Leren door te corrigeren. De leerwinst van de leerling-beoordelaar bij het nakijken van het werk van medeleerlingen in voortgezet bètaonderwijs

A. B. H. Bos , C. Terlouw en A. Pilot

Samenvatting

Vanwege de schaarste aan docenttijd is het relevant of tijdwinst voor docenten bereikt kan worden door leerlingen het werk van medeleerlingen te laten beoordelen. Maar ook is de vraag relevant of een dergelijke beoordeling leidt tot leerwinst voor de beoordelende leerling zelf. In een experiment in het voortgezet bètaonderwijs, waarbij iedere leerling een complete toets van een medeleerling nakeek, is gevonden dat er inderdaad een leerwinst optreedt bij de beoordelende leerlingen. Dit effect is nader onderzocht in een situatie, waarbij leerlingen antwoorden op toetsvragen beoordeelden met behulp van expliciete nakijkcriteria in een met ict ondersteunde opzet. De grootste leerwinst in deze digitale omgeving werd bereikt door leerlingen die een pretest maakten voordat ze de nakijkcriteria gebruikten bij het beoordelen van antwoorden van andere leerlingen.

1 Achtergrond en theoretisch kader

In de laatste decennia van de vorige eeuw is in Nederland de beschikbare tijd in het onderwijsprogramma van de bovenbouw van het voortgezet onderwijs voor de natuurwetenschappelijke vakken scheikunde, natuurkunde, en biologie gestaag verminderd. Als gevolg daarvan is ook de contacttijd voor deze vakken minder geworden. Te constateren valt, dat er ruwweg een halvering van de contacttijd is (Tweede Fase Adviespunt, 2002), waarbij door de bank genomen er sprake is van twee uur per week contacttijd met een vakdocent. Ook in de vernieuwde Tweede Fase, die in augustus 2007 van start is gegaan, is er in deze situatie geen verandering opgetreden. Schaarse, kostbare docententijd dient alleen al om deze reden zo doeltreffend en doelmatig mogelijk te worden ingezet met het oog op een zo hoog mogelijk rendement.

Formatief testen is volgens het overzichtsartikel van Black en William (1998) een manier om onderwijsprocessen meer rendement te geven. Er is hier evenwel sprake van een efficiëntieprobleem: de correctielast zal voor een Tweede Fasedocent, die op voltijdsbasis met 200 à 300 leerlingen te maken heeft, al snel een relatief groot aandeel in beslag nemen van de tijd die een docent in totaliteit beschikbaar heeft. Een voor de hand liggende oplossing vanuit het oogpunt van efficiëntie is het inzetten van ict voor het organiseren en afhandelen van de evaluatie van leerlingresultaten in alle fasen van het leerproces. Het gebruik van meerkeuzevragen is in een dergelijke ict-omgeving het gemakkelijkst implementeerbaar, maar het gebruik van meerkeuzevragen is niet altijd valide, gelet op de aard van de leerdoelen. Het is voor de toetsing van bepaalde natuurwetenschappelijke leerdoelen immers soms gewenst andersoortige toetsen te gebruiken, waarin grafieken, schetsen, structuurformules, uitgebreide berekeningen en stapsgewijze redeneringen van studenten worden gevraagd. Bij de beoordeling van dit type antwoorden is menselijke tussenkomst, die veel tijd kost, onontbeerlijk. Een docent natuurwetenschappen moet dus een keuze maken voor zijn tijdsbesteding: óf de tijd vooral besteden aan het uitleggen van lastige begrippen en relaties, het begeleiden van practica en dergelijke, óf vooral aan toetsing. Om tot een goede afweging te komen bij een dergelijke keuze, is het nodig een kwantitatief beeld te hebben van de bijbehorende effecten. Het is dus zaak om te zoeken naar doelmatige middelen die weinig tijd kosten. Formatief toetsen is een interessant en mogelijk effectief middel, maar is er ook een oplossing te vinden voor de correctietijd?

Een mogelijke oplossing voor het reduce-

ren van de correctietijd is het inzetten van studenten om eigen en andermans werk te beoordelen en er feedback op te geven, hetgeen met name in het hoger onderwijs steeds frequenter wordt toegepast (Dochy, Admiraal, & Pilot, 2003; Dochy, Segers, & Sluijsmans, 1999; Topping, 1998). Naast de bedrijfsmatige overwegingen (de hiervoor genoemde soort van studentinzet wordt in de regel niet betaald) en ondanks de soms bijbehorende negatieve reacties van studenten (zij moeten het werk doen, wat de docent normaliter doet; Clifford, 1999; Wen & Tsai, 2006) is er ook een meer theoretische overweging om studenten in te zetten bij beoordelingsactiviteiten. De overdracht van docenttaken naar studenten wordt hierbij niet alleen door economische motieven ingegeven, maar heeft ook theoretische wortels in het gedachtegoed van constructivisten (Bruner & Olson, 1973). De gedachte om de leerling meer verantwoordelijkheid te geven voor zijn eigen leerproces heeft de overdracht van traditionele docententaken naar de leerling als logische consequentie. In deze theoretische visie past het beoordelen door leerlingen van eigen werk of dat van medeleerlingen. Door deze werkwijze toe te passen komen vormen van formatief toetsen weer in beeld.

Er is nogal wat onderzoek gedaan naar het nakijken door studenten. Het brandpunt van het merendeel van de kwantitatieve onderzoeken ligt daarbij op het verband tussen de scores die studenten geven en die van de docent. De conclusies houden in dat:

- studenten aan andere studenten in het algemeen iets lagere cijfers geven dan de docent (Falchikov & Goldfinch, 2000);
- de minder presterende studenten zichzelf vaak te hoog becijferen (Boud & Falchikov, 1989);
- de beste studenten soms door anderen te laag worden becijferd (Sadler & Good, 2006), en
- het geslacht geen belangrijke rol meer blijkt te spelen, als de beoordeling 'blind' wordt uitgevoerd (Falchikov, 1997). Wel bericht Pope (2005) dat studentes tijdens het beoordelingsproces wat meer spanning ondervinden dan studenten.

Voorts adviseren Sadler en Good (2006) anonimisatie toe te passen bij het nakijken door studenten om juridische consequenties te vermijden.

Zoller, Tsaparlis, Fatsow en Lubezky (1997) maken een onderscheid tussen het beoordelen van hogere orde cognitieve vaardigheden (HOCV) zoals kritisch denken, vragen stellen, redeneren en slecht gedefinieerde problemen oplossen, en het beoordelen van lagere orde cognitieve vaardigheden (LOCV) zoals bijvoorbeeld memoriseren en het toepassen van algoritmen. Bij de laatste categorie (LOCV) treffen zij geen verschil aan tussen hoogleraren en studenten in diverse natuurwetenschappelijke disciplines. Voor prestaties waarvoor hogere orde vaardigheden nodig zijn, gaapt er een enorme kloof tussen de beoordelingen van hiervoor ongetrainde studenten en hun hoogleraren (Zoller, Tsaparlis, Fatsow, Lubezky, 1997). Het nakijken door studenten levert volgens Stefani (1994) voor de nakijkende studenten veel voordelen op in termen van een meer begripsvol leerproces en een hogere motivatie. Orsmond, Merry en Reiling (2002) zijn van mening dat nakijkende studenten een beter begrip krijgen van wat er geëist wordt in het leerproces.

