Beoordelen van docenten door studenten* Een literatuurstudie

S. J. M. BLOM en W. F. LANGERAK
Bureau Onderzoek van Onderwijs, Rijksuniversiteit Leiden

Samenvatting

Het artikel is een verslag van een literatuuronderzoek naar de kwaliteit van de oordelen die studenten geven over hun docenten. Een globale analyse van docenttaken geeft aan, dat studenten (mee) kunnen oordelen over het leermateriaal, de presentatie van het onderwijs en de rechtvaardigheid van de toets.

Het bleek, dat vrijwel uitsluitend literatuur uit de V.S. gevonden kon worden en dat deze literatuur vrijwel uitsluitend de presentatie behandelt.

De validiteit van de studentoordelen lijkt op het eerste gezicht niet slecht. Echter, bij nadere beschouwing blijkt een aanzienlijk deel van de variantie toegeschreven te kunnen worden aan verschillen tussen vakken in plaats van verschillen tussen docenten. Gezien de methodologische problemen die ontstaan als geprobeerd wordt om deze effecten te scheiden, lijkt in de praktijk een valide oordeel van studenten over de kwaliteit van hun docent nauwelijks haalbaar.

De betrouwbaarheid van de studentoordelen is zeer aanvaardbaar. Daarbij moet echter wel rekening gehouden worden met het feit, dat extra variantie geïntroduceerd wordt door verschillen tussen vakken.

De conclusie is, dat de oordelen van studenten van onvoldoende kwaliteit zijn om een belangrijke rol te spelen in beslissingen rond aanstelling en bevordering.

1. *Inleiding*

Aan alle universiteiten is een veelgehoorde klacht, dat te weinig aandacht wordt besteed aan het beoordelen van de doceerkwaliteiten van wetenschappelijk medewerkers, zowel bij aanstellingsprocedures als bij beslissingen rond vaste aanstelling en bevordering van medewerkers. Op een aantal plaatsen is men, zij het met enige voorzichtigheid, aan het proberen vaster omlijnde procedures voor deze beoordeling te ontwikkelen, onder andere aan de Universiteit van Leiden (zie bijstellingsnota meerjarenafspraken 1979–1982).

In het nu volgende artikel wordt, op basis van de beschikbare literatuur, nagegaan in hoeverre het oordeel van studenten over hun docenten opgenomen zou kunnen worden in deze procedures. Daarbij ging het ons vooral om de kwaliteit van de studentoordelen, zoals die tot uitdrukking komt in de betrouwbaarheid en validiteit.

In Nederland is bijzonder weinig ervaring opgedaan met het gebruik van oordelen die studenten geven over docenten. Hoewel incidenteel het oordeel van studenten gevraagd wordt (o.a. Van den Bergh e.a. 1978; Los, 1977; Geensen, 1969), hebben wij geen publikaties gevonden waarin onderzoek naar de kwaliteit van deze oordelen gerapporteerd wordt. In de Verenigde Staten is wel veel onderzoek verricht. Dit onderzoek blijft echter in grote trekken beperkt tot het beoordelen van het gedrag van docenten tijdens hoorcolleges.

De opzet van het artikel is als volgt. Eerst worden enige definities geformuleerd, die mede tot doel hebben om een overzicht te krijgen van de probleemstelling (hoofdstuk 2). Hoofdstuk 3 geeft een overzicht van gegevens uit de literatuur en in hoofdstuk 4 worden conclusies getrokken uit de beschikbare gegevens.

2. *Definities en begrippen*

Dit artikel handelt over de beoordeling van *onderwijskwaliteiten* van wetenschappelijke medewerkers. Dat de *onderzoekstaken* buiten beschouwing blijven houdt geen waardeoordeel over deze taken in, maar is slechts een gevolg van de beperking van het artikel. Een verdere beperking is, dat wij zullen spreken over oordelen van *studenten*. Om na te gaan

*Dit artikel is het verslag van een onderzoek, verricht in opdracht van het College van Bestuur van de Rijksuniversiteit Leiden.

welke plaats deze oordelen innemen in het geheel van beoordeling en promotie, dienen de volgende vragen gesteld te worden:
- wie beoordelen de docent;
- welke taken van de docent dienen beoordeeld te worden;
- wie nemen de beslissingen;
- welke gevolgen hebben deze beslissingen?

De beoordelaars

In het kader van de beoordeling van de onderwijskwaliteiten van docenten kunnen we vier groepen beoordelaars onderscheiden, naar de positie die zij innemen ten opzichte van het onderwijsproces.

1. *Studenten*. Met studenten wordt hier de groep 'consumenten' van het onderwijs bedoeld. Als in het vervolg over studenten gesproken wordt, bedoelen we daarmee alleen die studenten, die daadwerkelijk het onderwijs van de betrokken docent volgen, dan wel benaderd worden omdat zij in het verleden dat onderwijs gevolgd hebben. Studenten zijn om twee redenen belangrijk. Op de eerste plaats is het onderwijs voor hen bedoeld en moet op hen afgestemd zijn en op de tweede plaats zijn zij in het algemeen de enige regelmatige observatoren van de docent bij de uitoefening van een aantal doceertaken.
2. *Collega's*. In ruime zin de 'gelijken' van de docent wat betreft academisch niveau en werkkring. In engere zin diegenen, die op een voldoende niveau van het vakgebied van de betrokken docent op de hoogte zijn om zich een oordeel te kunnen vormen over de kwaliteit van de leerstof.
3. *Bestuurders*. Zij die verantwoordelijkheid dragen, zowel voor de inhoud als voor de kwaliteit van het werk van de betrokken docent. In de eerst kwaliteit zijn zij beoordelaars: komt de inhoud van het werk overeen met de gewenste inhoud; in de tweede kwaliteit zijn zij degenen die de uiteindelijke beslissing nemen.
4. *Externe beoordelaars*. Deze term wordt hier gebruikt voor een groep mensen die door hun taak en/of kennis een oordeel kunnen geven over bepaalde aspecten van de doceertaak. Hun oordeel zal in het algemeen op een ander abstractieniveau liggen dan dat van de andere beoordelaars. We kunnen bij deze groep bijvoorbeeld denken aan medewerkers van een onderwijsresearchbureau of soortgelijke diensten, dan wel aan facultaire medewerkers met een taak op dit gebied.

De hier genoemde groepen kunnen elkaar overlappen. Zo kan een student-assistent zowel collega als student zijn met betrekking tot een bepaalde docent. Ook kan een collega tevens bestuurder zijn. De onderscheidingen blijven echter geldig als we kijken naar de kwaliteit waarin iemand wordt aangesproken. Als we vragen om het oordeel van studenten, dan gaat het erom hoe de 'consument' het onderwijs ervaart. Daarbij gaat het dus alleen om de studenten die daadwerkelijk dat onderwijs volgen of gevolgd hebben; de vraag wordt niet gesteld aan de studentengroep in zijn algemeenheid. Zo kan de studentenfractie van de faculteitsraad niet het gevraagde oordeel geven. Zij zijn naar hun aard bestuurders en moeten, als zij oordelen, hun oordeel geven in de bestuurlijke sfeer.

