Onderzoek naar een instrument voor Toets Curriculum Overlap (TCO)

H. A. Moelands

Samenvatting

Opportunity to Learn (OTL) wordt als een belangrijke procesvariabele gezien bij de interpretatie van resultaten van leerlingen op toetsen. OTL kan gedefinieerd worden als de mate waarin leerlingen in de gelegenheid zijn gesteld zich de vereiste leerstof eigen te maken. Zo gedefinieerd, verwijst OTL naar de instructie die heeft plaatsgevonden en de hoeveelheid tijd die aan het leren is besteed. OTL kan ook gedefinieerd worden als de mate waarin het beoogde curriculum aansluit bij het geïmplementeerde curriculum zoals door de toets gemeten. Zo geformuleerd, krijgt het begrip OTL een wat specifiekere betekenis, wat tot uitdrukking gebracht wordt door te kiezen voor het begrip Toets Curriculum Overlap (TCO). Om een uitspraak te doen over de kwaliteit van het onderwijs dient de inhoud van de toets als operationalisatie van het beoogde curriculum in voldoende mate overeen te komen met het door leerkrachten geboden onderwijs, ofwel het geïmplementeerde curriculum. In dit artikel wordt verslag gedaan van een onderzoek naar de mogelijkheid een valide en betrouwbaar instrument te ontwikkelen voor het meten van TCO.

1 Inleiding

Uit diverse publicaties blijkt dat behaalde resultaten op toetsen een belangrijke indicatie zijn voor de kwaliteit van het onderwijs. Zo worden scholen op basis van resultaten op toetsen bijvoorbeeld in een rangorde geplaatst. De misvatting die daarbij kan ontstaan, is dat de plaats van een school op de ranglijst iets zegt over de kwaliteit van het onderwijs. Het zal duidelijk zijn dat de kwaliteit van het onderwijs op een school met lagere resultaten op een toets niet per se minder behoeft te zijn dan bij een school met hogere resultaten op die toets. Vanuit onderzoek is bekend dat vele factoren een rol spelen bij het vaststellen van de kwaliteit van het onderwijs. Dat geldt ook als deze kwaliteit bepaald wordt op basis van de resultaten van de leerlingen op toetsen. Twee componenten waaronder die factoren gecategoriseerd kunnen worden, zijn *input* en *proces* (zie Scheerens, 1989). Tot de component input behoren factoren als het beginniveau van leerlingen en beschikbare materiële en financiële middelen van een school. Onder de component proces vallen factoren als werkklimaat en schoolleiding. Ook factoren aangeduid met de Engelse termen 'time on task', 'direct instruction' en 'opportunity to learn' behoren tot de component proces. 'Opportunity to learn' (OTL) wordt in de literatuur over schooleffectiviteit als een belangrijke variabele gezien en is als zodanig ook terug te vinden in het door Scheerens (1989) beschreven CIPO-model, een acroniem dat staat voor Context, Input, Proces en Output. In de uitwerking van zijn model legt Scheerens de genoemde vier componenten uiteen in een aantal variabelen dat van invloed is op bereikte resultaten. Een van deze variabelen is OTL. Als nu bereikte resultaten een rol spelen bij het doen van uitspraken over de kwaliteit of de effectiviteit van het onderwijs, dan dient zoveel mogelijk rekening te worden gehouden met OTL. Voorwaarde is wel dat we OTL nader definiëren en meetbaar maken. Het onderhavige onderzoek laat zien dat OTL meetbaar is, en hoe met deze variabele omgegaan kan worden. Door met OTL rekening te houden, worden mogelijke conclusies over de kwaliteit of effectiviteit van onderwijs op basis van resultaten meer valide.

OTL wordt wel gedefinieerd als de mate waarin leerlingen in de gelegenheid zijn gesteld zich de vereiste leerstof 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. Aangenomen mag worden dat OTL een effect heeft op de resultaten van leerlingen op toetsen. Er moet vol-

doende sprake zijn van OTL om een uitspraak te mogen doen over deze resultaten en als afgeleide daarvan de kwaliteit of effectiviteit van het onderwijs, vooral als we daarbij een extern criterium hanteren als bijvoorbeeld de kerndoelen basisonderwijs. Willen we een uitspraak doen aan het einde van de basisschool in groep 8, dan moeten we ervan verzekerd zijn dat de leerlingen tijdens hun schoolloopbaan voldoende gelegenheid hebben gehad om de geformuleerde kerndoelen te bereiken. Zijn leerlingen daartoe niet in staat gesteld, dan geven de resultaten op toetsen die tot doel hebben na te gaan of de kerndoelen bereikt zijn, wel informatie over de mate waarin de leerlingen deze kerndoelen beheersen, maar vormen ze geen basis om scholen met elkaar te vergelijken of een oordeel uit te spreken over de kwaliteit van het gegeven onderwijs. Het gebruikte instrumentarium sluit in dat geval niet aan bij het gegeven onderwijs. Het is, met andere woorden, van belang vast te stellen in welke mate 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".