Het zelf laten construeren van de beoordelingscriteria zou een volgende stap kunnen zijn in de activiteiten van een nakijkende student. De daarover gerapporteerde effecten zijn evenwel wisselend. Langan e.a. (2005) rapporteren dat studenten die meehielpen met het ontwikkelen van criteria daardoor niet hoger scoorden. Sluijsmans, Brand-Gruwel en Van Merriënboer (2002) trainden toekomstige leraren voor het basisonderwijs in het kader van competentiegericht onderwijs voor de rol van de beoordelaar van de medestudent waarbij het construeren van beoordelingscriteria centraal stond. De studenten werden wel betere beoordelaars, maar scoorden soms wel, maar soms ook niet beter bij vakinhoudelijke opdrachten.

Het voorgaande geeft aan dat de empirische evidentie voor het nakijken door leerlingen of studenten wellicht nog niet een erg stevige basis heeft. Een dergelijk beeld komt ook uit andere literatuur naar voren. Boud en Falchikov (1989) zijn in hun overzichtsartikel over het beoordelen van eigen prestaties van mening dat een deel van het kwanti-

tatieve onderzoek van lage kwaliteit is. In een recenter overzichtsartikel van Falchikov en Goldfinch (2000) van studies over het nakijken van het werk van medeleerlingen in de periode 1959-1999 komt hetzelfde oordeel naar voren. Daarbij blijkt, dat de kwantitatieve studies van 'lage kwaliteit' uniform verdeeld zijn over alle decennia en deze eigenschap dus niet typisch lijkt te zijn voor oude of juist recente publicaties. Davies, Kumtepe en Aydeniz (2007), Minjeong (2005), en Chapman en Bloxham (2004) zijn daarentegen tamelijk stellig in het noemen van voordelen van het nakijken van werk van medeleerlingen. Zij baseren zich onder meer op de evidentie uit een studie van Bloxham en West (2004). Deze auteurs zijn echter veel bescheidener in hun uitspraken, noemen hun onderzoek semikwantitatief en geven aan, dat de uitkomsten vooral verder onderzoek verdienen (Bloxham & West, 2004). Ook Sadler en Good (2006) ten slotte stellen dat in de literatuur geen degelijk statistisch onderbouwd systematisch onderzoek van het leereffect op de nakijker zelf te vinden is en doen daarom zelf een experiment. Om het leereffect op de nakijker zelf systematisch en accuraat na te gaan, laten Sadler en Good (2006) een groep van ca. 100 leerlingen (leeftijd ca. 13 jaar) een biologietest maken. Een kwart van de leerlingen gaat vervolgens de eigen test nakijken. De helft van de leerlingen gaat de test van een andere leerling nakijken. Het resterende kwart van de leerlingen kijkt niets na. Een week later krijgen alle leerlingen onaangekondigd dezelfde test. Alle werken worden ook door een leraar nagekeken. De belangrijkste conclusie is dat alleen die leerlingen die het eigen werk hebben nagekeken een significant hogere score behalen bij de tweede test. Twee resultaten vallen bij dit onderzoek op: a) het 'pretesteffect' is afwezig en b) het nakijken van het werk van anderen heeft een veel geringer leereffect dan het nakijken van eigen werk. Een toelichting:

- ad a) Op grond van het overzichtsartikel over het effect van pretesten van Willson en Putnam (1982) zou er een effect van de pretest op de posttest kunnen zijn, omdat de deelnemers kort na elkaar *dezelfde* test maken.
- ad b) Er wordt geen sluitende verklaring gegeven voor het verschil tussen het effect van het nakijken van eigen en dat van andermans werk.

Voorts is de hoge gemiddelde score die bij de eerste testafname wordt gehaald een complicerende factor, omdat er dan een grote kans is op een plafondeffect. De auteurs maken in dit kader ook melding van een enigszins scheve frequentieverdeling die onder meer kenmerkend is voor een plafondeffect (Sadler & Good, 2006).

Samenvattend is onze conclusie dat er in de literatuur nog geen sterke basis van empirische evidentie aanwezig is voor de leerwinst van leerlingen of studenten door te corrigeren. Er is behoefte aan een nadere kwantitatieve onderbouwing met extra aandacht voor de methodische opzet.

Het belang van het nemen van verantwoordelijkheid van studenten voor eigen leren als ook de reductie van de correctielast voor de docent zijn twee goede redenen om leerlingen met name formatieve testen te laten nakijken. Het in de literatuur enigszins onderbelichte leereffect van het correctieproces op de nakijker zelf zou een derde reden kunnen zijn om leerlingen te laten nakijken. Het is vooralsnog niet duidelijk in hoeverre het nakijken van andermans werk ook een leerwinst bij de beoordelaar zelf geeft, en dit onderzoek zal zich dan ook voornamelijk richten op het meten hiervan. Met name in het kader van formatief testen is dit mogelijke leereffect belangrijker dan de precisie en accuraatheid van de door de leerlingen gegeven cijfers. Het resultaat is immers niet of hooguit zijdelings van belang voor plaatsing, selectie of certificering. In tegenstelling tot summatief testen ligt bij formatief testen de nadruk op terugkoppeling, reflectie, diagnose en bewaken van het leerproces (William & Black, 1996).

2 Vraagstellingen

In het eerste deelonderzoek zal onderzocht worden wat het leereffect bij de nakijker is als die andermans werk nakijkt, met een preposttestopzet waarbij ook een referentiegroep aanwezig is. Daarbij is de posttest niet gelijk aan de pretest. Dit laatste om een effect van

de pretest op de posttest te voorkomen. In een tweede pre-posttestexperiment zal worden nagegaan wat het leereffect bij de nakijker is van een door ict-ondersteund nakijkproces. Ook hierbij wordt gebruik gemaakt van een methodische opzet om het pretesteffect te controleren.

Uitgaande van het bovenstaande formuleren we meer specifiek de volgende twee onderzoeksvragen voor deze studie:

1. Hoe groot is de leerwinst door het nakijken van het werk van medeleerlingen bij de beoordelaar zelf in een conventionele onderwijsleersituatie?
2. Leidt het nakijken van het werk van medeleerlingen aan de hand van nakijkcriteria, al dan niet voorafgegaan door een pretest, tot een leerwinst bij de beoordelaar zelf in een met ict ondersteunde onderwijsleersituatie?

De vraagstellingen zullen in twee afzonderlijke experimenten worden beantwoord.

3 Experiment A

3.1 Methode

We bespreken in deze paragraaf eerst zowel de grote lijnen van de onderzoeksopzet als die van de onderwijsontwerpen. Vervolgens gaan we meer gedetailleerd in op de deelnemers. Daarna worden de leerstof en de bijbehorende testen beschreven. Ten slotte komt de gegevensverwerking aan de orde: correctieprocedure, statistische analyse en de bepaling van leerwinst.

