeel ‘uitsplitsing naar sekse’.
In figuur 6.3 valt op dat de trendlijn geen rechte lijn is, maar een grillig verloop kent. Natuurlijk wordt het patroon van deze lijn mede bepaald door het geringe aantal scholen waarop deze betrekking heeft. Toch zou het interessant zijn om ingeval de trendlijn gebaseerd zou zijn op een groot aantal scholen het verloop nader te bestuderen. Afwijkingen in het patroon geven mogelijk een indicatie van de ‘groeisnelheid’ waarmee de leerlingen de lesstof tot zich nemen. Uit het patroon zou kunnen blijken dat deze binnen een schooljaar en op momenten tussen schooljaren verschillend is. Ter illustratie wordt in figuur 6.3 verwezen naar de meetmomenten 3.6 en 4.1. Uit de grafiek blijkt duidelijk een terugval in vaardigheid. Meetmoment 3.6 verwijst naar de maanden juni en juli, wat het einde van het schooljaar is. Meetmoment 4.1 echter is de eerste meting aan het begin van het (nieuwe) schooljaar, de maanden augustus en september. Mogelijk is de terugval terug te voeren tot de vakantieperiode. Een bevestiging daarvoor is te vinden bij Waterreus (2002) die verwijst naar een review van Cooper, Charlton, Lindsay en Greathouse (1996), waaruit blijkt dat de progressie van leerlingen van ‘disadvantaged backgrounds’ tijdens de zomervakantie lager is dan die van andere leerlingen. In tabel 6.1 bleek dat de vier projectscholen veel 1,9-leerlingen hebben. Mogelijk behoren zij tot de leerlingen met ‘disadvantaged backgrounds’ waarnaar Cooper e.a. (1996) verwijzen.

Figuur 6.3 laat zien dat de resultaten van de vier scholen op de onderscheiden meetmomenten verschillen. Om de afwijking van de scholen ten opzichte van de

trendlijn beter zichtbaar te maken, kunnen de resultaten ook gepresenteerd worden in de vorm van een staafdiagram, zoals in figuur 6.4.



Figuur 6.4
Verschillen vaardigheidsscore scholen met trendlijn

In deze figuur komen de afwijkingen van de resultaten van de scholen tot de trendlijn en daardoor de ontwikkelingen binnen de school, beter tot uiting. Opvallend is wel dat figuur 6.4 laat zien dat de positie van de scholen (de meetmomenten 3.4 en 4.6 van school 3 vormen de enige uitzondering) ten opzichte van de trendlijn (positieve of negatieve afwijking) constant blijft. Blijkbaar geldt - althans voor deze scholen - dat de mate waarin scholen afwijken ten opzichte van de trendlijn gedurende de basisschooljaren niet verandert. Wel kan de afwijking ten opzichte van de trendlijn in de ene groep groter of kleiner zijn. Om nadere informatie te krijgen over mogelijke verklarende factoren, kan een school de resultaten uitsplitsen naar achtergrondvariabelen. In de volgende paragraaf wordt daar nader op ingegaan.

Samenvattend kan - onder de aanname dat de verschillen significant zijn - gezegd worden dat de resultaten uit figuur 6.4 scholen informeren over drie mogelijke (statistisch significante) soorten verschillen:

- algemene niveauverschillen tussen cohorten;
- trendverschillen (meer of minder groei);
- mate van variabiliteit (zijn de verschillen al of niet stabiel).

6.8 Uitsplitsing van de prestaties van scholen naar achtergrondkenmerken

In de analyse van de resultaten van de vier scholen in de voorgaande paragraaf is geen rekening gehouden met achtergrondkenmerken van leerlingen. Door achtergrondkenmerken van leerlingen in de analyse te betrekken, krijgt een school informatie over:

- de mate waarin afwijkende resultaten mogelijk toe te schrijven zijn aan groepen leerlingen met bepaalde achtergrondkenmerken binnen de school;
- hoe die bepaalde afwijkende groepen binnen een school presteren ten opzichte van vergelijkbare groepen leerlingen bij andere scholen.

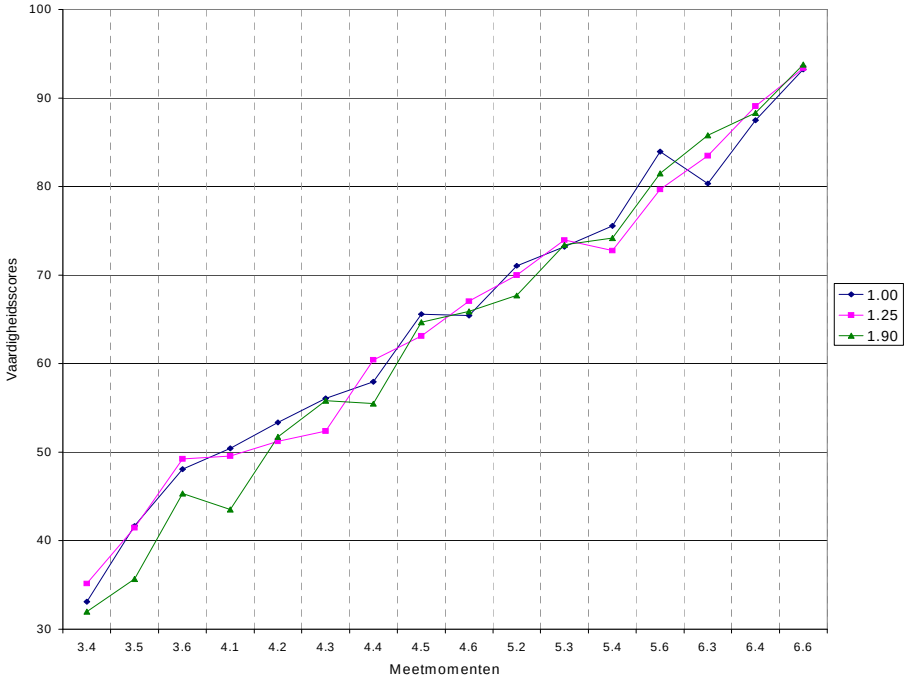
Met name het laatste punt is voor een eerlijke en voor een school meer informatieve vergelijking van belang. Zo is het mogelijk dat bepaalde groepen leerlingen binnen een school het minder goed doen op een vak in vergelijking met andere groepen. Dat wil dan niet zeggen dat het onderwijs op de desbetreffende school slecht is. Wanneer blijkt dat binnen een school een dergelijke groep vergelijkbaar presteert als vergelijkbare groepen binnen andere scholen, dan kan een school concluderen dat haar bijdrage aan de ontwikkeling van deze groep niet per definitie slecht is. Mogelijk zijn de achtergrondkenmerken van de leerlingen mede debet aan de afwijkende resultaten.

Om het als eerste genoemde punt te illustreren zullen de door de vier projectscholen behaalde resultaten uitgesplitst worden naar twee achtergrondkenmerken

van de leerlingen: het leerlinggewicht en de sekse. Merk op dat bij de twee uitgewerkte voorbeelden de analyses betrekking hebben op de vier scholen samen en niet op individuele scholen. Dat laatste is wel mogelijk, maar vanwege de beschikbaarheid van de data (en onder andere de geringe hoeveelheid gegevens op schoolniveau) is ervoor gekozen de uitsplitsing naar achtergrondkenmerken te laten zien aan de hand van de data van de vier scholen.

Uitsplitsing van de prestaties naar leerlinggewicht

Aan de hand van figuur 6.3 is geconstateerd dat de scholen een terugval kennen tussen de meetmomenten 3.6 en 4.1. Een interessante vraag is nu of deze terugval te herleiden is tot bepaalde groepen binnen de scholen. Om deze vraag te beantwoorden zijn in figuur 6.5 voor de scholen de resultaten uitgesplitst naar leerlinggewicht.



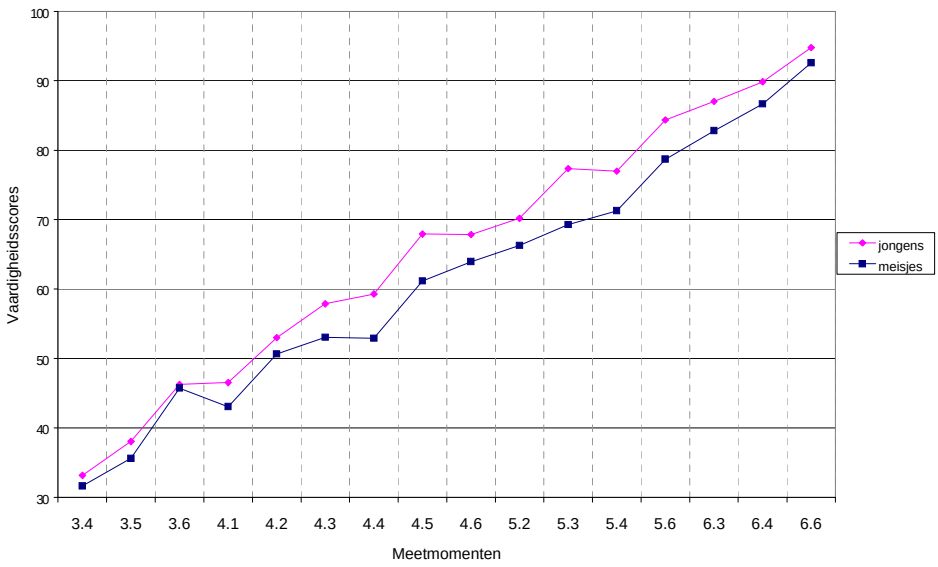
Figuur 6.5

Uitsplitsing resultaten scholen naar leerlinggewicht

Uit figuur 6.5 blijkt dat de geconstateerde terugval op meetmoment 4.1 ten opzichte van meetmoment 3.6 vooral toegeschreven moet worden aan de 1,9-leerlingen. Merk op dat de aantallen leerlingen waarop de grafiek betrekking heeft per categorie leerlinggewicht verschillen. De meeste leerlingen behoren tot de categorie 1,9 (zie tabel 6.2). Deze terugval is mogelijk in overeenstemming met de eerdere aangehaalde verwijzing van Waterreus (2002) naar Cooper e.a., (1996), die concluderen dat de progressie van leerlingen van 'disadvantaged backgrounds' tijdens de zomervakantie lager is dan die van andere leerlingen. Figuur 6.5 laat bovendien zien dat deze terugval niet blijvend is in de tijd en niet leidt tot een structurele achterstand van deze groep leerlingen ten opzichte van de andere groepen.

Uitsplitsing van de prestaties naar sekse

Een soortgelijke uitsplitsing als in figuur 6.5 kan ook gemaakt worden voor sekse (zie figuur 6.6).



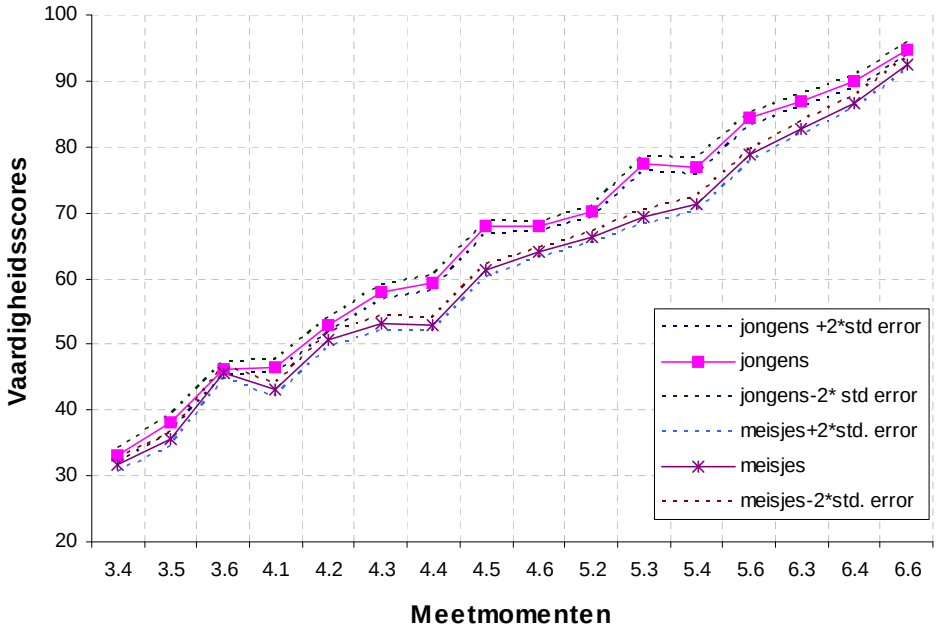
Figuur 6.6

Uitsplitsing resultaten scholen naar sekse

Uit deze weergave blijkt voornoemde terugval zich (bij deze vier scholen) meer voordoet bij de meisjes dan bij de jongens. Bovendien blijkt uit figuur 6.6. dat de meisjes over de hele linie lager scoren dan de jongens.

Hiervoor is reeds meerdere malen aangegeven dat de interpretatie van resultaten van analyses om zorgvuldigheid vraagt. Bij de interpretatie (en ook de weergave) van de resultaten dient rekening gehouden te worden met de betrouwbaarheidsintervallen. Uit figuur 6.5 zou geconcludeerd kunnen worden dat de meisjes van de vier projectscholen minder goed presteren dat de jongens en dat dit verschil in prestatie weliswaar minder wordt in groep 6, maar nog steeds in het voordeel is van de jongens. In figuur 6.7 zijn ook de betrouwbaarheidsintervallen (±2 maal de standaardfout) van de berekende score op de diverse meetmomenten opgenomen.

Figuur 6.7 laat zien dat bij een aantal meetmomenten de betrouwbaarheidsintervallen van de jongens en de meisjes elkaar overlappen en dat bij een aantal andere meetmomenten de intervallen elkaar niet overlappen. Alleen indien de intervallen elkaar niet overlappen, mag gesproken worden van een ‘echt’ verschil tussen de prestaties van beide groepen.



Figuur 6.7
Uitsplitsing resultaten scholen naar geslacht, inclusief standaardfouten

Ontwikkeling van de leerlingen binnen een school

In het voorgaande is aangetoond hoe scholen zich kunnen vergelijken met andere scholen. Op analoge wijze is het mogelijk om een school van informatie te voorzien die betrekking heeft op de ontwikkeling van groepen binnen de eigen school. Zie in dit verband figuur 6.2 waar de ontwikkeling van cohorten binnen een school ter sprake kwam. Een school vormt dan zijn eigen referentiekader, gebaseerd op door (groepen binnen) de school behaalde resultaten in voorgaande jaren. Als 'trendlijn' voor een school kan dan het gewogen gemiddelde dienst doen, dat gezien kan worden als de ontwikkeling van de 'gemiddelde leerling' binnen de groep die geanalyseerd wordt. Ook nu geldt dat de gevonden resultaten uitgesplitst kunnen worden naar achtergrondkenmerken van leerlingen. Door achtergrondkenmerken van leerlingen en de samenstelling van de populatie in de analyses te betrekken, is voor de school wellicht een 'logische' verklaring te

vinden voor eventuele afwijkende resultaten. Het is dan aan de school te besluiten of de mogelijke verklaringen al of niet leiden tot een wijziging in het beleid.

6.9 Opbrengsten van de analyses voor de school

In de paragrafen 6.7 en 6.8 is aangetoond hoe de ontwikkeling van scholen in de tijd zowel in een lijndiagram als in een staafdiagram beschreven kan worden. Hoewel de getoonde grafieken betrekking hebben op de vergelijking van een school met andere scholen, is het ook mogelijk om de ontwikkeling binnen een school op analoge wijze in beeld te brengen.
Met de aldus verkregen informatie kan een school nagaan:

- hoe zij het in een bepaald jaar doet in vergelijking met voorgaande jaren. De school is haar eigen referentiekader. Een school kan nagaan of er opvallende ontwikkelingen zijn die aandacht vragen;
- hoe de ontwikkeling verloopt van bepaalde groepen binnen de school. Zijn er bepaalde groepen waaraan mogelijke opvallende ontwikkelingen toegeschreven kunnen worden?
- of er een interactie-effect is. Bijvoorbeeld: worden geconstateerde verschillen tussen leerlingen met verschillende leerlinggewichten mogelijk beïnvloed door sekse?
- hoe groepen binnen een school presteren ten opzichte van andere, vergelijkbare, groepen in voorgaande jaren of bij andere scholen;
- hoe de school presteert in vergelijking met andere scholen (extern referentiekader), al of niet rekening houdend met de populatie van de school;
- of de geconstateerde verschillen tussen scholen of binnen een school wijzen in een bepaalde richting, zoals:
    - het zijn steeds dezelfde scholen die onder of boven het gemiddelde presteren;
    - de geconstateerde verschillen nemen in de tijd toe/af (trend) of de verschillen zijn fluctuerend: dan weer positief, dan weer negatief;

- of schoolbeleid leidt tot het gewenste resultaat of dat er aandachtsgebieden zijn die om (extra) aandacht vragen?

De hierboven genoemde informatie die gebaseerd is op de in paragraaf 6.7 en 6.8 geschetste analysemogelijkheid heeft met name een schoolinterne functie. Het primaire doel is de ontwikkeling van groepen van scholen in kaart te brengen met tot doel gepaste maatregelen te nemen. De informatie biedt de school aangrijpingspunten om het onderwijs zoals dat in het lopende jaar plaatsvindt nauwgezet te volgen en het onderwijs aan te passen (school improvement). Het doel ligt bij deze vorm van rapportage zeker niet bij het afleggen van verantwoording aan derden, maar doet een beroep op de eigen verantwoordelijkheid van een school om zorg te dragen voor een optimale kwaliteit van het geboden onderwijs, gegeven haar doelgroep. Door resultaten in het lopende jaar 'continue', te vergelijken met interne en met externe gegevens, kan een school steeds de 'vinger aan de pols houden'. Een school zou standaard voor een aantal leergebieden (bijvoorbeeld rekenen en spelling) de ontwikkeling van leerlingen gedurende het schooljaar kunnen volgen. Bij geconstateerde afwijkingen met de interne of externe referentie kan een school nadere analyses uitvoeren door achtergrondvariabelen van leerlingen in de analyses te betrekken (vergelijk figuur 6.5 en 6.6). Op basis van de aldus verkregen informatie, kan een school wellicht gerichtere acties ondernemen om de kwaliteit (ook voor de zittende leerlingen) te verbeteren. Door het verloop van het ontwikkelingsproces continu te volgen, kunnen maatregelen genomen worden en kunnen zo mogelijk bijstellingen plaatsvinden.

6.10 De bijdrage van de school

In het kader van kwaliteitszorg is de vraag van belang wat de bijdrage van de school is aan de ontwikkeling van de leerlingen. Scholen zullen ernaar moeten streven deze bijdrage zo hoog mogelijk te laten zijn. In de bijdrage of de toegevoegde waarde die een school heeft aan de ontwikkeling van groepen van leer-

lingen gaat het om factoren die aan een school zijn toe te schrijven, zoals schoolorganisatie, manier van lesgeven en ervaring van leerkrachten. Deze factoren kunnen per school verschillen.

Om de bijdrage van een school vast te stellen, maken we in dit hoofdstuk gebruik van de door Van den Bergh en Kuhlemeier (1997) in hoofdstuk 5 besproken multilevel modellen. We beperken ons in eerste instantie tot de data van de vier projectscholen. De te bespreken analyses op basis van deze data zijn in strikte zin genomen geen multilevel analyses. De scholen worden als dummy-variabele in de vergelijking opgenomen en er wordt geen variantiecomponent op het tweede niveau (tussen-scholen variantie) gedefinieerd. In feite heeft elke school haar eigen micro-model. Met deze in de literatuur genoemde 'slopes as outcomes model' (Kreft & De Leeuw, 1998) wordt het principe getoond hoe de bijdrage van een school te bepalen. In de praktijk betreft het in de regel niet een beperkt aantal scholen, maar zijn vele scholen betrokken wat het praktisch gezien ondoenlijk maakt om voor elke school de coëfficiënten te schatten. Voor dergelijke situaties lenen zich de 'echte' multilevel modellen zoals besproken in hoofdstuk 5. Aan de hand van een grotere dataset (in dit proefschrift de WOB-dataset genoemd, waarover later meer) zal de toepassing van deze modellen besproken worden.

In paragraaf 6.10.1 wordt aan de hand van de data van de vier projectscholen eerst een toepassing van het eerder besproken univariate variantie-analytisch model gepresenteerd. Met dit model kunnen we vaststellen in hoeverre geobserveerde prestatieverschillen tussen scholen op een bepaald moment in het onderwijs verschillen. Als tijdstip is het moment 4.6 (zie paragraaf 6.5) genomen. Ook de correctie voor de achtergrondvariabele 'sekse' komt aan bod. Omdat de dataset uit longitudinale data bestaat, is het mogelijk de resultaten van de leerlingen behaald op een eerder tijdstip als beginmeting in het model op te nemen. Deze correctie voor beginmeting, met als beginmeting het tijdstip 4.4, wordt in paragraaf 6.10.1 besproken. Deze correctie geeft de mogelijkheid om vast te stellen wat de bijdrage van de afzonderlijke scholen tussen de twee tijdstippen is (het univariate covariantie-analytisch model). Ook bij het laatst genoemde model

wordt de invloed van de achtergrondvariabele ‘sekse’ vastgesteld. Paragraaf 6.10.1 eindigt met de toepassing van het univariate leerwinst model. In dit model wordt de leerwinst gemodelleerd door het verschil tussen eind- en beginniveau als afhankelijke variabele in de analyse te betrekken.

Van den Bergh en Kuhlemeier beschrijven ook twee multivariate modellen: het multivariate variantie-analytische model en het multivariate leerwinstmodel. Ook met deze modellen zijn uitspraken te doen over de bijdrage van een school. De benaderingswijze van deze laatstgenoemde modellen wijkt in die zin af van de univariate modellen dat de multivariate modellen uitgaan van twee afhankelijke variabelen, namelijk de score op de beginmeting en de score op de eindmeting.
In het kader van dit proefschrift worden de resultaten van de univariate en multivariate modellen met elkaar vergeleken. Nagegaan wordt of toepassing van deze modellen tot verschillen leiden in de beoordeling van de bijdragen van scholen. Aangezien de dataset van de vier projectscholen erg beperkt van omvang is, is besloten de multivariate analyses uit te voeren op een andere dataset. Deze dataset is afkomstig uit een onderzoek naar een verkenning van de mogelijkheden de schoolspecifieke bijdrage aan de onderwijsopbrengst in kaart te brengen met behulp van het Cito-leerlingvolgsysteem en de Eindtoets basisonderwijs (Wijnstra, Ouwens en Béguin, 2003). Om een vergelijking te kunnen maken tussen de univariate en multivariate modellen, zijn op deze nieuwe dataset (in dit proefschrift aangeduid als de WOB-dataset) ook de univariate multilevel analyses nogmaals toegepast. De resultaten van deze analyses worden in paragraaf 6.10.2 besproken. In de paragrafen 6.10.3 tot en met 6.10.6 komen de uitgevoerde multilevel analyses aan bod en in paragraaf 6.10.7. wordt een vergelijking gemaakt tussen de resultaten van de univariate en multivariate analyses. In paragraaf 6.10.8 ten slotte wordt ingegaan op de betekenis van de uitgevoerde analyses voor de scholen.

6.10.1 Univariate analytische modellen

Toepassing van het univariate variantie-analytisch model

Voor de toepassing van de drie modellen in deze paragraaf is gebruik gemaakt van leerlingen waarvan de resultaten op twee tijdstippen beschikbaar waren. Dit betrof in totaal 193 leerlingen van de projectscholen 3, 4 en 5. Hoewel in principe bij de beschikbare data vier niveaus (school - cohort - leerling - tijdstip) onderscheiden kunnen worden, bevatte de dataset daarvoor te weinig scholen. De scholen 3 en 4 zijn in het model als dummies opgenomen en school 5 als referentieschool. Bovendien is er geen onderscheid gemaakt tussen leerlingen uit verschillende cohorten.

Als eerste is een model gepast met de scholen 3 en 4 als dummy-variabelen en geen achtergrondgegevens van leerlingen. Het model kan geschreven worden als:

$$toets2_{ij} = \beta_{0i} + \beta_1 dumsch3_j + \beta_2 dumsch4_j \qquad (6.1)$$

waarbij de coëfficiënt β_{0i} te schrijven is als: $\beta_{0i} = \beta_0 + e_{0ij}$, en $e_{0ij} \sim N(0, \sigma^2)$.

