Jansen, C.J.M. en M. Steehouder
1981 'Een taalverkeersprobleem: de voorlichting over individuele huursubsidie'. In: M. Steehouder en C.J.M. Jansen (red.): *Taalbeheersing 1981.* Enschede: Vereniging Interuniversitair Overleg Taalbeheersing (V.I.O.T.)/T.H. Twente.
Rijlaarsdam, G., K. de Glopper en H. Kreeft
1984 'Funktionele schrijfvaardigheidstoetsen: een voorstel voor nieuwe examens stelvaardigheid LBO/MAVO'. *Levende talen,* juni.
Wesdorp, H., H. van den Bergh en J.B. Hoeksma
1984 *De constructie van een aantal instrumenten ter meting van functionele taalvaardigheden ten behoeve van een periodiek peilingsonderzoek in het basisonderwijs.* Amsterdam: SCO.

Training van opstelbeoordelaars: effecten en problemen

Kees de Glopper

1. Inleiding

Voor de beoordeling van de produktief-schriftelijke taalvaardigheid zijn uiteenlopende evaluatiemethoden beschikbaar. Het meest gangbaar zijn de directe beoordelingsmethoden waarbij schrijfprodukten door beoordelaars op hun waarde worden geschat. Het langs directe weg beoordelen van de stelvaardigheid plaatst moedertaaldocenten en taalbeheersingsonderzoekers keer op keer voor lastige problemen. De betrouwbaarheid van opsteloordelen is immers vaak dubieus: het is eerder uitzondering dan regel dat een op één moment door één beoordelaar gegeven oordeel over een opstel overeenstemt met een ander, onafhankelijk verkregen oordeel. Met de betrouwbaarheid is ook de validiteit van de beoordeling van de produktief-schriftelijke taalbeheersing in het geding. Het is lang niet altijd duidelijk wat oordelen over opstellen nu eigenlijk meten: welke kwaliteiten van schrijfprodukten komen in de gevelde oordelen tot uitdrukking?

Verbeteringen van de kwaliteit van *opstelbeoordelingen* worden langs verschillende wegen nagestreefd. Vaak worden achteraf, ná beoordeling, maatregelen getroffen om de betrouwbaarheid van de beoordeling te verhogen. Dit gebeurt bijvoorbeeld door de oordelen van twee of meer onafhankelijk werkende beoordelaars door optelling te combineren tot een *jury-oordeel.* Zonder problemen is deze aanpak niet: er zijn duidelijke aanwijzingen dat beoordelaars consistent verschillend beoordelen; dit brengt met zich mee dat bij jury-vorming metingen van verschillende betekenis opgeteld worden.

In veel gevallen wordt ingegrepen in de beoordelingstaak, met de bedoeling de beoordelingstaak beter te omschrijven. Daarbij worden dan analytische beoordelingschema's verstrekt waarin te beoordelen deelaspecten van schrijfprodukten onderscheiden en gespecificeerd worden. Ook wordt wel gewerkt met een beoordelingsschaal van in kwaliteit oplopende voorbeeldopstellen. Om te garanderen dat beoordelaars analytische richtlijnen en/of schalen van voorbeeldopstellen op dezelfde en bedoelde wijze gebruiken wordt met name in taalbeheersingsonderzoek werk gemaakt van *training van beoordelaars.*

Met het gebruik van analytische schema's, voorbeeldschalen en training is ervaring opgedaan in een onderzoek naar de aanpak en resultaten van het stelvaardigheidsonderwijs in het Nederlandse voortgezet onderwijs.[1] Op de opzet van de trainingsprocedure, de effecten van de training en op de tijdens

W.K.B. Koning (red.), *Taalbeheersing in theorie en praktijk.* Dordrecht 1985: Foris Publications, 345–352.

de training gebleken problemen wordt in het onderstaande nader ingegaan. Daarbij zal de vraag of training effect lijkt te hebben centraal staan. Tevens zal aandacht worden besteed aan problemen in de beoordelingstaak die zich blijkens de tijdens de training gevoerde discussies voordoen.

2. Schrijfopdrachten, beoordelingsrichtlijnen en beoordelingsschalen

Voorafgaand aan de bespreking van de trainingsprocedure is het dienstig stil te staan bij een belangrijk aspect van het bovengenoemde onderzoek naar het stelvaardigheidsonderwijs: de gebruikte schrijfopdrachten, de analytische richtlijnen voor de beoordeling en de schalen van voorbeeldopstellen.