Experimenteel ontwerp en procedures

Leerlingen uit 4 vwo werden in twee equivalente groepen verdeeld (voor details, zie de volgende paragraaf). Alle leerlingen maakten dezelfde pretest (O_1). In de volgende stap kreeg de ene groep een correctievoorschrift van de pretest waarmee elke leerling een geanonimiseerde, willekeurig gekozen pretest van een medeleerling nakeek. Op hetzelfde moment dat groep 1 zich bezig hield met de evaluatie, maakte de tweede groep de posttest (O_2). Hoewel in de onderwijskunde een test terecht als interventie beschouwd wordt, maken we hier in navolging van Cook en Campbell (1979) een onderscheid tussen het testen en de daaropvolgende activiteiten. Het nakijken door een leerling van het werk van een ander zien we als een interventie (X). Na het nakijken maakte groep 1 de posttest (O_2). Omdat het een experiment in de onderwijspraktijk betrof waarbij na het experiment de uiteindelijke verschillen in leereffect tussen groepen zo klein mogelijk horen te zijn, kreeg de tweede groep na het maken van de posttest (O_2) ook een correctievoorschrift en een na te kijken werk van een andere leerling. Samengevat is het ontwerp voor groep 1 $O_1\ X\ O_2$, en voor groep 2 $O_1\ O_2\ (\ X\ )$.

Deelnemers

Voor het experiment werden die leerlingen uit het natuurprofiel geselecteerd die net gestart waren met lessen scheikunde in het 4e leerjaar vwo. Bij het samenstellen van de groepen werd gebruik gemaakt van de index *BX*. Deze index werd berekend met behulp van een procedure die gebruik maakte van de schoolresultaten over alle vakken uit het voorgaande halfjaar. Bij deze procedure werd een normaal verdeelde waarde opgelegd met een gemiddelde van 100 en een standaarddeviatie van 10.

De tweetraps gecomputeriseerde willekeurige indeling verliep als volgt. Uit de populatie werd willekeurig een leerling gekozen waarna een 'naaste buur' met hetzelfde geslacht en met een zo klein mogelijk verschil in *BX* werd gezocht. Vervolgens werd de eerste leerling willekeurig óf in groep 1 óf groep 2 geplaatst en de andere leerling in de andere groep. De groepen bleken niet van elkaar te verschillen in leeftijd, *BX*-score en geslacht, gelet op de resultaten van de *F*-testen en de Fisher Exact-toets (zie Tabel 1). Aan het experiment deden in totaal 33 leerlingen mee.

Instrumenten en materialen

Voor dit experiment werd een voor dit type onderwijs gebruikelijke papieren toets gebruikt bestaande uit 24 vragen en opgaven. In plaats van een naam werd een 6 cijferig nummer als identificatie gebruikt. Op het opgavenblad was ruimte voor het opschrijven van het antwoord. Voor de evaluatie was in het correctiemodel het antwoord op iedere vraag verdeeld in 1 tot 4 essentiële onderdelen. Voor ieder onderdeel mocht 1 punt worden

Tabel 1

Gegevens van participanten in experiment A

groep→	*1 (O_1 X O_2)*	*2 (O_1 O_2)*	*p*
Leeftijd ± s_d (jr)	15,9 ± 0,32	16,0 ± 0,47	0,46[a]
BX ± s_d	103,4 ± 10.4	104.0 ± 9.21	0,87[a]
% vrouwelijk	76	69	0,50[b]
aantal (*N*)	17	16	

[a] via *F*-toets ; [b] via Fisher Exact-toets

toegekend. Bij twijfel over het al of niet correct zijn van een vraag mocht een vraagteken worden geplaatst. De posttest bestond uit 15 korteantwoordvragen, verschillend van de pretest, maar uiteraard wel over dezelfde onderwerpen.

Voor de pretest waren ca. 25 minuten nodig, het nakijken kostte tussen de 12 en 15 minuten, en voor de posttest waren ongeveer 15 minuten nodig. De toets had betrekking op de inhoud van hoofdstuk 1 van deel 1 van het studieboek Chemie Overal (Franken, Kabel-van den Brand, & Korver, 1998) met de volgende onderwerpen: bouw en massa van atomen, het Periodiek Systeem, metalen – zouten – moleculaire stoffen, inter- en intramoleculaire wisselwerking, en waterstofbruggen. Een voorbeeld van een vraag staat in Figuur 1.

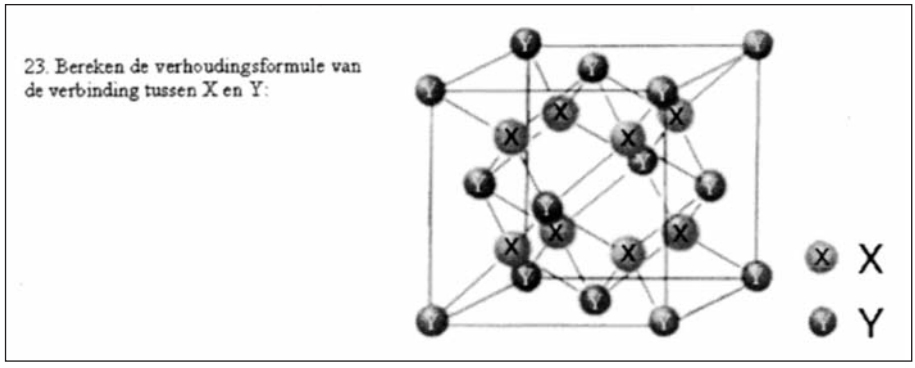
Correctieprocedure, statistische analyse en de bepaling van leerwinst

Het toekennen van punten aan de open vragen van de testen werd onmiddellijk na het experiment door twee docenten onafhankelijk uitgevoerd aan de hand van een gedetailleerd correctievoorschrift. Op één vraag na (over waterstofbruggen) bleek er geen systematisch verschil tussen deze twee correctoren. De correlatiecoëfficiënt tussen de twee beoordelingen was 0,99. Het gemiddelde van de scores van deze twee correctoren werd als de 'officiële score' genomen. De maximale scores voor posttest en pretest waren op 100 gesteld. De gemiddelde genormaliseerde winst G volgens Hake (1998) werd berekend met de formule

$$G = (\text{posttest}_{gem} - \text{pretest}_{gem}) / (100 - \text{pretest}_{gem}),$$

waarbij posttest_{gem} het gemiddelde van de posttestscores en pretest_{gem} het gemiddelde van de pretestscores van een bepaalde groep was (Hake, 1998).

