Tijdschrift voor Onderwijsresearch, 11 (1986), Nr. 6, pp. 312-317.

Notities en Commentaren

WOLTERS' ONGERECHTVAARDIGDE CONCLUSIES

Een kritisch commentaar op 'De functie van deel-geheel-schema's in het rekenonderwijs: een terugblik' (Wolters, 1984).

L.W.C. Tavecchio*, M. Beishuizen**, J.N. van den Berge*** en M.W. Bleek****

* *Vakgroep Wijsgerige en Empirische Pedagogiek, R.U.L.*

** *Vakgroep Onderwijskunde, R.U.L.*

*** *Werkzaam in het basisonderwijs te Leiden*

**** *Werkzaam in het vormingswerk te Den Haag.*

Door Wolters (1984) werden drie onderzoekingen naar redactieopgaven naast elkaar gezet en aan een heranalyse onderworpen. Deze onderzoekingen werden verricht tussen 1976 en 1979 en waren qua opzet en probleemstelling vergelijkbaar. Zij richtten zich op het leren oplossen van redactie-opgaven op een nieuwe, 'wiskundige' wijze, waarbij leertheoretische denkbeelden uit de Sovjet Unie een belangrijke rol spelen, met name de opvattingen van Davydov. Laatstgenoemde stelt dat leerlingen al in een veel vroeger stadium dan gebruikelijk in staat moeten worden geacht te kunnen abstraheren en opereren bij een 'gealgebraïseerde' instructie in het reken/wiskunde-programma. Het zogenaamde 'deel-geheel-schema' als oplossingsmodel speelde in elk van de drie onderzochte programma's een belangrijke rol.

Wolters' heranalyse gaat uit van nieuwere inzichten in de literatuur, die stellen dat naast mathematische structuur óók verschillen in *semantische structuur* van invloed zijn op het oplossen van redactie-opgaven. Door toepassing van deze semantische typologie op de toetsitems (hierover later meer) voerde Wolters haar heranalyse op de onderzoeksdata uit. Niet de totaalscores per toets werden nu als basis voor de effectmeting genomen, maar subscores voor afzonderlijke typen redactie-opgaven. Wolters' nieuwe conclusie luidt dat gesproken moet worden van een differentieel effect: alléén opgaven van het type 'deel-geheel' en 'puntsom' zouden verbetering te zien geven ten gevolge van de training met het 'deel-geheel-schema', en andere redactie-opgaven niet (o.c., p. 81, 82). Ook zou één van de drie onderzoekingen namelijk dat van Van den Berge en Bleek (1982) in het licht van deze heranalyse (achteraf) als irrelevant ter zijde geschoven moeten worden, omdat de toetsen te weinig 'gevoelige' opgaven van het bovengenoemde type zouden hebben bevat (o.c., p. 81).

Met Wolters zijn wij van mening dat dit meer recente theoretische interpretatiekader interessante verklaringsmogelijkheden biedt. Maar de wijze waarop Wolters haar heranalyse uitvoerde achten wij methodologisch zeer aanvechtbaar. Haar artikel bevat nogal wat voorbeelden van onzorgvuldig redeneren, inadequate analyses en, wat het belangrijkste is, ongerechtvaardigde conclusies. Onze indruk is dat Wolters te snel naar nieuwe verklaringen toe redeneert zonder voldoende onderbouwing. Daarop heeft onze reactie betrekking, die we kortheidshalve beperken tot drie illustraties.