In vergelijking (6.1) is $toets2_{ij}$ de score van leerling *i* van school *j* op tijdstip 4.6. β_0 is de gemiddelde score voor de leerlingen van school 5 op tijdstip 4.6 en e_{0ij} is de afwijking van een leerling *i* ten opzichte van het gemiddelde op dat tijdstip.

Passing van model (6.1) levert de parameterschattingen op zoals weergegeven in tabel 6.7.

Tabel 6.7

Schattingen van het univariate variantie-analytisch model op tijdstip 4.6

Fixed effects	Coëfficiënt	S.E.
β_0 = gemiddelde score op tijdstip 4.6	62,96	0,810
β_1 = coëfficiënt voor dumsch3	4,82	1,488
β_2 = coëfficiënt voor dumsch4	6,66	1,594

Random Effect	Variantiecomponent	S.E.
Niveau 1 variantie:		
$\sigma^2 = \text{var}\left(e_{0ij}\right)$	71,602	7,289
Deviantie	1372,036	

Uit tabel 6.7 blijkt dat de gemiddelde score van de leerlingen van school 5 op tijdstip 4.6 gelijk is aan 62,96. De coëfficiënten β_1 en β_2 geven aan dat de gemiddelde score van de leerlingen op school 3 en school 4 op dat tijdstip hoger zijn dan van school 5. Op basis van de ratio van de absolute waarde van de schattingen van de coëfficiënten en de bijbehorende standaardfouten $(|\beta|/se(\beta))$ kan een Z-toets uitgevoerd worden om te toetsen of de coëfficiënten significant afwijken van 0. Zowel β_1 als β_2 wijken significant ($\alpha = 0.05$) af van 0, waaruit geconcludeerd kan worden dat de scholen 3 en 4 hoger scoren dan school 5.

De resultaten uit tabel 6.7 zullen hierna als basis dienen om na te gaan of uitbreiding van het gehanteerde model tot een betere passing leidt.

Als eerste uitbreiding zijn in model (6.1) achtergrondvariabelen van leerlingen opgenomen, waarbij alleen sekse een significante bijdrage bleek te leveren.
Het model kan geschreven worden als:

$$toets2_{ij} = \beta_{0i} + \beta_1 dumsch3_j + \beta_2 dumsch4_j + \beta_3 sexe_{ij} \qquad (6.2)$$

waarbij de coëfficiënt β_{0i} te schrijven is als $\beta_{0i} = \beta_0 + e_{0ij}$, en $e_{0ij} \sim N(0, \sigma^2)$. De parameterschattingen van model (6.2) staan in tabel 6.8.

Tabel 6.8
Schattingen na opname van de achtergrondvariabele sekse in het model

Fixed effects	Coëfficiënt	S.E.
β_0 = gemiddelde score op tijdstip 4.6	65,67	1,002
β_1 = coëfficiënt voor dumsch3	4,40	1,426
β_2 = coëfficiënt voor dumsch4	5,92	1,534
β_3 = coëfficiënt voor sekse	-5,01	1,173

Random Effect	Variantiecomponent	S.E.
Niveau 1 variantie:		
$\sigma^2 = \text{var}(e_{0ij})$	65,437	6,661
Deviantie	1354,661	

De grootte van het verschil in deviantiescore tussen de modellen (6.1) en (6.2) (1372,036-1354,661) en het aantal coëfficiënten dat extra geschat moet worden (één), laat zien dat het model waarin ook de achter grondvariabele 'sekse' is opgenomen, beter past χ^2 = 17,38 met df = 1, p < 0,001.

Uit tabel 6.8 blijkt dat de scholen 3 en 4 ook na het opnemen van 'sekse' als verklarende variabele beide significant beter zijn dan school 5. In het model is 'sekse' als dummy-variabele opgenomen met 'jongens' als referentie. De variabele 'sekse' blijkt invloed te hebben op de prestaties van leerlingen op het tijdstip 4.6. De coëfficiënt is negatief, wat in dit geval betekent dat jongens het gemiddeld beter doen dan meisjes.

Toepassing van het univariate covariantie-analytisch model

De dataset waarop de besproken analyses gebaseerd zijn, bevatten de gegevens van dezelfde leerlingen op de tijdstippen 4.4 en 4.6. Door nu de gegevens op het

tijdstip 4.4 als beginmeting te beschouwen en deze in het model als covariaat op te nemen, wordt het eindniveau gecorrigeerd voor de bij aanvang aanwezige verschillen. Het voorgaande resulteert in een model dat het univariate covariantie-analytisch model genoemd is.

Voor het opnemen van de resultaten van de leerlingen op tijdstip 4.4. gaan we uit van model 6.1. Na opname van de resultaten van de leerlingen op tijdstip 4.4 als covariaat ziet dit model er als volgt uit:

$$toets2_{ij} = \beta_{0i} + \beta_1 dumsch3_j + \beta_2 dumsch4_j + \beta_3 toets1_{ij} \qquad (6.3)$$

waarbij de coëfficiënt β_{0i} te schrijven is als $\beta_{0i} = \beta_0 + e_{0ij}$, en $e_{0ij} \sim N\left(0, \sigma^2\right)$. β_0 is de gemiddelde score voor alle leerlingen op tijdstip 4.6 en e_{0ij} is de afwijking van leerling *i* van school *j* ten opzichte van de gemiddelde score β_0 op dat tijdstip.

De parameterschattingen van model (6.3) staan in tabel 6.9.

Tabel 6.9

Parameter schattingen met de resultaten op tijdstip 4.4 als covariaat

Fixed effects	Coëfficiënt	S.E.
β_0 = gemiddelde score op tijdstip 4.6	28,78	1,71
β_1 = coëfficiënt voor dumsch3	-0,41	0,86
β_2 = coëfficiënt voor dumsch4	2,49	0,91
β_3 = coëfficiënt voor toets 1 op tijdstip 4.4	0,65	0,03
Random Effect	**Variantiecomponent**	**S.E.**
Niveau 1 variantie:		
$\sigma_1^2 = \text{var}\left(e_{0ij}\right)$	22,124	2,253
Deviantie	1145,365	

De grootte van het verschil in deviantiescore tussen de modellen (6.1) en (6.3) (1372,036-1145,365), met het gegeven dat er slechts één coëfficiënt extra geschat moet worden, laat zien dat het model waarin de resultaten op tijdstip 4.4 als covariaat zijn opgenomen, beter past χ^2 = 226,67 met df = 1, p < 0,001.

De variantie σ_1^2 wijkt significant af van nul, wat betekent dat de resultaten van de leerlingen significant verschillen op het tijdstip 4.6. Een alternatief voor bovenstaand model is het univariate leerwinst model waarbij als afhankelijke variabele het verschil op de tijdstippen 4.6 en 4.4 wordt genomen (zie ook hoofdstuk 5).

Toepassing van het univariate leerwinstmodel

In dit model wordt de leerwinst gemodelleerd door de verschillen tussen de resultaten op de tijdstippen 4.6 en 4.4 als afhankelijke variabele in de analyse te betrekken. Het model kan als volgt worden weergegeven.

$$t(4.6-4.4)_{ij} = \beta_{0i} + \beta_1 dumsch3_j + \beta_2 dumsch4_j \qquad (6.4)$$

waarbij de coëfficiënt β_{0i} te schrijven is als $\beta_{0i} = \beta_0 + e_{0ij}$, en $e_{0ij} \sim N(0, \sigma^2)$.

Passing van model (6.4) levert de parameterschattingen zoals weergegeven in tabel 6.10 op.

Tabel 6.10

*Schattingen met het verschil op de tijdstippen 4.6 en 4.4 als afhankelijke variabe*le

Fixed effects	Coëfficiënt	S.E.
β_0 = gemiddelde score op tijdstip 4.6-4.4	10,01	0,58
β_1 = coëfficiënt voor dumsch3	-3,29	1,07
β_2 = coëfficiënt voor dumsch4	0,20	1,15

Random Effect	Variantiecomponent	S.E.
Niveau 1 variantie:		
$\sigma^2 = \text{var}\left(e_{0ij}\right)$	37,038	3,770
Deviantie	1244,817	

Uit de deviantiescore blijkt dat model 6.4 significant beter past dan model 6.1. Uit tabel 6.10 blijkt dat de gemiddelde verschilscore of leerwinstscore van de leerlingen op school 5 gelijk is aan 10,01. De leerlingen van school 3 kennen een lagere leerwinstscore. Deze is gelijk aan 10,01-3,29 = 6,72. Gemiddeld genomen leren de leerlingen van deze school in de periode 4.4 tot 4.6 minder dan de leerlingen van school 5. Voor school 4 kan geconcludeerd worden dat het gemiddelde verschil van de leerlingen van school 4 op de tijdstippen 4.6 en 4.4 niet significant afwijkt van die van school 5.

Ook model 6.4 kan uitgebreid worden door het opnemen van achtergrondvariabelen en ook bij dit model bleek alleen de variabele sekse een significante bijdrage te leveren.

Het model kan geschreven worden als:

$$t(4.6-4.4)_{ij} = \beta_{0i} + \beta_1 dumsch3_j + \beta_2 dumsch4_j + \beta_3 sexe_{ij} \qquad (6.5)$$

waarbij de coëfficiënt β_{0i} te schrijven is als: $\beta_{0i} = \beta_0 + e_{0ij}$, en $e_{0ij} \sim N\left(0, \sigma^2\right)$.

De parameterschattingen van model (6.5) staan in tabel 6.11.

Tabel 6.11

Schattingen na opname van de achtergrondvariabele sekse

Fixed effects	Coëfficiënt	S.E.
β_0 = gemiddelde verschilscore op 4.6-4.4	8,62	0,74
β_1 = coëfficiënt voor dumsch3	-3,06	1,05
β_2 = coëfficiënt voor dumsch4	0,57	1,14
β_3 = coëfficiënt voor sekse	2,58	0,86

Random Effect	Variantiecomponent	S.E.
Niveau 1 variantie:		
$\sigma^2 = \text{var}\left(e_{0ij}\right)$	35,413	3,605
Deviantie	1236,155	

Uit de afname van de deviantiescore van model (6.5) ten opzichte van model (6.4) gegeven het aantal extra opgenomen variabelen, mag geconcludeerd worden dat model (6.5) beter past dan model (6.4).

Uit tabel 6.11 blijkt dat de achtergrondvariabele sekse een significante bijdrage levert. We zien ook nu weer dat de gemiddelde leerwinst van de leerlingen op school 5 hoger is dan op school 3. Ook na opname van de achtergrondvariabele sekse wijkt de gemiddelde toename in leerwinst van de leerlingen op school 4 op de tijdstippen 4.6 en 4.4 niet significant af van school 5.

Merk de positieve waarde van coëfficiënt β_3 in tabel 6.11 op. In tabel 6.8 had de vergelijkbare coëfficiënt een negatieve waarde, wat betekende dat de jongens gemiddeld beter presteerden dan de meisjes op tijdstip 4.6. In model (6.5) waar de resultaten van tabel 6.11 betrekking op hebben, is de afhankelijke variabele niet het resultaat op tijdstip 4.6, maar de leerwinst over de periode 4.4 tot 4.6. De positieve waarde van de coëfficiënt β_3 in tabel 6.11 geeft aan dat de meisjes gemiddeld genomen meer leerwinst behaalden in voornoemde periode dan de jongens, wat zou kunnen betekenen dat de meisjes hun achterstand inlopen.

6.10.2 Univariate analytische modellen toegepast op de WOB-dataset[6]

De in de vorige paragraaf uitgevoerde analyses hadden betrekking op de data van de vier projectscholen. In paragraaf 6.10 is aangegeven dat om de resultaten van het toepassen van multivariate modellen te kunnen vergelijken met univariate modellen beide analyses op dezelfde dataset zijn uitgevoerd. Als dataset is gekozen voor wat werd aangeduid als de WOB-dataset. Alvorens de uitgevoerde analyses te bespreken, wordt eerst de gebruikte WOB-dataset kort beschreven.

De WOB-dataset bestaat uit de gegevens van 68 scholen. Van deze scholen zijn de resultaten beschikbaar op de Eindtoets Basisonderwijs in 2002 en 2003. Deze resultaten zullen als opbrengstmaat opgevat worden. Als beginmeting zijn de resultaten in jaargroep 4 op de volgende Cito-LVS-toetsen genomen:

- technisch lezen (Drie-Minuten-Toets, kaart 3);
- begrijpend lezen (Lezen met begrip; schaal Betekenisrelaties en schaal Verwijsrelaties);
- spelling (schaal Spellingvaardigheid);
- rekenen-wiskunde (schaal Rekenen-Wiskunde algemeen).

Bij de LVS-toetsen dient opgemerkt te worden dat deze in principe halverwege en aan het einde van groep 4 worden afgenomen. Bij een aantal scholen waren de gegevens halverwege groep 4 beschikbaar en bij een aantal aan het einde van groep 4. Gekozen is gebruik te maken van het toetsmoment met de meeste scores. Als voor dat moment een score ontbrak, maar er wel één voor het andere moment beschikbaar was, is op basis van de relatie tussen de scores op de twee

[6] De toepassing in dit proefschrift van de univariate analytische modellen op de WOB-dataset komt overeen met het door Wijnstra e.a. (2003) uitgevoerde onderzoek. Voor een deel zijn dezelfde analyses uitgevoerd op dezelfde dataset. Aangezien een aantal onafhankelijke variabelen in het kader van dit proefschrift anders gedefinieerd zijn, zijn de analyses opnieuw uitgevoerd. Bij de beschrijving van de analyses met de univariate analytische modellen is in dit proefschrift voor het overeenkomstig deel vrijelijk gebruik gemaakt van de tekst van Wijnstra e.a., 2003.

momenten een schatting gemaakt van de score voor het toetsmoment dat in de analyses werd opgenomen (zie Wijnstra e.a., 2003, p. 18).

Op basis van de beschikbare achtergrondgegevens van de leerlingen van de 68 scholen zijn de volgende achtergrondvariabelen in de analyses betrokken:

- het geslacht (code sekse);
- de etniciteit (code etniciteit);
- de schoolmaand (code: schmaand). Deze variabele geeft de geboortemaand van de leerling aan, waarbij de codering loopt van 1 - 12 en waarbij 1 overeenkomt met oktober en 12 met september;
- de vraag of een leerling één of meer klassen heeft overgedaan (code: lft_dubl);
- de vraag of een leerling één of meer klassen heeft overgeslagen (code: lft_vroeg).

In totaal bevat de dataset informatie over 5506 leerlingen afkomstig van 68 scholen, waarbij 2713 records betrekking hebben op het jaar 2002 en 2793 records op het jaar 2003. Aangezien er inhoudelijke verschillen zijn tussen de versies van de Eindtoets basisonderwijs in 2002 en 2003 konden deze niet worden gecombineerd en zijn de data voor beide jaren afzonderlijk geanalyseerd. De bespreking van de analyses heeft betrekking op het jaar 2003. De resultaten van de analyses op de dataset uit 2002 zullen deels geanalyseerd worden en zullen alleen gebruikt worden om deze met de resultaten uit 2003 te vergelijken. Voor zowel de dataset die betrekking heeft op 2003 als op 2002 geldt dat deze de resultaten bevatten van de leerlingen op de Eindtoets basisonderwijs uitgedrukt in standaardscores. Ook bevatten beide datasets de resultaten op drie onderdelen van de Eindtoets basisonderwijs, te weten: rekenen, taal en studievaardigheden.

In eerste instantie is een model gepast waarbij geen rekening is gehouden met het aanvangsniveau en de achtergrondkenmerken van leerlingen (het nulmodel). Op basis van deze analyse kunnen de scholen geplaatst worden in een rangorde. Dat zo mogelijk gecorrigeerd moet worden voor zowel achtergrondkenmerken van leerlingen als hun beginniveau om een eerlijkere vergelijking tussen scholen

te maken in hun bijdrage aan de toename in vaardigheid van leerlingen, is in hoofdstuk 3 besproken (vergelijk 'net achievement measures'). Om die reden is nagegaan in hoeverre opname van achtergrondvariabelen (model MA), respectievelijk de resultaten op de toetsen uit het Cito-LVS (model MT) als beginmeting en een combinatie van beide (model MV) als verklarende variabelen in het model, leiden tot een reductie van de variantie tussen scholen en leerlingen. Alle drie te bespreken modellen MA, MT en MV hebben de volgende vorm:

$$STD03_{ij} = \beta_o + \beta_1 X_1 + ...\beta_p X_p + \mu_{oj} + e_{ij}. \qquad (6.6)$$

In vergelijking (6.6) staat $STD03_{ij}$ voor de standaardscore van leerling *i* van school *j* op de Eindtoets basisonderwijs 2003. $i = 1, \ldots, I_j$, $j = 1, \ldots, J$. De *p* onafhankelijke variabelen $X_1, \ldots Xp$ hebben betrekking op de achtergrondkenmerken sekse, etniciteit, lft_dub, lft_vroeg, en schmaand, de schaalscores op de toetsen Technisch lezen (tl), Betekenisrelaties (sbr), Verwijsrelaties (svr), Spellingsvaardigheid (svs) en Rekenen-Wiskunde algemeen (rw), de variabelen naar aanleiding van de controle op lineariteit (subscript 'groot' of 'klein') en de dummy's die aangeven op welk toetsmoment de toetsen uit groep 4 zijn afgenomen. In het model is meegenomen dat de toetsen zowel een lineair verband als een kwadratisch verband kunnen hebben met de afhankelijke variabele. Zie bijvoorbeeld tl (lineair) en tl_2 (kwadratisch).

De resultaten van de analyses staan weergegeven in tabel 6.12. In de linkerkolom staan de variabelen die in het model zijn opgenomen. Vanzelfsprekend zijn bij de nulmodellen (aangeduid met 'nul') geen verklarende variabelen in het model opgenomen. Bij de andere modellen worden die variabelen vermeld die een statistisch significante bijdrage leverden aan het schatten van het schooleffect in vergelijking met het nulmodel. De niet-significante variabelen zijn niet in de tabel opgenomen. Tussen haakjes staat de standaardfout.

Tabel 6.12

Overzicht univariatie analyses WOB-dataset 2003

Model	MA-nul	MA	MT-nul	MT	MV-nul	MV
	1	2	3	4	5	6
Intercept	535,43 (0,57)	540,59 (0,99)	536,01 (0,75)	427,71 (5,20)	536,02(0,75)	435,08 (5,10)
sekse		-				-
etniciteit		-3,67 (0,77)				-
Lft_dubl		-6,72 (0,51)				-5,59 (0,61)
Lft_vroeg		3,97 (1,19)				-
schmaand		-				0,61 (0,06)
Tl				0,07 (0,01)		0,07 (0,01)
tl_groot				-		-
tl_2				-		-
tl_dum				-0,53 (1,25)		-1,00 (1,20)
sbr				0,53 (0,05)		0,48 (0,05)
sbr_groot				-		-
sbr_2				-2,83 (0,58)		-2,27 (0,56)
sbr_dum				0,27 (2,53)		-0,39 (2,44)
svr				0,07 (0,02)		0,06 (0,02)
svr_klein				-		-
svr_2				-		-
svr_dum				1,37 (1,91)		1,82 (1,85)
svs				0,06 (0,04)		0,06 (0,03)
svs_klein				-		-
svs_2				-		-
svs_dum				-2,13 (2,39)		-2,85 (2,23)
rw				0,48 (0,03)		0,46 (0,03)
Rw_klein				-		-
Rw_2				-0,54 (0,13)		-0,50 (0,13)
Rw_dum				2,33 (1,40)		2,58 (1,36)
Varianties						
school	18,78 (3,82)	14,91 (3,09)	15,51 (4,72)	8,46 (2,46)	15,49 (4,71)	8,25 (2,37)
leerling	82,56 (2,61)	74,41 (2,35)	82,83 (4,05)	34,99 (1,71)	82,57 (4,04)	31,91(1,56)
verbetering fit		221,9		745,04		818,96
aantal lln	2065/2793	2065/2793	873/2793	873/2793	871/2793	871/2793

Uit tabel 6.12 blijkt dat een aantal achtergrondvariabelen en een aantal toetsen een statistisch significante bijdrage leveren aan het schatten van het schooleffect in vergelijking met hun nulmodellen waarin alleen de standaardscores op de Eindtoets basisonderwijs zijn opgenomen. In combinatie (model MV) is de bijdrage zelfs groter. Onder dit volledige model MV wordt in totaal (15,49 + 82,57) - (8,25 + 31,91)/(15,49 + 82,57) = 59% van de variantie verklaard ten opzichte van het nulmodel (model MV-nul). Opname van de toetsen

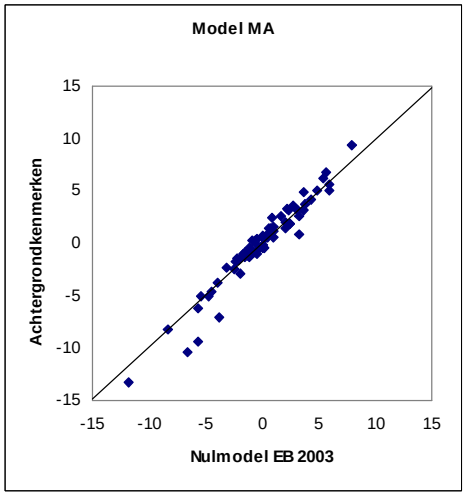
als beginmeting levert een reductie van de variantie tussen scholen en leerlingen op van (15,51 + 82,83)-(8,46 + 34,99)/(15,51 + 82,83) = 55,8%. De achtergrondvariabelen alleen verklaren (18,78 + 82,56) - (14,91 + 74,41)/ (18,78 + 82,56) = 11,9% van de variantie. Uit deze resultaten kan geconcludeerd worden dat er een sterke relatie is tussen de scores op de Eindtoets basisonderwijs en de toetsscores in groep 4. Deze is veel sterker dan de relatie tussen de scores op de Eindtoets basisonderwijs en de achtergrondkenmerken. De schooleffecten onder het nulmodel tonen daardoor een hoge correlatie (r = 0,97 over 68 scholen) met de schooleffecten onder het model met alleen de achtergrondkenmerken. De correlatie tussen de schooleffecten onder het nulmodel met de schooleffecten onder het model met alleen de toetsen bedraagt 0,62 (36 scholen). Bij opname van zowel achtergrondkenmerken en toetsscores bedraagt de correlatie 0,68 (36 scholen). Deze lage(re) correlaties hebben tot gevolg dat veel scholen op een andere positie terechtkomen dan onder het nulmodel. Dit effect is in figuur 6.8[7] onder D zichtbaar gemaakt.

In de grafieken A tot en met C van figuur 6.8 staat de relatie tussen de schooleffecten onder het nulmodel (horizontale assen) en onder het model met alleen de achtergrondkenmerken (MA), respectievelijk de toetsen (MT) en achtergrondkenmerken en toetsen (MV) (verticale assen) grafisch weergegeven. De schaalverdeling is gebaseerd op een transformatie naar eenheden van de standaardscoreschaal van de Eindtoets basisonderwijs met een standaarddeviatie van ongeveer 10 in de populatie van leerlingen. Een positieve waarde op de assen in de grafieken A tot en met C geeft aan dat de desbetreffende school gegeven het toegepaste model, een positieve bijdrage levert ten opzichte van de gemiddelde bijdrage van de aan het onderzoek deelnemende scholen. Een negatieve waarde geeft aan dat een school het relatief minder goed doet ten opzichte van de

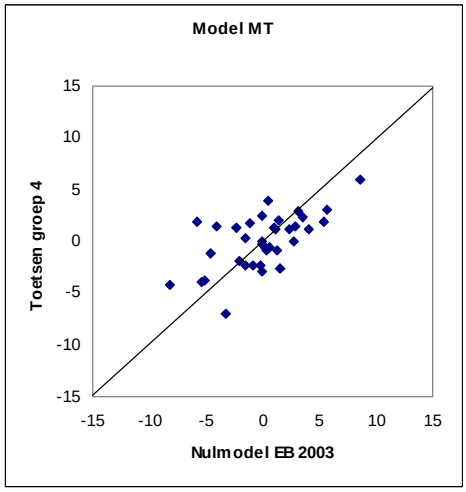
[7] Niet alleen voor figuur 6.8, maar ook voor de andere daarna te bespreken figuren geldt dat bij de weergave van de figuren geen rekening is gehouden met de onzekerheid van de uitkomsten. Daar moet bij de interpretatie van de gegevens rekening mee gehouden worden. De resultaten zijn in de regel niet zo absoluut als de figuren wellicht doen vermoeden. Later wordt uiteengezet hoe onzekerheid grafisch weergegeven kan worden (zie figuur. 6.15).

gemiddelde bijdrage. Bevindt een school zich onder de diagonaal dan geeft dat aan dat de positie van de school in de verdeling van scholen lager is na correctie voor achtergrondkenmerken en/of toetsresultaten. Ten opzichte van de andere scholen doet deze school het dan relatief slechter. Bevindt een school zich boven de diagonaal dan doet een school het relatief beter na correctie voor achtergrondkenmerken en/of toetsresultaten. Voor die scholen die zich op de diagonaal bevinden, heeft correctie voor achtergrondkenmerken en/of toetsresultaten geen effect op hun positie ten opzichte van de andere scholen. Grafiek D geeft de verschillen in rangorde bij correctie voor toetsresultaten en achtergrondkenmerken weer.