In het voorgaande is betoogd dat om valide uitspraken te kunnen doen over de kwaliteit van het onderwijs op basis van toetsprestaties, het belangrijk is dat de toetsen aansluiten bij het geboden onderwijs. 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 geïmplementeerde 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".

Internationaal zijn vele studies naar OTL verricht, waarbij als vertrekpunt toetsitems, of leerstofcategorieën worden gehanteerd. Op beide methoden wordt kort 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 e.a. (1995) maken melding van acht studies waarbij gebruik is gemaakt 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), verwijzend naar Oakes (1989) en McKnight en Curtis (1987, p. 19), 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 curricular 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 focuses 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 de 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 ge-

weest zich de vereiste leerstof eigen te maken. Deze omschrijving kan ruim geïnterpreteerd worden, door ook de instructiewijze en de hoeveelheid bestede tijd erbij te betrekken. In het voorgaande is het begrip 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 is aangegeven, ook aspecten als leertijd (time on task) en de instructiewijze van leerkrachten deel uit (kunnen) maken, maar van Toets Curriculum Overlap (TCO). Dit is ook het begrip dat in dit artikel gehanteerd wordt.

1.1 Doel van het onderzoek

Het doel van het onderzoek waarvan hier verslag wordt gedaan, was na te gaan of het mogelijk is een valide en betrouwbaar TCO-meetinstrument te ontwikkelen. Aan dit doel lagen de volgende 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?

In dit artikel wordt de ontwikkeling van een instrument TCO beschreven. De ontwikkeling 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.

In paragraaf 1 wordt de opzet van het onderzoek besproken. In paragraaf 2 komt de eerste onderzoeksvraag aan bod. Drie meetmethoden voor TCO, aangeduid met itemmethode, categoriemethode en lesmethode, worden onderzocht, en op basis van een aantal criteria wordt in paragraaf 3 een keuze voor één van deze meetmethoden gemaakt. Belangrijk is dat het te construeren meetinstrument valide en betrouwbaar is. In paragraaf 4 wordt onderzocht in hoeverre de gekozen meetmethode daaraan voldoet. Wanneer een meetmethode/-instrument beschikbaar is, dient onderzocht te worden hoe het instrument het beste ingezet kan worden in de onderwijspraktijk. Deze vraag wordt in paragraaf 5 beantwoord. In paragraaf 6 staat de vraag centraal wanneer de toets wel of niet meer af te nemen, gegeven de score op het TCO-instrument. Het artikel eindigt met een discussie in paragraaf 7.

2 Opzet van het onderzoek

Het instrument TCO is ontwikkeld aan de hand van een naar postcodegebied gestratificeerde steekproef van 450 scholen uit een bestand van 1490 gebruikers van de toets E3 Rekenen-Wiskunde. Voor deelname aan het onderzoek is 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 van 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 response. 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.

3 Toets Curriculum Overlap (TCO)

Bij de ontwikkeling van een instrument TCO zijn de volgende drie meetmethoden onderzocht:

1 het voorleggen van items aan leerkrachten (itemmethode);
2 het voorleggen van leerstofcategorieën (categoriemethode);
3 het vragen naar de gebruikte lesmethode (lesmethode).

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

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

Voor het onderzoek is een vragenlijst ontwikkeld die is voorgelegd aan de aan het onderzoek deelnemende 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 werd de leerkrachten naar de door hen gebruikte rekenmethode gevraagd, en naar 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 zich de bij de subcategorie behorende leerstof eigen hebben kunnen maken (of de subcategorie maakbaar is, hetgeen betekent dat de leerstof behandeld is en dat de leerlingen ermee hebben kunnen oefenen). Om leerkrachten te informeren waaruit de leerstof van de onderscheiden subcategorieën bestaat, is gebruikgemaakt van de in de handleiding van de toets E3 gebruikte omschrijvingen. Tot slot werd aan hen gevraagd ook per item aan te geven of het item, 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 leerstof, 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.