1. Het eerste onderzoek (Assink & Verloop, 1977) had als uitgangspunt de probleemstelling: 'Is het mogelijk het oplossen van redactie-opgaven te verbeteren door de leerlingen een algemeen rekenprogramma aan te bieden, waarin geprobeerd wordt inzicht te verschaffen in de mathematische structuur van de opgaven?'. Gemiddelden op voor- en natoets van twee experimentele groepen (beide kregen een zgn. 'Davydov-training', de een met lettersymbolen, de ander met cijfersymbolen) en van de controlegroep (die het standaard rekenprogramma volgde) werden eerst per groep vergeleken en daarna werden de verschilscores tussen voor- en natoets tussen de

Adres: Postbus 9507, 2300 RA Leiden.

drie groepen vergeleken. De naar de mening van de auteurs meest 'kritische' toets, de toets op de verschilscores, leverde in geen van de gevallen significante resultaten op (Assink & Verloop, 1977, p. 139/140 en 141). Voor de groepen afzonderlijk kon ook geen effect worden aangetoond, zij het dat het voortoets-natoets verschil in de letterconditie de significantiedrempel dicht benaderde: de gevonden t-waarde had een overschrijdingskans van .054 (de auteurs, o.c., p. 139, spreken over 'praktisch op 5%-nivo significant'). Wolters, destijds coördinator van het project redactiesommen, was aanmerkelijk enthousiaster zoals blijkt uit haar proefschrift (1978a, p. 86), waarin zij stelde dat 'Assink en Verloop hebben aangetoond dat een dergelijke werkwijze in het Nederlands onderwijs succesvol is', terwijl zij elders (1978b, p. 233) zelfs sprak van 'opzienbarende resultaten'.

Van den Berge & Bleek rapporteerden in 1982 echter over een door hen in 1979 uitgevoerde replicatie van het onderzoek van Assink & Verloop, waaruit bleek dat het experimentele rekenprogramma (wederom) geen positief resultaat opleverde: ook in dit onderzoek kon de hypothese dat leerlingen na het volgen van Davydov-training redactie-opgaven beter zouden oplossen *niet* worden bevestigd. Wat ons thans dan ook uitermate verbaast is de strekking van Wolters' terugblik. Wat een 'terugblik' heet te zijn, is in feite een doorgaan op de verkeerde weg van niet gerechtvaardigde conclusies.

In haar heranalyse van de onderzoeksresultaten van Assink & Verloop vervangt Wolters de (oorspronkelijke) t-toets procedure door een (ongetoetste) vergelijking tussen het 'percentage goede oplossingen' op de voor- en natoets. Bovendien wordt deze vergelijking nu uitgevoerd voor de diverse semantische structuurtypen afzonderlijk. Aldus tracht zij alsnog bewijsmateriaal aan te dragen dat tot de conclusie zou leiden dat het trainingsprogramma van Assink & Verloop wél succesvol is geweest. Wij gaan nu eerst nader in op de door Wolters gehanteerde maat, het verschil tussen het percentage goede oplossingen op voor- en natoets. Wolters komt in haar terugblik (1984, p. 80) tot de conclusie dat het door Assink & Verloop gebruikte trainingsprogramma een gunstig effect heeft op het oplossen van 'deel-geheel opgaven' en geen effect op het oplossen van 'vergelijkingsopgaven'. Zij komt tot deze conclusie door voor de experimentele groep en de controle groep het *percentage goede antwoorden* op de deel-geheel opgaven op respectievelijk voor- en natoets te vergelijken met het percentage goede antwoorden op respectievelijk voor- en natoets op de vergelijkingsopgaven. In Tabel 6 (p. 80) staan deze percentages vermeld. Laten we ons beperken tot de resultaten van de experimentele groep, d.w.z. de linkerhelft van de tabel:

	exp. groep (n = 8)	
	voortoets	natoets
deel-geheel opgaven	40%	63%
vergelijkingsopgaven	81%	88%