Ten overvloede nog dit: beslissingen worden uitsluitend door de bestuurders genomen. Zij kunnen daarbij echter gebruik maken van informatie van alle vier de groepen.

De taken van de docent

Bij het beschrijven van de taken van de docent wordt hier uitgegaan van de taken die zoal te vervullen zijn binnen het onderwijsproces. Daarmee is niet gezegd dat ook elke docent alle taken zal vervullen. Daarnaast zullen een aantal taken in groepsverband worden uitgevoerd, waarbij de inbreng van individuele docenten moeilijk te onderscheiden is. In hoeverre het al dan niet vervullen van taken van invloed is op een mogelijke promotie is een probleem op het gebied van de taakomschrijving (functiewaardering) en komt als zodanig in dit artikel niet aan de orde. Wij beperken ons tot de beoordeling van vervulde taken.

Wel zal dit punt later nog aan de orde komen, voorzover de beoordelingen van de verschillende taken elkaar beïnvloeden.

De volgende taken kunnen worden onderscheiden:
- Het kiezen van de *leerstof*. De docent moet de leerstof kiezen, aan de hand waarvan hij de gestelde leerdoelen wil verwezenlijken. Hieronder hoort ook de specificatie van de leerdoelen aan de hand van de eisen, gesteld door faculteit en vakgroep. De vraag in hoeverre de gekozen leerstof aan de eisen voldoet kan het best beoordeeld worden door deskundigen op het betrokken vakgebied.
- Het samenstellen van *leermateriaal*. In het vorige punt was aan de orde *welke* leerstof aan de studenten aangeboden wordt. In dit punt gaat het er om *hoe* deze leerstof wordt overgedragen. Op een of andere manier zal de leerstof aan de student bekend gemaakt moeten worden. Naast de directe presentatie, die in het volgende punt ter sprake komt gebeurt dat door leermateriaal, vastgelegd in geschriften, literatuuropgaven, bundelingen van

artikelen en eventueel film, videoband, geluidsband, e.d. Bedoeld is hier al het materiaal, dat door de student zelfstandig gebruikt kan worden. Op de eerste plaats kan gesteld worden dat de leerstof door het materiaal geheel gedekt moet worden: alle leerstof moet erin te vinden zijn. Op de tweede plaats moet het materiaal didactische waarde hebben: het moet de studenten helpen om de leerdoelen te bereiken. Op de derde plaats zou gesteld kunnen worden, dat het materiaal de student moet stimuleren bij zijn studie (waarbij overigens de grens tussen de tweede en de derde eis niet scherp getrokken kan worden). De dekking van de stof kan het best beoordeeld worden door deskundige collega's. Zij begrijpen wat de leerstof behelst en zij kunnen zien of alles in het materiaal voorkomt. Zij zullen echter niet altijd in staat zijn om de didactische waarde van het materiaal te beoordelen. 'The proof of the pudding is in the eating': studenten kunnen wel beoordelen of zij met het materiaal overweg konden. Uiteindelijk is het materiaal voor hen vervaardigd, zij zijn derhalve de beoordelaars bij uitstek. Ook de mate waarin het materiaal stimuleert kan het beste door de studenten beoordeeld worden. Bij het beoordelen van de didactische waarde zouden eventueel ook externe deskundigen ingeschakeld kunnen worden. Hun oordeel zal echter minder direct zijn dan dat van de studenten. Op grond van leerpsychologische overwegingen of vergelijkend onderzoek zal de deskundige op een aantal aspecten voorspellingen kunnen doen over de bruikbaarheid van het materiaal. Als het oordeel van de studenten beschikbaar is lijkt dat echter meer voor de hand te liggen.

- *De presentatie*. Op een of andere wijze zal de stof door een docent aan de studenten gepresenteerd worden via hoorcolleges, werkgroepen, practica, e.d., kortom alle vormen van contactonderwijs. Bij de beoordeling kunnen we een onderscheid maken tussen de structuur van de presentatie, de functie die de presentatie vervult in het leerproces en de presentatie zelf, de manier waarop de docent het brengt. De presentatie hoeft niet altijd dekkend te zijn voor de gehele stof. Zelfstudie speelt een belangrijke rol in de universitaire studie en hoeft niet altijd vooraf gegaan of gevolgd te worden door een mondelinge behandeling. Kwaliteit kan hier het best omschreven worden als het vervullen van een 'passende functie' in het gehele proces. Welke functie dat is, hangt voor een gedeelte af van de opbouw van het onderwijs en van de gebruikte methode (zie bijv. het eindrapport van de Commissie Kwaliteitsmaatstaven Onderwijs, 1976). In hoeverre de vervulde functie in de praktijk ook 'passend' blijkt te zijn kan het meest direct door de studenten beoordeeld worden. De presentatie moet immers passen in *hun* leerproces. Ook externe deskundigen kunnen een oordeel vormen over de opbouw van de presentatie. Hierbij geldt dezelfde redenering als beschreven bij het punt over leermateriaal: als goede studentoordelen beschikbaar zijn lijken die bruikbaarder. De presentatie zelf is bij uitstek het terrein van het studentoordeel. De studenten zijn de enige regelmatige observatoren van de docent. Dit is het gebied dat door de traditionele 'student-ratings' in de Verenigde Staten wordt bestreken. Punten ter beoordeling zijn onder andere de duidelijkheid van de presentatie en de mate waarin de student gestimuleerd wordt.
- *De toets*. De eisen, te stellen aan de toets, kunnen in twee groepen worden ingedeeld. Ten eerste moet de toets goed meten en ten tweede moet de toets rechtvaardig zijn ten opzichte van de studenten. De eerste eis betekent, dat de toets voldoende dekkend moet zijn voor de stof, hetgeen door deskundige collega's beoordeeld kan worden. Tevens moet de toets aan een aantal meettechnische eisen voldoen. Dit laatste kan beoordeeld worden door externe deskundigen. De rechtvaardigheid hangt samen met de zwaarte van de toets en met de mate waarin de student zich tijdens het onderwijs een juist beeld van de eisen heeft kunnen vormen. Uiteraard is dit weer ter beoordeling van de studenten.

Op het eerste gezicht lijkt het voor de hand liggend, ook de uitslag van de toets te hanteren als maat voor de kwaliteit van de docent. Echter: de uitkomst van de toets is geen maat voor de activiteiten van de docent, maar van de activiteiten van de student. Onderzoek naar de relatie tussen de kwaliteit van het onderwijs en studieresultaten, gemeten met een eindtoets, maken het meer dan aannemelijk, dat de kwaliteiten van de docent maar zeer ten dele in de resultaten van de studenten weerspiegeld worden. Veel factoren die niet onder controle van de docent zijn, zoals zelfstudie en invloed van andere vakken spelen een rol. Daarentegen heeft de docent aan een universiteit voor een belangrijk deel de uitslagen van de toets zelf in de hand.