De resultaten van de verwerking van de antwoorden op de drie methoden komen in de volgende twee paragrafen aan bod. In paragraaf 3.1 worden de item- en categoriemethode besproken en in paragraaf 3.2 de lesmethode.

3.1 Itemmethode en categoriemethode

Tabel 1 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

Tabel 1

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

Categorieën	Opvatting leerkrachten: Categorie (% maakbaar)	Opvatting leerkrachten: Item (% maakbaar)	Leerlingresultaten (% goed beantwoord)
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

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 gemiddelde percentage leerlingen weer dat de items van de subcategorieën correct heeft beantwoord.

Uit Tabel 1 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 leerlingscores (.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 1 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 worden teruggekomen.

3.2 Rekenmethode

In Tabel 2 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

Tabel 2

Overzicht van het aantal gebruikte rekenmethoden

Rekenmethode	Aantal leerkr.	%
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 rekenmethode	5	3
Onbekend	2	1

Tabel 3

Gemiddelde maakbaarheidsscore per rekenmethode

Rekenmethode	Gem. maakbaarheids-score over items	*SD*
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

Tabel 4

Overzicht maakbaarheid categorieën, uitgedrukt in procenten per rekenmethode

	% maakbare categorieën													
Rekenmethoden	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	96	96	89	89	100	91	98	96	85	93	93	91	91	57
Pluspunt	91	88	74	74	85	82	100	88	79	91	97	68	68	21

Noot. De kolommen 1 t/m 14 corresponderen met de in paragraaf 3 genoemde subcategorieën.

welke rekenmethode zij hanteren, vallen onder de categorie *onbekend*.

Uit Tabel 2 blijkt dat de meeste aan het onderzoek deelnemende leerkrachten gebruikmaken 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.

In Tabel 3 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.

Uit Tabel 3 blijkt dat er niet alleen verschillen zijn in maakbaarheidsscores tussen rekenmethoden ($F = 3.429$, $p = .006$), 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 gebruikmaken 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 3 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 zien.

De kolommen 1 tot en met 14 in Tabel 4 corresponderen met de in toets E3 onderscheiden 14 subcategorieën. In Tabel 4 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) gebruiken 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 laat zien dat de 14 subcategorieën niet allemaal evenveel aandacht krijgen in de methoden. Opvallend is de subcategorie *tijd* (kolom 14). Met name de leerkrachten die de rekenmethode "Pluspunt hanteren", maar ook de leerkrachten die gebruikmaken 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 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, is van een aantal factoren afhankelijk. 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 10 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 92% van de leerkrachten categorie 8, *aftrekken,* als maakbaar te beschouwen. Deze categorie bevat 7 items, met een gemiddeld 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.

5 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 model-'fit' opleverde. Ook het verwijderen van slecht fittende items of items met een extreem 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 vra-

genlijst 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 het feit dat Cronbachs α 0.88 is. 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.

6 Implementatie van het TCO-instrument

Bij de implementatie van het instrument TCO spelen praktische en psychometrische overwegingen een rol. In deze paragraaf komen drie mogelijkheden om het instrument TCO 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 vóór 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 de toets en het geboden onderwijs te groot blijft, dient de toets niet te worden afgenomen. In het vervolg van dit artikel wordt nader 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 0.17 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 0.34. 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 een invoerscherm op de computer of een formulier ontwikkeld moeten worden waarmee leerkrachten kunnen aangeven welke items maakbaar

zijn. Tevens zullen nieuwe omzettingstabellen voor de transformatie van ruwe scores naar schaalscores geconstrueerd moeten worden die aansluiten bij het volgens leerkrachten aantal maakbare items.

Ook bij de toepassing van 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 ook 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.

De voorkeur geniet TCO als entreemeting. 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 hun 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 het 80%-niveau als beheersingsniveau hanteert. Indien meer dan 20% van het aantal items niet maakbaar zou zijn, kunnen de leerlingen nooit meer dit beheersingsniveau halen. 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.

7 Effect beslisregel op niveau-indicatie Cito-LVS

Het Cito-LVS maakt bij haar rapportage gebruik van een vijftal niveaus. Welk niveau aan een leerling wordt toegekend, is afhankelijk van zijn vaardigheidsscore, die afhangt van het aantal goed beantwoorde items (53 items in totaal). Bij het indelen in niveaus hanteert het Cito-LVS voor de toets 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 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:

- 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 Tabel 5). In de bespreking van Tabel 5 worden de drie groepen aangeduid als *groep < 43*, *groep 43-48* en *groep > 48*.