Op 100 scholen voor voortgezet onderwijs verspreid door het hele land, hebben leerlingen van derde klassen LHNO, LTO, MAVO, HAVO en VWO gedurende drie lesuren schrijfprodukten vervaardigd. In totaal is gebruik gemaakt van 8 *schrijfopdrachten*; iedere leerling heeft geschreven naar aanleiding van 3 van deze 8 opdrachten: één 'funktionele' schrijfopdracht (uit een totaal van 4 opdrachten), één 'traditionele' opstelopdracht (uit een totaal van 3 opdrachten) en één briefopdracht. Dit artikel heeft betrekking op 5 van deze opdrachten: de 4 'funktionele' opdrachten en de briefopdracht.

De 4 *funktionele opdrachten* zijn korte, tot zeer korte schrijfopdrachten waarin beschrijving en informatieverstrekking centraal staan. Eén opdracht betreft de beschrijving van een fiets, ter informatie van een 'suikeroom' die een fiets cadeau wil geven; een andere opdracht betreft een zelfbeschrijving, bedoeld om herkenning door een nooit eerder ontmoete persoon te vergemakkelijken. Een derde opdracht bestaat uit een korte mededeling waarin een afspraak met een schoolhoofd wegens ziekte afgezegd moet worden. De vierde funktionele opdracht bestaat uit een kort briefje waarin gereageerd wordt op een advertentie voor vakantiewerk. De *briefopdracht* is een wat langere schrijfopdracht waarin het om het volgende gaat: aan een jongere medeleerling moeten aanwijzingen over opstelschrijven verstrekt worden, aanwijzingen die de kans op een goed cijfer van de docent(e) Nederlands verhogen.

Voor de schrijfopdrachten zijn *beoordelingsrichtlijnen* opgesteld in de vorm van een analytisch beoordelingsschema dat uit vier onderdelen bestaat. In de schema's worden eisen met betrekking tot de inhoud, de organisatie en de stijl van de opstellen onderscheiden en gespecificeerd. Daarnaast worden enkele, noodzakelijkerwijs vage richtlijnen voor een globale kwaliteitsindruk gegeven. Bij de *globale kwaliteitsindruk* wordt gevraagd om een oordeel waarin met name de inhoud, organisatie en stijl, maar ook spelling, interpunctie, grammatica en netheid afgewogen worden. Bij de beoordeling van *inhoud* gaat het bijvoorbeeld in het geval van de fietsopdracht om de vraag of het opstel voldoende informatie bevat om identificatie van de gewenste fiets te vergemakkelijken. Bij de beoordeling van *organisatie* gaat het om logische rangschikking van de gepresenteerde inhoudselementen; in het geval van de briefopdracht gaat het dan om het bijeenplaatsen van gelijksoortige of verwante adviezen en om de mate waarin de inhoudelijke opbouw geschraagd wordt door een corresponderende alinea-indeling. Bij de beoordeling van de *stijl* gaat het om zaken als variatie in zinsbouw en woordkeus, overgangen tussen zinnen en zinsverbindingen.

Als aanvulling op de analytische schema's zijn voor iedere schrijfopdracht voor de 4 beoordelingsdimensies *schalen van voorbeeldopstellen* geconstrueerd. Iedere schaal bestaat uit 3 opstellen die de extremen en het middelpunt van de bij het beoordelen te gebruiken 5-puntsschaal als het ware vastleggen. Deze voorbeeldopstellen lopen op in kwaliteit en vormen een schaal met ordinale eigenschappen. Deze voorbeeldopstellen zijn geselecteerd na een uitgebreide beoordeling van een 100-tal opstellen per opdracht door een drietal beoordelaars.

Analytische richtlijnen en voorbeeldopstellen leggen tezamen de kern van de beoordelingsprocedure vast. Bij het daadwerkelijke beoordelen dienen de beoordelaars ieder te beoordelen opstel te vergelijken met de voorbeeldschaal en met de richtlijnen. Om dit mogelijk te maken worden de opstellen 'dimensie-gewijs' beoordeeld: eerst allemaal globaal, dan op inhoud, dan op organisatie en tenslotte op stijl.