Volgens Wolters '... zien we in tabel 6 dat het trainingsprogramma ook hier weer een gunstig effect heeft op het oplossen van deel-geheel opgaven en geen effect op het oplossen van vergelijkingsopgaven' (o.c., p. 80-81). De vooruitgang op de deel-geheel opgaven (van 40% naar 63%) wordt vergeleken met de vooruitgang op de vergelijkingsopgaven (van 81% naar 88%). Wolters deelde ons mee[1] dat haar in Tabel 6 gepresenteerde resultaten betrekking hebben op vier deel-geheel opgaven en twee vergelijkingsopgaven. Dit betekent dat er in de betreffende groep van acht leerlingen een maximum aantal van 32 goede oplossingen op de vier deel-geheel opgaven en een maximum aantal van 16 goede oplossingen op de twee vergelijkingsopgaven kon worden behaald. In absolute aantallen ziet Tabel 6 (linkerhelft) er dan ook als volgt uit:

	voortoets	natoets
deel-geheel opgaven	13	20
vergelijkingsopgaven	13	14

Dus: 13 goede oplossingen op de voortoets en 20 op de natoets voor de deel-geheel opgaven en respectievelijk 13 en 14 goede oplossingen voor de vergelijkingsopgaven. Hieronder laten wij aan de hand van een drietal voorbeelden zien hoe deze resultaten tot stand hadden kúnnen

komen. De voorbeelden kunnen met vele andere worden aangevuld. Wij kozen voor drie 'sprekende' en beperkten ons tot de vier deel-geheel opgaven:

	Voorbeeld 1		*Voorbeeld 2*		*Voorbeeld 3*	
Leerling	Voortoets	Natoets	Voortoets	Natoets	Voortoets	Natoets
1	0	0	3	0	0	0
2	0	4	1	3	0	0
3	0	4	2	1	0	0
4	1	1	0	4	2	4
5	2	2	1	3	3	4
6	3	3	3	3	3	4
7	3	2	1	4	2	4
8	4	4	2	2	3	4
	—	—	—	—	—	—
Totaal:	13	20	13	20	13	20

Drie zeer uiteenlopende situaties die gemeen hebben dat het *percentage goede oplossingen* op de voortoets 40% bedraagt (13 van de 32) en op de natoets 63% (20 van de 32). In *voorbeeld 1* gaan alleen de zeer slechten op de voortoets (met score 0) er op vooruit, vier leerlingen blijven gelijk en één gaat er achteruit. In *voorbeeld 2* gaat de helft van de leerlingen vooruit, van de andere helft blijven er twee gelijk en twee gaan achteruit. Tot slot laat *voorbeeld 3* vooruitgang zien voor degenen die toch al goed waren: deze vijf leerlingen gaan vooruit, de drie met score 0 blijven ook op de natoets in gebreke. Uit deze beknopte illustratie moge blijken dat 'percentage goede oplossingen' een veel te onnauwkeurige maat is voor het aantonen van verschillen of het trekken van verantwoorde conclusies. De percentages in Tabel 6 van Wolters' heranalyse zeggen niets en zeggen alles: ze laten alle conclusies toe en sluiten er geen uit. Wat Wolters had moeten analyseren is het *aantal goede antwoorden per kind per opgave* en daarbij een adequate toets toepassen (bijvoorbeeld een t-toets voor verschillen tussen gecorreleerde steekproeven of een toets voor verschillen tussen afhankelijke proporties). De conclusie die zij op grond van de percentages uit Tabel 6 trekt berust op een volstrekt onjuiste analyse van de gegevens: slechts bij toeval zou ze het bij het rechte eind kunnen hebben en dat is precies het omgekeerde van hetgeen het geval had moeten zijn! Afgezien van de inadequaatheid van de door Wolters gehanteerde analysemethode kan men zich afvragen of 'oplossers' (d.w.z. de kinderen zelf) niet beter als analyse-eenheid gekozen hadden kunnen worden in plaats van 'oplossingen'. Immers, de effectiviteit van een methode behoort toch te worden bepaald aan de prestatieverbetering van meerdere 'typen' leerlingen, bijvoorbeeld goeden én slechten. Zoals uit het hierboven vermelde 'voorbeeld 3' blijkt versluiert een maat als 'percentage goede oplossingen' mogelijk het feit dat er slechts door goede leerlingen vooruitgang wordt geboekt, terwijl de 'slechten' er niets aan hebben. Wat is in zo'n geval de effectiviteit van een methode? Precies hetzelfde betoog als hiervoor met betrekking tot Wolters' heranalyse van de gegevens van Assink & Verloop werd gehouden kan worden opgebouwd met betrekking tot de heranalyse van haar eigen onderzoeksgegevens (1984, p. 79-80, zie vooral Tabel 5). Ook in dit geval onderwerpt Wolters gedeelten van de gegevens uit haar dissertatie aan een 'heranalyse' op basis van semantische structuurtypen met behulp van percentage-gewijze vergelijkingen. Wij kozen het materiaal van Assink & Verloop in Tabel 6 ter illustratie, omdat het daarbij ging om 8 personen en 4 opgaven, zodat de (fictieve) voorbeelden zowel getalsmatig als inhoudelijk op overzichtelijke wijze konden worden uitgewerkt. Wolters' conclusie op basis van de percentages in Tabel 5 over '... een verband tussen type (semantische structuur) opgave en trainingsprogramma-effect' (o.c., p. 80) blijft op grond van het naar aanleiding van Tabel 6 gehouden betoog evenzeer in het luchtledige hangen. Vergelijk ook de typologische analyse van de toetsen hierna. Kortom: door de volstrekt ongerechtvaardigde conclusies met betrekking tot de percentages uit de tabellen 5 en 6 valt in feite iedere grond onder het betoog op p. 81 weg, waardoor de conclusie van Wolters 'Het lijkt aannemelijk op grond van het voorgaande te veronderstellen dat het trainingspro-