Voor de bespreking van Tabel 5 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 schattingsfout 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 niveautoekenning 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 5 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 maak-

Tabel 5

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.
Maakbaarheidsscore < 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
Maakbaarheidsscore 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
Maakbaarheidsscore > 48	A	586	20				606
	B	20	334	8			362
	C		9	162	3		174
	D			3	43	2	48
	E				2	16	18
	Tot.	606	363	173	48	18	1208

Noot. A = 46 of meer items goed beantwoord, B = 40 t/m 45 items goed beantwoord, C = 32 tot 40 items goed beantwoord, D = 24 tot 32 goed beantwoord, E = minder dan 24 items goed beantwoord.

bare 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 er 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 er 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 alle 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.

Gegeven de voorgaande resultaten kan geconcludeerd worden dat het verantwoord is 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.
- 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.

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 subcategorieën bevraagd zullen worden, wat betekent dat delen van het curriculum niet in de toets aan bod komen. 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 leerstof te behandelen, behoren zij ook tot de groep > 48. Van deze groep is vastgesteld dat het aantal misclassificaties erg beperkt is.

8 Discussie

Uit het onderzoek kan geconcludeerd worden dat het concept *maakbaarheid* meetbaar is. Ook heeft het onderzoek laten zien dat het mogelijk is een instrument TCO te ontwikkelen. Met dit instrument is het mogelijk om vóór afname van een toets vast te stellen of de afname verantwoord is, in die zin dat de inhoud van de toets in voldoende mate overeenkomt met het geboden onderwijs. Mocht dat niet zo zijn, dan kan afname van de toets leiden tot een onaanvaardbaar hoog percentage misclassificaties. Het voorleggen van toetsen met een (te) groot aantal niet-maakbare items kan leiden tot een onderschatting van de vaardigheid van leerlingen. Uit het onderzoek dat uitgevoerd is voor de toets Rekenen-Wiskunde E3 van het Cito-LVS, kan geconcludeerd worden dat het instrument TCO leerkrachten een goede indicatie geeft wanneer de toets E3 afgenomen kan worden. Of anders gezegd: het instrument TCO geeft aan in welke mate er sprake is van overlap tussen beoogd curriculum en geïmplementeerd curriculum. Het instrument TCO biedt de mogelijkheid meer valide uitspraken te doen op basis van de door leerlingen behaalde resultaten en daarmee indirect over de kwaliteit of effectiviteit van het geboden onderwijs. Het instrument TCO is exemplarisch ontwikkeld voor de toets Rekenen-Wiskunde E3. De uit het onderzoek voortgekomen beslisregel is niet zonder meer toepasbaar op de andere toetsen uit het Cito-LVS. Voor andere toetsen zal deze opnieuw vastgesteld moeten worden. Wel lijkt de toegepaste procedure om te komen tot een beslisregel ook voor andere toetsen toepasbaar.

Het onderzoek laat zien dat, gegeven de variëteit aan lesmethoden in het onderwijs, bij het gebruik van methode-onafhankelijke toetsen TCO een belangrijke rol speelt. Wat betekent deze constatering nu voor leerkrachten en toetsconstructeurs? Leerkrachten dienen in hun oordeel of een item maakbaar is, een goed onderscheid te (kunnen) maken tussen maakbaarheid en moeilijkheid. In de praktijk zal dat soms lastig zijn. Enerzijds, omdat zij wellicht vinden dat bepaalde leerstofonderdelen op een andere wijze in hun onderwijs aan bod zijn geweest dan door de items gerepresenteerd. Een dergelijke constatering kan leiden tot een oordeel "niet maakbaar". Mogelijk ligt het accent dan niet zo zeer op het construct dat getoetst wordt, maar meer op de specifieke bevraging van de leerstof door het item. Ook kan het gevoel "afgerekend" te worden op basis van de prestaties van leerlingen een negatief effect hebben op het oordeel. In dat geval kan het zijn dat niet de maakbaarheid van een item het uitgangspunt is, maar de moeilijkheid. Een mogelijk gevaar dat ook speelt, is het fenomeen 'teaching-to-the-test'. Door vooraf inzage te hebben, kunnen leerkrachten het onderwijs doelbewust afstemmen op de inhoud van de toets. Niet het onderwijsprogramma staat dan centraal, maar de inhoud van de toets. Een gevaar dat zich eerder zal voordoen als de resultaten op de toetsen een rol gaan spelen in een "afrekencultuur". Het spreekt voor zich dat een TCO-instrument daar niet voor gebruikt mag worden; een TCO-instrument doet in die zin een groot beroep op de professionaliteit van de leerkracht. Daar komt bij dat de toetsen uit het Cito-LVS formatieve toetsen zijn, en dat de afname van deze toetsen in de regel tweemaal per jaar plaatsvindt. Het instrument TCO vraagt per keer maximaal 10 minuten van de leerkracht. Allemaal redenen die aangeven dat de kans op 'teaching-to-the-test' niet zo groot is.