3. Opzet en effecten van de training van opstelbeoordelaars

De bij het onderzoek ingeschakelde beoordelaars zijn getraind in het gebruik van de analytische richtlijnen en van de voorbeeldopstellen. Bij de beoordeling van de funktionele opdrachten en de briefopdracht waren in totaal 12 docenten betrokken, afkomstig uit het LBO, MAVO, HAVO en VWO. Bij ieder van de 4 funktionele opdrachten beoordeelden 3 docenten. Bij de briefopdracht beoordeelden alle 12 docenten in 4 jury's van 3 beoordelaars ieder een deel van de opstellen.

De training ging als volgt in zijn werk. Voorafgaand aan de trainingsbijeenkomst beoordeelden de docenten thuis een 12-tal opstellen, na bestudering van de richtlijnen en voorbeeldopstellen. Op de trainingsbijeenkomst werden de oordelen van de beoordelaars èn van de 2 uitvoerders van het onderzoek met elkaar vergeleken. Vervolgens werden de gevallen waarin niet alle oordelen geheel of bijna geheel overeenkwamen besproken. Dat waren er vele . . ., bij ieder van de 5 opdrachten werden zo alle 12 trainingsopstellen wel één of meerdere malen besproken: soms voor inhoud, soms voor organisatie, dan weer voor het globale oordeel of de stijl. Bij de bespreking kwamen tal van problemen bij het gebruik van de richtlijnen en de voorbeeldschalen ter sprake; problemen waarop in de discussie teruggekomen wordt.

Na de bespreking van de vooraf beoordeelde opstellen, die meestal een hele ochtend of een groot deel van een ochtend in beslag nam, werden ter plekke nog eens een klein aantal (tussen de 5 en 9) opstellen beoordeeld. Opnieuw werd dan gesproken over verschillen in oordelen en over eventueel nieuw

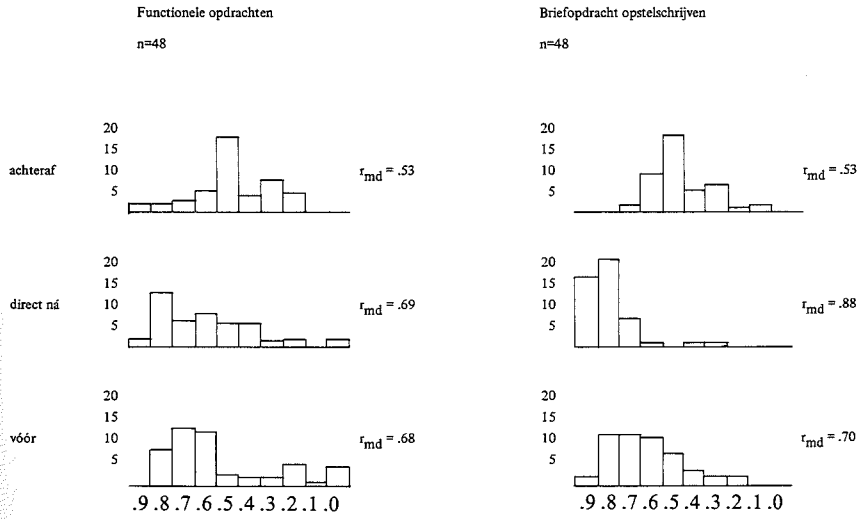
gerezen problemen. Na afloop van de training kregen de beoordelaars als 'huiswerk' ieder ruim 300 opstellen mee.

De hamvraag bij al deze inspanningen is natuurlijk de volgende: *heeft training van beoordelaars effect?* Wordt de overeenstemming tussen de beoordelaars door training verhoogd? En: blijft een eventuele verhoging ook na verloop van tijd gehandhaafd? Een waterdicht antwoord op deze vragen kan gezien het gehanteerde design niet geleverd worden: de overeenstemming tussen de beoordelaars kan weliswaar tweemaal op drie momenten (vóór training, direct ná training en enige tijd later) middels een a-selecte steekproef van 36 opstellen uit het eigenlijke beoordelingswerk bepaald worden, een controlegroep ontbreekt evenwel (we hebben geen beoordelaars zonder training dezelfde opeenvolging van beoordelingsactiviteiten laten verrichten). Er is dus sprake van een variant van het 'one-group pretest- posttest design'. In dit design zijn voor de gehanteerde 'treatment' tal van alternatieve verklaringen denkbaar. Heel ernstig is dit probleem niet: op een experimentele toetsing van de effectiviteit van training waren we niet uit; wel wilden we nagaan of training met een verhoging van de overeenstemming tussen beoordelaars gepaard gaat. In *tabel 1* is de opzet van het onderzoek nader uitgewerkt.