Deze positie wordt niet door iedereen ingenomen. Veel onderzoekers maken, ons inziens ten onrechte, geen duidelijk onderscheid tussen activiteiten van de docent en activiteiten van de student. Integendeel: regelmatig wordt de effectiviteit van het onderwijs en van de docent geoperationaliseerd in studieprestaties van de student. Een bekend voorbeeld hiervan

is de studie van Dubin en Taveggia (1968), waarin onderwijsmethoden vergeleken worden met het examencijfer als criterium. Een uitgebreidere behandeling van de problemen die met deze operationalisatie samenhangen wordt onder andere gegeven in Blom (1978) en Vos (1978).

Het oordeel: formatief of selectief?

Informatie over de taakvervulling van docenten kan in grote trekken op twee manieren gebruikt worden. In de eerste plaats kan de informatie voor de docent zelf aanleiding zijn om zijn werkwijze te veranderen, aan te vullen of te verbeteren. De informatie staat dan uitsluitend ter beschikking van de docent, die zelf beslist wat er verder mee gedaan wordt. Deze wijze van gebruik zullen we in het vervolg formatief noemen.

Een tweede manier is de selectieve wijze van gebruik. De informatie staat dan ter beschikking van de bestuurders die beslissingsbevoegdheid hebben ten aanzien van de docent. Op grond van de informatie besluiten zij tot aanstelling, ontslag, promotie, e.d.

De manier waarop de informatie gebruikt wordt is van belang bij het vaststellen van de eisen waaraan de meetinstrumenten moeten voldoen. Bij de selectieve meting staat de nauwkeurigheid en de vergelijkbaarheid van oordelen voorop. Een oordeel is nooit absoluut, maar relatief ten opzichte van vergelijkbare docenten. Nauwkeurigheid is vereist om recht te doen aan de belangen van de betrokkenen. Gebrek aan nauwkeurigheid kan misschien deels worden opgevangen door aan de docent een recht op wederwoord toe te staan, waarbij hij nieuwe informatie kan toevoegen.

Bij formatief gebruik, als er geen dwingende consequenties aan het oordeel verbonden worden, is de nauwkeurigheid minder doorslaggevend. Hèt criterium is hier de bruikbaarheid. De docent zal alleen iets aan de informatie hebben, als hij op grond daarvan tot actie kan overgaan. Naast interpreteerbaarheid is daarvoor van belang, dat informatie loskomt over die punten, waarover de docent voldoende controle kan uitoefenen om zijn werkwijze te veranderen. Een docent die hoort dat zijn bordschrift slecht leesbaar is kan daar wat aan doen. Als een docent echter hoort dat hij niet geslikt wordt omdat zijn gezicht de studenten niet aanstaat, doet hij daar weinig mee.

Beoordeling van docenten kan ook in twee fasen geschieden. Als de prestaties van de docent onvoldoende worden geacht kan hem een bepaalde tijd worden gegund om hieraan iets te doen. In die periode kan de informatie gezien worden als formatief. Slechts als binnen de periode geen afdoende verbetering is opgetreden komt de selectieve functie naar voren. In feite werken natuurlijk veel personeelsbeoordelingssystemen op deze manier. Zelden zal een werknemer onmiddellijk op straat worden gezet als zijn werk niet bevalt. Variatie bestaat wel in de mate waarin de formatieve functie expliciet wordt gemaakt, dat wil zeggen, de mate waarin wordt aangegeven op welke terreinen de tekorten liggen en hoe ze kunnen worden weggewerkt.

Of de beschikbare informatie selectief gebruikt kan worden is afhankelijk van de kwaliteit van de informatie. Dat neemt niet weg, dat bestuurders in de situatie verkeren, waarin selectieve beslissingen genomen zullen worden, ongeacht de kwaliteit van de beschikbare informatie. Dat kan dilemma's opleveren, die echter meer politiek dan wetenschappelijk van aard zijn. Wij zullen ons dan ook beperken tot een oordeel over de kwaliteit van de informatie en niet ingaan op het beslissingsproces.

In dit hoofdstuk hebben we een overzicht gegeven van de taken van de docent, de informatie die van belang kan zijn met betrekking tot de vervulling van die taken en het gebruik van deze informatie.

In het tweede deel van dit artikel gaan we nader in op de informatie die studenten kunnen leveren over de wijze waarop de docenten hun taak vervullen. Zoals hiervoor vermeld, komen daarvoor de volgende aspecten in aanmerking:
- de didactische waarde van het leermateriaal
- de structuur en uitvoering van de presentatie
- de rechtvaardigheid van de toets.

Bij het bestuderen van de literatuur is de meeste aandacht gegeven aan de bruikbaarheid van de studentoordelen in het kader van personeelsbeoordeling. Daarbij zijn zowel de inhoud als de kwaliteit van het oordeel van belang.

3. *Oordelen van studenten*

3.1. *Inleiding*

In dit hoofdstuk wordt een overzicht gegeven van wat er uit de literatuur bekend is over de betrouwbaarheid (hoe nauwkeurig is het studentoordeel?) en validiteit (waarover zegt het studentoordeel iets?) van studentoordelen. We beginnen met een bespreking van de validiteit van studentoordelen.

3.2. *Validiteit van studentoordelen*

Deze paragraaf valt in drie delen uiteen. Ten eerste zullen we aandacht besteden aan de inhoud van

beoordelingsvragenlijsten: op wat voor soort zaken hebben vragen uit een beoordelingslijst betrekking. Vervolgens geven we een kort overzicht van de richtingen waarin men studentoordelen heeft trachten te valideren: onderzoek naar de relatie van studentoordelen met diverse soorten variabelen.

Tenslotte zullen we op enkele studies, die zich qua vraagstelling en/of opzet van de massa onderscheiden, wat nader ingaan. Deze studies doen nog al wat afbreuk aan de vermeende validiteit van studentoordelen.

3.2.1. *Inhoud van beoordelingsvragenlijsten*

Een van de manieren waarop de validiteit van studentoordelen kan worden onderzocht bestaat hieruit dat beoordelingsvragenlijsten op impliciet aanwezige dimensies of factoren kunnen worden nagezocht. In een overzichtsartikel van Kulil & McKeachie (1975) worden de resultaten van 11 van dergelijke studies, betrekking hebbend op verscheidene vragenlijsten, besproken. Naar Kulik & McKeachie concludeerden, zijn de volgende vier factoren meestal in vragenlijsten aanwezig:

a) skill: bekwaamheid/doceervaardigheid van een docent;
b) structure: organisatie en voorbereiding van zowel de kursus als geheel als van afzonderlijke colleges;
c) work load/difficulty: zwaarte van een kursus voor de studenten, veeleisendheid van docent;
d) rapport: relatie van docent tot individuele student, vriendelijkheid van docent, aandacht voor de individuele student.

De eerste factor (skill) behelst een brede dimensie; de overige drie zijn meer specifiek. Behalve deze vier min of meer vaste factoren zijn er nog twee herhaaldelijk gevonden:

e) instructor-group interaction: mate waarin docent erin slaagt een atmosfeer te scheppen, waarin studenten hun meningen durven te ventileren, zowel tegenover de docent als tegenover elkaar;
f) feedback: de door de docent verstrekte informatie over de kwaliteit van door studenten gegeven antwoorden/geleverd werk.

