Instituut ELAN
Instituut voor
Expertise-ontwikkeling in het VO
Lerarenopleiding
Aansluiting VO-HO
Nascholing in het VO

Universiteit Twente
de ondernemende universiteit

Het leereffect op de peerassessor van conventioneel en gecomputeriseerd Peer Assessment in voorgezet bèta-onderwijs

Floris A.B.H. Bos & Cees Terlouw
Faculteit Gedragswetenschappen, Instituut ELAN, Universiteit Twente
Albert Pilot
Centrum voor Didactiek van Wiskunde en Natuurwetenschappen, Universiteit Utrecht
(2006)

ELAN doc 2006 – 03
Mei 2006
GW-ELAN.06.315

Universiteit Twente
GW - ELAN
Postbus 217
7500 AE ENSCHEDE

tel.: 053 - 489 3560
fax: 053 - 489 4755
elan@edte.utwente.nl
http://www.utwente.nl/elan

Het leereffect op de peerassessor van conventioneel en gecomputeriseerd Peer Assessment in voorgezet bèta-onderwijs

Floris A.B.H. Bos, Cees Terlouw & Albert Pilot (2006). Enschede, Nederland: Universiteit van Twente.

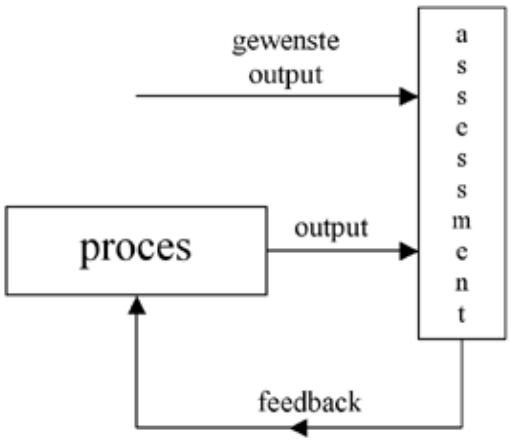
Inhoudsopgave

Samenvatting.
De toepassing van 'peerassessment' (PA) leidt tot een forse leerwinst voor leerlingen, omdat deze verleid worden tot diepe verwerking van de leerstof (onderwijseffectiviteit). Voorts is er tijdwinst voor docenten (onderwijsefficiëntie). Bij sterkere leerlingen is de leerwinst groter dan bij zwakkere leerlingen. Het interventie-onafhankelijke effect bij het gecomputeriseerde assessment-experiment B kan wellicht worden toegeschreven aan diffusie, een aspecifiek effect, te onderscheiden van een specifieke sensitisering en specifiek leren op conceptniveau. Een optimaal leereffect in een digitale omgeving wordt bereikt als er a) een pretest wordt afgenomen ter sensitisering b) een gecomputeriseerd peerassessment wordt afgenomen.

1 Introductie en theoretisch kader.

Het beoordelen of de assessment van leerlingproducten is een complexe taak, en nog steeds een van de essentiële taken van het docentschap. Saddler (Saddler, 1989) plaatst 'assessment' centaal door het een bepaalde *functie* in het leerproces te geven: het via feedback actief verkleinen van het verschil tussen de gewenste en behaalde leerlingprestatie waarbij hij de overdracht van het deels impliciete beoordelingsrepertoire van docent naar leerling met een adequate selfmonitoring als een nevendoel ziet.
Het geven van een assessment in de vorm van een kaal cijfer kan volgens Saddler (1989) amper als feedback beschouwd worden. Een leerling moet:
1) een idee hebben van wat er van hem verwacht wordt,
2) de mogelijkheid krijgen om in te zien in hoeverre zijn prestaties of producten daarmee overeenkomen of afwijken, en tenslotte
3) in staat zijn dat verschil weg te werken.

In figuur 1 hebben wij getracht een en ander te schematiseren.



Figuur 1: cyclus van assessment en feedback op de leerprestatie.