opdracht	aantal beoordelaars	aantal beoordeelde opstellen		
		vóór	direct ná	achteraf
beschrijving fiets	3	12	5	36
zelfbeschrijving	3	12	9	36
mededeling schoolhoofd	3	12	9	36
sollicitatiebrief	3	12	9	36
brief aan medeleerlingen	4 X 3	12	5	36

Tabel 1: Opzet vergelijk overeenstemming vóór training, direct ná training en achteraf, tijdens het eigenlijke beoordelingswerk

Bij het vergelijken van de overeenstemming richten we de blik op de overeenstemming in de rangordeningen van de opstellen die de beoordelaars aanbrengen. Op de 3 momenten vóór, direct ná en achteraf hebben we de 'Spearman rangorde-correlaties' tussen alle verschillende beoordelaars uitgerekend, voor de oordelen op alle dimensies van het analytische schema. Voor de bovenste helft van *tabel 1* levert dat 3X48 rangorde-correlaties op. (N.B. 4 opdrachten × 3 beoordelaarsparen × 4 correlaties (globaal, inhoud, organisatie, stijl)). Voor de onderste helft van de tabel geldt hetzelfde: ook hier vallen 3X48 rangorde-correlaties te berekenen. In *figuur 1* geven we de verdelingen van de correlaties op de verschillende tijdstippen weer, voor de funktionele opdrachten tezamen en voor de briefopdracht.



Figuur 1: Verdelingen en mediane rangorde-correlaties vóór training, direct ná training en achteraf

In *figuur 1* zijn de correlaties ingedeeld in intervallen van .00 – .09, .10 – .19, .20 – .29, enzovoort. Bij iedere verdeling is de mediane rangorde-correlatie weergegeven. De data vertonen op het oog een interpreteerbaar patroon. De correlaties vóór training in de laagste regionen (rechts op de balk) verschuiven naar links direct ná de training. De meeste hoge correlaties (>.80) zijn te vinden direct ná training. Achteraf tijdens het uiteindelijke beoordelen – dat plaatsvond in een tijdsverloop van ± één week na training – concentreren de correlaties zich in het midden van de schaal, in het interval van .50 – .59.

Langs meer formele weg zijn we nagegaan of de verschillen in correlaties vóór, direct ná en achteraf statistisch *significant* zijn. Daartoe is een toetsingsprocedure in twee stappen uitgevoerd. Eerst is een 'twee-wegs variantieanalyse' op rangordeningen van correlaties uitgevoerd, volgens de procedure van Friedman (Siegel 1956). Daarmee wordt getoetst of de correlaties onder de 3 condities significant verschillen. Deze procedure komt er heel in het kort op neer dat voor ieder beoordelaarspaar nagegaan wordt onder welke conditie – hier: op welk moment – de rangorde-correlatie het grootst is. Aan deze correlatie wordt het hoogste ranggetal (3) toegekend. De één na hoogste correlatie krijgt het ranggetal (2) toegekend, de laagste het getal (1). Per

conditie worden de ranggetallen gesommeerd en gekwadrateerd. Vervolgens wordt bepaald of deze kwadratensom significant verschilt van de onder kans te verwachten som. Toetsing vindt plaats op het gebruikelijke significantieniveau ($\alpha = .05$). Pas als deze 'nulhypothese van geen verschil' verworpen is voor de drie condities tezamen, kunnen de condities paarsgewijs met elkaar vergeleken worden, bij $\alpha = .05$. Wordt de 'nulhypothese van geen verschil' niet verworpen, dan dient bij paarsgewijze vergelijking het significantie-niveau aangepast te worden om a posteriori kapitalisatie op kans te vermijden. Hier kiezen we dan voor een α van .025. Voor de paarsgewijze vergelijking gebruiken we de 'sign test' of tekentoets (Siegel 1956), waarmee vastgesteld kan worden of de verdeling van hogere en lagere correlaties per beoordelingspaar voor de verschillende condities aan kans toegeschreven kan worden. Met behulp van de procedure volgens Friedman kan bij tweezijdige toetsing de 'nulhypothese van geen verschil' tussen de 3 condities voor beide opdrachten verworpen worden (χ^2_r is gelijk aan 7.88 respectievelijk 62.04 bij df= 2, in beide gevallen is $p<.05$).