A Model met achtergrond Kenmerken

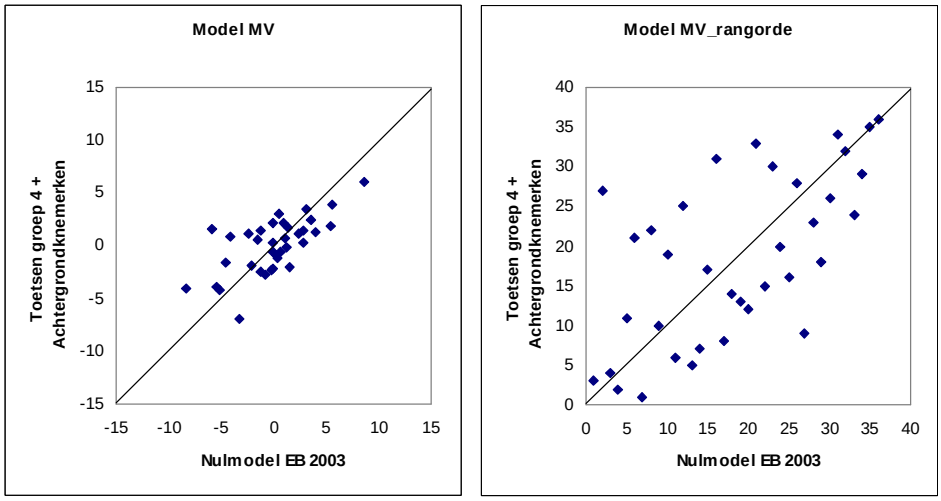


B Model met toetsen groep 4



C Model met toetsen groep 4 en achtergrondkenmerken

D Vergelijking rangorde scholen modellen MV-nul en MV



Figuur 6.8
Relatie tussen de schooleffecten op de standaardscores van de Eindtoets basisonderwijs in drie modellen, in vergelijking met de effecten onder het nulmodel.

Uit grafiek A kan men aflezen dat bij controle voor alleen achtergrondkenmerken de scholen nagenoeg niet van rangorde veranderen. De meeste scholen bevinden zich op of nabij de diagonaal. De veranderingen in de bijdrage van de scholen zijn in de grafieken B en C aanzienlijker. Hoewel bij de modellen MT en MV de spreiding van de schooleffecten wel afneemt, blijft deze toch nog aanzienlijk. Grafiek D laat zien hoe de relatieve positie van de scholen in de verdeling verandert na correctie voor achtergrondkenmerken en toetsresultaten. Op de assen staat de plaats van de scholen in de verdeling van scholen onder het nulmodel (horizontale as) en onder het model MV (verticale as) weergegeven. Voor drie scholen blijkt deze niet te veranderen, 15 scholen verschuiven in positieve richting en 18 scholen nemen na correctie een lagere positie in.

In de zojuist besproken analyses was de afhankelijke variabele de score van de leerlingen op de volledige Eindtoets basisonderwijs (exclusief wereldoriëntatie dat facultatief is voor scholen) uitgedrukt in standaardscores. De berekening van de standaardscores is gebaseerd op de resultaten van de leerlingen op de onderdelen rekenen, taal en studievaardigheden. Ook voor deze drie onderdelen afzonderlijk is op vergelijkbare wijze nagegaan in hoeverre correctie voor achtergrondkenmerken (MA), toetsen (MT) en een combinatie van achtergrondkenmerken en toetsen (MV) van invloed is op de schooleffecten. De correlaties tussen de schooleffecten onder de modellen MA, MT en MV met hun respectievelijke nulmodellen staan in tabel 6.13. Voor de vergelijkbaarheid zijn ook de correlaties voor de standaardscores in de eerste rij van tabel 6.13 opgenomen.

Tabel 6.13
Correlatie schooleffecten onder het nulmodel in vergelijking met de modellen MA, MT en MV

	MA	MT	MV
Standaardscores	0,97	0,62	0,67
Rekenen	0,96	0,70	0,75
Taal	0,97	0,56	0,73
Studievaardigheden	0.98	0,72	0,76

Uit tabel 6.13 blijkt dat de orde van grootte van de diverse correlaties vergelijkbaar is. Wel laat de opname van een voormeting (MT) de grootste verschuiving in schooleffecten zien bij het onderdeel taal.

Schooleffecten op basis van de gegevens Eindtoets basisonderwijs 2002

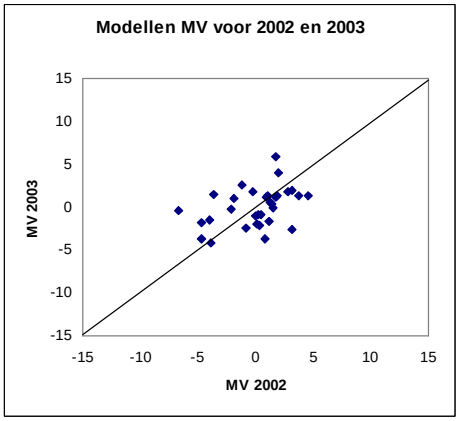
De tot nu toe verrichte analyses hebben betrekking op de gegevens van de Eindtoets basisonderwijs 2003. Om de vraag te beantwoorden hoe stabiel deze gegevens in de tijd zijn, zijn voorgaande analyses ook uitgevoerd op de resultaten van de leerlingen op de Eindtoets basisonderwijs 2002. De resultaten van 2002 wijken in het algemeen niet af van de resultaten van 2003. De correlaties van de schooleffecten onder het nulmodel in vergelijking met de modellen MA,

MT en MV voor 2002 staan weergegeven in tabel 6.14. Ter vergelijking zijn in de tabel 6.14 ook de resultaten over 2003 opgenomen.

Tabel 6.14
Correlatie schooleffecten onder het nulmodel in vergelijking met de modellen MA, MT en MV voor 2002 en 2003

	MA		MT		MV	
	2003	2002	2003	2002	2003	2002
Standaardscores	0,97	0,98	0,62	0,52	0,67	0,57
Rekenen	0,96	0,98	0,70	0,68	0,75	0,66
Taal	0,97	0,97	0,56	0,59	0,73	0,56
Studievaardigheden	0.98	0,98	0,72	0,63	0,76	0,62

Ook voor 2002 geldt dat de modellen MT en MV leiden tot een verschuiving in de rangorde van scholen. De verschuivingen voor 2002 zijn in het algemeen groter dan voor 2003. Ook een correctie voor alleen achtergrondkenmerken laat voor beide jaren nagenoeg hetzelfde effect zien. Opgemerkt moet worden dat de verschuiving in rangorde voor het model MV voor de jaren 2002 en 2003 niet betrekking heeft op *dezelfde* scholen (zie daarvoor figuur 6.9).



Figuur 6.9
Relatie tussen de schooleffecten in het volledige model voor 2002 en 2003

De resultaten zoals vermeld in figuur 6.9 hebben wel betrekking op dezelfde scholen uit 2002 en 2003. De relatie uitgedrukt als correlatiecoëfficiënt is 0.40 (berekend over 32 scholen), wat in het algemeen betekent dat de relatieve rangorde van een school in de verdeling van scholen met betrekking tot zijn schooleffect verschillend is tussen beide jaren. De gegevens van het ene schooljaar zijn dus niet maatgevend voor de gegevens van het andere schooljaar.

6.10.3 Vaststellen schooleffect met behulp van multivariate analytische modellen

In paragraaf 6.10.2 zijn de schooleffecten bepaald met univariate analytische modellen. In deze modellen wordt slechts één afhankelijke variabele onderscheiden. In de eerder besproken modellen waren dat de behaalde scores op de totale Eindtoets basisonderwijs en de behaalde scores op de onderdelen rekenen, taal en studievaardigheden. Schooleffecten kunnen ook bepaald worden met multivariate analytische modellen. Deze modellen onderscheiden zich van de univariate modellen door het opnemen van meer afhankelijke variabelen. In de volgende paragraaf worden de resultaten beschreven van de analyses naar schooleffecten met multivariate modellen. Ook worden de resultaten uit deze analyses vergeleken met de resultaten verkregen met univariate modellen. In de eerste plaats zijn multivariate modellen gepast met als afhankelijke variabelen de resultaten op de drie onderdelen van de Eindtoets basisonderwijs (multivariate modellen naar inhoudsdomein). Deze analysemethode wordt hierna aangeduid met de term ‘MV-inhoudsdomein’. In de tweede plaats zijn als afhankelijke variabelen de scores op de beginmeting en de scores op de eindmeting genomen. Deze laatstgenoemde methode wordt door Van den Bergh en Kuhlemeier (1997) het multivariate variantie-analytisch model genoemd (zie hoofdstuk 5). Tijdens de bespreking van de resultaten van de analyses wordt deze laatstgenoemde methode aangeduid met de term MUVA-model. Exemplarisch zijn bij het MUVA-model als beginmeting alleen de scores van de leerlingen op de toets Rekenen-Wiskunde algemeen genomen. Deze toets verklaarde de meeste

variantie in vergelijking met de andere toetsen uit groep vier. Als eindmeting is de score op de Eindtoets basisonderwijs uitgedrukt in standaardscores genomen.

6.10.4 MV-inhoudsdomein modellen

De afhankelijke variabelen bij dit type multivariate modellen zijn de resultaten van de leerlingen op de drie onderdelen rekenen, taal en studievaardigheden van de Eindtoets basisonderwijs 2003. Alle drie de modellen hebben de volgende vorm:

$$Y_{ij} = \beta_o + \beta_1 X_1 + ... \beta_p X_p + \mu_{oj} + e_{ij}. \qquad (6.7)$$

In vergelijking (6.7) staat Y_{ij} voor het behaalde resultaat van leerling *i* van school *j* op het onderdeel rekenen, respectievelijk taal en studievaardigheden. $i = 1, \ldots, I_j$, $j = 1, \ldots, J$. De *p* onafhankelijke variabelen $X_1, \ldots Xp$ hebben betrekking op de onafhankelijke variabelen zoals deze in de modellen zijn opgenomen. In totaal zijn drie analyses uitgevoerd waarbij bij elke analyse de schooleffecten onder het nulmodel en onder het model met alleen achtergrondkenmerken (MA), met toetsen (MT) en met achtergrondkenmerken en toetsen (MV) met elkaar zijn vergeleken. Exemplarisch zijn voor het model MV de resultaten in tabel 6.15 opgenomen. In de linkerkolom staan de variabelen die in het model zijn opgenomen. In de overige kolommen staan de resultaten voor de drie afhankelijke variabelen rekenen, taal en studievaardigheden. In tabel 6.15 staan alleen die variabelen vermeld die een statistisch significante bijdrage leverden aan het schatten van het schooleffect.

Tabel 6.15

Resultaten multivariate analyse WOB-dataset 2003 model MV

	Model MV		
	Rekenen	**Taal**	**Studievaardigheden**
Intercept	-71,109 (6,55)	- 29,855 (4,91)	-27,052 (3,22)
Sekse	1,629 (0,46)	- 2,017 (0,43)	-
Etniciteit	-	-	-
lft_dubl	-7,058 (0,84)	-5,633 (0,74)	-2,995 (0,46)
lft_vroeg	-	-	-
Schmaand	0,140 (0,07)	0,170 (0,06)	-
Tl	0,125 (0,02)	0,041 (0,02)	0,048 (0,01)
Tl_groot	-	-	-
Tl_2	-	-	-
t5_dum	-2,123 (1,67)	-0,120 (1,37)	-0,631 (0,80)
Sbr	0,866 (0,07)	0,244 (0,05)	0,307 (0,04)
sbr_groot	-	-	-
sbr_2	-4,368 (0,69)	-	-1,283 (0,37)
sbr_dum	-1,239 (3,23)	2,251 (2,36)	2,109 (1,56)
Svr	0,070 (0,02)	-	0,041 (0,01)
svr_klein	-	-	-
svr_2	-	-	-
svr_dum	-2,042 (1,90)	-	-1,138 (0,85)
Svs	0,095 (0,04)	-	-
svs_klein	-	-	-
svs_2	-	-	-
svs_dum	-0,751 (2,28)	-	-
Rw	0,396 (0,04)	0,671 (0,03)	0,259 (0,02)
rw_klein	-	-	-
rw_2	-0,410 (0,18)	-0,785 (0,16)	-0,283 (0,10)
rw_dum	4,691 (1,88)	1,894 (1,56)	0,790 (0,95)
Varianties			
School			
rekenen	16,745 (4,79)		
taal	10,546 (3,27)	9,071 (2,75)	
studievaardigheden	7,248 (2,09)	4,761 (1,51)	3,182 (0,99)
Leerling			
rekenen	63,007 (3,08)		
taal	26,188 (2,13)	49,320 (2,41)	
studievaardigheden	21,140 (1,44)	17,099 (1,24)	20,317 (0,99)

Uit de waarden van de coëfficiënten in de drie kolommen is af te leiden of de desbetreffende variabele een positieve of een negatieve bijdrage levert aan de waarde van de afhankelijke variabele. Uit de covariantiematrix kan afgeleid worden of de resultaten op de afhankelijke variabelen op de twee onderscheiden niveaus hoog of laag met elkaar correleren.

In tabel 6.16 staan de correlatiecoëfficiënten voor de drie afhankelijke variabelen voor beide niveaus opgenomen.

Tabel 6.16
Overzicht correlatiecoëfficiënten covariantiematrix

	Niveau 1	Niveau 2
rekenen - taal	0,47	0,86
rekenen - studievaardigheden	0,59	0,99
taal - studievaardigheden	0,54	0,89

Tabel 6.16 laat zien dat de onderscheiden drie afhankelijke variabelen hoog correleren op niveau 2 (school). Dat wil zeggen dat in het algemeen geldt dat als een school goed presteert op het ene onderdeel (bijvoorbeeld rekenen) dat gemiddeld genomen ook het geval is voor de andere onderdelen (bijvoorbeeld taal en studievaardigheden). Op het eerste niveau (de leerling) gaat deze vergelijking minder op. De correlatiecoëfficiënten op dit niveau zijn beduidend lager. In het algemeen geldt dat een leerling die hoog scoort op bijvoorbeeld rekenen, niet per definitie ook hoog scoort op een van de andere onderdelen.

Op basis van de resultaten van de uitgevoerde multivariate analyses is nagegaan in hoeverre correctie voor achtergrondkenmerken (MA), toetsen (MT) en een combinatie van achtergrondkenmerken en toetsen (MV) van invloed is op de schooleffecten. In tabel 6.17 staan de correlaties tussen de schooleffecten onder de modellen MA, MT en MV met hun respectievelijke nulmodellen weergegeven. Ter vergelijking zijn in tabel 6.17 ook de correlaties uit tabel 6.13 opgenomen die betrekking hebben op de vergelijkbare univariate modellen MA, MT en MV.

Tabel 6.17

Correlatie schooleffecten onder het nulmodel in vergelijking met de modellen MA, MT en MV voor zowel de multivariate als de univariate modellen

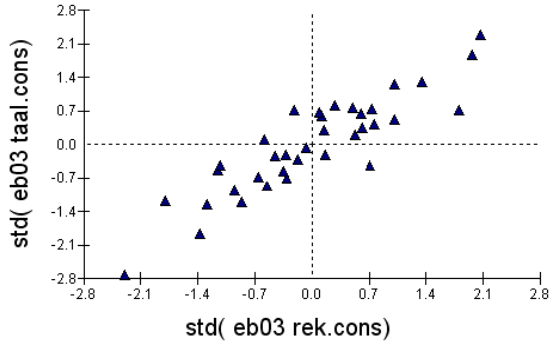
	MA		MT		MV	
	Multi	Uni	Multi	Uni	Multi	Uni
Rekenen	0,97	0,96	0,67	0,70	0,72	0,75
Taal	0,97	0,97	0,42	0,56	0,50	0,73
Studievaardigheden	0,97	0,98	0,66	0,72	0,71	0,76

Uit tabel 6.17 blijkt dat bij de uitgevoerde multivariate analyses op het onderdeel taal na, de orde van grootte van de diverse correlaties vergelijkbaar is. Ook bij de univariate analyses laat het onderdeel taal het grootste effect zien. Het opnemen van toetsresultaten uit groep 4 (model MT) leidt tot de grootste verschuiving in de rangorde van scholen. In het algemeen geldt dat het effect bij de univariatie analyses wat kleiner is in vergelijking met de multivariate analyses.

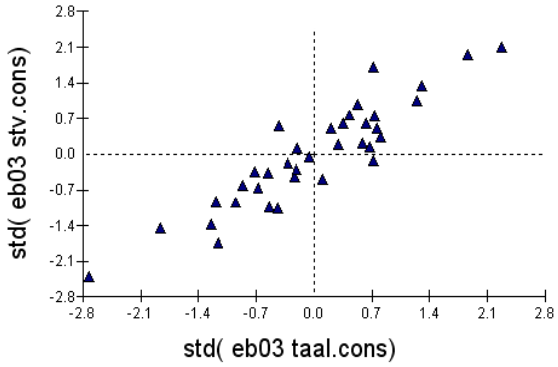
De multilevel analyses zijn uitgevoerd met het programma MlWin (Rasbash e.a., 2000). Dit programma geeft de mogelijkheid bij multivariate analyses de 'standardized residuals'[8] van de afhankelijke variabelen tegen elkaar uit te zetten. Om deze gestandaardiseerde schooleffecten te bepalen, worden de berekende schooleffecten gedeeld door de geschatte standaardfouten, waardoor een maat ontstaat die onafhankelijk is van de variantie. Deze maat is gemakkelijker te interpreteren en geeft de relatieve positie van het schooleffect ten opzichte van de andere schooleffecten weer. In figuur 6.10 zijn de gestandaardiseerde schooleffecten paarsgewijs afgebeeld. De analyses die leiden tot figuur 6.10 zijn gebaseerd op de gegevens van 36 scholen.

[8] De term 'residual' is een term die gebruikt wordt in de handleiding van het programma MlWin. Aangezien de 'residuen' in de uitgevoerde analyses schooleffecten zijn, zal in het vervolg van dit proefschrift in plaats van de term residu de term (schatting van het) schooleffect gebruikt worden.

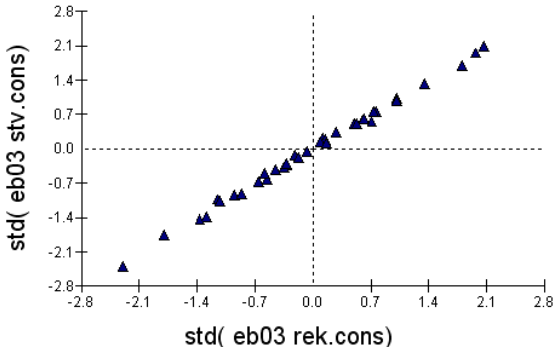
A Rekenen versus Taal



B Taal versus Studievaardigheden



C Rekenen versus Studievaardigheden



Figuur 6.10

Verband tussen de gestandaardiseerde schooleffecten multivariate analyse model MV EB03

Op de horizontale en verticale assen staan de onderdelen van de Eindtoets basisonderwijs die met elkaar gecorreleerd worden. In grafiek A is dat rekenen met taal, in grafiek B taal met studievaardigheden en in grafiek C ten slotte rekenen met studievaardigheden. De assen geven de afwijking van een school op het betreffende onderdeel weer ten opzichte van de verwachte gemiddelde score van die school gegeven haar kenmerken, uitgedrukt in standaarddeviaties. Elk driehoekje in de figuren correspondeert met een school. In totaal zijn per grafiek de resultaten van 36 scholen afgebeeld. Uit elke grafiek is af te lezen of, gegeven haar verwachte gemiddelde score, de mate waarin een school beter of minder goed presteert, op beide onderdelen hetzelfde is. Trekken we een denkbeeldige diagonaal vanuit de 'oorsprong' dan zijn de resultaten van die scholen die zich beneden deze diagonaal bevinden relatief beter op het onderdeel dat op de horizontale as staat afgebeeld ten opzichte van het onderdeel op de verticale as. Voor de scholen die zich boven de diagonaal bevinden geldt het omgekeerde.
De grafieken A en B laten zien dat scholen aanzienlijk van positie kunnen verschuiven bij een vergelijking van zowel rekenen als studievaardigheden met taal. De correlatiecoëfficiënt voor grafiek A is gelijk aan 0,893 en voor grafiek B 0,916. De verschuivingen in grafiek C zijn beperkt (r = 0,998). Blijkbaar verandert in het algemeen de positie van de scholen niet of nauwelijks uitgaande van de onderdelen rekenen en studievaardigheden. Natuurlijk kan de positie voor een individuele school wel wat wijzigen, maar deze wijziging zal gezien de hoge correlatiecoëfficiënt beperkt zijn.

6.10.5 Vergelijking multivariate en univariate analyses voor de drie onderdelen van de Eindtoets basisonderwijs

In de paragrafen 6.10.2 en 6.10.4 zijn de resultaten van univariate en multivariate analyses op (onderdelen van) de Eindtoets basisonderwijs besproken. Voor beide typen zijn voor de drie onderdelen rekenen, taal en studievaardigheden de schooleffecten bepaald voor de modellen MA, MT en MV en hun respectievelijke nulmodellen. In tabel 6.18 worden de resultaten van de analyses met elkaar vergeleken.

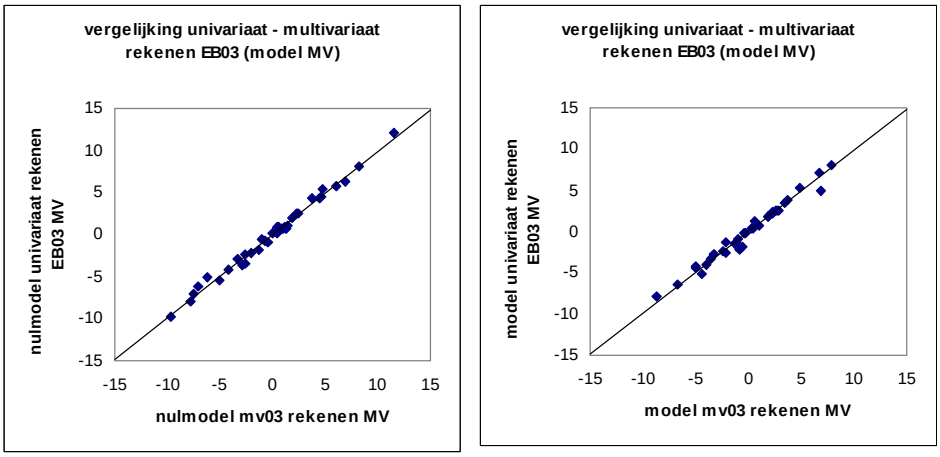
Tabel 6.18

Vergelijking schooleffecten voor de onderdelen rekenen, taal en studievaardigheden op basis van multivariate en univariate analyses voor de modellen MA, MT en MV en hun respectievelijke nulmodellen

Onderdeel	Model		Correlatie schooleffect Multivariaat - Univariaat
Rekenen	MA	Nulmodel	0,996
		Model	0,995
	MT	Nulmodel	0,995
		Model	0,988
	MV	Nulmodel	0,995
		Model	0,988
Taal	MA	Nulmodel	0,946
		Model	0,965
	MT	Nulmodel	0,887*
		Model	0,979*
	MV	Nulmodel	0,885*
		Model	0,977*
Studievaardigheden	MA	Nulmodel	0,971
		Model	0,959
	MT	Nulmodel	0,981
		Model	0,906
	MV	Nulmodel	0,981
		Model	0,924

* Bij de correlaties voor taal zijn twee scholen buiten beschouwing gelaten waarvoor wel resultaten op alle relevante onafhankelijke variabelen beschikbaar waren. Om een onduidelijke reden werden deze scholen in de MLWin software als missend geclassificeerd in het multivariate model.