Een tweede leerwinstmeting verliep via bepaling van de kennisgroeiexponent *B* (Bos, Terlouw, & Pilot, 2007). Het is experimenteel aangetoond dat er in diverse onderzoeksontwerpen en diverse vakgebieden een verband is tussen pretest en posttest, dat met een machtsfunctie kan worden beschreven. Uitzetten van *log*(posttest/pretest) tegen *log*(pretest/100) geeft een rechte lijn door de oorsprong. De helling van deze lijn is de kennisgroeiexponent *B*. Voor de deelnemers van een bepaalde groep werd voor iedere deelnemer de pretest- en de posttestscore in een speciaal computerprogramma ingevoerd. Met deze gegevens werd een niet-lineaire kleinste kwadratenaanpassing uitgevoerd. Op basis van deze berekeningen kon exponent *B* worden geschat alsook de fout in deze para-



Figuur 1. Voorbeeld van een pretestvraag, experiment A.

meter. Via een *t*-test werd ook de significantie van verschillen in *B* bepaald. Via simulaties is aangetoond dat deze methode veel accuratere resultaten en een veel kleinere kans op een type II fout (een foutnegatieve uitspraak) geeft dan de gebruikelijke werkwijze met effectgroottes. Als er voor een bepaalde categorie op grond van het experimenteel ontwerp geen pretestwaarden bekend waren, werd als alternatief de gemiddelde pretestscore van een vergelijkbare groep of categorie gebruikt. Voor de berekening van *B* werd dan de formule gebruikt

$$B = -log(\text{posttest}_{gem}/\text{pretest}_{gem}) / log(\text{pretest}_{gem}/100) \text{ gebruikt,}$$

waarbij met posttest_{gem} en pretest_{gem} de gemiddelde post- en pretestscores van een groep of categorie wordt bedoeld.

Vergeleken met de methode, waarin pre- en posttestscoreparen per deelnemer gebruikt worden, is komen vast te staan dat *B*-waarden die met groeps- of categoriegemiddelden berekend worden wat conservatiever zijn. In een dergelijk geval is het ook niet mogelijk om de fout in *B* te schatten. Als voor de berekening van de *B*-waarde een groeps- of categoriegemiddelde wordt gebruikt, staat achter de *B*-waarde de aanduiding *(cat)* vermeld (Bos, Terlouw, & Pilot, 2007). Een nadere typering van waarden van *B* staat in Tabel 2.

Voor empirisch bepaalde parameters van functies is het gebruikelijk de standaardafwijking van het gemiddelde (S_e) op te geven. Het verband tussen de standaardafwijking S_d en S_e is

$$S_e = S_d / \sqrt{v},$$

Tabel 2

Nominale schaal voor de kennisgroei-exponent B (Bos & Terlouw, 2005)

exponent	*leerwinsttypering*
$B \leq 0{,}40$	laag
$0{,}40 < B < 0{,}60$	gemiddeld
$B \geq 0{,}60$	hoog

waarin v het aantal vrijheidsgraden is.
Voor de leerwinst werd voorts de effectgrootte volgens Cooper (1998) berekend.

3.2 Resultaten

Onderzoeksvraag 1. Hoe groot is de leerwinst door het nakijken van het werk van medeleerlingen bij de beoordelaar zelf in een conventionele onderwijsleersituatie?

In Tabel 3 staat de leerwinst van experiment A weergegeven. De pretestwaarden van Tabel 3 verschilden niet significant van elkaar: de twee groepen waren gelijkwaardig. Er was sprake van betrouwbare toetsen: Cronbach's α voor de pretest was 0,67 en voor de posttest 0,78. Tussen de uitslagen op de pre- en posttest werd een sterk lineair verband gevonden (groep 1 $R = 0{,}81$; groep 2: $R = 0{,}88$).

Uit het verschil in scores op pre- en posttest volgt (Tabel 3, regel 2) dat het nakijken van andermans werk tot een significant leereffect bij de beoordelaar zelf leidde. De effectgrootte volgens Cooper (1998) was 1,07. Uit de leerwinst *G* volgens Hake (1998) viel voor groep 1 een classificatie *gemiddelde leerwinst* af te leiden die volgens Hake typisch is voor het zogenaamde *interactive engagement* onderwijs. Gelet op de gevonden waarde van de kennisgroei-exponent (*B*) was hier sprake van een *lage* leerwinst. De gevonden negatieve leerwinst bij groep 2 heeft te maken met het feit dat de twee toetsen niet geheel gelijkwaardig waren: de posttest was iets moeilijker dan de pretest. Wanneer dit verdisconteerd werd in de berekening, was er in dit experiment sprake van een *gemiddelde* leerwinst.

4 Experiment B

4.1 Methode

Experimenteel ontwerp en procedures
Om het effect van het nakijken te onderscheiden van het pretesteffect, en om alle testpersonen hetzelfde te laten nakijken, alsook de controle op het toepassen van criteria te verbeteren, werd een tweede, ict-ondersteunde opzet ontworpen. Uit een eerder door verge-

Tabel 3

De resultaten en leerwinst van experiment A

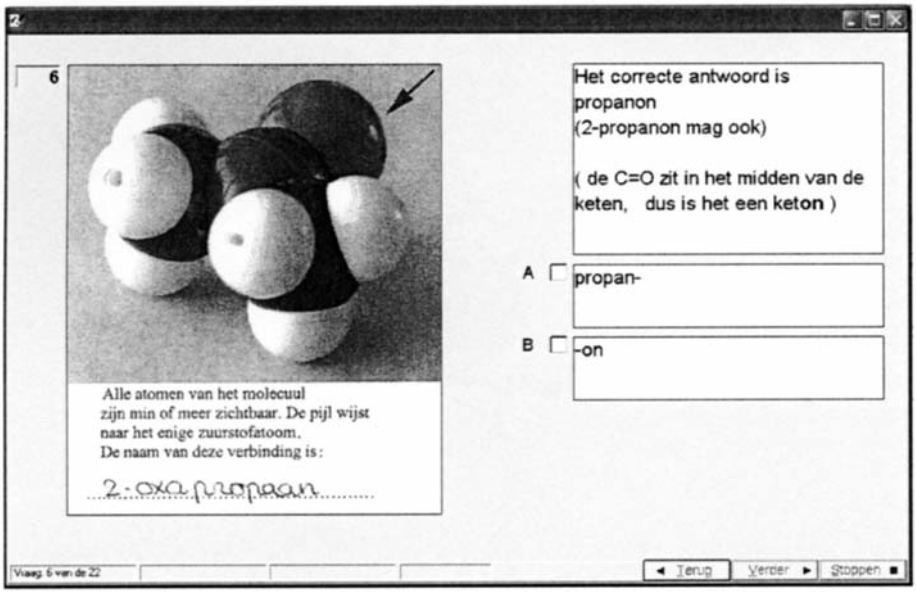
groep →	*1 (O_1 X O_2)*	*2 (O_1 O_2)*	*p*
Pretest (max= 100) gem ± s_d	52,7 ± 15,9	48,9 ± 15,0	0,483[a]
Posttest (max= 100) gem ± s_d	69,1 ± 20,1	46,6 ± 22,0	0,004[a]
Leerwinst *G* gem ± s_e	0,39 ± 0,067	0,00 ± 0,069	≤0,001[t]
Kennisgroei-exponent *B* ± s_e	0,35 ± 0,17	-0,17 ± 0,17	≤0,001[t]

[a] via *F*-toets ; [b] via *t*-toets

lijkbare leerlingen gemaakte schriftelijke toets van 22 vragen werd bij iedere vraag uit alle gemaakte werken één *geschikt* antwoord gekozen: in een pilootonderzoek bleek er een kwadratisch verband te zijn tussen leereffect en score van het na te kijken werk. Op grond van deze bevinding gold als selectiecriterium dat een *geschikt* antwoord niet geheel goed, maar ook niet geheel fout was.