Met de tekentoets kan nu dus bij $\alpha= .05$ via eenzijdige toetsing nagegaan worden of de overeenstemming direct ná training hoger is dan de overeenstemming voor training. De corresponderende hypothese kan voor de funktionele opdrachten *niet* verworpen worden, voor de briefopdracht *wel*. (De z-waarden zijn -.433 respectievelijk -4.47.) Dit resultaat vertoont duidelijke overeenkomst met het verloop in de waarden van de mediane correlaties in *figuur 1*. Alleen bij de briefopdracht gaat training dus gepaard met een significante verbetering van de overeenstemming. Bij de funktionele opdrachten is de op het oog al zwakkere trend niet significant.

Via tweezijdige toetsing kan nagegaan worden of het overeenstemmingsniveau dat direct na training bereikt is ook achteraf gehandhaafd blijft. De corresponderende 'hypothese van geen verschil' moet helaas in beide gevallen verworpen worden (de z-waarden zijn -3.32, respectievelijk 6.78). Blijkbaar wordt het niveau van overeenstemming niet gehandhaafd. Hoewel dit niet erg verrassend is, mag hier toch van een tegenvallend resultaat gesproken worden.

Tenslotte valt nog de vergelijking te maken tussen het overeenstemmingsniveau vooraf en het overeenstemmingsniveau achteraf. Verwacht mag worden dat de overeenstemming achteraf beter is dan de overeenstemming voorafgaand aan training. Ook hier vallen de resultaten tegen. Bij de funktionele opdrachten is geen sprake van verschil ($z = -1.29$), bij de briefopdracht is de overeenstemming achteraf lager dan vooraf ($z = 3.61$)! Een resultaat dat tegengesteld is aan wat redelijkerwijs verwacht mag worden. Een resultaat ook dat om verklaring vraagt, hoe speculatief die ook moge zijn.

5. Discussie

Er zijn duidelijke aanwijzingen dat training van opstelbeoordelaars op korte termijn vruchten af kan werpen. Blijvend zijn de verbeteringen evenwel niet. Het lijkt er zelfs op dat de beoordelaars hun taak na verloop van tijd slechter gaan uitvoeren dan ze vooraf ongetraind deden. Vooral dit laatste is een vreemd resultaat. Eén mogelijke verklaring kan snel terzijde geschoven worden: de training heeft de beoordelaars op het verkeerde been gezet. Als de training eerder verwarring dan verduidelijking had opgeleverd zou dat ook direct ná training moeten blijken. De deelnemende docenten waren bovendien zelf positief gestemd over de trainingsbijeenkomsten, hetgeen een andere contra-indicatie is voor deze verklaring. Bij de aanduiding 'ongetraind' is overigens nog wel een kanttekening op zijn plaats. 'Ongetraind' is niet 'ongeïnstrueerd'. Vooraf hebben de beoordelaars wel de uitvoerige richtlijnen en voorbeeldopstellen bestudeerd.

Een meer plausibele verklaring zou de volgende kunnen zijn. De eerste 12 trainingsopstellen en ook de 5 à 9 opstellen die direct ná training beoordeeld zijn, waren zo gekozen dat ze een flinke spreiding in kwaliteit bevatten. Misschien was het nodig geweest om de training ook meer op het midden van de schaal te richten. In het uiteindelijk te beoordelen opstelmateriaal zitten natuurlijk meer middelmatige dan extreem goede of slechte schrijfprodukten. Het waarschijnlijkst lijkt het echter dat de complexiteit van de beoordelingstaak de beoordelaars gaandeweg parten is gaan spelen. Tijdens de trainingen kwamen in de discussie over de uiteenlopende oordelen tal van problemen naar voren. Daarbij waren bekende problemen, zoals het *signifisch effect* (vaagheid van de richtlijnen) en het *halo-effect* (overschaduwing van een te beoordelen aspect door een opvallende andere eigenschap). Ook waren verschillen in strengheid duidelijk merkbaar, verschillen die voor een deel teruggevoerd leken te kunnen worden op ervaringen uit de eigen beoordelingspraktijk van de deelnemende docenten.