Omdat deze cyclus naar onze opvatting in alle fasen van het leerproces moet voorkomen, ligt adequate selfmonitoring aan de basis van het leerproces. Het lijkt daarbij een goede gedachte om peers in te schakelen voor formatief assessment, omdat er impliciet sprake lijkt te zijn van leren door de peerassessor. Door te kijken naar de producten van anderen krijgt hij een indringend beeld van wat er ook van hemzelf wordt verlangd.
Een dergelijke functionele benadering van Sadler (1989) sluit aan bij het onderwijsleertheoretische kader van Gal'perin. Controle behoort tot de essentiële kenmerken van de handeling volgens Gal'perin's handelingstheorie, en is daardoor ook één van de onderdelen van de hoofdonderwijsfuncties (Terlouw, 2004): er is sprake van formatieve toetsing bij het oefenen en er wordt vervolgens terugkoppeling gegeven; en voorts wordt er teruggekoppeld na de toetsing (zie ook

tabel 1) In tabel 1 staat aangegeven welke onderwijsfuncties naast de hiervoor al genoemde door peerassessment in meerdere of mindere mate zouden kunnen worden gerealiseerd.

A Voorwaardelijke onderwijsfuncties		
A1	Motiveren	X
A2	Aansluiten bij de beginsituatie	
A3	Inzicht geven in de leerdoelen	X
B Hoofdonderwijsfuncties		
B1	Oriënteren	
	ontdekken/verwerven kenniselementen en probleemaanpak (kennis/probleemaanpak)	X
	operationeel maken van de kennis / probleemaanpak	X
B2	Oefenen	
	-oefenen met de kennis / probleemaanpak	
	-geven van terugkoppeling	X
	-gelegenheid geven tot reflecteren	X
B3	Toetsen	
	-welke resultaten zijn er bereikt	X
	-komen de resultaten overeen met de normen	X

Tabel 1: *De mogelijke realisatie door peerassessment van onderwijsfuncties voorvloeiend uit de onderwijsleertheorie van Gal'Perin*

Vooral in het hoger onderwijs geniet peer assessment (PA) een toenemende populariteit onder docenten én onderzoekers in diverse disciplines (Dochy, Segers, & Sluijsmans, 1999; Topping, 1998). De transfer 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).
Er is met name in de assessment-literatuur van voor 1995 is er veel aandacht geweest voor de (summatieve) test als meetinstrument in het kader van certificering. In de nieuwere literatuur worden in diverse bronnen de voordelen gemeld van een bredere (formatieve) toepassing van diverse typen van assessment in alle stadia van het leerproces en het inschakelen daarbij van de studenten zelf (Dochy & McDowell, 1997).
Welke uitkomsten aangaande peerassessment worden zoal in de literatuur gevonden?
In een design van hoge kwaliteit blijken de uiteindelijke waarderingen door peers niet eens zoveel af te wijken van de officiële waardering. In een meta-analyse van 48 kwantitatieve studies op gebied van peer assessment in een breed spectrum van disciplines (Falchikov 2000) werd met name de overeenkomst tussen de hoogte van de scores die door de professionals werd gegeven (*faculty grading*) vergeleken met die van door medestudenten gegeven beoordelingen (*peer grading*). Door deze auteurs werd de overeenkomst met enige reserve in verband gebracht met de validiteit van de beoordeling. In de analyse werd gebruikt gemaakt van 2 maten, een simpele versie van de klassieke *effect size* en daarnaast de correlatiecoëfficiënt. Er werd een gewogen gemiddelde van *effect sizes* van d = -0.02 berekend. Dat wil zeggen, dat de leerlingen elkaar iets lagere cijfers geven dan de officials, maar gezien de enorme spreiding van d (van -0.75 tot +1.25), waarbij ook nog een paar extremen uit de berekening zijn weggelaten, is dit verschil gemiddeld verwaarloosbaar. Het gewogen gemiddelde van de correlatiecoëfficiënten was 0.69 (van 0.14 tot 0.99). De hoogste overeenkomst werd gevonden waar één globale beoordeling van het leerlingproduct moest worden gegeven en waar goed begrepen criteria toegepast werden. Het beoordelen van typische academische producten (testen etc.) bleek een hogere overeenkomst te geven dan het beoordelen van praktijkactiviteiten. Kwalitatief goede onderzoeksdesigns bleken hogere correlaties te geven dan minder goed opgezette onderzoeken. Eerstejaars studenten bleken even goede assessors te zijn als ouderejaars en er bleek ook geen verschil tussen de waarderingen door groepen in vergelijking met waarderingen door enkelingen.
In een verwante, oudere "zuster"-analyse van 57 kwantitatieve self-assessments (Boud & Falchikov, 1989) benadrukten Boud & Falchokov de wens om studenten meer invloed te geven op hun eigen leren. Hier werd echter geconstateerd, dat bij toename van de leeftijd de studenten