gramma een positief effect heeft op het oplossen van deel-geheel opgaven' dus *niet* getrokken kan worden.

2. Overigens roept ook het onderwijsprogramma als zodanig twijfels op ten aanzien van de vergelijkbaarheid van de drie door Wolters in haar heranalyse betrokken onderzoekingen. Het Davydovprogramma werd door Assink & Verloop omgezet in een Nederlands programma bestemd voor *tweede klassers.* In het replicatie-onderzoek van Van den Berge & Bleek werd dit programma verbeterd en ook aan tweede klassers voorgelegd. Wolters (1978; 1984) gaf in haar onderzoek hetzelfde programma aan *derde en vierde klassers.* Bovendien werd het programma door haar verlengd met een deel c. Men kan zich echter afvragen of eenzelfde programma voor zo'n grote (en gedifferentieerde) groep leerlingen, afkomstig uit verschillende klassen, geschikt kan zijn. Bovendien rijst ook onmiddellijk de vraag of überhaupt onderzoeksresultaten van tweede klassers mogen worden vergeleken met die van derde en vierde klassers. Speelt de factor 'leeftijd' daar niet een grote rol in evenals verschil in rekenkennis? Zeker wanneer we de oorspronkelijke probleemstelling er op na slaan.

3. Tot slot willen we nader ingaan op de kwestie van de *typologie* van redactiesommen. Deze typologie speelde in Wolters' artikel een belangrijke rol als uitgangspunt voor haar differentiële heranalyse. Zoals aan het begin opgemerkt concludeert zij nu dat het zogenaamde deel-geheel schema niet geschikt zou zijn voor álle typen redactie-opgaven. Het onderzoek van Van Berge & Bleek zou in dit licht niet langer relevant zijn, want in de natoetsen zouden 'geen deel-geheel opgaven opgenomen zijn' (o.c. p. 81). In deze onderzoekingen (ook Assink & Verloop) zou 'niet gevarieerd (zijn) volgens semantische structuurtypen', omdat dergelijke typologieën toen nog niet in de literatuur bekend waren (o.c. p. 79).