Andere, minder bekende problemen deden zich ook voor. Beoordelingsdimensies blijken bijvoorbeeld niet altijd onafhankelijk te zijn: opstellen met bijzonder weinig inhoud zijn lastig op organisatie te beoordelen. Beoordelingsdimensies zijn verder niet altijd uni-dimensionaal: bij een dimensie zijn meerdere aspecten van belang, bij organisatie bijvoorbeeld logische opbouw en alinea-gebruik. Hierdoor ontstaan combinatie- en wegingsproblemen.

De veronderstelling lijkt gewettigd dat door meer intensieve en vooral meer herhaalde training een hoger niveau van overeenstemming bereikt en gehandhaafd kan worden. En dat opent perspectieven op acceptabele schrijfvaardigheidsmetingen. Want ook bij de onderhavige, onvolmaakte aanpak is door de combinatie van analytische richtlijnen, beoordelingsschalen en training al in veel gevallen een acceptabele tot goede jury-betrouwbaarheid bereikt. De homogeniteits-coëfficienten α voor de 4 beoordelingsdimensies en de 4 jury's bij de funktionele opdrachten en bij de briefopdrachten belopen een mediane waarde van .78. Hopeloos is het dus met de beoordeling van de stelvaardigheid zeker niet gesteld.

Noten

1. Dit onderzoek wordt uitgevoerd aan de Stichting Centrum voor Onderwijsonderzoek van de Universiteit van Amsterdam met subsidie van SVO.

Literatuur

Siegel, S.
1956 *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Analytische beoordeling van de samenvattingsopdracht C.S.E. Nederlands voor het VWO

Ton Hendrix en Piet Sanders

1. Inleiding

"Geef van de onderstaande tekst, die ongeveer 2200 woorden telt, een samenvatting in maximaal 500 woorden".

Zo luidt de opdracht boven een betogende tekst die VWO-leerlingen in $2^1/_2$ uur tijd voor hun C.S.E. Nederlands moeten samenvatten. Het is een eenvoudig geformuleerde opdracht voor een complexe taakstelling, zowel voor de examenkandidaat als voor de beoordelaars. In de 35 jaar dat de samenvattingsopdracht deel uit maakt van het C.S.E. voor het VWO (vroeger Gymnasium en HBS-b) is er al het nodige gezegd over de pedagogische waarde, over vakinhoudelijke, vakdidactische en – de laatste jaren – de toetstechnische problemen en onduidelijkheden van dit examenonderdeel.

In deze bijdrage zal verslag worden gedaan van een op het Cito verricht onderzoek naar de mogelijkheden de beoordelaarsovereenstemming bij de samenvattingsopdracht te verbeteren door middel van een daartoe ontwikkeld analytisch beoordelingsschema. Eerst komt de ontwikkeling van het tekstafhankelijk, analytisch beoordelingsschema dat bij het onderzoek gebruikt is aan de orde, daarna volgt een korte weergave van de opzet en de resultaten van het onderzoek.

2. De huidige situatie

Allereerst geef ik een overzicht van de huidige stand van zaken bij de beoordeling van de samenvatting op het VWO-eindexamen. De leerlingen krijgen een betogende tekst met de genoemde opdracht om daarvan een samenvatting te maken van maximaal 500 woorden. Ofschoon in de opdracht niet is aangegeven welk type samenvatting verlangd wordt, is zo langzamerhand wel duidelijk geworden uit het correctievoorschrift aan docenten, dat het gaat om wat Van Eemeren et al. (1975) een 'informatieve samenvatting' noemen: een complete tekst die aanzienlijk korter is dan de oorspronkelijke tekst en die de hoofdzaken en de hoofdlijnen daarvan weergeeft.

De docent-beoordelaar wordt geacht zijn oordeel over de samenvatting te baseren op de door de Centrale Examencommissie verstrekte *Aanwijzingen voor de beoordelaar.* De aanwijzingen komen op het volgende neer:

W.K.B. Koning (red.), *Taalbeheersing in theorie en praktijk.* Dordrecht 1985: Foris Publications, 353–361.