Uit tabel 6.18 blijkt in het algemeen dat voor het vaststellen van de rangorde van de schooleffecten het niet uitmaakt of de analyses uitgaan van univariate of van multivariate modellen. De correlaties zijn min of meer gelijk en erg hoog. Gegeven de hoogte van de correlaties zal de rangorde van de scholen op basis van beide type analyses nagenoeg vergelijkbaar zijn. Ter illustratie zijn in figuur 6.11 voor het onderdeel rekenen de resultaten van bovengenoemde analyses naar de schooleffecten voor het model MV voor beide type analyses weergegeven. De linker grafiek uit figuur 6.11 heeft betrekking op de nulmodellen en de rechter grafiek betreft de modellen met de onafhankelijke variabelen.



Figuur 6.11
Vergelijking schooleffecten multivariate en univariate analyses voor model MV voor het onderdeel rekenen

De assen geven de afwijking van een school weer ten opzichte van de verwachte gemiddelde score van die school gegeven haar kenmerken. Is de afwijking positief dan doet de school het beter dan verwacht zou mogen worden, is de afwijking negatief dan presteert de school minder. Voor de scholen die zich onder de diagonaal bevinden zijn de effecten positiever wanneer deze berekend zijn met het multivariate model ten opzichte van het univariate model. Voor de scholen boven de diagonaal geeft het univariate model een gunstiger beeld ten opzichte

van het multivariate. Zoals op basis van de correlatiecoëfficiënten uit tabel 6.18 te verwachten is, liggen in figuur 6.11 de schooleffecten op basis van de multivariate en de univariate analyses nagenoeg op een rechte lijn. Dat betekent dat de verwachte afwijkingen voor de scholen op basis van de univariate analyses (verticale as) en op basis van de multivariate analyses (horizontale as) nagenoeg gelijk zijn. Dat geldt zowel voor het nulmodel (r = 0,995) als voor het model waarin de achtergrondvariabelen en de toetsen als covariaten zijn opgenomen (r = 0,998).

6.10.6 Het multivariate variantie-analytisch model (MUVA)

In paragraaf 6.10.3 is aangegeven dat de beschikbare data ook geanalyseerd zouden worden met het multivariate variantie-analytische model (MUVA-model) zoals dat beschreven is door Van den Bergh en Kuhlemeier (zie hoofdstuk 5). In paragraaf 6.10.3 is aangegeven dat ook het MUVA zich onderscheidt van het univariate model door het opnemen van meer afhankelijke variabelen. Het MUVA-model onderscheidt zich echter ook van het in paragraaf 6.10.4 besproken multivariate model door in het onderhavige geval als afhankelijke variabelen niet de drie onderdelen van de Eindtoets basisonderwijs te nemen, maar een score op de beginmeting en een score op de eindmeting. Als eindmeting geldt de score op de Eindtoets basisonderwijs uitgedrukt in standaardscores. Zoals in paragraaf 6.10.3 al is aangegeven, zijn exemplarisch bij het MUVA-model als beginmeting alleen de scores van de leerlingen op de toets Rekenen-Wiskunde algemeen genomen. Deze toets verklaarde de meeste variantie in vergelijking met de andere toetsen uit groep vier.

De resultaten na het passen van het MUVA-model staat weergegeven in tabel 6.19. Merk op dat in het model alleen die variabelen opgenomen zijn die een significante bijdrage leveren aan het schatten van het schooleffect. Het MUVA-model kent dezelfde opbouw zoals weergegeven in vergelijking (6.7) met het verschil dat de afhankelijke variabelen nu betrekking hebben op de

beginmeting, te weten de toets Rekenen-Wiskunde (t8), en de eindmeting, te weten de Eindtoets basisonderwijs 2003 uitgedrukt in standaardscores (eb03_std).

Tabel 6.19

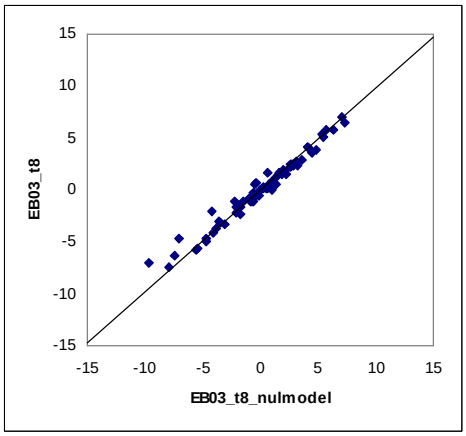
MUVA-model met de resultaten op de toets Rekenen-Wiskunde (t8) als beginmeting en de resultaten op de Eindtoets basisonderwijs (eb03_std) als eindmeting

	MUVA-Model	
	t8	**eb03_std**
Intercept	82,740 (1,10)	539,232 (1,03)
Sekse	-2,926 0,33)	-
Etniciteit	-3,407 (0,78)	-3,615 (0,75)
lft_dubl	-3,371 (0,53)	-7,347 (0,51)
lft_vroeg	-	3,695 (1,00)
Schmaand	-	0,139 (0,05)
Varianties		
School		
T8	13,098 (2,78)	
eb03_std	7,173 (2,34)	16,012 (3,27)
Leerling		
T8	83,225 (2,55)	
eb03_std	48,496 (2,06)	76,435 (2,40)

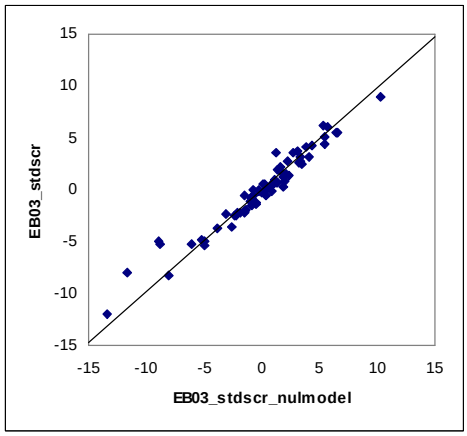
Uit de waarden uit de covariantiematrix is af te leiden dat de correlatie tussen de resultaten op de beginmeting met de eindmeting voor niveau 2 (de school) gelijk is aan 0,50 en voor niveau 1 (de leerling) aan 0,61. Dit betekent dat voor beide niveaus geldt dat een hoge (of lage) score op het ene onderdeel niet automatisch betekent dat ook op het andere onderdeel een hoge (of lage) score behaald wordt.

In figuur 6.12 is de relatie tussen de schooleffecten op zowel de beginmeting als de eindmeting weergegeven voor het model MA en de daarbij behorende nul-modellen.

A Beginmeting



B Eindmeting



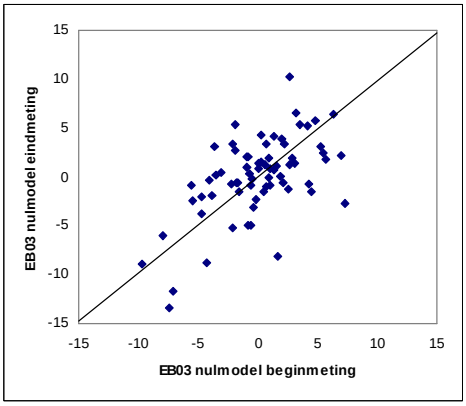
Figuur 6.12
Relatie tussen de schooleffecten beginmeting en eindmeting in het model MA en het nulmodel

In de grafieken A en B van figuur 6.12 staan op de horizontale assen de schooleffecten onder de nulmodellen voor de beginmeting (grafiek A) en de eindmeting (grafiek B). Op de verticale assen staan de schooleffecten na opname in de modellen van de (significante) achtergrondvariabelen (model MA). De grafieken A en B laten zien dat de relatie tussen de schooleffecten groot is. Voor de beginmeting is deze 0,984 (69 scholen) en voor de eindmeting 0,971 (69 scholen). De spreiding van de schooleffecten is bij de eindmeting groter dan bij de beginmeting. Dat geldt zowel voor het nulmodel als voor model MA. Zowel bij de beginmeting als de eindmeting heeft correctie voor achtergrondkenmerken gemiddeld genomen geen grote invloed op de rangorde van de scholen wat betreft hun schooleffecten. Dit resultaat is op zich niet verrassend omdat dit ook bleek uit de eerder besproken univariate en multivariate analyses.

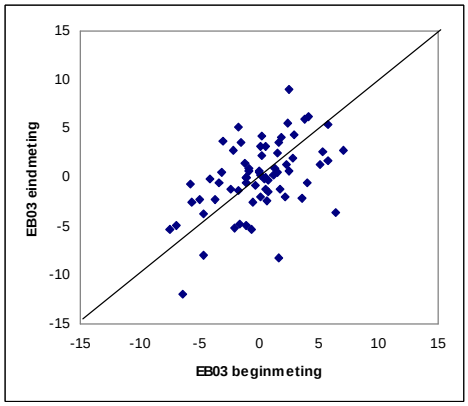
Ook op de dataset van de Eindtoets basisonderwijs voor 2002 zijn bovengenoemde analyses uitgevoerd. De resultaten voor 2002 zijn vergelijkbaar met die van 2003. De correlatie tussen het model MA en het nulmodel voor de beginmeting bedraagt 0,985 (69 scholen) en voor de eindmeting is deze 0,980 (69 scholen).

De relatie tussen de schooleffecten op de beginmeting en eindmeting is minder sterk in vergelijking met de relatie tussen de schooleffecten beginmeting en eindmeting in het model MA en het nulmodel zoals figuur 6.13 laat zien.

Grafiek A: Nulmodel



Grafiek B: Model MA



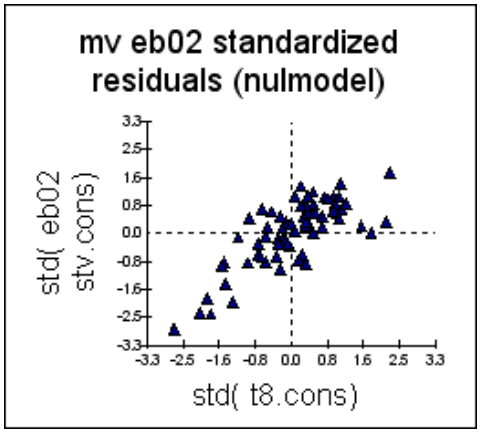
Figuur 6.13
Relatie tussen de schooleffecten 'beginmeting' en 'eindmeting' voor het nulmodel en het model MA (69 scholen)

Voor de nulmodellen (grafiek A) bedraagt de correlatie tussen de beginmeting en de eindmeting 0,587 en voor de modellen MA (grafiek B) is de correlatie 0,493. De hoogtes van de correlaties geven aan dat er verschuivingen plaatsvinden in de volgorde van scholen op basis van de schooleffecten. Uit figuur 6.13 blijkt ook dat de spreiding in schooleffecten toeneemt bij de eindmeting in vergelijking met de beginmeting. Dat geldt zowel voor grafiek A als voor grafiek B.

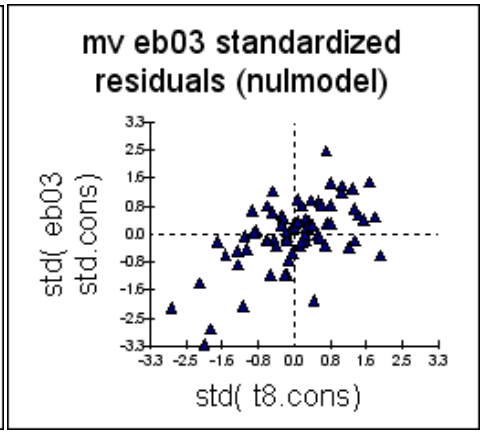
De vergelijkbare correlaties voor de Eindtoets basisonderwijs 2002 tussen de beginmeting en de eindmeting zijn onder het model MA 0,726 (69 scholen) en onder het nulmodel 0,785 (69 scholen). De verschuiving in rangorde op basis van de schooleffecten is over het geheel genomen in 2002 wat minder dan in 2003.

In figuur 6.14 zijn deze effecten voor 2002 en 2003 gevisualiseerd door grafisch de gestandaardiseerde schooleffecten (zie paragraaf 6.10.4) van de afhankelijke variabelen tegen elkaar uit te zetten. Dat is gedaan voor zowel de nulmodellen (grafieken A en B) als voor de modellen MA (grafieken C en D). De assen geven de afwijking van een school weer ten opzichte van de verwachte gemiddelde score van die school gegeven haar kenmerken, uitgedrukt in standaarddeviaties.

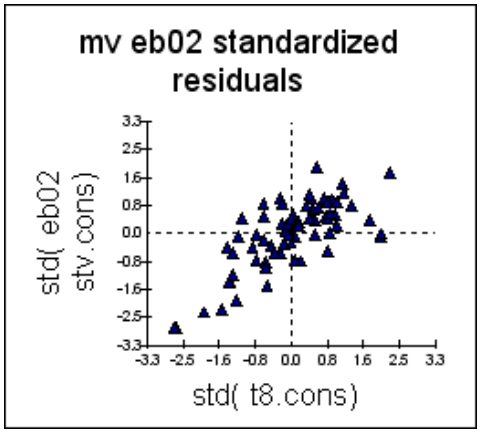
A EB02 nulmodel



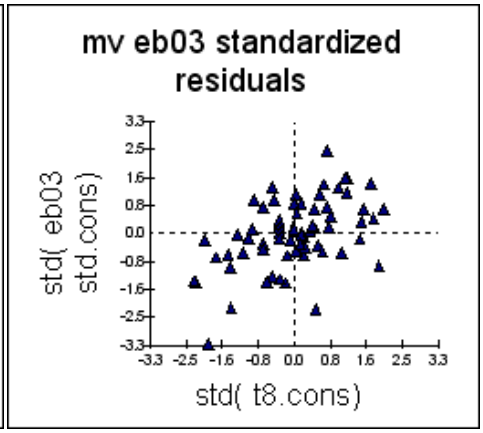
B EB03 nulmodel



C EB02 nulmodel MA



D EB02 nulmodel MA

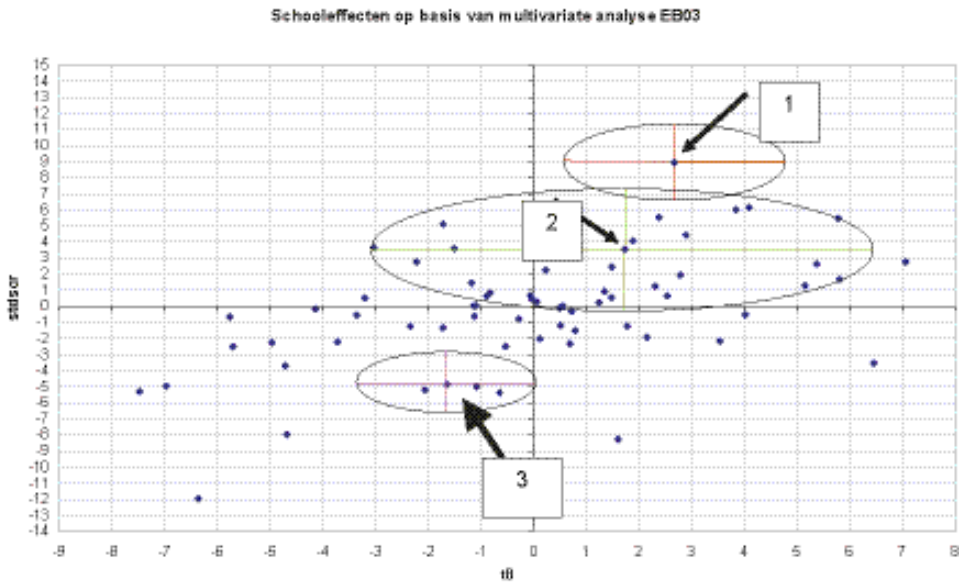


Figuur 6.14

Relatie gestandaardiseerde schooleffecten multivariate analyse MUVA-model EB02 en EB03

Nader onderzoek laat zien dat het niet per definitie dezelfde scholen zijn die zowel in 2002 als in 2003 hoog of laag in de verdeling van scholen staan op basis van de grootte van het schooleffect. Of met andere woorden: de positie van een individuele school in de verdeling van scholen die deel uitmaken van de steekproef, verschilt in de jaren 2002 en 2003 en wellicht mag wel gesteld worden dat deze positie wisselt van jaar tot jaar.

Dat uitspraken over schooleffecten omzichtig gedaan moeten worden, laat figuur 6.15 zien. In deze figuur zijn grafisch de schooleffecten voor 2003 weergegeven voor de beginmeting ‘t8’ en de eindmeting ‘stdscr’. De assen in figuur 6.15 geven de afwijking van een school weer ten opzichte van de verwachte gemiddelde score van die school gegeven haar kenmerken. Op de horizontale as staan de afwijkingen (schooleffecten) met als afhankelijke variabele de beginmeting ‘t8’ en op de verticale as de afwijkingen met als afhankelijke variabele de eindmeting ‘stdscr’.



Figuur 6.15
Betrouwbaarheidsintervallen schooleffecten

In figuur 6.15 zijn twee punten (schooleffecten) gemarkeerd door de pijlen 1 en 2. Het eerste punt dat gemarkeerd wordt door pijl 1 heeft als waarden (+2,53; +8,98). Voor het tweede punt zijn de waarden (+1,59; +3,56). Om te onderzoeken of beide punten significant afwijken van elkaar kunnen betrouwbaarheidsintervallen opgesteld worden. Hierbij worden intervallen bepaald met ± 1,39 standaardfout. Overlap van zulke intervallen geeft aan dat de verschillen niet significant zijn op het 5% niveau (Goldstein, 1999, p. 36). Beide punten nemen duidelijk een andere positie in de verdeling van scholen in. Kijken we echter naar de 95% betrouwbaarheidsintervallen van beide punten zoals weergegeven door de twee ovalen, dan zien we dat deze elkaar overlappen. Uit figuur 6.15 valt af te leiden dat de twee genoemde punten statistisch gezien (95% betrouwbaarheidsinterval) niet van elkaar afwijken. Het zal duidelijk zijn dat ook voor andere waarden in de grafiek geldt dat deze statistisch gezien niet van elkaar afwijken. Dat dat niet voor alle waarden opgaat, laat het punt dat gemarkeerd wordt door pijl 3 zien. Dit punt wijkt statistisch gezien wel significant af van de twee besproken punten. De 95% betrouwbaarheidsgrenzen overlappen elkaar niet.

6.10.7 Samenvatting univariate en multivariate analyses

Op basis van de uitgevoerde univariate en multivariate analyses blijkt dat correcties voor achtergrondvariabelen (MA), toetsen (MT) en zowel achter grondvariabelen als toetsen (MV) invloed hebben op de positie van individuele scholen in de verdeling van scholen uitgaande van de grootte van de schooleffecten. Correctie voor toetsen laat de grootste verschuiving zien. De verschuiving bij alleen een correctie voor achtergrondvariabelen is beperkt. Mogelijk is dat te wijten aan de steekproef en het soort achtergrondvariabelen dat beschikbaar is. De verschuivingen doen zich ook voor op de onderdelen rekenen, taal en studievaardigheden van de Eindtoets basisonderwijs.

Een vergelijking van de resultaten van de analyses op basis van de ‘univariate modellen’ (paragraaf 6.10.2) en de ‘MV-inhoudsdomein’modellen’ (paragraaf

6.10.4) laat zien dat de correlaties tussen de berekende schooleffecten hoog is (zie tabel 6.16 en figuur 6.13). Het maakt dus in het algemeen geen verschil of de schooleffecten bepaald worden op basis van de 'univariate modellen' of op basis van de 'MV-inhoudsdomein' modellen. Merk op dat voor de individuele school wel een verschil kan optreden en ook optreedt.

Passing van het MUVA-model laat zien dat scholen een heel andere positie in kunnen nemen in de verdeling van scholen bij een 'beginmeting' in vergelijking met een 'eindmeting' (zie figuur 6.13). Opgemerkt moet worden dat het bij de begin- en de eindmeting om andere vaardigheden gaat. De beginmeting had betrekking op de resultaten op een toets Rekenen-Wiskunde voor groep 4 uit het Cito-LVS en de eindmeting betrof de resultaten op de Eindtoets basisonderwijs uitgedrukt in standaardscores.

In het algemeen geldt dat de positie van scholen op basis van hun schooleffecten in de verdeling van scholen omzichtig geïnterpreteerd moet worden. Figuur 6.15 laat zien dat voor vele scholen geldt dat de 95% betrouwbaarheidsintervallen van hun schooleffecten een (grote) overlap vertonen, waardoor de schooleffecten van deze scholen statistisch gezien zich niet van elkaar onderscheiden.

Uit een vergelijking van de analyses voor 2002 en 2003 kan geconcludeerd worden dat het patroon zoals beschreven voor 2003 ook geldt voor 2002. Wel blijkt dat de onderlinge relatie van scholen in 2002 afwijkt van die van 2003. Een school die goed 'presteert' in 2002 doet dat niet per definitie ook in 2003. Dit betekent dat op basis van één onderzoeksjaar geen algemeen geldende conclusies over de prestaties van individuele scholen getrokken kunnen worden. Merk bovendien op dat de inhoud van de Eindtoets basisonderwijs 2002 afwijkt van die van 2003.

6.10.8 Betekenis analyses voor scholen

Voor het beantwoorden van de vraag wat de betekenis van de besproken analyses voor de scholen is, maken we een onderscheid tussen 'accountability' (verantwoording afleggen) en 'school improvement' (schoolverbetering).
De analyses tonen aan dat de resultaten op een beginmeting (in de uitgevoerde analyses waren dat de resultaten op een aantal LVS-toetsen in groep 4) in combinatie met andere kwalitatieve kenmerken van scholen en/of leerlingen (in de uitgevoerde analyses beperkten de kenmerken zich tot het niveau van de leerlingen) leidt tot een nuancering in de beoordeling van de resultaten op een later moment (in de uitgevoerde analyses betrof dat groep 8 van het basisonderwijs). We hebben laten zien dat de positie van een school in de verdeling van scholen kan wijzigen. Voor de ene school zal deze wijziging positief zijn en voor de andere school negatief. Neemt men de effectiviteit van een school als invalshoek dan is een vergelijking tussen scholen na correctie voor een beginmeting en achtergrondkenmerken de meest eerlijke.
Nu scholen weliswaar op een reëlere basis met elkaar vergeleken kunnen worden, zijn concrete aangrijpingspunten voor scholen om tot betere resultaten te komen daarmee nog steeds niet te geven. Daar komt bij dat het vaststellen van de bijdragen van scholen op basis van de resultaten op de Eindtoets basisonderwijs geen effect zal hebben op de leerlingen die in het onderzoek betrokken zijn. Deze leerlingen verlaten immers de school. Bovendien hebben de analyses laten zien dat de resultaten in het ene jaar niet zonder meer gelden voor het andere jaar. Weliswaar laten de uitgevoerde analyse voor 2002 en 2003 over het geheel genomen dezelfde patronen zien, maar dat geldt niet op het niveau van de individuele scholen.
Een bijkomend probleem vormt mogelijk de implicatie van geconstateerde verschillen. Uit oogpunt van efficiëntie en effectiviteit van het onderwijs zouden ´slecht´ presterende scholen daarop aangesproken kunnen worden. Dit zou dan wel moeten gebeuren op basis van in de tijd stabiele resultaten, daar er anders mogelijk ten onrechte gekapitaliseerd wordt op de resultaten in een bepaald jaar. Of het aanspreken zal leiden tot betere resultaten is natuurlijk op voorhand niet te zeggen evenmin als wat dan mogelijke consequenties daarvan zijn. Zo zouden

bijvoorbeeld scholen in de wat minder bevolkte gebieden niet zomaar gesloten kunnen worden vanwege hun streekfunctie. Wel zou in samenspraak met de school gericht gezocht kunnen worden naar verbetermogelijkheden. Ook zou een school 'verplicht' kunnen worden contact op te nemen met vergelijkbare 'goed' presterende scholen, om te komen tot een vorm van 'twinning' of collegiale visitatie waarmee we terecht gekomen zijn op het gebied van schoolverbetering.