Ook hier blijkt Wolters onzorgvuldig in haar analyse en redenering. In de voortoetsen van Van den Berge & Bleek vond Wolters wél een deel-geheel-opgave *(som 2),* maar in de natoets zou deze verdwenen zijn (o.c. p. 79). Een nogal haastige conclusie, want de vergelijkbare deel-geheel-opgave stond in de natoets gewoon wat verderop als *som 10!* Het argument dat oudere onderzoekingen – zonder gebruik van semantische typologie – geen variatie van redactiesommen in dit opzicht zouden bevatten, is naar onze mening aanvechtbaar. Immers wanneer dergelijke toetsen representatief zijn samengesteld – hetgeen onderzoekers meestal nastreven – mag men (impliciet) een soortgelijke variatie verwachten als deze typologieën nu (expliciet) in kaart brengen.

Een typologische heranalyse van de besproken toetsen kon hier meer duidelijkheid verschaffen, die in het artikel van Wolters ontbreekt. Zij verwijst slechts fragmentarisch naar de aanwezigheid van de verschillende typen redactie-opgaven in de toetsen (vgl. hierboven), maar zij geeft geen volledig overzicht. Dit achten wij een ernstige omissie, want de feitelijke basis voor haar nieuwe conclusie blijft daardoor oncontroleerbaar. Daarom lieten we *twee onafhankelijke beoordelaars* buiten deze discussie[2] alle items classificeren in de volgende categorieën (vgl. Riley, Greeno & Heller, 1983; Wolters, 1984): 1. 'Change' (erbij-eraf), 2. Idem maar 'change' of 'start' onbekend (puntsom), 3. 'Combine' en 'Part-part-whole' (deel-geheel), 4. 'Compare' (vergelijking). Conform Wolters' analyse (o.c., p. 80) werd bovendien onderscheid gemaakt tussen opgaven met méér of één rekenkundige bewerking(en). Kortheidshalve volstaan we met de gemiddelde weergave over voor- en natoetsen, aangezien deze als parallelversies vrijwel identiek waren samengesteld (zie Tabel 1).

Twee conclusies kunnen uit deze tabel getrokken worden: (1) In de toetsen van Assink & Verloop en Van den Berge & Bleek was het aandeel 'gevoelige' items voor het deel-geheel-schema, namelijk de typen 2 en 3, *minstens even groot* als in Wolters' eigen toetsen; (2) Wolters' beperking tot alleen de opgaven met één rekenkundige bewerking gaf een nog smallere basis aan haar heranalyse, namelijk slechts *één, twee of drie items* per subscore, zoals in het rechterdeel van nevenstaande tabel kan worden nagegaan. Laatstgenoemde conclusie onderstreept de bezwaren die hiervoor reeds tegen Wolters' statistische heranalyse werden gemaakt. De eerste conclusie betekent dat de onderzoeksresultaten van Van den Berge & Bleek, die wel enige maar

Tabel 1: *Aanwezigheid van 4 typen redactie-opgaven in de voor- en natoetsen (gemiddeld).*

besproken onderzoekingen:	totaal opgaven per toets:	in àlle toetsopgaven: 1. erbij-eraf	2. punt-som	3. deel-geheel	4. verge-lijking	opgaven met één bewerking: 1. erbij-eraf	2. punt-som	3. deel-geheel	4. verge-lijking	beoorde-laars-overeen-stemming:
-Assink & Verloop	10	0	3	4	3	0	1	2	3	k[1]) = 0.85
-Wolters	10	2	3	1	4	0	1	0	2	k = 69
-Van den Berge & Bleek	14	1	5	6	2	1	1	2	2	k = .88

[1]) k = Cohen' kappa

geen significante vooruitgang vonden na training met het deel-geheel-schema, wel degelijk relevant genoemd moeten worden.