De hierboven genoemde zes factoren hebben, met uitzondering van factor b, alle te maken met concreet doceergedrag tijdens contacturen. Dit betekent nog niet dat studenten over andere zaken geen oordeel zouden (kunnen) hebben. Het hangt er maar net vanaf welke vragen in beoordelingsvragenlijsten worden opgenomen. Zo zijn er enige vragenlijsten met aanvullende, facultatieve vragen over practica; ook vragen over de kwaliteit van gebruikte tekst- of studieboeken komen in sommige vragenlijsten voor (zie Werdell, 1966). Verder is er voor verscheidene vragenlijsten een factor gevonden, die te maken heeft met het oordeel van studenten over de rechtvaardigheid/kwaliteit van toetsen (Frey, 1973; Granzin & Painter, 1973; Haslett, 1976). Met andere woorden, opname van anderssoortige vragen dan die met concreet doceergedrag hebben te maken, leidt ertoe of kan ertoe leiden dat andere dan de gebruikelijke factoren worden gevonden. Dat betekent dat het aantal en soort dimensies, dat een vragenlijst bestrijkt, naar believen kunnen worden uitgebreid.

3.2.2. *Relatie van studentoordelen met diverse soorten variabelen*

Van diverse soorten variabelen is onderzocht of zij van invloed zijn op, dan wel samenhangen met studentoordelen over docenten. De literatuur is dusdanig omvangrijk, dat we ons tot enige algemene opmerkingen en conclusies moeten beperken. Hierbij verlaten we ons in belangrijke mate op het eerder genoemde overzichtsartikel van Kulik & McKeachie (1975), alsmede op een overzichtsartikel van Costin et al. (1971) en op een monografie van Page (1974) over studentoordelen.

Tamelijk veel onderzoek is er verricht naar de samenhang tussen studentoordelen en oordelen afkomstig van collega's en supervisoren. Over het algemeen vertoont het oordeel van deze beoordelaarscategorieën enige overeenkomst. Verder is gebleken dat doceervaardigheid, zoals die met studentoordelen wordt gemeten, niets zegt over de researchkwaliteiten van een docent. Ook is gebleken dat meer ervaren docenten over het algemeen iets betere beoordelingen krijgen dan minder ervaren docenten. Er zijn voorts enige aanwijzingen dat training van docenten, of dat nu via een algemene kursus gebeurt (Costin, 1968) dan wel meer specifiek plaatsvindt voor die punten die door studenten als matig of zwak zijn beoordeeld (Centra, 1973; McKeachie, 1969), tot gunstiger studentoordelen leidt. Het verstrekken van studentoordelen zonder dat daaraan een training wordt gekoppeld lijkt daarentegen niet tot gunstiger studentoordelen te leiden (Aleamoni, 1978).

In veel studies is de vraag onderzocht of het tentamencijfer, dat men verwacht te behalen of behaald heeft, van invloed is op het oordeel over een docent. Over het algemeen is er zowel voor afzonderlijke docenten als voor groepen docenten geen sterke samenhang tussen tentamencijfer en studentoordeel

aantoonbaar. Voor groepen docenten is in enkele gevallen een sterk verband, hetzij positief (Frey, 1973; Gessner, 1973) hetzij negatief (Bendig, 1953; Rodin & Rodin, 1972), tussen gemiddeld tentamencijfer en klasseoordeel over een docent gevonden, maar deze studies hebben over het algemeen betrekking op kleine groepen docenten (5–20). Dat er geen duidelijk verband is tussen tentamencijfer en studentoordeel is, zoals we reeds in het vorige hoofdstuk hebben gesteld, niet zo verwonderlijk. Veel factoren, waarop een docent geen of slechts ten dele een greep heeft, zijn medebepalend voor het tentamencijfer dat een student behaalt.

Betrekkelijk veel onderzoek is ook gedaan naar variabelen die iets te maken hebben met de omstandigheden waaronder een docent lesgeeft. Zo is aangetoond dat verplichte vakken (required courses) over het algemeen iets lager worden beoordeeld dan niet-verplichte vakken (elective courses). Verder lijken ouderejaarsstudenten geneigd iets betere beoordelingen te geven dan jongerejaarsstudenten, alhoewel deze tendentie lang niet uit alle studies spreekt. Iets soortgelijks is dat meer geavanceerde kursussen over het algemeen iets betere beoordelingen krijgen dan meer elementaire kursussen. Studies naar de mogelijke invloed van de sexe van de beoordelaar op de aan een docent gegeven beoordeling geven over het algemeen geen verschil te zien tussen de oordelen van vrouwelijke en mannelijke studenten; ook lijkt er, als het om globale beoordelingen gaat, geen verschil te bestaan in de beoordeling van mannelijke en vrouwelijke docenten. Verder is herhaaldelijk gevonden dat klassen met meer dan 40 studenten iets lagere beoordelingen aan hun docenten geven dan kleinere klassen; bij klassen met minder dan 40 studenten is er geen systematisch verschil gevonden tussen kleinere en grotere klassen.

In nogal wat studies is onderzocht of persoonlijkheidskenmerken van studenten van invloed zijn op hun oordelen over docenten. De onderzochte variabelen waren nogal verschillend van aard: attitudes, emotionele en motivationale factoren, alsmede cognitieve variabelen zijn onderwerp van studie geweest. Veelal kon niet worden aangetoond dat persoonlijkheidsverschillen tussen studenten doorwerken in het oordeel over een docent. Iets soortgelijks geldt voor studies die gericht waren op de vraag of persoonlijkheidsverschillen tussen docenten tot uitdrukking komen in het studentoordeel. Voor de meeste persoonlijkheidsvariabelen bleek dit niet het geval te zijn. Enkele studies evenwel die zich bezighielden met communicatie variabelen (Bendig, 1955; Guthrie, 1954; Isaacson et al., 1963) leverden wel iets op. Uit deze studies komt naar voren dat die docenten als het meest doeltreffend door studenten worden gezien, die verbaal begaafd, enthousiast en expressief zijn, en verder een brede, in het bijzonder een culturele belangstelling aan de dag leggen.

Bedenkend dat het soort studies, als hierboven besproken, veelal tot doel hebben na te gaan of bepaalde variabelen een storende invloed hebben op het studentoordeel over een docent, kunnen we concluderen dat, alhoewel er soms zwakke verbanden aantoonbaar zijn, tentamencijfers, lesomstandigheden, persoonlijkheidsverschillen tussen studenten, alsook persoonlijkheidsverschillen tussen docenten geen echt storende invloed hebben op het studentoordeel over een docent.