zichzelf steeds preciezer maar wel te laag gingen inschatten. De beste studenten waren het meest realistisch en het meest kritisch voor zichzelf.
In deze meta-analyses van peer assessment en self assessment ligt de primaire focus op precisie en validiteit van scoren door studenten. Veel interessanter vinden wij de potentiële leereffecten van assessment in het algemeen en peerassessment in het bijzonder. De overdracht van een deel van docentverantwoordelijkheid aan leerlingen wordt in dezen niet alleen ingegeven door motieven van efficiëntie in onderwijs, maar ook door die voor effectiviteit van leerprocessen, hetgeen spoort met het constructivistische, actieve leren. Het is al langer bekend dat studenten in de rol van tutors meer en dieper leren dan tutees (Annis, 1983). Als studenten iets leren om over het desbetreffende onderwerp een les te verzorgen, nemen de studenten een actievere houding aan dan wanneer zij leren met het oog op een test. Dit verschil komt met name tot uiting in plezier in het leren, interesse en betrokkenheid (Benware & Deci, 1984). In de literatuur wordt de opzet van een Peer Assessment wel gekarakteriseerd een algemene typologie op basis van 17 relevante variabelen (Topping, 1998). Wij zullen die karakterisering ook gebruiken (zie de paragraaf over *Design*). Deze typologie wordt ook gebruikt in een dissertatie over Peer Assessment in universitair onderwijs (Berg, 2003). Van den Berg, Admiraal, & Pilot benadrukken dat bij het ontwerp van peerassessment in het academische schrijfonderwijs, er voor een adequate feedback zowel een formatief als summatief aspect moet zijn, er met kleine groepen gewerkt moet worden en het commentaar mondeling moet worden toegelicht (Berg, Admiraal, & Pilot, 2006).
Wetenschappelijk en praktisch is de aandacht voor efficiëntie en effectiviteit van het onderwijs van belang. In eerder onderzoek werden gecomputeriseerde toetsen gebruikt om leerlingen te sensitiseren voor effectieve leerprocessen (Bos & Terlouw, 2005). In combinatie hiermee kan wellicht leerlingbeoordeling (‘peerassessment’) – zo mogelijk ook in een gecomputeriseerde setting - enerzijds een docent ontlasten, en anderzijds het leren bevorderen

2 Vraagstelling

1. Wat is het leereffect van ‘peerassessment’ op de peerassessor in een conventionele en een gecomputeriseerde onderwijssituatie?
2. Wat is de rol hierbij van (pretest) sensitisering?

3 Methode

3.1 Design en procedure.

Experiment A.
Voor een exploratief experiment werd voor een eenvoudige onderzoeksopzet gekozen. Via een zorgvuldige tweetrapsrandomisatie (Bos, Terlouw, & Pilot, 2006) werden 36 leerlingen uit 4 VWO in twee groepen verdeeld. Alle leerlingen maakten een pretest ('O_1'). In de volgende stap kreeg de ene groep een correctievoorschrift van de pretest waarmee een geanonimiseerde willekeurig gekozen pretest van een andere leerling kon worden nagekeken. Hoewel in de onderwijskunde een test terecht als interventie beschouwd wordt, maken we hier in navolging van Cook en Campbell (Cook & Campbell, 1979) onderscheid tussen het testen en daaropvolgende activiteiten. Het nakijken van het werk van een andere leerling (peerassessment) zullen we hier aanduiden met 'X'. Na het peerassessment maakte groep 1 de posttest 'O_2'. Op hetzelfde moment, dat groep 1 zich bezig hield met het assessment, maakte de tweede groep de posttest 'O_2'. Omdat het een ecologisch experiment betrof, waarbij 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. Een en ander is samengevat in tabel 2.