De gebruikte antwoorden hoorden dus niet bij één leerling, maar waren afkomstig van 22 verschillende leerlingen. In deze 22 antwoorden kwamen 23 verschillende deelonderwerpen (stukjes leerstof) pregnant aan de orde. In één antwoord kwamen twee verschillende deelonderwerpen aan de orde. De 22 antwoorden werden tezamen met de bijbehorende vraag gescand (zie Figuur 2). Bij ieder gescand antwoord (met daarbij de oorspronkelijke vraag) werd een nieuwe, homologe korteantwoordvraag ten behoeve van pre- en posttest gemaakt, zodat 23 nieuwe vragen ontstonden.

In plaats van een indeling in experimentele en controlegroepen werd gebruik gemaakt van een variant van het quasi-experimentele *Seperate Sample Pretest-Posttest Control Group Design* (Campbell & Stanley, 1963, p. 55), waarbij in dit geval zowel de pretest als de behandeling (*peer*evaluatie) willekeurig werd toegewezen aan de leerlingen. De posttest werd bij alle proefpersonen afgenomen. De opzet is verwant aan een *Solomon Four Group Design* (Campbell & Stanley, 1963), waar in het onderhavige geval de vier groepen vanuit één groep worden gevormd door willekeurige toewijzing van pretest en behandeling aan de proefpersonen. De techniek in dit experiment zou ook *orthogonale randomisatie* genoemd kunnen worden, waarbij er in gedachten twee dimensies zijn: die van de proefpersonen en haaks (orthogonaal) hierop, die van de diverse behandelelementen. Normaliter vindt er randomisatie



Figuur 2. Voorbeeld van een scherm, waarin de leerling nagaat of een gegeven antwoord voldoet aan de criteria in experiment B.

van proefpersonen plaats. In dit experiment vindt er randomisatie in de andere dimensie plaats, namelijk van behandelelementen.

De werkwijze hierbij was als volgt. Uit een pool van 23 korteantwoordvragen werden voor iedere proefpersoon 12 willekeurig gekozen vragen ter beantwoording aangeboden (O_1). Vervolgens werden uit 22 gescande antwoorden van leerlingen op sterk verwante vragen er 12 willekeurig gekozen die tegelijkertijd met de bijbehorende correctiecriteria werden aangeboden. Wanneer de criteria goed werden toegepast, werd dit gemeld, maar als het correctievoorschrift door de participant onjuist werd toegepast, volgde er onmiddellijk een uitgebreide terugkoppeling (X). Ten slotte werd de complete set van 23 korteantwoordvragen als posttest (O_2) aangeboden.

Er is een kans p = 0,522 (12/23) dat een bepaalde vraag in de pretest (O_1) voorkomt en een kans p = 0,545 (12/22) dat een bepaald deelonderwerp ter correctie wordt aangeboden (X). Op onderwerpsniveau ontstaan dan vier mogelijkheden (zie Tabel 4).

Het voordeel van de opzet is dat iedereen een uniek experiment doet, maar dat de volledige groep gemiddeld hetzelfde experiment doet. Iedere testpersoon krijgt uit dezelfde set pretestvragen en evaluatie-opdrachten willekeurig een subset aangeboden. De gemiddelde deelnemer doet dus gemiddeld hetzelfde. Zowel de interne validiteit als de externe validiteit lijken met deze opzet gewaarborgd te zijn (Campbell & Stanley, 1963, p. 56).

Bij dit ontwerp in experiment B komt iedere deelnemer in alle groepen (categorieën) voor. Het is zeer wel mogelijk, dat de activiteiten door een proefpersoon bij het ene deelonderwerp ook van invloed zijn op de prestaties bij een willekeurig ander deelonderwerp. In dat geval is er sprake van een vorm van transfer die vanuit educatief oogpunt zeker niet ongewenst is. Het is te verwachten, dat de verschillen in score op de posttest tussen de categorieën door dit effect wel kleiner worden.

Tabel 4

Experimenteel ontwerp van experiment B met de kansen dat een bepaalde vraag in de pretest voorkomt (O_1), of dat een bepaald onderwerp ter correctie wordt aangeboden en zonodig van terugkoppeling voorzien (X)

	Niet beoordeeld		*Wel beoordeeld*	*Totaal*
niet in pretest O_2	0,217	XO_2	0,261	0,478
wel in pretest O_1O_2	0,237	O_1XO_2	0,285	0,522
totaal	0,455		0,545	

Het moet nogmaals worden benadrukt, dat het beoogde product van deze nakijkactiviteiten géén cijfer voor een individuele leerling kon zijn, omdat de 22 te corrigeren antwoorden niet van één leerling maar van 22 verschillende leerlingen afkomstig waren. Het was uitdrukkelijk de bedoeling, dat onder gecontroleerde condities nakijkcriteria werden toegepast.

Deelnemers

De leerlingen waren afkomstig uit hetzelfde cohort als de leerlingen in experiment A. Alleen leerlingen bij wie ten tijde van het experiment het onderwerp koolstofchemie in het curriculum voorkwam werden uitgekozen. Aan het experiment namen 44 leerlingen deel, (leeftijd 16,2 ± 0,4 jaar en 25% jongens). Bij dit experiment werden, zoals gezegd, groepen samengesteld door niet de leerlingen, maar de pretestvragen en de onderdelen van de behandeling willekeurig toe te wijzen.

Instrumenten en materialen

Voor experiment B werd met standaard toetssoftware (Wintoets 3.0) een pre-posttest alsook het beoordelingsinstrument geconstrueerd. Voor het dichotoom toekennen van punten aan onderdelen van het antwoord werd het zogenaamde meer-meerkeuzevraagtype als vraagmodel gebruikt. Bij dit type vraag kunnen één of meer items als correct worden aangevinkt. Een voorbeeld van een dergelijke vraag staat in Figuur 2. Bij ieder getoond antwoord werd een correctievoorschrift voor de desbetreffende vraag gegeven.

Bij oppervlakkige beschouwing lijkt deze voorbeeldvraag naar een elementair feit te vragen. Bedacht moet echter worden, dat myriaden van dit soort modellen kunnen worden geconstrueerd. Ook voor afgestudeerde scheikundigen is het herkennen van een der-

gelijk model niet iets wat zij rechtstreeks uit het geheugen kunnen oproepen: ook zij moeten eerst een (voor hen in dit geval simpel) bewust cognitief proces doorlopen (tellen van het aantal C-atomen, kijken waar de H-atomen aan vast zitten, afleiden waar pi-banden zitten enz.). Door vragen te baseren op homologe verbindingen en daarenboven plastic modellen van 3 fabricaten te fotograferen, alsook specifieke grafische programmatuur (*Molecular Modelling software*) te gebruiken, werd gepoogd telkens een dergelijk cognitief proces uit te lokken.