3.2.3. *Enkele probleemgebieden*

Hieronder gaan we nader in op enkele studies, waarin vanuit verschillende gezichtspunten een minder gunstig oordeel wordt geveld over de validiteit van studentoordelen. Verscheidene onderzoekers hebben zich beziggehouden met de vraag of de inhoud van een kursus en de doceerstijl van een docent onafhankelijk van elkaar kunnen worden beoordeeld. Naftulin et al. (1973) lieten een nepdocent, in casu een acteur, opdraven voor een gehoor van psychiaters, psychologen, sociaal werkers en dergelijke. Voor dit gehoor hield de nepdocent een lezing over een niet-vertrouwd, maar voor de toehoorders toch wel interessant onderwerp. De lezing was nogal oppervlakkig van aard en stak niet al te best in elkaar; de nepdocent was geïnstrueerd het verhaal met veel enthousiasme en humor te brengen. Na de lezing gaven de toehoorders aan de hand van een vragenlijst hun oordeel over allerlei zaken, zowel de docent betreffende als de lezing zelf. Ondanks het feit, dat de lezing niet zo veel voorstelde, werd zowel over de docent als over de kwaliteit van het verhaal gunstig geoordeeld. Kennelijk laten zelfs vaklui zich gemakkelijk door het enthousiasme van een docent tot een gunstig oordeel – oordeel over de docent – verleiden, dat zich bovendien, en ten onrechte tot de inhoud van de kursus uitstrekt.

Iets soortgelijks is gevonden door Williams & Ware (1977). Binnen het kader van een algemene kursus werden een tweetal speciale lezingen gehouden, ook in dit geval door een acteur. De inhoud van de lezingen, dat wil zeggen het aantal punten dat ter sprake werd gebracht, kende drie gradaties: het volledige verhaal en onvolledige verhalen met respectievelijk ongeveer 50% en 15% van de punten uit het volledige verhaal. De duur van de verschillende verhalen werd zo goed als mogelijk gelijkgeschakeld door de niet-volledige verhalen op te vullen

met niet ter zake doend materiaal. Verder had de doceerstijl van de acteur twee gradaties: hij moest zich òf expressief/enthousiast gedragen òf juist niet. De zes mogelijke combinaties van informatiedichtheid en doceerstijl werden aan evenzovele groepen van studenten voorgeschoteld. Bij de eerste en tweede lezing werd voor iedere groep studenten eenzelfde combinatie van informatiedichtheid en doceerstijl aangehouden. Na ieder van de lezingen werd van de betrokken studenten gevraagd een oordeel te geven over een reeks van zaken: kennis van de docent, zijn presentatie, humor en enthousiasme, of men wat geleerd had, of men de lezing interessant vond, en zo meer. Deze gegevens werden per student in één totaalskore uitgedrukt. Uit het onderzoek bleek dat het oordeel van studenten in de expressieve conditie niet de informatiedichtheid van de lezing weerspiegelde. Dat was wel het geval in de niet-expressieve conditie; bij een volledig verhaal viel het oordeel gunstiger uit dan bij de meer oppervlakkige verhalen. In deze effecten kwam bij de tweede lezing (over een aanverwant onderwerp) geen verandering. Dus ook de gegevens van Williams & Ware wijzen erop dat expressiviteit en enthousiasme van een docent tot een misplaatst gunstig oordeel kunnen leiden. Oftewel, inhoud van een kursus en doceerstijl worden niet onder alle omstandigheden uit elkaar gehouden.

Een met de studies van Naftulin et al. en Williams & Ware vergelijkbaar probleem is door Zelby (1974) bestudeerd. Hij vroeg zich af of het studentoordeel over een docent afhankelijk is van de methode van lesgeven. Voor een tweetal kursussen (Technology in society; Electromagnetic fields), die hijzelf gaf, hield hij er, wisselend van jaar tot jaar, twee methodes van lesgeven op na. Aan het niveau van de kursussen en de te bespreken onderwerpen, alsmede aan de door studenten te investeren tijd werd zo goed als niets veranderd. Bij de ene methode van lesgeven lag de nadruk op het memoriseren van feiten en het routinematig kunnen oplossen van problemen: een detail-gerichte benadering. De andere methode van lesgeven was er daarentegen veel meer op gericht studenten te leren hoe zij verschillende problemen zouden kunnen aanpakken. Bij deze laatste methode vielen de studentoordelen ten aanzien van een groot aantal aspecten, zowel de docent als de kursus betreffende, aanzienlijk negatiever uit dan bij de eerste methode. Illustratief is in dit verband het feit, dat Zelby, bij vergelijking met docenten van zijn eigen faculteit, bij de meer probleemgerichte benadering terugviel van het eerste kwartiel – de topdocenten (25%) – naar het derde. Uit deze gegevens valt geen andere conclusie te trekken dan dat wensen of verwachtingen van studenten omtrent methodes van lesgeven het oordeel over een docent kunnen beïnvloeden.

Wij gaan tenslotte op één type onderzoek nog nader in. Het gaat om studies waarin de vraag centraal staat of het oordeel over een docent onafhankelijk is van het gedoceerde vak. Met andere woorden, het gaat bij dit soort studies om de vraag of de 'aardigheid' of aantrekkelijkheid van een vak het oordeel over de kwaliteit van een docent beïnvloedt. Of een vak als aardig of aantrekkelijk door studenten wordt gezien, zal waarschijnlijk, behalve van het vak zelf, van allerlei factoren afhangen: studiekeuze, plaats van het vak binnen het curriculum, zwaarte van het vak en dergelijke. Kortom, het valt niet te verwachten dat bepaalde vakken altijd en door iedereen als aardig en aantrekkelijk, of juist als onaardig en onaantrekkelijk, worden gezien. Niettemin is de vraag gerechtvaardigd of docenten van verschillende vakken verschillend worden beoordeeld. Welnu, hier zijn zeer sterke aanwijzingen voor. Zo toonde Rayder (1968) aan dat van docentkenmerken als sexe, leeftijd, rang en vakgebied ('department'), waarin iemand werkzaam is, het vakgebied de beste voorspelling geeft van doceervaardigheid, als gemeten met behulp van studentoordelen (te vinden bij Kulik & McKeachie, 1975). Verder vond Solomon (1966) dat docenten uit verschillende vakgebieden (Social Sciences; Humanities; Mathematics + Natural Sciences; Practical (Applied) subjects) significant verschillend werden beoordeeld door studenten (te vinden bij Page, 1974). Ook Remmers (1963) vond dat er aanzienlijke verschillen bestaan tussen de beoordelingen van docenten afkomstig uit verschillende vakgebieden. Welke docenten – docenten van α- dan wel van β-vakken – over het algemeen betere beoordelingen van studenten krijgen is onduidelijk. De monografie van Page vermeldt dienaangaande een tweetal studies:

- Walker (1969) vond dat docenten in de exacte vakken (Sciences + Mathematics) over het algemeen hogere beoordelingen krijgen dan docenten in andere vakken of vakgebieden;
- Clark & Keller (1954) vonden daarentegen dat docenten die een α-vak doceren (Social Sciences; Humanities) over het algemeen een betere beoordeling krijgen dan docenten die een β-vak doceren (Natural Sciences).

Niettemin wijzen de studies van Walker en Clark & Keller erop dat in de beoordeling van een docent het oordeel over een vak contaminerend kan werken. In een bijzonder aardige studie van Romney (1976) is dit onlangs nog eens aangetoond. Voor verschillende

eerstejaarskursussen, die ieder integraal door verschillende docenten aan verschillende groepen studenten ('sections') gegeven waren, ging hij variantie-analytisch na of er een significant kursuseffect bestond in de door studenten gegeven oordelen over docenten. In een tweetal elkaar qua gegevens gedeeltelijk overlappende analyses – er waren teveel gegevens om in een keer door de beschikbare computer te laten verwerken – kwam hij tot de slotsom dat er in studentoordelen over docenten zowel een significant docent- als een significant vakeffect aanwezig is. Deze conclusie van Romney doet nogal wat afbreuk aan de vermeende validiteit van studentoordelen en dit des te meer, omdat de in de twee analyses gebruikte gevens (studentoordelen) op een ruime verscheidenheid aan vakken betrekking hadden (scheikunde; computerkunde; psychologie; maatschappelijk werk; filosofie; Engels; politieke wetenschappen; sociologie; natuurkunde).