Voor schoolverbetering lijken de in het eerste deel van dit hoofdstuk besproken LVS-gegevens beter in aanmerking te komen dan de gegevens op basis van de Eindtoets basisonderwijs. Inhoudelijk beter omdat het LVS over toetsen beschikt die betrekking hebben op meerdere onderdelen van een bepaald vakgebied. Zo leveren deze toetsen betere informatie op over mogelijke aangrijpingspunten. Bovendien verkrijgt men met de LVS-toetsen informatie tijdens de schoolloopbaan van leerlingen in het basisonderwijs en kunnen dezelfde leerlingen (nog) profiteren van eventuele interventies. Daar komt bij dat aan de hand van dezelfde groep leerlingen mogelijk de resultaten van de interventies bepaald kunnen worden. Een probleem dat zich ook bij het LVS voordoet, is de stabiliteit en de (on)nauwkeurigheid van de gegevens. Ook de schoolgrootte kan van invloed zijn op de verkregen resultaten. De resultaten van de kleine scholen zullen meer tenderen naar het algemeen gemiddelde. En de vraag is of de betrouwbaarheidsintervallen van de diverse metingen elkaar niet teveel gaan overlappen. Zie bijvoorbeeld de figuren 6.6 en 6.7 waar in figuur 6.7 ter illustratie betrouwbaarheidsintervallen zijn opgenomen. De resultaten in figuur 6.7 hebben betrekking op een beperkt aantal scholen. Mochten betrouwbaarheidsintervallen elkaar overlappen dan kunnen we statistisch gezien geen onderscheid meer tussen scholen maken.

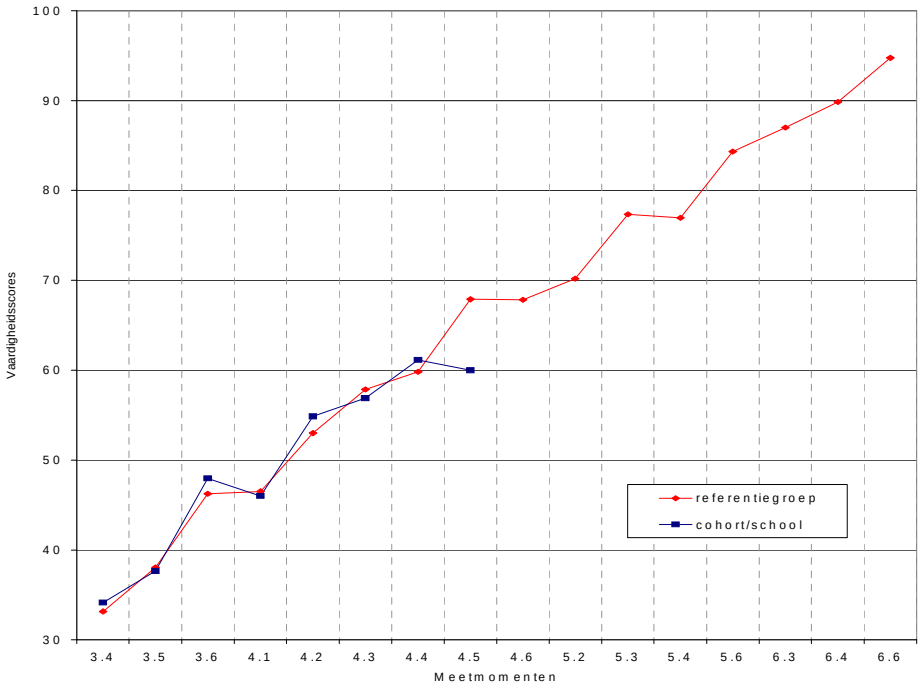
Voorbeeld van een rapportage aan scholen

Op basis van de uitgevoerde analyses volgt hier een voorbeeld van een rapportage aan scholen. Het voorbeeld gaat uit van de analyses zoals besproken in de paragrafen 6.7 en 6.8, waar aandacht besteed is aan het volgen van de ontwikkeling van groepen van leerlingen in de tijd en aan de analyses naar schooleffecten, zoals besproken in paragraaf 6.10. In het proefschrift is in paragraaf 6.4 in de

vorm van beschrijvende statistieken aangegeven hoe scholen geïnformeerd kunnen worden over de populatie van de school en over wijzigingen daarvan. Hoewel informatie over (wijzigingen in) de populatie zinvol voor scholen kan zijn, zal in deze afsluitende paragraaf dit onderdeel niet terugkomen, maar wordt volstaan met te verwijzen naar de besproken tabellen in paragraaf 6.4.

- Het volgen van groepen van leerlingen in de tijd

In paragraaf 6.7 en 6.8 is aan de hand van de data van vier projectscholen aangetoond dat het mogelijk is de resultaten van leerlingen binnen een school in de tijd te volgen. Ook is het mogelijk uitsplitsingen naar achtergrondkenmerken te maken. Figuur 6.16 geeft een fictief voorbeeld van een rapportage die aan scholen aangeboden kan worden. In de figuur staan op de horizontale as de meetmomenten afgebeeld en op de verticale as de vaardigheidsscores. De rode lijn in figuur 6.16 stelt een referentiegroep voor. Dat kunnen diverse groepen zijn. Zo kan gedacht worden aan een referentielijn die aangeeft wat voor de school gebruikelijk is gezien voorgaande jaren. Ook is het mogelijk de resultaten van (eventueel tot een zinvol cluster samengevoegde) andere scholen op te nemen. Wat uiteindelijk als referentie genomen wordt, is aan de school om te bepalen.



Figuur 6.16

Het volgen van de resultaten van een cohort of school in de tijd

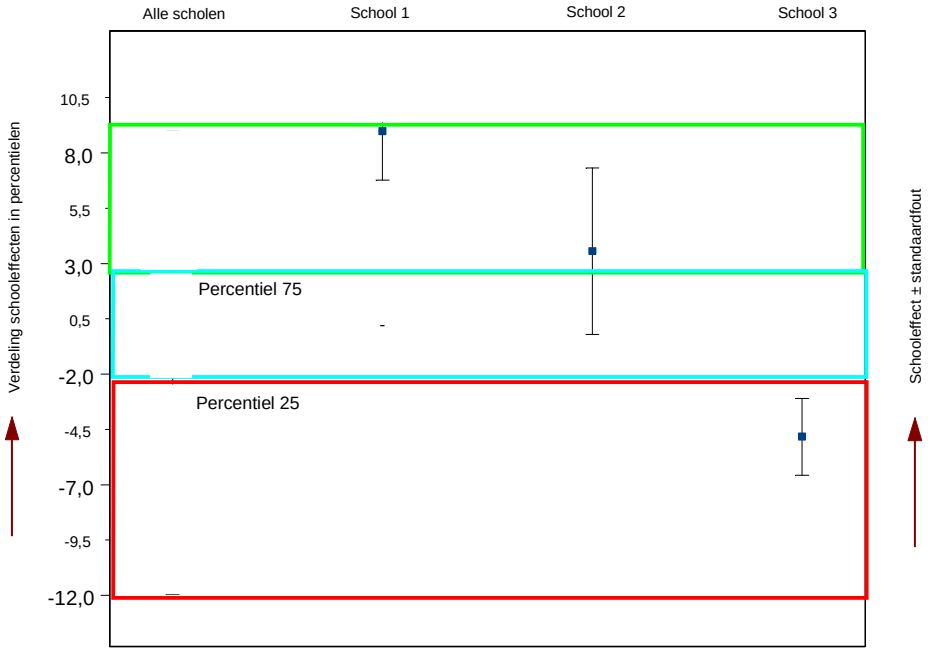
Met behulp van figuur 6.16 krijgt een school op ieder meetmoment informatie over de vorderingen van het cohort of de school als geheel. Na elk meetmoment kan de school aflezen hoe de prestaties van het cohort of de school zich op dat moment verhouden tot die van de referentiegroep. Zodra zich een opvallende wijziging in de ontwikkeling voordoet, kan een school nadere analyses uitvoeren. Als voorbeeld is in figuur 6.16 het meetmoment 4.5 opgenomen. Duidelijk is te zien dat de resultaten van het cohort of de school achterblijven in vergelijking met de referentiegroep. Op basis van deze rapportage kan een school direct actie ondernemen. Een school kan door het uitsplitsen van de resultaten naar achtergrondkenmerken van leerlingen onderzoeken of de afwijking toegeschreven kan worden aan bepaalde groepen. Mogelijk is de groepssamenstelling gewijzigd of ligt ´de oorzaak´ in het onderwijsaanbod zelf. Om te zien of de afwijking verband houdt met de desbetreffende groep, kan een school nagaan of

het geconstateerde beeld zich ook bij andere leergebieden voordoet. Na het vaststellen of en welke acties gewenst zijn, geeft deze vorm van rapporteren de school de mogelijkheid op een volgend moment na te gaan wat het mogelijk effect ervan is.
Omdat we ook bij deze vorm van rapportage te maken hebben met betrouwbaarheidsintervallen zou nog onderzocht kunnen worden bij welke terugval of stijging in resultaten er sprake is van een signaal dat om aandacht vraagt. Dat is in het kader van dit proefschrift niet gedaan. Aangezien het in het onderhavige voorbeeld gaat om een signaalfunctie en er zeker geen sprake is van modeltoetsing, zou overwogen kunnen worden een afwijking van één standaardfout als uitgangspunt te nemen.

- Het vaststellen van het schooleffect

Bij het vaststellen van het schooleffect gaat het meer dan bij het volgen van de ontwikkeling van groepen van leerlingen in de tijd om de vraag ‘hoe doe ik het als school’. De vraag ‘levert mijn school een voldoende bijdrage aan de ontwikkeling van de leerlingen’ is veel meer een verantwoordingsvraag dan de vraag hoe leerlingen zich binnen de school ontwikkelen. Dat laatste heeft primair tot doel het onderwijs zo optimaal mogelijk lopende het schooljaar aan te passen aan de leerlingen. Bij het vaststellen van het schooleffect is de informatie die een school krijgt meer algemeen van aard, waarbij andere scholen als referentie dienen. Natuurlijk kan de rapportage over het schooleffect voor een school ook aanleiding zijn om het onderwijs eens nader onder de loep te nemen.

Een voorbeeld van een rapportage over het schooleffect staat in figuur 6.17. In deze figuur zijn de drie scholen waarvan in figuur 6.15 de betrouwbaarheidsintervallen besproken zijn, als voorbeeld genomen. Wanneer in figuur 6.17 gesproken wordt over ‘verdeling van schooleffecten in percentielen’, dan heeft dat betrekking op de 68 scholen die deel uitmaken van de WOB-dataset.



Figuur 6.17
Voorbeeldrapportage schooleffect

In figuur 6.17 staat aan de linkerkant de verdeling van de schooleffecten van de 68 scholen afgebeeld. In de grafiek is van de verdeling van de scholen het 75e en het 25e percentiel afgebeeld. Aan de hand van deze grafiek kan een school vaststellen wat haar positie is in de verdeling van scholen. Aan de rechterkant van figuur 6.17 staan de geschatte schooleffecten van alle individuele scholen afgebeeld, inclusief de 95% betrouwbaarheidsintervallen. Op basis van deze informatie kan een school nagaan of haar schooleffect significant op het 5%-niveau verschilt van dat van een andere school[9] . Zo is in figuur 6.17 te zien dat het schooleffect van school 1 niet significant afwijkt van het schooleffect van school 2, maar wel van het schooleffect van school 3. Ook het schooleffect van

[9] Merk op dat het betrouwbaarheidsinterval voor school 1 slechts gedeeltelijk weergegeven is. De bovengrens van de betrouwbaarheidsintervallen wordt bepaald door het hoogst behaalde schooleffect. Hetzelfde geldt voor de ondergrens. Deze wordt bepaald door het laagst behaalde schooleffect.

school 2 wijkt significant af van dat van school 3. Door nu de beide weergaven met elkaar te combineren, kan een school ook vaststellen waar zij staat in de verdeling van schooleffecten van alle betrokken scholen. Op basis van een vergelijking van de schooleffecten blijkt school 1 tot de betere scholen te behoren en school 3 tot de mindere. Door nu in de verdeling drie gebieden te onderscheiden en deze met een kleur aan te geven, kunnen scholen op een eenvoudige wijze te weten komen waar zij staan in de verdeling. In figuur 6.17 beperkt de indeling zich tot drie gebieden. Het spreekt voor zich dat elke andere indeling ook mogelijk is.
In figuur 6.17 is het gebied vanaf het 75e percentiel groen omrand. Het gebied tussen het 75e en het 25e percentiel is blauw omrand en het onderste gebied rood. Deze driedeling en kleurcodering zou tot de volgende rapportage kunnen leiden.

School 1: groen
School 2: groen - blauw
School 3: rood

De volgorde van de codering groen - blauw geeft aan dat het geschatte schooleffect zich in de groene zone bevindt, maar dat rekening houdend met de onzekerheid van de schatting ook de blauwe zone tot de mogelijkheid behoort. In feite geeft dus de eerste kleur aan waar het geschatte schooleffect zich bevindt, terwijl de tweede of eventueel derde kleur aangeeft tot hoever het 95% betrouwbaarheidsinterval zich uitstrekt. Indien slechts één kleur vermeld staat, dan betekent dat dat het schooleffect, ook indien rekening gehouden wordt met het 95% betrouwbaarheidsinterval, zich in het desbetreffende gebied bevindt. Het spreekt voor zich dat ook andere kleurcombinaties mogelijk zijn, elk met hun eigen betekenis. Zo geeft de combinatie rood - blauw aan dat de school met dit geschatte schooleffect tot de groep '=< percentiel 25' behoort, maar dat gezien het betrouwbaarheidsinterval ook 'percentiel 25 < groep < percentiel 75' mogelijk is.

Uit het onderzoek blijkt dat voor een school de resultaten van het ene jaar niet maatgevend behoeven te zijn voor het andere jaar (zie de bespreking van figuur 6.9). Wel zou een school op analoge wijze als in figuur 6.17 jaarlijks

geïnformeerd kunnen worden over haar positie in de verdeling van scholen. Een school kan dan vaststellen of zij over de jaren heen eenzelfde positie inneemt. Met name wisselende en lage posities kunnen voor een school een signaal zijn om nadere analyses uit te voeren.

7 Schoolzelfevaluatie in de toekomst

EVADOS is ontwikkeld om de vraag 'Hoe doe ik het als school?' te beantwoorden. Met EVADOS moeten scholen in staat zijn conclusies te trekken over de mate waarin men tevreden kan zijn met het effect van het geboden onderwijs. Op deze manier levert EVADOS een bijdrage aan de kwaliteit van het onderwijs. Bij de ontwikkeling van EVADOS is een zestal uitgangspunten geformuleerd:

1 leerresultaten zijn een indicatie voor de kwaliteit van het onderwijs;
2 leerresultaten dienen in de tijd gevolgd te worden;
3 leerresultaten dienen gerelateerd te (kunnen) worden aan input- en proceskenmerken;
4 het gebruik van EVADOS mag geen extra belasting voor scholen betekenen;
5 het gebruik van EVADOS dient te leiden tot verantwoorde uitspraken;
6 EVADOS dient scholen (of andere aggregatieniveaus) van zowel een intern als een extern referentiekader te voorzien.

In dit hoofdstuk wordt eerst nagegaan in hoeverre EVADOS recht doet aan de uitgangspunten. Uit de bespreking zal blijken dat nog een aantal punten waaronder het verzamelen van data om aandacht vraagt. In het vervolg van dit hoofdstuk wordt Data Warehousing als een mogelijke oplossing voor deze problematiek aangedragen. Ook zullen nieuwe ontwikkelingen die ook een positieve bijdrage kunnen hebben aan schoolzelfevaluatie aan bod komen. Tot slot wordt het Expertisecentrum Kwaliteitszorg geïntroduceerd. We laten zien dat een dergelijk centrum positief kan bijdragen aan de kwaliteitszorg van het onderwijs.

Leerresultaten als indicatie voor de kwaliteit van het onderwijs

EVADOS is bedoeld als een procedure die informatie aanlevert over de kwaliteit van het onderwijs. Voor een omschrijving van het begrip kwaliteit is in dit proefschrift (zie hoofdstuk 2) aangesloten bij de opvatting van Scheerens die

kwaliteit opvat als het geheel van wenselijke hoedanigheden, zoals deugdelijkheid, gezondheid of effectiviteit van organisaties (Scheerens, 1996a). Kwaliteitszorg omschrijft hij als 'de actieve gerichtheid, zo men wil het 'beleid' van de organisatie om die wenselijke hoedanigheden ook inderdaad te manifesteren'. In zijn optiek vraagt kwaliteitszorg om een actieve benadering, waarin 'kwaliteit' als een probleem wordt gezien en er actief gestreefd wordt naar registratie, handhaving of liefst verbetering van kwaliteit. Scheerens merkt op dat het belangrijk is te operationaliseren wat als indicator voor de kwaliteit wordt aangemerkt. In dit proefschrift is ervoor gekozen de leerresultaten van leerlingen op toetsen als indicatie te zien voor de kwaliteit van het onderwijs. Het meten van leerresultaten kan op een valide en betrouwbare wijze plaatsvinden, waardoor een goede basis aanwezig is om veranderingen in het onderwijs te signaleren. Tevens zijn voor het (basis)onderwijs kerndoelen geformuleerd die als een objectief criterium kunnen gelden waartegen de leerresultaten afgezet kunnen worden. En voor zover geen objectief criterium aanwezig is, kunnen referentiegegevens verzameld worden. Met deze referentiegegevens kan een school vaststellen in hoeverre haar resultaten afwijken van andere (vergelijkbare) scholen. De keuze voor leerresultaten als indicator voor de kwaliteit van het onderwijs impliceert niet dat input- en procesfactoren geen invloed (kunnen) hebben op de kwaliteit van onderwijs. Beide factoren dienen ook betrokken te worden bij het uitspreken van een oordeel over de kwaliteit. Hoewel vele procedures voor schoolzelfevaluatie (vooral) procesfactoren als evaluatiedoel zien, worden in dit proefschrift zowel input- als procesfactoren gezien als mogelijke verklarende factoren voor behaalde leerprestaties. Leerresultaten hebben een signaalfunctie om na te gaan of - gegeven input en proces - de bereikte resultaten als bevredigend gezien mogen worden. Afwijkende resultaten, zowel in positieve als in negatieve zin, kunnen een aanzet zijn voor het nader bestuderen van input- en procesfactoren met een daartoe geschikt instrumentarium. Voor zover van input- en procesfactoren bekend is dat deze van invloed (kunnen) zijn op de leerresultaten, kunnen deze in de analyse van deze resultaten betrokken worden. Het onderwerp Toets Curriculum Overlap zoals nader uitgewerkt in hoofdstuk vier van dit proefschrift is daar een voorbeeld van.

Leerresultaten dienen in de tijd gevolgd te worden

Voor EVADOS geldt dat leerresultaten van leerlingen als indicatie gezien worden voor de kwaliteit van het onderwijs. Kwaliteitszorg werd omschreven als een actieve gerichtheid (het beleid) om die kwaliteit te manifesteren. Voor een school betekent dit dat zij moet nagaan welke leerresultaten zij als uitgangspunt gaat nemen voor het nemen van beleidsbeslissingen. Het zal duidelijk zijn dat een éénmalige meting niet zal volstaan. Door omstandigheden hoeft een enkele meting niet representatief te zijn voor dat wat voor de school gebruikelijk is. Te denken valt daarbij aan een afwijkende samenstelling in populatie en het aantal leerlingen dat aan een toets deelneemt. In het geval dat het een school met een gering aantal leerlingen betreft, is de bijdrage van iedere afzonderlijke leerling aan de schoolprestatie relatief groot. Een goede leerling trekt het school-gemiddelde onevenredig veel omhoog en een minder goede leerling omlaag. Om deze reden is het niet wenselijk beleid op éénmalige metingen te baseren. Het volgen van de resultaten van leerlingen in de tijd geeft een stabielere indicatie van de kwaliteit van het geboden onderwijs. Dat wat voor een school gebruikelijk is, wordt dan zichtbaar(der). Uitschieters vallen op en kunnen mogelijk verklaard worden.

De keuze om de resultaten van leerlingen in de tijd te volgen, brengt wel een aantal andere keuzes met zich mee ten aanzien van:

1 Het te gebruiken toetsinstrumentarium

   De resultaten op de te gebruiken toetsen dienen met elkaar vergeleken te kunnen worden. Niet iedere toets biedt deze mogelijkheid. Op de een of andere manier zal een transformatie van de resultaten op verschillende toetsen naar eenzelfde vaardigheidsschaal of scoreschaal mogelijk moeten zijn. Alleen dan is een vergelijking te maken.

2 De frequentie van toetsafname

Als aan punt 1 voldaan is, blijft de vraag nog open hoe vaak een toets afgenomen dient te worden? Kan volstaan worden met één meting per jaar of zijn meerdere metingen noodzakelijk? Een voorbeeld van een toets die één keer per jaar afgenomen wordt, is de Eindtoets basisonderwijs. Op basis van de resultaten van deze toets ontvangen de deelnemende scholen een schoolrapport, waarin zij kunnen aflezen wat de gemiddelde schaalscore voor dat jaar is. Deze schaalscore kunnen zij vergelijken met de behaalde scores in voorgaande jaren en met de scores van een landelijk referentiekader. De Eindtoets basisonderwijs heeft echter als nadeel dat de leerlingen zelf niet meer kunnen profiteren van eventuele beleidswijzigingen naar aanleiding van de resultaten. Zij verlaten immers de school. Een ander nadeel is dat met deze toets alleen informatie verzameld wordt aan het einde van de basisschoolperiode. Hoe het effect van een beleidsmaatregel doorwerkt in voorafgaande leerjaren (groepen) is dan niet vast te stellen.

Een andere toets waarvan de resultaten door toepassing van itemresponstheorie getransformeerd kunnen worden naar eenzelfde vaardigheidsschaal is de Entreetoets. Een voordeel van de Entreetoets ten opzichte van de Eindtoets basisonderwijs is de frequentie van afname. Kent de Eindtoets basisonderwijs slechts één afnamemoment aan het einde van het basisonderwijs, de Entreetoets kent meerdere afnamemomenten die bovendien plaatsvinden in meerdere groepen. Hierdoor krijgt een school de mogelijkheid de ontwikkeling van het onderwijs in de tijd te volgen en wel tijdens de schoolloopbaan van leerlingen. Leerlingen kunnen dan mogelijk nog profiteren van eventuele beleidswijzigingen.

Toetsen die het ook mogelijk maken om de ontwikkeling van leerlingen in de tijd te volgen, zijn de in hoofdstuk vier besproken toetsen van het Cito-LVS. Met deze toetsen is het mogelijk leerlingen in de tijd te volgen gedurende hun gehele verblijf op de basisschool. Het grote voordeel hiervan is dat de op school aanwezige populatie leerlingen optimaal kan profiteren van doorgevoerde beleidswijzigingen naar aanleiding van geconstateerde ontwikkelingen.

3 Volgen van dezelfde leerlingen in de tijd

In het voorgaande werd kort ingegaan op het volgen van leerlingen in de tijd. Opgemerkt werd dat de populaties van scholen jaarlijks wisselen, deels omdat ieder jaar leerlingen na groep acht de school verlaten en ook ieder jaar nieuwe leerlingen hun onderwijsloopbaan beginnen in groep één, en deels omdat leerlingen tussentijds instromen en tussentijds de school verlaten. Dit betekent dat een school bij het vergelijken van resultaten van groepen van leerlingen in de tijd, steeds na zal moeten nagaan in hoeverre de groepen qua samenstelling nog vergelijkbaar zijn en in hoeverre eventuele verschillen toegeschreven kunnen worden aan de wisselende populaties. Om verschillen niet toe te hoeven te schrijven aan verschillen in populaties, zou ervoor gekozen kunnen worden de resultaten van dezelfde leerlingen in de tijd te volgen. Voorwaarde is wel dat leerlingen dan uniek te identificeren zijn. Invoering van het onderwijsnummer voor het basisonderwijs maakt dit in principe mogelijk.