De leerlingen bleken in deze opzet goed te kunnen werken, hoewel ze het in het begin nieuw en verwarrend vonden om over te schakelen van een toets waarbij je zelf het antwoord moet geven naar een opdracht waarbij moet worden nagegaan of er door een ander wel het juiste antwoord gegeven is. De pretest (12 vragen) kostte 13,0 ± 3,9 minuten, de evaluatie (12 opdrachten) 8,5 ± 3,1 minuten en de posttest (23 vragen) 15,0 ± 3,9 minuten.

De leerstof bestond uit een inleiding op koolstofchemie met de volgende onderwerpen: typen molecuulmodellen, notatievarianten, alkanen/isomeren, alkylgroepen, radicalen, carbokationen en carbanionen, onverzadigde verbindingen, (cyclo-)alkanen, alkenen en alkynen, (cis/trans) isomeren, aldehyden, ketonen, en carbonzuren. De leerwinst werd berekend volgens Hake (1998) en Bos e.a. (2007). Daarnaast werd de effectgrootte volgens Cooper (1998) berekend.

4.2 Resultaten

Onderzoeksvraag 2. Leidt het nakijken van het werk van medeleerlingen aan de hand van nakijkcriteria, al dan niet voorafgegaan door een pretest, tot een leerwinst bij de beoordelaar zelf in een met ict-ondersteunde onderwijsleersituatie?

In Tabel 5 staan de pre- en posttestresultaten en de leerwinst voor de vier categorieën in het onderzoeksontwerp. Het verschil tussen het totale gemiddelde van alle pretestantwoordscores in categorie 2 en categorie 3 was in de variantieanalyse niet significant ($F(1, 526) = 1{,}92$; $p = 0{,}17$). De moeilijkheidsgraad van deze categorieën verschilden niet. Aangezien de andere categorieën op dezelfde wijze tot stand zijn gekomen (zie voorgaande paragraaf), lijkt de conclusie gerechtvaardigd dat de moeilijkheidsgraad van alle categorieën gelijk was. De gemiddelde pretestwaarde voor categorie 2 en 3 samen was 23,0 ± 14. Deze waarde werd gebruikt als referentiewaarde voor de leerwinstberekening voor de categorieën, waarin geen pretest werd afgenomen.

De betrouwbaarheid van de posttest was bevredigend (Cronbach's $\alpha = 0{,}85$).

Het effect van de pretest

Om na te gaan of de gemeten verschillen op de posttest van betekenis waren, werd een Bonferroni-analyse uitgevoerd. De resultaten staan in de Tabel 6. Het hoogste resultaat op de posttest werd bereikt in categorie C_3, als dus eerst een pretest werd afgenomen en de leerling daarna een aanverwante vraag van een andere leerling nakeek. Het resultaat was in die categorie significant hoger vergeleken met categorie C_2 (alleen afnemen van een pretest) en vergeleken met de blanco setting (geen ingreep, alleen posttest, categorie C_0). Het resultaat van nakijken en pretest (catego-

Tabel 5

De resultaten in experiment B voor vier designcategorieën

Designcategorie	C_0 Alléén posttest	C_1 Alléén peer-evaluatie	C_2 Alléén pretest	C_3 pretest + *peer*evaluatie
Pretest ± s_d (max= 100)	-	-	20,7 ± 14,0	24,9 ± 14,0
Posttest ± s_d (max= 100)	39,5 ± 17,4	47,9 ± 16,6	37,9 ± 17,3	55,7 ± 16,0
Gain G ± s_e	0,21 (cat)	0,32 (cat)	0,31 ± 0,053	0,41 ± 0,039
Kennisgroei-exponent B ± s_e	0,37 (cat)	0,50 (cat)	0,32 ± 0,13	0,61 ± 0,076

Tabel 6

Significantie (p) volgens Bonferroni van verschillen tussen designcategorieën in posttestresultaten

Designcategorie		Designcategorie	C_0 Alléén posttest	C_1 Alléén peer-evaluatie	C_2 Alléén pretest
C_0	Alléén posttest		-		
C_1	Alléén *peer*evaluatie		0,14	-	
C_2	Alléén pretest		1	0,035	-
C_3	pretest + *peer*evaluatie		$4{,}6\ 10^{-5}$	0,14	$3{,}4\ 10^{-6}$

rie C_3) was echter niet significant hoger dan het nakijken zonder pretest (categorie C_1). Ook was het verschil tussen categorie C_2 en categorie C_0 statistisch niet significant. Het afnemen van een pretest met het oog op een leereffect heeft dus alleen zin, als er een nakijkactiviteit op volgt. De effectgroottes ten opzichte van categorie C_0 (Cooper, 1998) waren voor categorie C_1 0,49, voor categorie C_2 -0,09 en voor categorie C_3 0,97. Met de gemiddelde pretestwaarde voor categorie C_2 en C_3 werd de *gain* **G** en groei-exponent *B* voor categorie C_0 en C_1 berekend. De leerwinst was zowel volgens Hake (1998) als volgens onze criteria (Bos, et al., 2007) voor de categorieën C_0 en C_2 *laag*, en voor de categorie C_1 *gemiddeld*. De leerwinst voor categorie C_3 viel in de categorie *hoog*.

5 Conclusies en discussie

Samengevat zijn er voor de twee vraagstellingen de volgende conclusies:

a) Experiment A. Het toepassen van conventioneel ('papieren') nakijken van andermans werk leidde bij de beoordelaar zelf tot een significant hoger leereffect. Er was sprake van een *gemiddelde* leerwinst.
b) Experiment B. Zowel het toepassen in een ict-ondersteunde onderwijsleersituatie van een combinatie van een pretest met nakijken als het nakijken alléén leidde bij de beoordelaar zelf tot een significant hoger leereffect. De leerwinst voor het nakijken alléén was *gemiddeld*, die voor de combinatie pretest met nakijken *hoog*. Ook werd duidelijk dat het geen zin heeft om alléén een pretest te geven zonder een daarop aansluitende leeractiviteit.

Uit de experimenten zou de conclusie kunnen worden getrokken dat het beoordelen van het werk van een medeleerling effectief is voor het leren van de beoordelaar zelf. Een alternatieve verklaring voor de leerwinst in experiment A lijkt te zijn dat bij het nakijken sprake is van extra oefening met de relevante kennis en de probleemaanpak. Vanuit het perspectief van de lerende beoordelaar is dat inderdaad het geval. Echter, vanuit het perspectief van een docent – en dit is een ander perspectief – is er sprake van een overname van de beoordelingsfunctie door leerlingen, waarmee naar ons oordeel staande kan worden gehouden dat nakijken effectief én efficiënt is. Aangaande experiment B zouden de hogere gemiddelde scores voor de conditie *alleen nakijken van werk van medeleerlingen* en de conditie *pretest en beoordelen van medeleerlingen* (zie Tabel 5) ten opzichte van de conditie *alléén pretest* of *alléén posttest* wellicht ook kunnen worden verklaard uit het feit dat in de eerste twee genoemde condities de proefpersonen al eenzelfde of verwante opgave in de behandeling hebben gezien als in de posttest (interactie van de behandeling met de posttest). Een onafhankelijke posttest bij experiment B had tot een meer sluitende conclusie kunnen leiden.