Eén studie (Hogan, 1973) is tenslotte nog vermeldenswaard. In deze studie werden validiteit en betrouwbaarheid van studentoordelen min of meer rechtstreeks tegenover elkaar gesteld; aldus vormt een bespreking van deze studie een aardige overgang naar de volgende paragraaf over betrouwbaarheid van studentoordelen (hoe nauwkeurig of onnauwkeurig is het studentoordeel over een docent?).

Hogan ontleende zijn gegevens aan een uitgebreid projekt, uitgevoerd aan de Universiteit van Wisconsin-Green Bay in het studiejaar 1971/72. In de herfst van 1971 werden gegevens (studentoordelen) verzameld over zo'n 260 kursussen; in de lente van 1972 werd dit nog eens gedaan voor zo'n 320 kursussen. Uitgaande van dit gegevensbestand ging Hogan na welke docenten in de herfst en daaropvolgende lente dezelfde kursus hadden gegeven (30 docenten); naar aan te nemen valt, aan verschillende groepen studenten. Verder ging hij na welke docenten in de lente een andere kursus hadden gegeven dan in de herfst (45 docenten).

Tenslotte spoorde hij kursussen op, waarvan de docent in de herfst een andere was dan in de lente (39 paren docenten). Voor ieder van deze drie groepen docenten ging Hogan na hoe het gesteld was met de stabiliteit van door studenten gegeven oordelen. De gegevens werden verzameld aan de hand van de 'Course Comments Questionnaire'. Deze vragenlijst omvat, naar eerder uit factoranalyse was gebleken (Hartley & Hogan, 1972), een 7-tal schalen met elk 5 items:

- Global Rating – an overall, summative judgement about the course and instructor.
- Responsiveness – a rating of the instructor's concern for and interaction with students.
- Difficulty – students' estimate of how difficult or demanding the course and instructor were.
- Organization – a rating of organization, clarity of procedures, and preparation.
- General Cognitive Development – students' perception of how the course affected their development of cognitive abilities not directly tied to the course, e.g., 'I developed my ability to identify main points or central issues.'
- Specific Cognitive Development – students' perception of how the course affected their development in the cognitive area related rather directly to course content, e.g., 'I became able to analyze new and complicated material in the field.'
- Relevance – students' perception of the effect of the course on development of interest, concern, and appreciation, e.g., 'I became aware of ways the subject is involved in my own life.'

De eerste vier factoren of schalen zijn juist weer die, welke gewoonlijk in beoordelingsvragenlijsten voorkomen; de laatste drie daarentegen zijn betrekkelijk specifiek voor de 'Course Comments Questionnaire'. Voor iedere schaal afzonderlijk werd de stabiliteit bepaald, en dit per groep van docenten.

De door Hogan gepresenteerde gegevens voor de drie groepen docenten laten twee dingen zien. Ten eerste, de stabiliteit van studentoordelen is aanzienlijk hoger voor docenten die dezelfde kursus in verschillende semesters gaven dan voor docenten die in de 'opeenvolgende' semesters twee verschillende kursussen gaven. Ten tweede, dit verschil in stabiliteit bestaat, naar blijkt uit een berekeningsprocedure ('path analysis') die Kulik & Kulik (1974) op de gegevens van Hogan hebben toegepast, voor een deel uit een vakeffect. Met andere woorden, ook de gegevens van Hogan duiden erop dat het studentoordeel over een docent voor een deel is ingegeven door de aantrekkelijkheid of 'aardigheid' van een vak.

De resultaten van het laatst besproken type studies samenvattend, kunnen we niet anders concluderen dan dat het studentoordeel over een docent voor een belangrijk deel een oordeel over de aardigheid/aantrekkelijkheid van een vak inhoudt. Docent- en vakeffect zijn in principe van elkaar onderscheidbaar (vgl. Romney, 1976), maar dit valt in de praktijk alleen te bereiken als hetzelfde vak door meer docenten aan verschillende groepen studenten wordt gegeven. Deze situatie doet zich evenwel op te beperkte schaal voor om er een beoordelingssysteem op te kunnen baseren. Met andere woorden, het gebruik van studentoordelen voor de aanstelling en bevordering van medewerkers is in de praktijk moeilijk uitvoerbaar.

3.3. *Betrouwbaarheid van studentoordelen*

De betrouwbaarheid van studentoordelen heeft te maken met de vraag hoe nauwkeurig studenten van een docent kunnen aangeven of zij hem/haar goed, dan wel minder goed of zelfs slecht vinden. De betrouwbaarheid van studentoordelen is over het algemeen op de volgende twee manieren onderzocht:

a) door bepaling van de mate waarin studenten consistent zijn in hun oordeel over een docent (eenmalige afname van een vragenlijst, waarbij studenten een docent op een aantal aspekten beoordelen; bij dit type onderzoek gaat het meestal om de vraag in hoeverre doceervaardigheid, opgevat als een één-dimensionaal kenmerk, betrouwbaar is te meten);
b) door bepaling van de mate waarin het oordeel van studenten over meerdere docenten stabiel is (test-hertest methode, waarbij docenten met een zeker tijdsinterval twee maal worden beoordeeld).

Onder 'validiteit' is uiteengezet dat het oordeel over een docent voor een belangrijk deel een oordeel over het gedoceerde vak inhoudt. Dit contaminerende vakeffect zal met name bij methode b verhogend werken op de betrouwbaarheidsschattingen; methode b is meestal toegepast op docenten die verschillende vakken gaven. Betrouwbaarheidsschattingen volgens deze methode zijn dus naar alle waarschijnlijkheid kunstmatig hoog. Of een dergelijk artefact ook voor methode a geldt – slechts één docent wordt beoordeeld –, is onduidelijk. Niettemin valt aan te nemen, voorzover het oordeel over de inhoud van een vak en het oordeel over een docent met elkaar worden verward en voorzover het vak in zijn te onderscheiden aspecten consistentie vertoont, dat ook consistentieschattingen kunstmatig hoog kunnen uitvallen. De hierna te presenteren gegevens, met name die voor methode b, dienen dan ook met de nodige terughoudendheid te worden bekeken.

Naar uit het overzichtsartikel van Costin et al. (1971) blijkt, zijn studenten zeer consistent in hun oordeel over een docent (methode a: een 10-tal studies). Daarentegen geven betrouwbaarheidsschattingen volgens methode b een iets minder gunstig beeld te zien, maar de gevonden betrouwbaarheden zijn nog alleszins aanvaardbaar (een 15-tal studies, te vinden in Costin et al. (1971), Kulik & McKeachie (1975) en Page (1974)).