Ook voor toetsconstructeurs is TCO van belang. In eerste instantie voor het maken van een valide toets. Het heeft geen zin leerlingen een toets voor te leggen waarvan op voorhand al bekend is dat leerlingen bepaalde items niet kunnen maken, omdat de benodigde leerstof en vaardigheden niet onderwezen zijn, of niet verwacht mag worden dat leerlingen daarover beschikken. Dat laatste behoeft enige nuancering. Voor toetsconstructeurs is met name de functie van de toets van belang. Betreft het een toets die tot doel heeft de opbrengsten van het geboden onderwijs te meten, of gaat het om een toets die tot doel heeft de kennis en vaardigheden van een leerling te meten, ongeacht de vraag of de leerling zich de toegepaste kennis tijdens het onderwijs in dat vak of op andere momenten heeft eigen gemaakt? Bovendien kunnen toetsconstructeurs leerlingen de mogelijkheid bieden om op "alternatieve/eigen" wijze

(bijvoorbeeld afwijkend van dat wat gangbaar is in methoden of in de klas steeds geoefend wordt) items op te lossen. Het gaat er immers om, vast te stellen waar een leerling staat in zijn ontwikkeling. Dat mag en kan afwijken van de klassenpraktijk. De ene leerling zal verder zijn in zijn ontwikkeling dan de andere. Het is vervolgens aan het onderwijs om daarop in te spelen. Met deze "afwijkende" items worden meer facetten van het ontwikkelings-/vaardigheidsniveau van een leerling zichtbaar.

Ten slotte: In het onderwijs vindt 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 -afname 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.

Literatuur

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

Haan, D. M. de. (1992). *Measuring test-curriculum overlap*. Academisch proefschrift, Universiteit 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* (pp. 1-21). Trowbridge: Redwood Books.

McKnight, W., & Curtis, C. (Eds.). (1987). *The underachieving curriculum: assessing US school mathematics from an international perspective*. Illinois: Stipes Publishing Company.

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

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 research. In Organisation for Economic Co-operation and Development, *Measuring the quality of schools* (pp. 81-102). Paris: OECD.

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

Schmidt, W. H., & McKnight, C. C. (1995). Educational opportunity in mathematics and science: an international perspective. *Educational Evaluation and Policy Analysis*, *17*, 337-353.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). Psychometrische en statistische aspecten van peilingsonderzoek. (PPON-rapport nr. 4). Arnhem: Cito.

Wiley, 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*, 355-370.

Manuscript aanvaard: 10 maart 2004

Auteur

Henk Moelands is als onderwijskundige werkzaam bij het Psychometrisch Onderzoek- en Kenniscentrum van de Citogroep.

Correspondentieadres: H. A. Moelands, Citogroep, POK, Postbus 1034, 6801 MG Arnhem, e-mail: henk.moelands@citogroep.nl

Abstract

The development of an instrument Test Curriculum Overlap

Opportunity To Learn (OTL) is considered an important process variable in explaining the results of students on tests. OTL can be defined as the extent to which pupils have been given the opportunity to master a subject. Defined in this way, OTL refers to the instruction that has been taken place and the amount of time spent to learn a subject. OTL can also be defined as the extent to which the intended curriculum matches the implemented curriculum as measured by the test. With this definition, OTL has a more specific meaning that is expressed by the concept Test Curriculum Overlap (TCO). In order to evaluate the quality of education, the content of the test - as a specification of the intended curriculum - must align with subject matter provided by the teacher, that is the implemented curriculum. In this article the development of a TCO-instrument is presented.