4 Centraal of decentraal verwerken van leerresultaten

Een volgende keuze betreft de verwerking van de leerresultaten. Als de leerresultaten centraal verwerkt worden, heeft dat als voordeel dat de data op één plaats aanwezig en daar ook beschikbaar zijn. Zowel de Eindtoets basisonderwijs als de Entreetoets kennen een centrale verwerking van de leerlingantwoorden. De resultaten op de LVS-toetsen echter worden door de scholen zelf verwerkt. Voor zover men geïnteresseerd is in (landelijke) referentiegegevens, zullen de gegevens ter verwerking centraal verzameld moeten worden. Het spreekt voor zich dat daarvoor een systeem ontwikkeld dient te worden. Hierbij zou gedacht kunnen worden aan e-mail of een webbased-applicatie.

Het probleem wat betreft dataverzameling dat zich voordoet bij afnamen met papieren toetsen komt te vervallen zodra gebruik gemaakt gaat worden van beeldschermtoetsen. Door leerlingen toetsen achter het beeldscherm te laten maken, zijn de antwoorden automatisch in elektronische vorm beschikbaar, waardoor gemakkelijker centraal over de data beschikt kan worden. Ont-

wikkelingen in deze richting bij het Cito-LVS zijn gaande. Beeldscherm-toetsen komen later in dit hoofdstuk aan bod.

5 Afname op vaste of op variabele tijdstippen

Onderzoek bij een aantal scholen heeft laten zien dat scholen de toetsen uit het Cito-LVS afnemen op momenten die niet overeenkomen met de in de handleiding aanbevolen afnamemomenten. Leerkrachten hebben daar ongetwijfeld goede redenen voor. Zo kan het zijn dat een leerkracht van mening is dat een bepaalde groep leerlingen nog niet toe is aan de toets en er voor kiest de toets op een ander moment af te nemen. Voor zover referentie-gegevens gebonden zijn aan afnamemomenten, kan dit een probleem zijn omdat een goede vergelijking dan niet meer mogelijk is. Ook in vergelijking met andere scholen doet zich dan een probleem voor. In dit kader kan verwezen worden naar figuur 6.1. Omdat de bij het project betrokken vier scholen op vele verschillende tijdstippen de toetsen afnamen bij hun leer-lingen, ontstond een onvolledige dataset. Voor een individuele leerkracht kan het gegeven zijn onderwijs zinvol zijn toetsen op meerdere of op andere dan in de handleiding aangegeven tijdstippen af te nemen. Het ingeschatte niveau van zijn leerlingen kan daarvoor bepalend zijn. Het gaat immers om zicht te krijgen op de ontwikkeling van zijn leerlingen. Als echter het doel (ook) is om de resultaten van zijn leerlingen te kunnen vergelijken met de resultaten van andere leerlingen, dan is het wellicht wenselijk om te kiezen voor vaste afnamemomenten. Bovendien kunnen referentiegegevens op die tijdstippen verzameld worden. Als scholen de resultaten van groepen van leerlingen gebruiken voor het maken van beleidskeuzes dan is er wellicht ook niets op tegen om de gegevens slechts enkele malen (of zelfs één maal per jaar) te verzamelen. Het effect van beleid is niet altijd direct zichtbaar en vraagt vaak wat meer tijd. Het ‘verplichten’ van scholen de toetsen (ook) op vooraf aangegeven tijdstippen af te nemen, geeft de mogelijkheid om de resultaten met elkaar te vergelijken. En, hoewel voor de dagelijkse onderwijspraktijk wellicht minder relevant, zijn vanuit het oogpunt van de onderzoeker ook goede redenen te noemen, waarom het juist wenselijk is om toetsen op vaste tijdstippen af te nemen. Belangrijk is dus te kijken naar

potentiële gebruikers en hun wensen. In het vervolg van dit hoofdstuk zal op deze potentiële gebruikers nog ingegaan worden.

Leerresultaten relateren aan input- en proceskenmerken

Het belang van input- en proceskenmerken bij de interpretatie van resultaten van leerlingen is in het proefschrift al meerdere malen onderstreept. De resultaten op zich zijn een ontoereikende indicatie voor de kwaliteit van het geboden onderwijs. Steeds zullen de omstandigheden waaronder het onderwijs heeft plaatsgevonden in het oordeel betrokken dienen te worden. EVADOS biedt scholen die mogelijkheid om bij de interpretatie van de opbrengsten met deze omstandigheden rekening te houden. Het spreekt voor zich dat deze dan wel beschikbaar dienen te zijn. Schooladministratiepakketten bevatten vooral inputgegevens: informatie over leerlingen, leerkrachten en de school. Informatie over de in hoofdstuk vier besproken procesfactoren is niet in de pakketten aanwezig. Informatie daarover dient apart verzameld te worden. Het besproken instrument Toets Curriculum Overlap is daar een voorbeeld van.

Bij de interpretatie van proceskenmerken dienen de door scholen te manipuleren variabelen niet meegenomen te worden. Correctie dient alleen plaats te vinden op door scholen niet te manipuleren variabelen. Correctie voor manipuleerbare variabelen beloont een school die slecht 'scoort' op deze variabelen. Een school A die zorgdraagt voor een slecht schoolklimaat krijgt een relatief grotere correctie op de bereikte resultaten dan een school B met een beter schoolklimaat, waardoor ogenschijnlijk de prestaties van school A beter lijken. En die resultaten zijn nu mogelijk zo laag vanwege dat slechte schoolklimaat.

Het gebruik van EVADOS mag geen belasting voor scholen zijn

EVADOS is in eerste instantie ontwikkeld voor de scholen. Wil EVADOS gebruikt worden dan moet het voor scholen ten eerste zinvolle informatie opleveren en mag ten tweede het gebruik geen (grote) belasting voor scholen zijn. Om die reden is in het onderzoek gebruik gemaakt van informatie die scholen toch al opslaan in schooladministratie- en toetsregistratiepakketten. Het onderzoek heeft aangetoond dat de informatie in deze pakketten bruikbaar is

voor schoolzelfevaluatie, maar dat de informatie niet zonder meer gebruikt kan worden. Zowel de integriteit als de volledigheid van de data behoeven aandacht. Informatie van leerlingen die de school verlaten wordt niet bewaard en de pakketten laten de gebruikers zoveel ruimte toe dat dat ten koste kan gaan van de volledigheid van de data of de consistentie waarmee de data opgeslagen worden. Geconcludeerd is dat een andere wijze van dataopslag de voorkeur verdient, waarbij overeind blijft dat deze opslag niet belastend voor een school mag zijn. In het vervolg van dit hoofdstuk komt dit onderwerp nader aan de orde.

Uitspraken dienen verantwoord te zijn

(Beleids)beslissingen mogen niet gebaseerd zijn op incidenten. Waar mogelijk dient de informatie betrekking te hebben op meerdere jaren. Maar ook in rapportages dient duidelijk tot uitdrukking te worden gebracht op welke jaren deze precies betrekking hebben, hoe groot de aantallen leerlingen zijn waarop de uitspraak betrekking heeft en hoe betrouwbaar de resultaten zijn. In dat verband is het rapporteren van betrouwbaarheidsintervallen gewenst. In hoofdstuk zes is daar aandacht aan besteed.

Het verstrekken van zowel een intern als een extern referentiekader

Door te rapporteren aan scholen over hoe zij het in een bepaald jaar doen, weten de scholen nog niet hoe goed zij het doen. Om die vraag te kunnen beantwoorden zijn referentiegegevens nodig. Deze kunnen gevormd worden door de resultaten van de school te vergelijken met de behaalde resultaten in voorgaande jaren (intern referentiekader) of door de resultaten van de school te vergelijken met de resultaten van andere scholen (extern referentiekader). EVADOS biedt beide mogelijkheden. Hoe groot het extern referentiekader is, wordt bepaald door de mate waarin gegevens van andere scholen beschikbaar zijn.

7.1 Het Data Warehouse

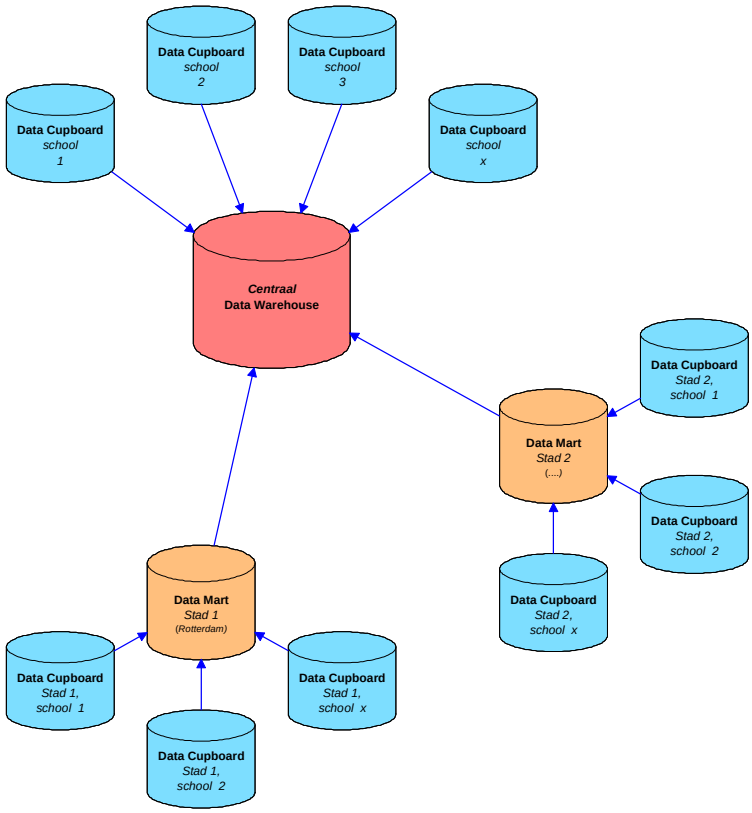
Uit de bespreking tot nu toe kan geconcludeerd worden dat in de uitwerking van EVADOS recht gedaan is aan de uitgangspunten. Wel is bij het onderdeel 'volgen van de leerresultaten in de tijd' een vijftal aandachtspunten geformuleerd. Deze staan echter een verdere uitwerking of toepassing van EVADOS niet in de weg. Een punt van aandacht echter is het komen tot een adequate dataverzameling. Hoewel de verwachting was dat voor schoolzelfevaluatie uitgegaan zou kunnen worden van de op scholen aanwezige schooladministratie- en toetsregistratiepakketten, blijkt in de praktijk dat de op scholen aanwezige informatie toch niet zonder meer geschikt is voor schoolzelfevaluatie. Naast het probleem van de integriteit van de data is ook het ontbreken van historische gegevens over leerlingen in de diverse pakketten voor schoolzelfevaluatie een groot gemis. De gegevens zoals deze opgeslagen zijn, zijn niet verzameld voor het vaststellen van schooleffecten. Het werken met de gegevens uit schooladministratiepakketten zoals uitgevoerd in het kader van dit proefschrift was echter erg informatief. Duidelijk is geworden dat de functionaliteit van deze pakketten en de wijze waarop de data in de pakketten worden opgeslagen, niet aansluiten bij beleidsgerichte analyses zoals EVADOS dat voorstaat. Een andere wijze van dataopslag is dan ook gewenst. Wij stellen voor om bij de opslag van data gebruik te maken van methoden en technieken die aangeduid worden met Data Warehousing.

Een Data Warehouse (DHW) kan letterlijk opgevat worden als een 'gegevenspakhuis', dat wil zeggen een database waarin relevante informatie op een centrale plaats is opgeslagen. Als gegevenspakhuis is het DWH een uitstekende voorziening om grote hoeveelheden gegevens op te slaan en ter beschikking te stellen voor analyse en rapportage. De gegevens in een DWH kunnen daartoe uit één of meer bronsystemen komen. Hierbij is het niet noodzakelijk dat deze bronsystemen fysiek op dezelfde locatie staan. Aan een DWH worden op vaste tijdstippen data toegevoegd, waarbij de in het DWH al aanwezige gegevens niet gemuteerd worden. Op deze manier wordt een historie op-

gebouwd die de gebruiker in staat stelt over een bepaald onderwerp in de tijd te rapporteren.

Een speciale vorm van een DWH is een Data Mart (DM). Een DM is gerelateerd aan een specifieke groep van gebruikers die allen dezelfde specifieke set van gegevens gebruiken of tot hetzelfde 'business proces' behoren. Voor school-zelfevaluatie zijn dit in eerste instantie de scholen verenigd in bijvoorbeeld een samenwerkingsverband of in een gemeente, die hun eigen specifieke informatie-behoeften hebben met hun eigen processen en procedures. DM's kunnen op het eigen niveau een bepaald aspect van locale informatievoorziening verzorgen en kunnen bovendien doorverbonden zijn met een centraal DWH. Indien een dergelijk situatie zich voordoet, ontstaat een combinatiestructuur waarbij alle betrokkenen (scholen en/of samenwerkingsverbanden/gemeenten) gebruik kunnen maken van alle aanwezige gegevens. In de literatuur spreekt men dan van een 'multi-tiered data warehouse', waarvan in figuur 7.1 een schematisch weergave staat.

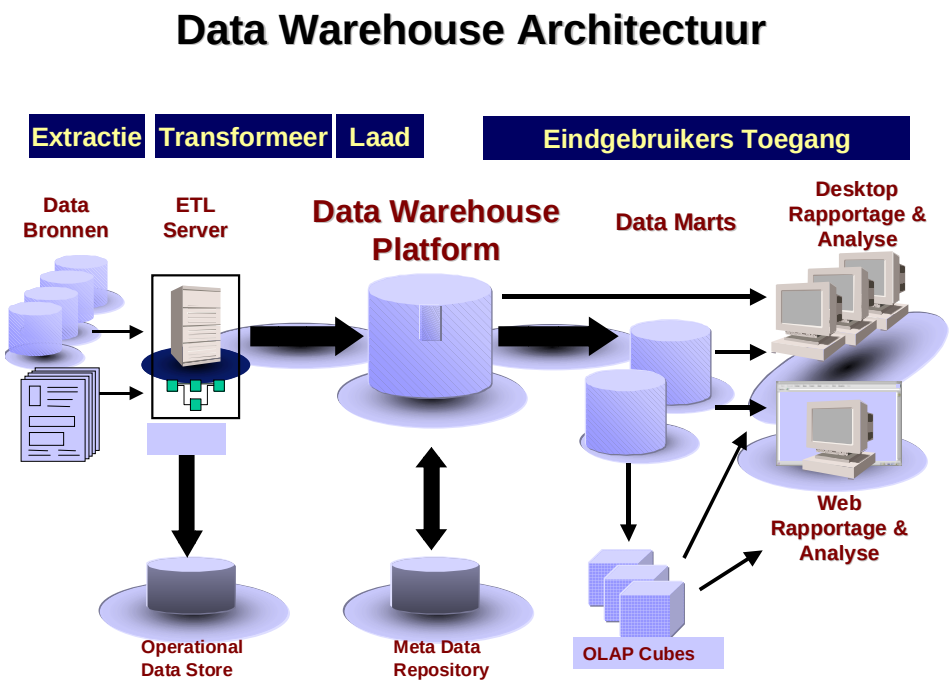
De term 'Data Cupboard' uit figuur 7.1 verwijst naar een situatie dat een individuele school rechtstreeks aangesloten is op het DWH en niet via een tussenlaag (bijvoorbeeld een gemeente) zoals dat bij een DM wel het geval is. Vertaald naar het onderwijs zouden de twee getekende DM's in figuur 7.1 gemeenten of samenwerkingsverbanden kunnen voorstellen. Deze DM's betrekken de informatie wel van het centrale DWH, maar rapporteren specifiek en alleen aan die organisatorische eenheden die deel uitmaken van het betreffende DM.



Figuur 7.1
Multi-tiered Data Warehouse: Centraal DWH met Lokale Data Marts (bijvoorbeeld steden) en Data Cupboards (individuele scholen)

Een schematische weergave van een DWH staat weergegeven in figuur 7.2. Belangrijke elementen in een DWH zijn de databronnen waar de vereiste informatie in opgeslagen is en de ETL server. Deze server zorgt voor het Extraheren, het Transformeren en het Laden van de gewenste informatie uit de databronnen in het DWH. Tijdens de transformatie wordt ervoor gezorgd dat de gegevens in de juiste samenstelling en hoedanigheid worden geprepareerd alvorens deze worden geladen, waarbij de gegevens tevens dienen te voldoen aan vooraf vastgestelde kwaliteitseisen. Het DWH in figuur 7.2 verstrekt informatie aan een aantal DM's, dat gericht is op groepen van gebruikers met specifieke wensen.

Figuur 7.1 laat zien dat scholen (als data cupboards) ook rechtstreeks aangesloten kunnen zijn op een DWH. Volgens bepaalde procedures en voorzieningen vinden op basis van de gegevens in het DM rapportages naar gebruikers plaats. Dit kunnen zowel standaard-, als op speciaal verzoek samengestelde rapporten zijn en ze kunnen schriftelijk of via het internet verspreid worden.



Figuur 7.2

Data Warehouse Architectuur (Bron: Van de Zand, 1999)

7.2 De meerwaarde van een DWH voor EVADOS

De op scholen aanwezige schooladministratie- en toetsregistratiepakketten voldoen niet aan de eisen die gesteld kunnen worden aan een voor EVADOS gewenst informatiesysteem. De pakketten kunnen gekenschetst worden als operationele systemen die ontwikkeld zijn ten behoeve van het uitvoeren van transacties op individuele scholen (administratie). Voor schoolzelfevaluatie is echter een systeem gewenst dat het onderwijsbeleid op een school ondersteunt en evalueert, wat omschreven kan worden als een beleidsinformatiesysteem. Beide soorten systemen stellen aparte eisen aan de opzet ervan, én aan de datastructuur. Voor operationele systemen geldt dat zij:

- zijn gericht op het verwerken van grote aantalen eenvoudige transacties, met name de invoer van gegevens en het uitvoeren (verwerken) van administratieve handelingen;
- veel gedetailleerde informatie bevatten;
- gegevens bevatten die continu aan verandering onderhevig zijn (mutaties zijn steeds mogelijk en worden in de regel ook vaak toegepast);
- niet gericht zijn op de integratie van gegevens uit verschillende operationele systemen of bronnen;
- niet of nauwelijks historische gegevens verzamelen, dat wil zeggen dat zij bij mutaties van kenmerken geen - of gedurende slechts een beperkte periode - historische gegevens bijhouden;
- zeer performance-gevoelig zijn, dat wil zeggen dat ze in staat moet zijn om - gegeven de beschikbare gegevens - op vraag allerlei informatie snel op te leveren;
- een statische structuur hebben met een variabele inhoud.

De ervaringen met de operationele systemen zoals onderzocht en gebruikt in het ZEBO-project laten zien dat deze operationele systemen adequaat functioneren gegeven het doel waarvoor ze ontwikkeld zijn, maar dat voor EVADOS deze systemen:

- Teveel vrijheid geven voor de mutatie van gegevens

Gegevens over leerlingen, leerkrachten en de school kunnen in de tijd variëren en blijken dat in de praktijk ook te doen. De geanalyseerde school- en administratiepakketten kennen voor schoolzelfevaluatie geen adequate voorziening om hiermee om te gaan. Voor schoolzelfevaluatie echter is het erg belangrijk rekening te (kunnen) houden met gegevens die in tijd kunnen variëren.

- Historische gegevens onvoldoende bijhouden
  De voor een school relevante gegevens over leerlingen zijn in de school-administratie- en toetsregistratiepakketten opgeslagen, zolang de leerling zich op school bevindt. Zodra een leerling de school verlaat, worden de gegevens weggeschreven naar een historisch bestand. Dit bestand bevat echter niet alle in de loop der tijd verzamelde gegevens van de leerlingen. Een aantal voor schoolzelfevaluatie belangrijke gegevens (zoals toets-resultaten) gaat daarbij verloren;
- De gegevens onvoldoende controleren of valideren
  In geval er vergissingen of fouten gemaakt worden bij het invoeren van gegevens, laten de systemen dit toe. Interne controles worden niet uit-gevoerd;
- onvoldoende mogelijkheden bieden tot integratie van de gegevens
  De gegevens van de leerlingen in de pakketten zijn niet opgehangen aan een unieke leerlingcode. Slechts door het uitvoeren van arbeidsintensieve hande-lingen is het mogelijk gegevens uit de diverse pakketten aan elkaar te koppelen.

Voor EVADOS is een systeem gewenst dat inspeelt op de genoemde problemen met de op scholen aanwezige operationele systemen. Bovendien dient het systeem in staat te zijn over een aantal onderwerpen informatie te verstrekken op basis waarvan de gebruiker beleidsbeslissingen kan nemen. Ook moet het systeem historische gegevens bevatten die desgewenst geaggregeerd kunnen worden. Tot slot dient het systeem niet primair gericht te zijn op het verrichten van administratieve handelingen (transacties), maar dient het systeem het beleid op een school of voor een groep scholen te ondersteunen en te evalueren. Dergelijke systemen worden omschreven als beleidsinformatiesystemen. Het

DWH is zo'n beleidsinformatiesysteem. Met het DWH kunnen volgens vaststaande procedures en voorzieningen rapporten worden samengesteld die aansluiten bij beleidsvragen. Ook bestaat de mogelijkheid gebruik te maken van 'data mining' en 'knowledge discovery', technieken die schijnbaar niet bestaande relaties tussen verschillende gegevenselementen in het DWH of DM zichtbaar kunnen maken. De rapportage van de gegevens kan schriftelijk zijn of via het internet verspreid worden. In paragraaf 7.4 worden enkele voorbeelden van rapportages gepresenteerd.

In algemene zin zijn er twee manieren om de ontwikkeling en implementatie van een DWH te benaderen. Dit zijn de 'coporate' data warehouse benadering en de 'incrementele' benadering. De coporate data warehouse benadering gaat uit van een alomvattend data warehouse dat in een keer wordt ontwikkeld en geïmplementeerd. Een zeer arbeidsintensieve en een niet erg flexibele benadering als er zich veranderingen in de organisatie voordoen. De incrementele benadering daarentegen gaat uit van de snelle ontwikkeling van een datamart met een beperkte set van gegevens. Voor schoolzelfevaluatie leent zich de incrementele benadering het beste. De opzet van een DM zou zich in eerste instantie kunnen richten op het primair onderwijs, om vervolgens op een later tijdstip uitgebreid te worden naar andere sectoren van het onderwijs. Zie voor meer informatie over beide benaderingen Inmonn (1996) en Kimball (1996).

Met een technische voorziening als een DWH kunnen op scholen aanwezige gegevens in diverse administratieve pakketten beter vertaald worden naar informatie ten behoeve van beleidsbeslissingen. Gegeven de huidige situatie in het Nederlands onderwijs waar bij de meeste toetsen de afname nog via het papier plaatsvindt, blijft het laden van de gegevens uit de schooladministratiepakketten naar het DWH organisatorisch een punt van aandacht. Via de exportfunctie van de diverse pakketten zullen de gewenste gegevens aangeleverd moeten worden. Nu echter de computer steeds meer ingezet wordt ter ondersteuning van het onderwijsleerproces, dienen zich nieuwe mogelijkheden aan. Zo krijgen leerlingen steeds meer mogelijkheden om aan het beeldscherm toetsen te maken, wat als groot voordeel heeft dat de resultaten direct na afloop van de afnamen

beschikbaar zijn. Bovendien zijn de resultaten van de leerlingen direct in elektronische vorm opgeslagen en kan op relatief eenvoudige wijze over deze data beschikt worden. Beeldschermtoetsen leveren op deze manier een bijdrage aan de oplossing van het dataverzamelingsprobleem. Door gebruik te maken van bijvoorbeeld een webbased-applicatie kunnen scholen na afname van beeldschermtoetsen de resultaten van de leerlingen rechtstreeks doorsturen naar het DWH. Voor EVADOS brengt deze werkwijze het grote voordeel met zich mee dat referentiegegevens continu geactualiseerd kunnen worden en dat daarvoor geen grootschalig (en kostbaar) onderzoek hoeft plaats te vinden. Voor scholen betekent deze werkwijze dat zij steeds over de meest recente referentiegegevens kunnen beschikken.