In de diverse studies verricht volgens methode b zijn enige regelmatigheden, alhoewel niet al te sterke, te ontdekken. Ten eerste, om een redelijke betrouwbaarheid te verkrijgen moet een docent door minstens 25 studenten zijn beoordeeld (Remmers, 1959; Braunstein & Benston, 1971).

Ten tweede, lage betrouwbaarheidsschattingen doen zich soms voor in die gevallen waarin één en dezelfde kursus door meer docenten is gegeven (Morsh et al., 1956; Greenwood et al., 1976).

Ten derde, als bij test en hertest verschillende groepen studenten uit dezelfde studentengeneratie of uit verschillende generaties als beoordelaars worden gebruikt (Heilman & Armentrout, 1936; Drucker & Remmers, 1950; Braunstein & Benston, 1971; Hogan, 1973), vallen de betrouwbaarheidsschattingen over het algemeen wat lager uit dan bij studies waarbij een docent door dezelfde groep herhaald is beoordeeld.

Ten vierde, en dit sluit aan bij het vorige punt, er zijn geen duidelijke aanwijzingen dat met een toenemend tijdsinterval tussen test en hertest de stabiliteit van studentoordelen zou afnemen. Zelfs bij tijdsintervallen van meer dan 5 jaar (Heilman & Armentrout, 1936; Drucker & Remmers, 1950) zijn nog alleszins aanvaardbare betrouwbaarheidsschattingen gevonden. Uit het voorgaande valt eigenlijk alleen maar te concluderen, dat betrouwbaarheidsschattingen, indien vakeffecten worden veronachtzaamd, op een alleszins aanvaardbaar niveau liggen. Maar hiermee is niet alles gezegd. Betrouwbaarheidsschattingen, als deze meer docenten betreffen, zeggen iets over de globale betrouwbaarheid van de totale groep van oordelen over docenten. Nu zijn er, zoals onder andere uit de door Romney (1976) gepresenteerde gegevens valt af te leiden, erg veel docenten die een goede tot zeer goede beoordeling krijgen, terwijl slechts weinig docenten als slecht worden beoordeeld (zie ook Widlak e.a., 1973). Bij het berekenen van betrouwbaarheid dragen in het algemeen de uitschieters, in ons geval dus slechte docenten, het meeste bij, als we tenminste de bijdrage per docent bekijken. Dit betekent in feite, dat er in de bovenste regionen, tussen goede en zeer goede docenten, minder makkelijk onderscheid te maken valt dan men zou verwachten op grond van een betrouwbaarheidsschatting, berekend voor een totale groep docenten. Kortom, de in de literatuur gerapporteerde betrouwbaarheden zeggen ons inziens weinig over de vraag of er in de hogere regionen echt betrouwbaar geselecteerd zou kunnen worden. Wij vermoeden evenwel, dat deze vraag ontkennend beantwoord moet worden.

4. *Konklusies*

Naar onze indruk schieten de betrouwbaarheid en de validiteit van studentoordelen op een aantal punten ernstig tekort. Het oordeel over de docent is niet

onafhankelijk van de gebruikte onderwijsmethode (detail- versus probleemgerichte benadering). Verder is er in het oordeel over de docent een duidelijk vakeffect aanwezig: de aantrekkelijkheid of aardigheid van een vak heeft een contaminerend effect. Verder is het aannemelijk dat dit vakeffect leidt tot kunstmatig hoge betrouwbaarheidsschattingen voor groepen docenten die verschillende vakken doceren. Nog een ander probleem vormt het feit, dat studentoordelen zich, door de scheve verdeling van de oordelen over doceervaardigheid, beter lenen voor het opsporen van slechte docenten dan voor het opsporen van echt goede docenten.

Alles bij elkaar genomen moeten we stellen, dat, als het gaat om de aanstelling of bevordering van wetenschappelijke medewerkers, het studentoordeel over de doceeraspecten van hun taak van onvoldoende kwaliteit is om in de selectieprocedure in te bouwen.

Deze conclusie kan ook bij een aantal andere schrijvers gevonden worden. McKeachie (1969) adviseert om studentoordelen uitsluitend in de formatieve sfeer te gebruiken. Page (1974) die de bruikbaarheid van studentoordelen voor Engelse universiteiten onderzocht stelt, dat de ervaringen in de Verenigde Staten het hoogst onwaarschijnlijk maken dat in Engeland ooit wordt overgegaan tot een selektief gebruik van studentoordelen.

De meningen over dit onderwerp zijn binnen de Amerikaanse universiteiten zeer verdeeld. In het algemeen raadt men aan, grote terughoudendheid te betrachten en in ieder geval naast studentoordelen ook andere bronnen van informatie te gebruiken (Kulik en McKeachie, 1975). Overigens worden de meeste studentoordelen verzameld door de studenten zelf. De belangrijkste functie is dan het verschaffen van informatie aan studenten, terwijl er geen sprake is van invloed op het personeelsbeleid. Een onderzoek van Gustad (geciteerd door Page, 1974) toont aan, dat het gebruik van studentoordelen voor promotie en aanstelling aan het afnemen is, terwijl de nadruk steeds meer op formatief gebruik komt te liggen.

Tot slot nog een enkele opmerking over zaken die met dit probleem samenhangen. Zoals reeds eerder is opgemerkt, wordt in de literatuur vrijwel uitsluitend geschreven over oordelen met betrekking tot het doceergedrag van docenten bij hoorcolleges. Elders (Blom, 1978) is betoogd, dat dat gedrag naar alle waarschijnlijkheid niet de belangrijkste onderwijstaken van de docent representeert. In hoeverre studentoordelen over andere aspecten van de doceertaak van hogere kwaliteit is, valt uit de literatuur niet op te maken. Het type problemen dat wij in het voorgaande gesignaleerd hebben, maakt echter, dat wij ons daarover weinig illusies maken. Ook is weinig bekend over de kwaliteit van de oordelen van andere betrokkenen. Hoewel een van de belangrijkste problemen, de contaminantie van vak- en docenteffecten, bij andere betrokkenen waarschijnlijk een duidelijk mindere rol zal spelen, is het niet zeer waarschijnlijk, dat in deze sfeer erg valide oordelen gevonden zullen worden. Een probleem dat dan zeker sterk naar voren zal komen is, dat geen 'harde' definitie te geven is van een 'goede docent'. Onderzoek naar effectiviteit van docenten heeft tot nu toe erg weinig opgeleverd.

Wel zijn er mogelijkheden om oordelen over docenten formatief te gebruiken. Er zijn aanwijzingen dat docenten, althans in de ogen van hun beoordelaars, verbeteren als ze de beschikking krijgen over oordelen over hun doceerkwaliteit. Alleen het verstrekken van deze oordelen is echter niet voldoende. Er dient gelegenheid tot training of een andere vorm van begeleiding aan gekoppeld te worden (Centra, 1973; Aleamoni, 1978; McKeachie, 1969).

Literatuur

Aleanomi, L. M., The usefullness of student evaluations in improving college teaching. *Instructional Science*, 1978, (7), 95–105.