Wat betreft de evenmin significante 'opzienbarende resultaten' van Assink & Verloop (vgl. Wolters, hiervoor) verschaft de tabel ook andere informatie dan Wolters' suggestieve taalgebruik doet vermoeden. In tegenstelling tot de bevindingen van Van den Berge & Bleek betrok Wolters deze onderzoeksresultaten wél in haar heranalyse, waarin zij het eerder besproken differentiële effect constateerde. Want zo schrijft Wolters in haar artikel (o.c., p. 81): 'In het onderzoek van Assink & Verloop bestond de helft van de opgaven in de natoets uit deel-geheel-opgaven en het effect van het trainingsprogramma was dan ook waarneembaar'. Volgens Wolters' latere mededeling[1] betrof haar heranalyse echter *vier* 'deel-geheel-opgaven' (met één rekenkundige bewerking, waartoe zij haar analyse beperkte). Zoals in de tabel kan worden nagegaan kwamen onze beoordelaars echter niet verder dan *drie* van de tien opgaven d.w.z. één 'deel-geheel-opgave en twéé 'puntsommen'. Daarnaast classificeerden zij drie items als 'vergelijkingsopgaven', terwijl Wolters twee items van dit type onderscheidde bij Assink & Verloop (vgl. hiervoor). De waarheid zal wel ergens in het midden liggen, rekening houdend met beoordelings-variantie. Belangrijker is dat deze classificatie-details opnieuw de kernvraag oproepen of de heranalyse van Wolters statistisch wel toelaatbaar was? In dit geval met het argument dat de (sub) aantallen items per itemtype wel *erg gering* waren om daarop subscores (en effectverschillen) te baseren.

Samenvattend bestrijden wij niet de wenselijkheid van nieuwe en meer genuanceerde benaderingen in het onderzoek van redactie-opgaven, zoals Wolters in navolging van meer recente literatuur bedoelt. Maar daarbij is dan meer zorgvuldigheid gewenst, bijvoorbeeld ten aanzien van de classificatie van redactie-opgaven volgens semantische structuur (waaraan nog wel enige haken en ogen zitten waarop wij nu niet ingaan). Het is te hopen dat toekomstige pogingen minder slordig en onzorgvuldig zijn dan Wolters naar onze mening in haar 'heranalyse' etaleert. En wat betreft het inhoudelijke discussiepunt van de 'nuttige functie' van het deel-geheel-schema in het rekenonderwijs, moeten de eerste *relevante* empirische bouwstenen naar onze mening nog steeds worden aangedragen.

NOTEN

1. Persoonlijke communicatie met Wolters, 1985. Overigens kwamen onze beoordelaars tot een iets andere classificatie, vgl. Tabel 1 met typen redactie-opgaven. Overigens doet het exacte aantal opgaven van beide typen voor de illustratie van de inadequaatheid van de door Wolters verrichte heranalyses niet ter zake.
2. Het betrof hier de studenten Daudt-De Jong en Gulden, die een doctoraalscriptie over (andere aspecten van) redactiesommen maakten, en die bereid waren hun medewerking aan de beoordelingen te verlenen

LITERATUUR

Assink, E.M.H., & Verloop, N. (1977). Het aanleren van deel-geheel relaties in het aanvankelijk rekenonderwijs. *Pedagogische Studiën, 54,* 130-142.

Berge, J.N. van den & Bleek, M.W. (1982). Het oplossen van redactie-opgaven. *Pedagogische Studiën, 59,* 71-80.

Riley, M.S., Greeno, J.G., & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In H.P. Ginsburg (ed.). *The development of mathematical thinking,* (p. 153-200). New York: Academic Press.

Wolters, M.A.D. (1978a). *Van rekenen naar algebra. Een ontwikkelingspsychologische analyse.* Rijksuniversiteit Utrecht (dissertatie).

Wolters, M.A.D. (1978b). Algebra op de basisschool – ja of nee. *Pedagogisch Tijdschrift/Forum voor Opvoedkunde, 3,* 227-235.

Wolters, M.A.D. (1984). De functie van deel-geheel schema's in het rekenonderwijs: een terugblik. *Tijdschrift voor Onderwijsresearch, 9,* 71-83.

Ontvangen 30-10-1985.