De meest eenvoudige vorm van beeldschermtoetsen zijn de lineaire toetsen. Bij deze variant van beeldschermtoetsen is er sprake van een vaste toets en de leerlingen moeten alle opgaven uit de toets maken. Een speciale groep beeldschermtoetsen zijn de computer adaptieve toetsen (CAT's), toetsen die de nieuwste psychometrische technieken combineren met het gebruik van de computer voor de afname van toetsen. Bij CAT's krijgen de leerlingen niet een vaste toets voorgelegd maar opgaven die zo goed mogelijk zijn afgestemd op hun vaardigheid. Elke keer als een leerling een vraag beantwoordt, schat de computer afhankelijk van zijn antwoord zijn vaardigheid en zoekt vervolgens een nieuwe opgave die aansluit bij het geschatte vaardigheidsniveau van de leerling. Door de nauwkeurige afstemming van de opgaven op het niveau van de kandidaat is het mogelijk met een relatief gering aantal opgaven nauwkeurig te meten. Omdat de keuze van een volgende opgave bepaald wordt door het geschatte vaardigheidsniveau van de leerling, krijgt elke leerling in principe een andere 'toets' voorgelegd. Omdat de opgaven echter op één schaal liggen zijn de toetsprestaties van de leerlingen vergelijkbaar.

Het toepassen van CAT's vraagt om een uitgebreide opgavenbank die natuurlijk onderhouden dient te worden. Nieuwe opgaven zullen eraan toegevoegd moeten worden en verouderde opgaven verwijderd. Als een dergelijke opgavenbank ter beschikking komt van het onderwijs vinden de evaluaties plaats op basis van

opgavenbanken die volledig up-to-date zijn. En dat geldt dan ook voor de referentiegegevens die leerkrachten kunnen gebruiken om de resultaten van hun leerlingen te evalueren en voor de school de kwaliteit van het geboden onderwijs.

CAT's leveren indirect ook een bijdrage aan de verhoging van de effectiviteit van het onderwijs. CAT's bieden scholen de mogelijkheid de effectiviteit van interventies - zoals het inzetten van onderwijshulpprogramma's - te onderzoeken. Een leerling maakt een toets, behaalt een onvoldoende resultaat, wat vervolgens leidt tot een onderwijskundige maatregel. Na deze maatregel krijgt de leerling 'eenzelfde' toets voorgelegd om te zien of het gewenste effect bereikt is en of wellicht andere maatregelen gewenst zijn. Een toename in effectiviteit van het onderwijs op leerlingniveau leidt ook tot een toename in effectiviteit op schoolniveau.

CAT's sluiten ook goed aan bij een aantal uitgangspunten zoals deze in het begin van dit hoofdstuk zijn geformuleerd. CAT's bieden de mogelijkheid om toetsen vaker bij leerlingen af te nemen, waardoor meer gegevens beschikbaar komen om de ontwikkeling van hen in de tijd te volgen. Doordat de afnamen achter het beeldscherm plaatsvinden en van daaruit de mogelijkheid bestaat de data rechtstreeks te sturen naar een DWH, vraagt dit van de school geen extra belasting. Tevens zorgt deze technische voorziening ervoor dat de data op een centrale plaats terechtkomen waar ook de analyses en de rapportages verzorgd kunnen worden. De constante stroom aan data zorgt ervoor dat de referentiegegevens steeds up-to-date kunnen zijn. Ook aan een eventuele wens om toetsen op vaste of variabele tijdstippen af te nemen, vormt voor CAT's geen probleem. 'Dezelfde' toetsen kunnen op meerdere momenten afgenomen worden.

7.3 Expertisecentrum Kwaliteitszorg

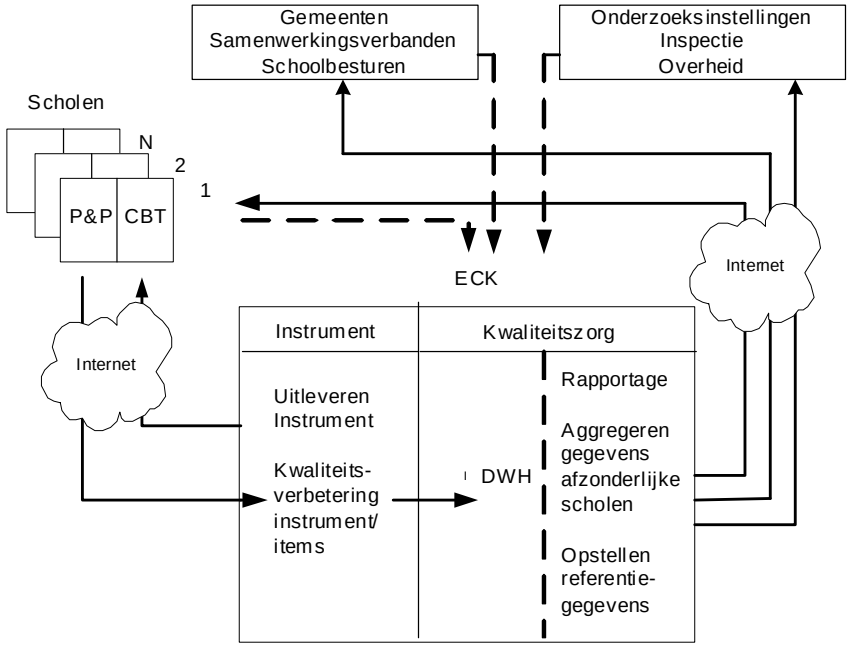
In de voorgaande paragraaf zijn ontwikkelingen besproken die van groot belang zijn voor de kwaliteitszorg van het onderwijs. Elke ontwikkeling draagt op zijn eigen manier bij aan de verhoging van de kwaliteit van het onderwijs. Enerzijds door met meer geavanceerde technieken toetsen in te zetten waarmee nog efficiënter en nauwkeuriger uitspraken gedaan kunnen worden over de vaardigheden van leerlingen en daaraan gekoppeld de kwaliteit van het onderwijs. Anderzijds door de verbetering van infrastructurele voorzieningen waardoor het datatransport op een efficiënte(re) wijze kan plaatsvinden en de beschikbaarheid van de gewenste gegevens verbetert.

De overheid streeft naar een grotere bestuurlijke afstand, waaruit voor scholen een vergroting van de zelfstandigheid en meer eigen verantwoordelijkheid voortvloeit. Het besproken DWH en de wijzen waarop scholen geïnformeerd kunnen worden over de effectiviteit van hun onderwijs leveren daaraan een bijdrage. Scholen zullen wel steeds een actief beleid van kwaliteitszorg moeten blijven voeren waarbij het van cruciaal belang is dat zij kunnen beschikken over adequate informatie met betrekking tot de kwaliteit van het geboden onderwijs en het effect van ingezette beleidsmaatregelen. Toepassing van informatie- en communicatietechnologie (ICT) kan bij het beschikbaar maken van deze informatie een belangrijke rol spelen en zo stimulerend werken ten aanzien van de kwaliteitszorg in het onderwijs. Het ontbeert scholen aan middelen en kennis om de voor kwaliteitszorg relevante gegevens op een juiste wijze voor de gewenste doeleinden te kunnen gebruiken. Om de bij scholen beschikbare gegevens toepasbaar te maken voor kwaliteitszorg is een systeem nodig waarmee deze direct kunnen worden omgezet in voor scholen bruikbare informatie. Een dergelijk systeem zou ondergebracht kunnen worden in een Expertisecentrum Kwaliteitszorg (ECK) met als doel het zorgdragen voor een permanent monitorsysteem. De doelgroep voor een dergelijk ECK hoeft zich in de praktijk niet te beperken tot de scholen. Ook gemeenten, samenwerkingsverbanden en andere organisaties die op enigerlei wijze betrokken zijn bij het onderwijs (zoals onderzoeks-

instellingen of inspectie) zouden van de diensten van een dergelijk centrum gebruik kunnen maken. Een uitbreiding met meer doelgroepen brengt ook de mogelijkheid met zich mee om het DWH dat in het ECK is ondergebracht ook te voeden met gegevens uit andere bronnen, wat tot nieuwe informatie kan leiden. Zo zou een gemeente gegeven haar taakstelling geïnformeerd kunnen worden over de doelmatige besteding van financiële middelen in het kader van het achterstandsbeleid. Onderzoekers zouden een beroep op het ECK kunnen doen voor het verkrijgen van data voor het doen van onderzoek. Hetzelfde geldt voor de inspectie. Door de resultaten van leerlingen te aggregeren en de informatie te vertalen naar macro niveau, is mogelijk ook de overheid een belanghebbende.

Het ECK levert niet alleen een bijdrage aan het monitoren van de kwaliteit van het onderwijs door diverse geledingen van informatie te voorzien, maar het kan ook een belangrijke schakel zijn om te komen tot kwaliteitsverbetering van het toetsmateriaal. Hoewel in de huidige praktijk er met name nog schriftelijk getoetst wordt, is de verwachting dat in de nabije toekomst meer gebruik gemaakt zal worden van toetsen aan het beeldscherm. Door de toename van het aantal computers in de klas wordt de weg geopend om leerlingen een individuelere leerweg te laten volgen, waaraan de toetsing zich zal moeten aanpassen. De hiervoor besproken CAT's zijn daar voorbeelden van. Door nu vanuit een ECK zowel de toetsen te beheren als de resultaten van leerlingen op die toetsen, kunnen de resultaten ook gebruikt worden om te komen tot kwaliteitsverbetering van het toetsmateriaal. Met deze laatste toepassing hebben we niet alleen meer te maken met een informatiesysteem, maar ook met een toetssysteem.

Wat de plaats van een ECK in een toets- en informatieststeem zou kunnen zijn, laat figuur 7.3 zien.



Figuur 7.3
Positie ECK in een toets- en informatiesysteem

Scholen kunnen via een webbased applicatie inloggen in het ECK waar zij (on-line) toetsen kunnen opvragen. Deze toetsen worden vervolgens naar de scholen toegestuurd. Een school kan beslissen de toetsen in de vorm van een papieren toets (P&P) af te nemen of aan het beeldscherm (CBT). De resultaten van de leerlingen gaan via het net weer terug naar het ECK, waar zij gebruikt kunnen worden om (1) de kwaliteit van de items en de toetsen te verbeteren en (2) informatie over de kwaliteit van het onderwijs te genereren. In het toets- en informatiesysteem kan het zinvol zijn ook te kunnen beschikken over gegevens die bij scholen of bij andere instanties (bijvoorbeeld gemeenten) aanwezig zijn. Via een bepaalde procedure moet deze informatie aan het ECK aangeleverd kunnen worden. Deze gegevensstroom wordt in figuur 7.3 voorgesteld door de onderbroken lijnen. De ononderbroken lijnen aan de rechterkant geven aan dat de informatie vanuit het ECK zowel voor scholen als voor andere belang-hebbenden bestemd kan zijn. Het spreekt voor zich dat de aard van de informatie aansluit bij de informatiebehoefte van de afnemer. Zo zullen scholen geïnfor-

meerd worden over hun functioneren in vergelijking met een relevant referentiekader. Gemeenten, samenwerkingsverbanden en schoolbesturen ontvangen informatie na bewerking (aggregeren) van de gegevens op schoolniveau. Dat geldt mogelijk ook voor de inspectie en de overheid. De te verstrekken informatie zal in de regel uit kant-en-klare rapporten bestaan die aansluiten bij de informatiebehoeften. Onderzoekers zullen niet altijd geïnteresseerd zijn in kant-en-klare rapporten, maar zullen meer geïnteresseerd zijn in het kunnen beschikken over data om eigen onderzoek uit te voeren. De mogelijkheid om data aan te leveren voor nader onderzoek, moet ook één van de doelstellingen van het ECK zijn.

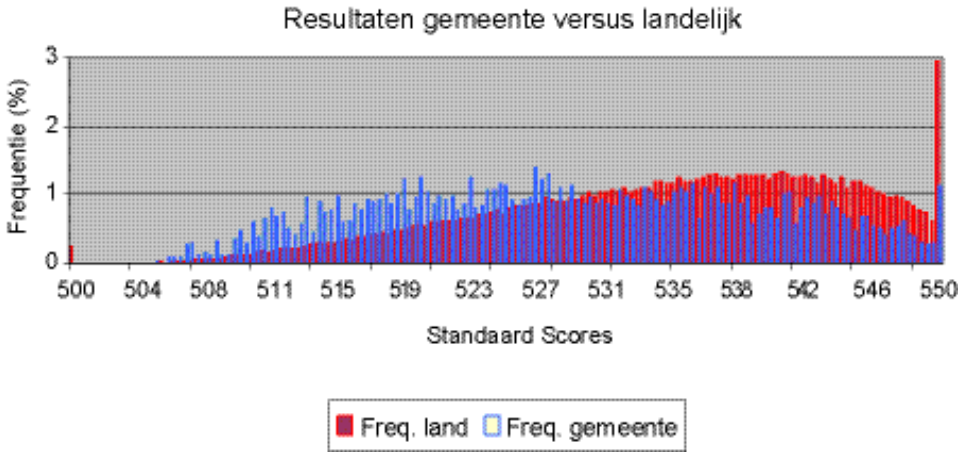
Het ECK kan als spil fungeren bij het verstrekken van informatie over de kwaliteit van het onderwijs aan allerlei geledingen binnen het onderwijs. In 2001/2002 is in een pilot project een eerste versie van een DWH ontwikkeld, waarin vervolgens gegevens van leerlingen die deelgenomen hebben aan de Eindtoets basisonderwijs (EB) over een periode van vijf jaren zijn opgeslagen. Doel van dit experiment was te onderzoeken in hoeverre het inderdaad mogelijk blijkt te zijn om een DWH als uitgangspunt te nemen voor het verstrekken van referentiegegevens en het doen van nadere analyses. Het DWH kan gezien worden als een essentieel onderdeel van het ECK. In hoofdstuk 6 hebben we laten zien hoe individuele scholen geïnformeerd zouden kunnen worden over de kwaliteit van het geboden onderwijs door de resultaten van (groepen van) leerlingen in de tijd te volgen. In de volgende paragraaf worden enkele voorbeelden gepresenteerd van informatie die gebaseerd zijn op de resultaten van leerlingen gerelateerd aan achtergrondkenmerken, maar waarbij gerapporteerd wordt op een hoger aggregatieniveau. Bij de voorbeelden is uitgegaan van de gemeente als doelgroep.

7.4 Voorbeeldrapportages op gemeentelijk niveau

In deze paragraaf worden vier voorbeelden gepresenteerd. In figuur 7.4 staan de resultaten van alle scholen binnen een gemeente afgezet tegen een landelijk referentiekader. In figuur 7.5 wordt een relatie gelegd tussen de woonwijken in de gemeente en de behaalde resultaten op de EB. Figuur 7.6 geeft een overzicht van de behaalde scores van naar schoolgewicht ingedeelde scholen. En in figuur 7.7 ten slotte wordt een overzicht gepresenteerd van de behaalde scores op de vier onderdelen van de EB. Merk op dat de te tonen figuren exemplarisch zijn in die zin dat zij aangeven wat mogelijkheden zijn. Afhankelijk van de wens van de afnemer kunnen (ook) andere grafieken gemaakt worden. De in deze paragraaf opgenomen grafieken zijn wel gebaseerd op reële wensen van een gemeente in de praktijk.

Vergelijking van de resultaten op gemeentelijk niveau met het landelijk beeld

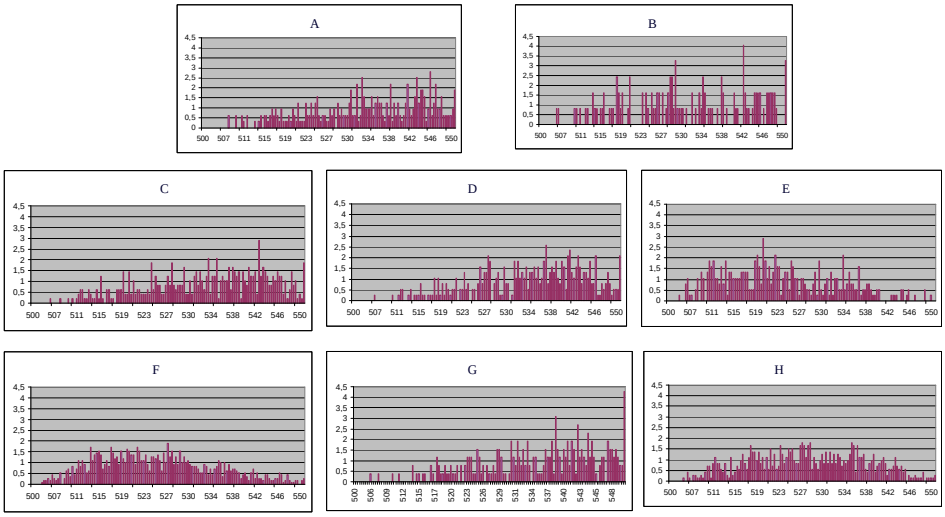
In figuur 7.4 zijn de resultaten van alle leerlingen uit een gemeente op de EB afgezet tegen de landelijke resultaten.



Figuur 7.4
Vergelijking resultaten gemeente versus landelijk beeld

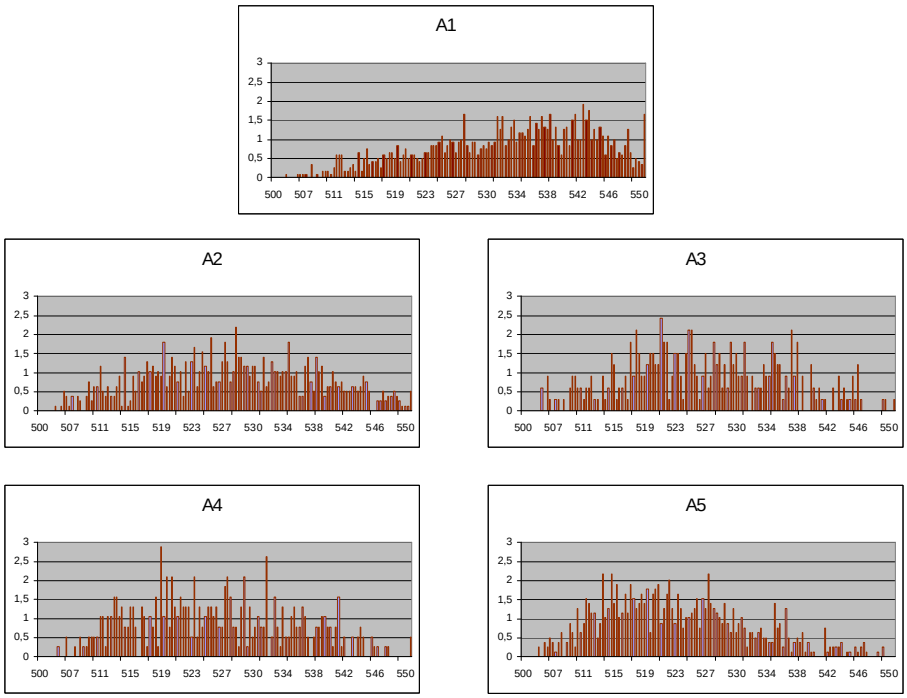
Op de horizontale as staat de scoreschaal. Bij de EB loopt deze van 501 tot 550 met als gemiddelde 535. Op de verticale as staat het percentage scholen met de desbetreffende scoreschaal. Uit figuur 7.4 kan een gemeente afleiden in hoeverre zijn resultaten afwijken van het landelijk beeld. In het gegeven voorbeeld blijkt dat de resultaten voor de betreffende gemeente achterblijven. Er zijn beduidend meer leerlingen die lager scoren op de EB in vergelijking met de landelijke populatie.

Voor een gemeente kan een dergelijke constatering aanleiding zijn zich af te vragen of het met name bepaalde wijken binnen de gemeente zijn die bijdragen aan dit beeld. Daartoe kunnen op gemeentelijk niveau de behaalde resultaten uitgesplitst worden naar woonwijken. In figuur 7.5 zijn de gegevens van de onderzochte gemeente uitgesplitst naar 8 wijken (A tot en met H).



Figuur 7.5
Vergelijking resultaten (woon)wijken binnen een gemeente op de EB

Figuur 7.5 geeft aan dat de resultaten van de diverse wijken in de gemeente verschillend scoren op de EB. Voor een gemeente kan dit een reden zijn nader onderzoek te doen naar de populatie van de wijken om te zien of bepaalde (onderwijskundige) beleidsmaatregelen gewenst zijn. En andere uitsplitsing die mogelijk informatief is voor een gemeente is te kijken hoe scholen ingedeeld naar schoolgewicht presteren op de EB. Figuur 7.6 is een voorbeeld van een dergelijke uitsplitsing. In het voorbeeld van figuur 7.6 zijn de scholen ingedeeld in vijf groepen. Elke groep representeert een gemiddeld schoolgewicht. Zo zijn in groep A1 alle scholen ondergebracht met een schoolgewicht dat loopt van 1,0 tot 1,25 en in groep A5 bevinden zich alle scholen met een schoolgewicht van 1,8 tot 1,9.

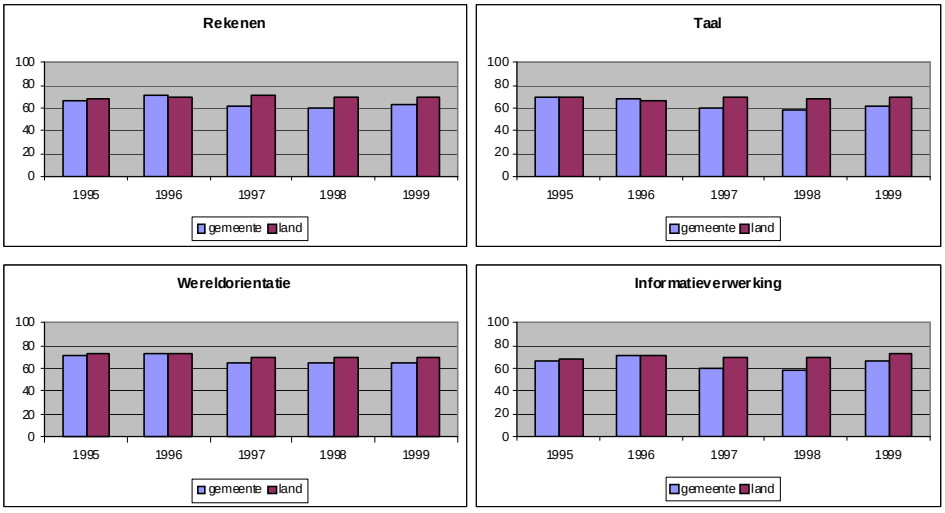


Figuur 7.6
Vergelijking resultaten scholen met verschillend schoolgewicht op de EB

Uit figuur 7.6 kan geconcludeerd worden dat scholen met een hoger schoolgewicht in het algemeen lagere resultaten behalen op de EB dan scholen met een lager schoolgewicht. Mogelijk hangen de resultaten op de EB samen met de populatie leerlingen die een school bezoekt (zie figuur 7.5).

Een gemeente zou ook kunnen besluiten om de resultaten op de EB meer inhoudelijk te analyseren. Zo zou een gemeente kunnen nagaan wat de door scholen behaalde scores zijn op de vier onderdelen van de EB gedurende een aantal jaren. Figuur 7.7 geeft informatie op deze vraag.

In figuur 7.7 staan per onderdeel van de EB het percentage goed beantwoorde opgaven (verticale as) weergegeven over een periode van vijf jaren. Als referentie zijn de landelijke resultaten genomen. Figuur 7.7 laat zien dat voor deze gemeente de scholen op alle onderdelen lager scoren dan het landelijk gemiddelde. Alleen 1996 vormt daarop een uitzondering. In dat jaar waren op twee onderdelen de resultaten beter dan het landelijk gemiddelde. Wat de reden(en) daarvoor is of zijn, kan niet uit figuur 7.7 afgeleid worden. Wel is het informatie die de gemeente kan doen besluiten na te gaan wat mogelijke verklarende factoren (geweest) kunnen zijn.