Bendig, A. W., Comparisons of psychology instructors and national norms on the Purdue Rating Scale. *Journal of Educational Psychology*, 1953, (44), 435–439.

Bendig, A. W., Ability and personality characteristics of introductory psychology instructors rated competent and empathetic by the students. *Journal of Educational Research*, 1955, (48), 705–709.

Bergh, I. van den, *Verslag projectgroep evaluatie van eigen onderwijs*. Leiden: Subfaculteit Psychologie, 1978.

Blom, S. J. M., *Docententrainingen aan de Rijksuniversiteit Leiden. Voorstellen voor een beleid*. Leiden: Bureau Onderzoek van Onderwijs, Memorandum 453–78, 1978.

Braunstein, D. N. and Benston, G. J., Student and Departmental chairman view of the performance of university professors, *Proceedings of the 17th international congress of applied psychology*, Liege, Belgium, 1971, *Bijstellingsnota Meerjarenafspraken 1979–1982*.

Centra, J. A., Do student ratings of teachers improve teaching? *Change*, 1973, (5), 12–13.

Clark, K. E. and Keller, R. J., Studentratings of college teaching. In: R. Eckert, et.al. *A university looks at its program*, Minneapolis: University of Minnesota, 1954.

Commissie Kwaliteitsmaatstaven Onderwijs, *Eindrapport*. Leiden: Rijksuniversiteit, 1976.

Costin, F., A graduate course in the teaching of psychology: Description and evaluation, *Journal of Teacher Education*, 1968, (19), 425–432.
Costin, F., Greenough, W. T. and Menges, R. J., Student rating of college teachers, Reliability, validity and usefullness. *Review of Educational Research*, 1971, (41), 511–535.
Drucker, A. J. and Remmers, H. H., Do alumni and students differ in their attitudes towards instructors? *Purdue University Studies in Higher Education*, 1950, (70), 62–64.
Dubin, R. and Taveggia, Th. *C., The Teaching Learning Paradox*, Eugen (Oregon): Center for the Advanced Study of Educational Administration, 1968.
Frey, P.W., Student ratings of teaching: Validity of several rating factors, *Science*, 1973, (182), 83–85.
Geensen, M., *Evaluatie hoorcolleges* ..., Leiden, een serie memoranda over de evaluatie van diverse hoorcolleges, Bureau Onderzoek van Onderwijs, 1969.
Gessner, P. K., Evaluation of Instruction, *Science* 1973, (180), 566–570.
Granzin, K. L. and Painter, J. J., A new explanation for students' course evaluation tendencies, *American Educational Research Journal*, 1973, (10), 115–124.
Greenwood, G. E., Hazelton, A., Smith, A. B., Ware, W. B., A study of the validity of four types of student ratings of college teaching assessed on a criterion of student achievement gains, *Research in Higher Education*, 1976, (5), 171–178.
Guthrie, E. R., *The evaluation of teaching: A progress report*, Seattle: University of Washington, 1954.
Hartley, E.L. and Hogan, T. P., Some additional factors in student evaluation of courses. *American Educational Research Journal*, 1972, (9), 241–250.
Haslett, B. J., Student knowledgeability, student sex, class size and class level: Their interaction and influence on student ratings of instruction, *Research in Higher Education*, 1976, (5), 39–65.
Heilman, J. and Armentrout, W. D., The ratings of college teachers on ten traits by their students, *Journal of Educational Psychology*, 1936, (27), 197–216.
Hogan, T. P., Similarity of student ratings across instructors, courses and time, *Research in Higher Education*, 1973, (1), 149–154.
Isaacson, R. L., McKeachie, W. J. and Milholland, J. E., Correlation of teacher personality variables and student ratings, *Journal of Educational Psychology*, 1963, (54), 110–117.
Kulik, J. A. and McKeachie, W. J., The Evaluation of teachers in higher education. In: F. N. Kerlinger (Ed), *Review of Research in Education*, Itasca (I11): Peacock Publishers Inc., 1975, 210–240.
Kulik, J. A. and Kulik, C. C., Student ratings of instruction. *Teach. Psych.*, 1974, (1), 51–57.
Los, F., *Vragenlijst systematische inleiding*, Leiden, subfaculteit Psychologie, niet gepubliceerd.
McKeachie, W. J., *Teaching Tips*, Lexington (Mass): Heath & Co., 1969.
Morsh, J. E., Burgess, G. G. and Smith, P. N., Student achievement as a measure of instructor effectiveness, *Journal of Educational Psychology*, 1956, (47), 79–88.
Naftulin, D. H., Wave, J. E. Jr., and Donelly, F. A., The Doctor Fox lecture: A paradigm of educational seduction, *Journal of Medical Education*, 1973, (48), 630–635.
Page, C. F., *Student evaluation of teaching: the American experience*. London: Society for Research into Higher Education, 1974.
Rayder, N. F., College student ratings of instructors, *Journal of Experimental Education*, 1968, (37), 76–81.
Remmers, H. H., Rating methods in research on teaching. In: N.L. Gage, (Ed.), *Handbook of research on teaching*, Chicago: Rand-McNally, 1963, 329–378.
Remmers, H. H., *The appraisal of teaching in large universities*, Ann Arbor: University of Michigan, 1959.
Rodin, M. and Rodin, B., Student evaluations of teachers. *Science*, 1972, (177), 1164–1166.
Romney, D., Course effect versus teacher effect on students' ratings of teaching competence, *Research in Higher Education*, 1976, (5), 345–350.
Solomon, D., Teacher behavior dimensions, course characteristics and student evaluation of teachers, *American Educational Research Journal*, 1966, (3), 35–47.
Vos, P., Het woord bij de daad. In: H. F. M. Crombag en T. M. Chang (Red.). *Een kleine zoölogie van het onderwijs*, Leiden: Universitaire pers, 1978.
Walker, B. D., An investigation of selected variables relative to the manner in which a population of junior college students evaluate their teachers, *Dissertation abstracts*, 1969, (29/9-b), 3474.
Werdell, Ph. R., *Course and teachers evaluation*, Washington D. C.: United States National Student Association, 1969.
Widlak, F. W., McDaniel, E. D., Feldhusen, J. F., *Factor Analysis of an Instructor Rating Scale*, Paper presented at the Annual meeting of the American Educational Research Association, 1973.
Williams, R. G. and Ware, J. E., An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecturer, *American Educational Research Journal*, 1977, (14), 449–457.
Zelby, L. W., Student faculty evaluation, *Science*, 1974, (183), 1267–1270.

Curriculum vitae:

S. J. M. Blom (geb. 1947) was na zijn opleiding electrotechniek (T.H. Delft) werkzaam als H.T.S.-docent. Sinds 1976 als medewerker Bureau Onderzoek van het onderwijs aan de universiteit te Leiden belast met het opzetten en uitvoeren van cursussen over onderwijs t.b.v. docenten.
Adres: Boerhaavelaan 2, Leiden.

W. F. Langerak (geb. 1943) studeerde psychologie met als specialisatie funktieleer (GU, Amsterdam). Sinds 1971 als stafmedewerker verbonden aan het Bureau Onderzoek van het Onderwijs (RU, Leiden).
Adres: Boerhaavelaan 2, Leiden.