Ons resultaat lijkt in strijd met de resultaten Sadler en Good (2006). Zij vonden wel een leereffect ($B \approx 0{,}50$) bij het controleren van het eigen werk en *geen* effect bij het controleren van andermans werk. Een mogelijke verklaring zou het optreden van een type II fout kunnen zijn, een onterechte rapportage

van ‘geen effect’. Uit de gerapporteerde gegevens valt af te leiden, dat de kans daarop aanzienlijk is. Een tweede verklaring voor het verschil tussen het effect van het nakijken van eigen werk in tegenstelling tot het nakijken van andermans werk zou ook gevonden kunnen worden in een grotere interesse van leerlingen in het eigen werk dan in dat van een anonieme medeleerling, zeker als de beoordelaar geen voordeel ziet in het moeizame corrigeren van andermans werk. Boud en Falchikov (1989) stellen terecht dat het nakijken ook *beloond* moet worden.

De experimenten A en B in de onderhavige publicatie betroffen het leren van nieuwe natuurwetenschappelijke begrippen in het voortgezet onderwijs waarbij in de kern de beoordelaars gebruik maakten van door de docent verstrekte criteria. De resultaten kunnen niet zonder meer worden gegeneraliseerd naar of worden vergeleken met andere arrangementen in bijvoorbeeld het hoger onderwijs. Zo kunnen bij het opleiden van leraren in het hoger onderwijs componenten als criteria definiëren, een kwalitatief beoordelingsrapport schrijven, en feedback geven effectief en relevant zijn (Sluijsmans & Prins, 2006). Maar dit sluit het nut van andere vormen van controle-activiteiten niet bij voorbaat uit, zoals de in dit onderzoek gebruikte vorm van het toepassen van beoordelingscriteria die door de docent zijn vastgelegd en de controle op het correct toepassen daarvan. Wij zien deze vorm als een eerste stap in het ontwikkelen van hogere orde cognitieve vaardigheden in natuurwetenschappelijke vakken (Zoller, 1999; Zoller et al., 1997). Leerlingen die net zijn begonnen met een natuurwetenschappelijk vakonderdeel zelf beoordelingscriteria te laten ontwikkelen, lijkt ons in dit kader prematuur.

Omdat de accuraatheid en precisie van de beoordelingen door leerlingen vergeleken met die van docenten vanuit onze optiek van ondergeschikt belang waren, hebben wij geen nadere gegevens daarover in de paragraaf met resultaten van experiment A vermeld. Zoals aangegeven in het literatuuroverzicht geven de leerlingen in het algemeen wat lagere scores dan de beroepskrachten. Wij hebben ook dergelijke verschillen gevonden. Deze verschillen tussen de officiële scores en de *peer*evaluaties zijn weliswaar statistisch significant, maar op zich zeer gering (in de orde van 2% van de totaalscore). Ook in de eerder genoemde meta-analyse van 48 studies (Falchikov & Goldfinch, 2000) blijken leerlingen gemiddeld iets lagere cijfers te geven. Falchikov geeft een gewogen gemiddelde effectgrootte van –0,02. Observatie van het evaluatieproces geeft een mogelijke verklaring hiervoor. Wanneer een antwoord iets afwijkt van het correctiemodel zijn leerlingen geneigd het antwoord fout te rekenen, terwijl een docent snel de merites van een alternatieve oplossing doorziet én waardeert.

Op grond van de literatuur mag worden gesteld, dat de verschillen tussen *peer*beoordelaars en officiële beoordelaars worden bepaald door het onderwerp, het type vraag, het correctiemodel en door de kwaliteit van de *peer*beoordelaars. Niet alleen verschillen de *peer*beoordelaars onderling, maar ook de kwaliteit van het na te kijken werkstuk is van invloed. Als een leerling bijna geen vraag heeft beantwoord, valt er niet veel na te kijken. Het leereffect op de beoordelaar zal dan waarschijnlijk beperkt zijn. Zoals reeds aangegeven, leek er in een pilootonderzoek een kwadratisch verband te zijn tussen leereffect en score van het na te kijken werk, maar dit verdient gericht en grootschaliger onderzoek.

In de onderhavige studie was de correlatie tussen de scores gegeven door de *peers* en de officiële scores 0,96. Deze correlatie is gezien de literatuur zéér hoog. Met de aangeleverde criteria voor correctie van toetsen in dit deel van de scheikunde en met dit type leerlingen kan naar ons oordeel een redelijk precies en accuraat beoordelingsresultaat worden bereikt. De echte winst zit echter naar onze mening in de forse leerwinst voor de leerlingbeoordelaar zelf met als bonus een lastenverlichting voor de docent.

In een metastudie over pretesteffecten vonden Willson en Putnam (1982) een verhogend effect van de afname van een pretest op de scores van de posttest met effectgroottes in de orde van 0,30-0,50. Zij concludeerden dat in onderwijskundig, psychologisch en sociologisch onderzoek “there is a general pretest effect which cannot be safely ignored” (Willson & Putnam, 1982, p. 256). Ook wij vonden eerder duidelijke pretesteffecten en

gebruikten deze om de onderwijswinst te verhogen (Bos & Terlouw, 2005). Opmerkelijk genoeg vinden wij in experiment B dergelijke krachtige effecten niet terug. Hoe is dit te verklaren?

Een eerste mogelijke verklaring is het optreden van *transfer,* omdat iedere deelnemer in alle vier de categorieën voorkomt. Weliswaar worden bij het nakijken vragen beoordeeld waarin één bepaald concept aan de orde komt of een bepaalde competentie is vereist, maar het beoordelen van een vraag kan ook een uitstraling hebben naar andere onderwerpen. Het type kennis dat bij deze inleiding van koolstofchemie aan de orde komt, is immers zeer wendbaar. Stukjes kennis die verworven zijn bij het beoordelen van de ene vraag kunnen zeer goed worden gebruikt bij andere onderdelen. De gevonden verschillen tussen de diverse condities (de designcategorieën in de Tabellen 5 en 6) kunnen dan waarschijnlijk ook als ondergrens worden beschouwd van effecten die bij een 'echt' Solomon Four-ontwerp zouden worden gevonden. Een praktisch bezwaar van dit alternatief bij eenzelfde, relatief klein groepsaantal is de dreiging van een type II-fout, omdat de groepsgroottes dan nog maar een kwart zouden zijn; een grotere groepsgrootte is dan noodzakelijk. Wij vermoeden dat de invloed van de pretest bij experiment B waarschijnlijk niet duidelijk uit de verf komt door dit transfereffect. Er wordt weliswaar een lager resultaat behaald bij het nakijken zonder pretest, echter het verschil met de combinatie pretest met nakijken is niet significant.