Figuur 7.7
Percentage goed beantwoorde opgaven op de vier onderdelen van de EB

7.5 Tot slot

In dit proefschrift is getracht te laten zien hoe op basis van op scholen (en eventueel ook bij andere instanties) aanwezige gegevens, informatie over de kwaliteit van het onderwijs verkregen kan worden. Met deze informatie zijn scholen in staat gerichter naar de kwaliteit van het geboden onderwijs te kijken en beschikken zij over informatie op basis waarvan (beleids-)beslissingen genomen kunnen worden. Met de ontwikkelde procedure worden scholen niet alleen in staat gesteld hun prestaties te vergelijken met de resultaten in voorgaande jaren, maar ook met een extern referentiekader. Om voor scholen relevante referentiegegevens te kunnen aanbieden, is het van belang dat deze ook verzameld worden. Het besproken DWH-concept kan daar een goede bijdrage aan leveren. Zoals paragraaf 7.4 heeft laten zien, geeft het DWH tevens de mogelijkheid om naast scholen ook samenwerkingsverbanden en gemeenten van informatie te voorzien, als ook de inspectie en de overheid.

Het op structurele basis leveren van informatie over de kwaliteit van het onderwijs aan verschillende belanghebbenden vraagt om diverse expertises die verenigd zouden kunnen worden in het besproken ECK. In een dergelijk centrum zijn de volgende drie expertises van belang:

- inhoudelijke expertise (materiedeskundigen);
- methodologische expertise;
- informatie-technologische expertise.

De eerste twee genoemde expertises zijn voor een deel uitgewerkt in het onderhavige proefschrift bij de bespreking van EVADOS. De uitgevoerde pilotstudie naar de ontwikkeling van een DWH heeft laten zien dat inmiddels ook voldoende ervaringen en ontwikkelingen op het gebied van de ICT aanwezig zijn om een dergelijk centrum verantwoord te kunnen vormgeven.

Dit proefschrift gaat over een procedure waarmee scholen de kwaliteit van hun onderwijs op basis van resultaten op toetsen kunnen evalueren. Hoewel EVADOS de mogelijkheid biedt naar de kwaliteit van scholen te kijken vanuit

een extern controleperspectief, is dat niet de doelstelling ervan. Vooropstaat het interne managementperspectief zoals verwoord door de subtitel van dit proefschrift 'het evalueren van en door scholen'. EVADOS dient scholen van informatie te voorzien waarmee gericht aan schoolverbetering kan worden gedaan. Externe referentiegegevens hebben met name tot doel om een school te informeren wat 'maatschappelijk' verwacht mag worden (hoe doen andere scholen het). Voor zover geaggregeerde gegevens op bijvoorbeeld gemeentelijk niveau gebruikt worden, zouden deze een signaalfunctie moeten hebben. Doen wij het als gemeente 'goed'? Zijn wij tevreden met de resultaten van de scholen? Ook deze vragen dienen wat betreft de in dit proefschrift beschreven procedure vanuit het oogpunt van schoolverbetering (beleidsondersteuning cq. beleidsvoorbereiding) gesteld te worden. Geaggregeerde resultaten op bijvoorbeeld gemeentelijk niveau informeert een gemeente over dat wat met het onderwijs bereikt wordt. Vanuit de verantwoording van de gemeente kunnen bij tegenvallende resultaten gerichte beleidsacties ondernomen worden. Daarbij kan dan bijvoorbeeld gedacht worden aan mogelijkheden die het onderwijsachterstandenbeleid biedt. Als blijkt dat de resultaten op gemeentelijk niveau positief gewaardeerd worden, ontslaat dat de individuele school niet van de verplichting om het eigen functioneren na te gaan.

In dit proefschrift stond het basisonderwijs centraal. Dit was om praktische redenen. In het basisonderwijs is het gewenste toetsinstrumentarium om leerlingen in de tijd te volgen beschikbaar. Bovendien is bij een aantal toetsen sprake van een centrale dataverwerking en kent het basisonderwijs een onderwijsadministratiepakket met een hoge penetratiegraad op de Nederlandse onderwijsmarkt. De scope kan natuurlijk verbreed worden naar andere sectoren van het onderwijs. Door deze verbreding ontstaat niet alleen de mogelijkheid om uitspraken te doen over de kwaliteit van het onderwijs van deze sectoren, maar zijn ook gegevens gedurende een langere periode bekend. Met deze gegevens is het beter mogelijk om bijdragen van het onderwijs in kaart te brengen en bestaat bovendien de mogelijkheid onderzoek te doen naar de ontwikkelingen van leerlingen over een langere periode (onderwijsloopbaan).

Literatuur

Abswoude, A.A.H. (1999). *De ontwikkeling van een instrument voor 'Toets Curriculum Overlap'*. (OPD Memorandum 99-1). Arnhem: Cito.

Agerbeek, M., Hageman, E., & Lakmaker, H. (1997) Trouw-onderzoek school prestaties. Trouw, 25 oktober 1997.

Bergh van den, H., & Kuhlemeier, H. (1997). Multiniveau modellen voor de analyse van leerwinst vergeleken. *Tijdschrift voor onderwijsresearch*, 22, 2, 54-75.

Bokhove, J., Schoot, F. van der, & Eggen, T. (1996). *Balans van het reken onderwijs halverwege de basisschool 2*. Arnhem: Cito.

Bosker, R.J. (1990). *Extra kansen dankzij de school?* Nijmegen: ITS (dissertatie Rijksuniversiteit Groningen.

Bosker, R.J., & Scheerens, J. (1995). A self-evaluation procedure for schools using multilevel modelling. *Tijdschrift voor Onderwijsresearch*, 1, 58-68.

Bosker, R.J., Houtveen, A.A.M., & Meijen, G.W. (1998). Programma voor Beleidsgericht Onderzoek Primair Onderwijs.

Bosker, R., Béguin, A., & Rekers, L. (2001). Hoe meten we de prestatie van een school? In A.B. Dijkstra, S. Karsten, R. Veenstra en A.J. Visscher (red.). *Het oog der natie: scholen op rapport* (pp. 121-135) Assen: Van Gorcum.

Brandsma, H.P. (1993). *Basisschoolkenmerken en de kwaliteit van het onderwijs*. Groningen: RION. Proefschrift Groningen.

Brookover, W., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). *School social systems and student achievement. Schools can make a difference*. New York: Bergin Publishers Book.

Bryk, A.S., & Raudenbusch, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Coleman, J.S., Campbell, E.Q., Hobson, C.F., McPartland, J., Mood, A.M., Weifeld, F.D., & York, R.L. (1966). *Equality of educational opportunity*. Washington D.C.: Department of education and welfare.

Cooper, H.N.B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: a narrative and meta-analytic review. Review of educational research, vol 66 (3), p. 227-268.

Creemers, B.P.M. (1995). *School improvement and school effectiveness: sustaining links*. Paper presented at the 1995 ICSEI congress, Leeuwarden.

Creemers, B.P.M. (1994). *The effective classroom*. London: Cassell.

Creemers, B.P.M. (1995). Implementatie van het onderwijsaanbod. In: B.P.M. Creemers & G.J. Reezigt (Eds.), *Onderwijsaanbod*. (pp. 27-40). Alphen a/d Rijn: Samson J.H.D. Tjeenk Willink.

Cremers-van Wees, L.M.C.M., Rekveld, I.J., Brandsma, H.P., & Bosker, R.J. (1995). *Instrumenten voor zelfevaluatie. Inventarisatie en beschrijving. Enschede*: Universiteit Twente, Onderzoek Centrum Toegepaste Onder wijskunde (VO-94606).

Cremers-van Wees, L.M.C.M., Rekveld, I.J., Brandsma, H.P., & Bosker, R.J. (1995). *Instrumenten voor zelfevaluatie. Beschrijving van 31 instrumenten*. Enschede: Universiteit Twente, Onderzoek Centrum Toegepaste Onder wijskunde (VO-94606).

Edmonds, R. (1979). Effective schools for the urban poor. Educational Leadership, 37, 15-24.

Eeden, P. van den, & Meijnen, G.W. (red.) (1990). Multi-niveauonderzoek: uit gangspunten en toepassingen. Themanummer. *Tijdschrift voor Onderwijs research*, 15(5).

Fahrmeier, L., & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. New York: Springer.

Fitz-Gibbon, C.T. (1997). *From Value Added Indicators to Evidence Based Education: The Task for the Next Decade*. Paper presented at an IARTV Morning Forum Value Added Measures in Education on 17 March 1997, in Melbourne. Seminar series: July 1997, No.65.

Gezamenlijke richtinggevende uitspraken van de vertegenwoordigers van de koepelorganisaties en de minister van onderwijs en wetenschappen (1993/1994). Den Haag: SDU.

Gillijns, P., & Moelands, F. (1992). Toepassing van itemresponstheorie ten behoeve van een leerlingvolgsysteem. *Tijdschrift voor Orthopedagogiek*, 4, 162-176.

Goffree, F., & Frowijn, R. (1996). *Op weg naar zelfevaluatie in het basis onderwijs. Katern 1: Toetsen en kerndoelen*, Enschede: SLO.

Goldstein, H. (1999). *Multilevel statistical models*. Kendall's Library of Statistics. Internet Edition april 1999.

Groot, A.D. de. (1983). Is de kwaliteit van het onderwijs te beoordelen? In B.P.M. Creemers et al. (red). *De kwaliteit van het onderwijs*. Groningen: RION/Wolters-Noordhoff.

Haan, D.M. de. (1992). *Measuring Test-Curriculum Overlap*. Doctoral thesis, University of Twente, Enschede.

Husén, T., & Tuijnman, A. (1994). Monitoring standards in education: why and how it came about. In: A.C. Tuijnman, & T.N. Postlethwaite (Eds.). *Monitoring the standards of education*. Trowbridge: Redwood Books.

Hendriks, M. (1997). Constructie van een instrumentarium 'School- en klaskenmerken'. In: Engelen, R. e.a. *Schoolzelfevaluatie in het basisonderwijs* SVO-prpoject 95405). Enschede: Universiteit van Twente.

Hill, P.W. (1995). *Value added measures of achievement*. Paper resented at an IARTV seminar in October 1994, and further developed at Assessing and Reporting Students' Educational Progress, an ACER/NCD Seminar held at the Melbourne Business School on March 1, 1995.

Hill, P.W., & Rowe, K.J. (1996). Multilevel modelling in school effectiveness research. *School effectiveness and school Improvement*, 7, 1-34.

Inmon, W.H. (1996). Building the Data Warehouse. New York: Wiley.

Jencks, C.S., Smith, J., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heys, B. & Michelson, S. (1972). *Inequality: a reassessment of the effect of family and schooling in America*. New York: Basic Books.

Kimball R., The Data warehouse Toolkit. New York: Wiley.

Kreft, I., & De Leeuw, J. (1998). *Introduction multilevel modeling*. London: Sage Publication.

Levine, D.K., & L.W. Lezotte. (1990). *Unusually effective schools: a review and analysis of research and practice*. Madison, WI: National Center for Effective Schools Research and Development.

McKnight, W., & Curtis, C. (eds.) (1987). *The Underachieving Curriculum: Assessing US School Mathematics from an International Perspective*. Illinois: Stipes Publishing Company.

Ministerie van Onderwijs Cultuur en Wetenschappen (1998). Kerndoelen basisonderwijs 1998. *Over de relatie tussen de algemene doelen en kerndoelen per vak.* 's-Gravenhage: SDU

Moelands, H.A., & Ouborg, M.J. (1995). *School self-evaluation in primary education in the Netherlands*. Paper presented at the conference for Senior European Community Officials in Brussels, held on November 30th and December 1st, 1995.

Moelands, H.A., & Sanders, P.F. (1996). Onderwijskundige meetinstrumenten. In: J. Scheerens (red). *Kwaliteitszorg in het onderwijs*. Onderwijskundig Lexicon. Deel Centrale Onderwijsthema's. p. 51-73. Alphen aan den Rijn: Samson H.D. Tjeenk Willink.

Moelands, H.A., Ouborg, M.J., & Engelen, R.J.H. (1996). *Schoolzelfevaluatie in het basisonderwijs. Output gerelateerd aan input*. Arnhem: Cito.

Moelands, H.A., Ouborg, M.J., & Engelen, R.J.H. (1997). Output gerelateerd aan input. In: Scheerens, J. (red.). *Schoolzelfevaluatie in het basisonderwijs*. Interimrapportage van het gezamenlijke project van CITO, SLO en OCTO (periode maart 1995 - december 1996). Enschede: Universiteit Twente, Onderzoek Centrum Toegepaste Onderwijskunde (SVO-95405).

Moelands, H.A. (2004). Onderzoek naar een instrument voor Toets Curriculum Overlap (TCO). *Pedagogische Studiën*, 81, p. 214-227.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters*. The Junior Years. Sommerset: Open Books.

Oakes, J. (1989). 'What educational indicators? The case for assessing the school context'. *Educational Evaluation and Policy Analysis*, Vol 9l. 11 (2), p. 181-199.

Onderwijs Automatiseringsbureau (1994). *Handboek ESIS-A en –B.* Nieuwegein: Onderwijs Automatiseringsbureau (OAB).

Oosterbeek, H., & Webbink, D. (2001). Risico's van indicatoren voor school kwaliteit. In A.B. Dijkstra, S. Karsten, R. Veenstra en A.J. Visscher (red.). *Het oog der natie: scholen op rapport* (pp. 111-120) Assen: Van Gorcum.

Pelgrum, W.J. (1989). *Educational Assessment: Monitoring, Evaluation and the Curriculum*. De Lier: ABC

Pelgrum, W.J., Voogt, J., & Plomp, T. (1995). Curriculum indicators in international comparitive rearch. In Organisation for Economic Co-operation and Development, *Measuring the quality of schools* (pp. 81-102). Paris: OECD.

Petegem van, P. (1994). Spiegeltje, spiegeltje aan de wand ... Effectieve scholenonderzoek als reflectiebasis voor zelfevaluatie. *Tijdschrift voor Onderwijsrecht en Onderwijsbeleid*, 1994, nr. 1., p. 15-24.

Petegem van, P. (1997). *Scholen op zoek naar hun kwaliteit. Effectieve scholenonderzoek als inspiratiebron voor de zelfevaluatie van scholen*. (Academisch proefschrift). Universiteit Gent.

Quinn, R.E., & J. Rohrbaugh. (1983). Spatial model of effectiveness criteria towards a competing values approach to organizational analysis. *Management science*, jaargang 29, 1983, pag. 363-377.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., & Lewis, T. (2000). *A user's guideto MlWinI, London: Multilevel project, Iniversity of London*.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, S. (1979). *Fiftheen thousand hours. Secondary schools and their effects on children*. London: Open Books, Publishers Ltd.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., & Lewis, T. (2000). *A user's guide to MlWin*, London: Multilevel project, University of London.

Schaffer, E.G., & Nesselrodt, P.S. (1992). *The development and testing of the special strategies observation system*. Paper presented at the American Education Research Association Annual Meeting in San Francisco, California on April 20.

Scheerens, J. (1989). *Wat maakt scholen effectief?* Den Haag: SVO, balansreeks nr. 1.

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School effectiveness and school improvement*, 1 (1), pp. 61-80.

Scheerens, J. (1996a). Beoordeling en evaluatie in het kader van kwaliteitszorg in het onderwijs. In J. Scheerens (red). *Kwaliteitszorg in het onderwijs*. Onderwijskundig Lexicon. Deel Centrale Onderwijsthema's. p. 51-73. Alphen aan den Rijn: Samson H.D. Tjeenk Willink

Scheerens, J. (1996b). Towards an integrated instrumentation of school self-evaluation. In: Engelen, R. e.a. *Schoolzelfevaluatie in het basisonderwijs* (SVO-project 95405). Enschede: Universiteit van Twente.

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

Schmidt, W.H., & McKnight, C.C. (1995). Educational Opportunity in Mathematics and Science: An International Perspective. *Educational evaluation and policy analysis*, Fall 1995, Vol 17, No. 3, pp. 337-353.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modelling*. London: Sage Publication.

Stringfield, S.C., & Slavin, R.E. (1992). A hierarchical longitudinal model for elementary schools effects. In B.P.M. Creemers & G.J. Reezigt (Eds.), *Evaluation of educational effectiveness*. (pp. 35-68). Groningen: ICO.

Tuin van der, A.C., & Werf van der, M.P.C. (1996). *Effects from differences in instruction characteristics between mathematics teachers on mathematics achievement*. Paper presented at the American Educational Research Association Annual Meeting in New York.

Tuin van der, A.M.C. (1997). *Differences between effective and less effective teachers within school. A multilevel approach*. Paper presented at the International Congress for School Effectiveness and Improvement (ICSEI '97) in Memphis USA on January 6.

Tweede kamer, Vergaderjaar 1994-1995, 24248, nr2, bp5. 's-Gravenhage: SDU Uitgeverij.

Van de Zand, I. (1999). Voorstudie data warehouse. Breda: Newcom information systems N.V.

Veenstra, D.R., Dijkstra, A.B., Peschar, J.L., & Snijders, T.A.B. (1998). Scholen op rapport. Een reactie op het Trouw-onderzoek naar schoolprestaties. *Pedagogische Studiën*, 75, 121-134.

Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. (PPON-rapport, nr. 4). Arnhem: Cito.

Verhelst, N.D. (1992). *Het eenparameter logistisch model (OPLM). Een theoretische inleiding en een handleiding bij het computerprogramma*. (OPD-memorandum 92-3). Arnhem, Cito Instituut voor Toetsontwikkeling.

Verhelst, N.D. (1993). Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometerie in de praktijk* (pp. 83-178). Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: Computer program and manual*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N., Staphorsius, G., & Kleintjes, F. (2001). Scholen langs de meetlat. *De psycholoog*. Vol. 36 (2001), nr. 12 ; p. 658-664.

Visscher, A., Dijkstra, A.B., Karsten, S., & Veenstra, R. (2001). Standaarden en aanbevelingen voor de publicatie van schoolprestatie-indicatoren. In A.B. Dijkstra, S. Karsten, R. Veenstra en A.J. Visscher (red.). *Het oog der natie: scholen op rapport* (pp. 239-250) Assen: Van Gorcum.

Voogt, J.C. (1995). Schooldiagnose. In: Cremers, H.P.M. et al (red.) Onderwijskundig Lexicon. Alphen aan den Rijn: Samson Tjeenk Willink.

Vos. T. de. (1992). *Tempo-Test-Rekenen: Test voor het vaststellen van het vaardigheidsniveau der elementaire bewerkingen (automatisering) voor het basis- en voortgezet onderwijs*. Nijmegen: Berkhout.

Waterreus, J.M. (2002). *School performance measurement and teacher mobility in the Netherlands*. Paper presented at the ORD in Antwerpen, held on May 19-21, 2002.

Whiley, D.E., & Yoon, B. (1995). Teacher reports on Opportunity To Learn: analyses of the 1993 California Learning Assessment System (CLASS). *Educational Evaluation and Policy Analysis*, 17, 255-370.

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). *De toegevoegde waarde van de basisschool. Verkenning van de mogelijkheden de schoolspecifieke bijdrage*

aan de onderwijsopbrengst in kaart te brengen met behulp van het Cito Leerlingvolgsysteem en de Eindtoets basisonderwijs. Arnhem: Citogroep.

Willms, J.D. (1992). *Monitoring school performance. A guide for educators*. London: The Falmer Press.

Willms, J.D., & Kerckhoff, A.C. (1995). The Challange of Developing New Educational Indicators. *Educational Evaluation and Policy Analysis*. Vol. 17, No. 1, pp 113-131.

Summary

The Dutch educational decentralization policy, aimed at giving schools more freedom and making them more responsible for the quality of their education, has resulted in a growing interest in quality (control) procedures at the school level. Schools are held more responsible for the quality of the education they provide and therefore have to carry out an active policy of quality control.

In this dissertation, a procedure for schools to evaluate the quality of their education is presented. The procedure is called EVADOS, a Dutch acronym for 'evaluation of and by schools'. Quality of education can be defined in several ways. In EVADOS, student achievement, in particular, achievement measured by test results, is considered to be the most important indicator of the quality of education. However, since the educational input and processes of schools are so different, it is not fair to determine the quality of a school on the basis of test results alone. The test results obtained by students have to be adjusted for differences in input and processes, and EVADOS makes it possible for schools to make this adjustment.

Factors that contribute to the quality of a school have been identified by research on school effectiveness. These factors can be categorized according to input, process, and output (CIPO model). In this dissertation, the CIPO model is used as a conceptual framework to summarize the results of school effectiveness research. Examples of relevant context factors from school effectiveness research are school size, school category, and urban or rural environment. Examples of relevant input variables are teacher experience and pupil background characteristics. Process factors can be divided into school level factors and classroom level factors. Examples of process factors are educational leadership, orderly atmosphere, structured teaching, and opportunity to learn. Output can be cognitive or non-cognitive. In school effectiveness research,

cognitive outcomes have dominated. In this dissertation, the construction of a valid and reliable instrument to measure the process variable 'opportunity to learn' and research with this instrument is presented. The instrument determines the overlap between what is being tested by an arithmetic test in grade 1 of primary school and the arithmetic content taught to the students.

In this dissertation, student achievement is seen as an indicator of the quality of the school. Because so many factors have an effect on the quality of a school, one single measurement of achievement is not sufficient to determine this quality. Therefore tests that can determine achievement over time, such as tests from the Cito pupil monitoring system, should be used.

EVADOS provides schools with three kinds of sources of information on how they are performing. First, on the basis of information stored in school administration packages, characteristics of the school population are described. Second, EVADOS enables schools to monitor the progress of groups of students over time. By combining the results on tests with the data stored in the school administration packages, schools can break down the results into relevant subgroups. Moreover, EVADOS enables schools to compare their results with internal and external reference data. In this dissertation, several examples of how to inform schools about the progress of groups of students are presented. Third, EVADOS informs schools about their added value on the basis of test results at the end of primary education.
Univariate and multivariate multilevel models have been used to determine the added value of a school. It is shown how school effects can be estimated according to three types of univariate multilevel models and two types of multivariate multilevel models. The analyses with these models show that controlling for background variables, for tests, and for background variables and tests, has an effect on the position of individual schools in the distribution of schools based on the size of the school effects. Controlling for tests results in the largest shift, whereas the shift when controlling for only background variables is minimal.

Because a high correlation was found between the school effects estimated by the univariate models and one of the multivariate models, it was concluded that it does not matter whether the school effects are estimated with the univariate models or the multivariate model applied. The research presented in this dissertation shows that the position of a school in the distribution of schools changes over time. A school performing well in 2002 will not necessarily perform well in 2003. The conclusion therefore is that it is not possible to draw valid conclusions about the performance of individual schools on the basis of the results of one year. Finally, it is necessary to be cautious with the interpretation of the position of schools on the basis of their school effects, since the 95% confidence intervals of the school effects of many schools show a (tremendous) overlap.

The research presented in this dissertation shows that it is possible to provide schools with relevant information about characteristics of the school population, about the progress of groups of students over time and about the added value of the school. With this information, a school can determine its performance compared to other schools and to its own performance in previous years. However, our research shows that the storage of data in school administration packages does not comply with the requirements of EVADOS. The Data Warehouse concept was presented as a possible solution for this lack of compliance.

The dissertation ends with the description of an Expertise Centre for Quality Assurance. If schools wish to have an active policy regarding quality assurance, it is crucial for them to have at their disposal adequate information with regard to the education being offered and the effect of policy measures. Information and communication technology can play a very important role in disclosing this information. However, schools lack the means and knowledge to use the data that are relevant for quality assurance. An Expertise Centre for Quality Assurance could help and support schools to achieve this goal and maintain a permanent monitoring system. The target group for such an Expertise Centre is not limited to schools. Municipalities, groups of schools, school boards, and

other kinds of organizations involved in education (e.g., educational research institutes and the inspectorate) could also benefit from the services of such a centre. Some examples of how to report to the municipalities are presented. Finally, it is concluded that the educational, methodological, and ICT expertise necessary to establish an Expertise Centre for Quality Assurance is available in the Netherlands.