Een tweede mogelijke verklaring zou het gebruik van de vergelijkingsmethode volgens Bonferroni kunnen zijn: het conservatieve karakter van deze vergelijkingsmethode verkleint waarschijnlijk behoorlijk de kans dat er significante verschillen worden gevonden voor de posttestwaarden van de conditie *alléén pretest* in experiment B (zie Categorie C_2 in de Tabellen 5 en 6).

Uit de resultaten van dit onderzoek willen we dan ook geen negatieve conclusies ten aanzien van positieve effecten van pretesten trekken, zeker niet – gelet op de onderzoeksresultaten – als de pretest onmiddellijke terugkoppeling geeft. Wij vermoeden dat transfer hier het vermogen om onderscheid te maken tussen de diverse effecten verminderd heeft en dat in een echt Solomon Four-ontwerp met voldoende groepsgrootte wel degelijk een significant verschil constateerbaar zal zijn. Het nadeel van het gebruik van effectgroottes komt in het experiment B sterk naar voren. Er wordt vergeleken met een referentie (categorie C_0) en niet met de afzonderlijke pretestscores. Omdat in categorie C_0 er waarschijnlijk door de hiervoor genoemde transfer ook een leereffect optreedt, lijken de effecten kleiner dan de winstmetingen via de groei-exponent B, waarbij de (individuele) posttestresultaten worden vergeleken met individuele pretestresultaten. Het leereffect in de referentiecategorie zou zonder pretest via de gebruikelijke effectmetingen onopgemerkt blijven. We treffen hier een sterk argument aan vóór het O_1XO_2-ontwerp. Anders gezegd, in een onderzoeksontwerp dat gericht is op het vaststellen van de leerwinst zou er -zo mogelijk- altijd een pretest moeten plaatsvinden (Hake, 2001).

Literatuur

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education., 5*(1), 7-74.

Bloxham, S., & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education, 29*, 721-733.

Bos, A. B. H., & Terlouw, C. (2005). *Symposium ORD 2005 ICT-gebruik in bètavakken*. Opgehaald op 21 januari 2007, van http://www.utwente.nl/elan/onderzoek/publicaties/elandoc/2005/2005-01.pdf.

Bos, A. B. H., Terlouw, C., & Pilot, A. (2007). *A Pretest-corrected learning gain*. Opgehaald op 21 januari 2007, van http://www.utwente.nl/elan/onderzoek/publicaties/elandoc/2007/2007-004.pdf.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*, 529-549.

Bruner, J. S., & Olson, D. R. (1973). Learning through experience and learning through media. *Prospects, 3*(1), 20-38.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Chapman, A., & Bloxham, S. (2004). Improving student achievement in a multidisciplinary context. *Learning and Teaching in the Social Sciences, 1*, 181-188.

Clifford, V. A. (1999). The development of autonomous learners in a university setting. *Higher Education Research & Development, 18*, 115-128.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues designs for field settings.* Chicago: Rand McNally College Publishing Company.

Cooper, H. (1998). *Synthetizing research, A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Davies, N. T., Kumtepe, E. G., & Aydeniz, M. (2007). Fostering Continuous Improvement and Learning Through Peer Assessment: Part of an Integral Model of Assessment. *Educational Assessment, 12*, 113-135.

Dochy, F., Admiraal, W., & Pilot, A. (2003). Peer- en co-assessment als instrument voor diepgaand leren : bevindingen en richtlijnen. *Tijdschrift voor Hoger Onderwijs, 21*, 220-229.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education, 24*, 331-350.

Falchikov, N. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment & Evaluation in Higher Education, 22*, 385-396.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis. *Review of Educational Research, 70*, 287-322.

Franken, P. W., Kabel-van den Brand, M. A. W., & Korver, E. J. (1998). *Chemie Overal* (Vol. vwo NG/NT 1). Houten, Nederland: Educatieve Partners Nederland B.V.

Hake, R. R. (1998). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics., 66*, 64-74.

Hake, R. R. (2001). *Pre/Post Paranoia*. Geraadpleegd op 21 januari 2007, van http://lists.asu.edu/cgi-bin/wa?A2=ind0105&L=aera-d&P=R19884.

Langan, A., Wheater, C., Shaw, E., Haines, B., Cullen, W., Boyle, J., et al. (2005). Peer assessment of oral presentations: effects of student gender, university affillation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education, 30*, 21-34.

Minjeong, K. (2005). *The effects of the assessor and assessee's roles on preservice teachers' metacognitive awareness, performance, and attitude in a technology-related design task.* Tallahassee, FL: Florida State University.

Orsmond, P., Merry, S., & Reiling, K. (2002). The Use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*, 309-323.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30*, 51-63.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31.

Sluijsmans, D., Brand-Gruwel, S., & Merriënboer, J. van. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*, 443-454.

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32*, 6-22.

Stefani, L. (1994). Peer, self and tutor assessment: relative abilities. *Studies in Higher Education, 19*, 69-75.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249-276.

Tweede Fase Adviespunt. (2002). *De implementatie van de vernieuwingen in de Tweede Fase van Havo en Vwo*. Geraadpleegd op 21 januari 2007, van www.tweedefase-loket.nl.

Wen, L. M., & Tsai, C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*, 27-44.

William, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*, 537-548.

Willson, V. L., & Putnam, R. R. (1982). A meta-

analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*, 249-258.

Zoller, U. (1999). Scaling-up of higher-order cognitive skills-oriented college Chemistry teaching: An action-oriented research. *Journal of Research in Science Teaching, 36*, 583-596.

Zoller, U., Tsaparlis, G., Fatsow, M., & Lubezky, A. (1997). Student self-assessment of higher order cognitive skills in college science teaching. *Journal of College Science Teaching, 27*, 99-101.

Manuscript aanvaard: 15 mei 2008

Auteurs

Floor Bos is wetenschappelijk onderzoeker bij het Instituut ELAN, Universiteit Twente.

Cees Terlouw is lector Instroommanagement en aansluiting bij Saxion Hogescholen.

Albert Pilot is hoogleraar Didactiek van het curriculum en hoogleraar Chemiedidactiek bij de Universiteit Utrecht.

Correspondentieadres: Floor Bos, Oerdijk 2b, 7433 AA, Schalkhaar. E-mail: abh.bos@home.nl.

Abstract

Learning by marking: The learning gains of the peer assessor by peer marking in pre-university science education

Since teacher time tends to become a scarce commodity, it is relevant to investigate whether transfer of assessment tasks to students can relieve teacher tasks. It is also relevant to investigate a possible learning gain to the peer assessor himself when performing an peer assessment. Learning gain by marking a complete paper-and-pencil test of a fellow student was investigated in a conventional experimental setting in pre university science education. An average learning gain was found. This effect was further investigated in a modelised, computer assisted setting. Maximum learning gain is found when a) a pre-test is made in combination with b) a computerised application of marking criteria.