groep 1	O_1 X O_2
groep 2	O_1 O_2 (X)

Tabel 2: *Experimenteel design van experiment A*

Experiment B.
Om het effect van het peerassessment te onderscheiden van het pré-testeffect en verschillen in kwaliteit van na te kijken werk in te perken werd een tweede, gecomputeriseerde opzet ontworpen. Uit een schriftelijke toets werden antwoorden van verschillende leerlingen ingescand. In de 22 geselecteerde antwoorden kwamen 23 verschillende deelonderwerpen aan de orde. In één antwoord kwamen 2 verschillende deelonderwerpen expliciet aan de orde. De gegeven antwoorden waren niet geheel correct, maar ook niet geheel fout. Bij ieder antwoord werd een nieuwe, homologe kort-antwoordvraag gemaakt, zodat 23 nieuwe vragen ontstonden. Uit de pool van 23 pré-testvragen werden voor iedere proefpersoon 12 willekeurig gekozen vragen ter beantwoording aangeboden ('O_1'). Vervolgens werden er uit de 22 ingescande antwoorden 12 willekeurig gekozen en tegelijkertijd met een correctievoorschrift aangeboden. Wanneer de correctie goed werd uitgevoerd werd dit gemeld, maar als het correctievoorschrift door de participant onjuist werd toegepast, volgde er onmiddellijke uitgebreidere feedback ('X'). Tenslotte werd de complete set van 23 kort-antwoordvragen als posttest ('O_2') aangeboden.
Deze opzet is sterk verwant aan het *Solomon Four Group Design* (Campbell & Stanley, 1979), waarin het product van wel/geen pretest en wel/geen treatment 4 groepen oplevert. Er is 12/23 = 52,2% kans dat een bepaalde vraag in de pretest ('O_1') voorkomt en 12/22 = 54,5% kans, dat een bepaald deelonderwerp ter correctie wordt aangeboden ('X'). Op onderwerpsniveau ontstaan dan 4 mogelijkheden (zie tabel 3).

	non-assessed	assessed	totaal
niet in pretest	'O_2' 21.7%	'XO_2' 26.1%	47.8%
wel in pretest	'O_1O_2' 23.7%	' O_1XO_2' 28.5%	52.2%
totaal	45.5%	54.5%	

Tabel 3: *Experimenteel design 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 ('X')*

Normaliter wordt een randomisatieprocedure 1 maal toegepast op de testpersonen. Als we het aantal proefpersonen op N stellen, wordt er in de geschetste opzet N maal gerandomiseerd met de pré-testvragen en N maal met de treatment-items. Als we in een tabel op iedere regel (de y-richting) de diverse proefpersonen zetten en in de kolommen (de x-richting) de diverse taken, dan wordt er in de regel in de y-richting gerandomiseerd. De door ons gebruikte procedure, staat eigenlijk haaks hierop. We zouden hierom van *orthogonale randomisatie* willen spreken, een soort *Pseudo-Solomon Four-design*. Het voordeel is, dat iedereen een uniek experiment doet, maar dat de volledige groep gemiddeld hetzelfde experiment doet. Iedere testpersoon krijgt uit dezelfde set pré-testvragen en assessmentopdrachten willekeurig een subset aangeboden dus de gemiddelde deelnemer doet gemiddeld hetzelfde. De groep hoeft niet in 4 subgroepen verdeeld te worden. Wanneer er in een design door uitwisseling van informatie tussen de personen van verschillende groepen wederzijdse beïnvloeding optreedt, speekt men van *diffusie*. In de door ons geschetste procedure zit diffusie als impliciete bedreiging van de interne validiteit ingebakken (zie ook paragraaf Discussie).

Een typologie voor beide experimenten volgens Topping staat in tabel 4 (Topping, 1998).

#	Variabele in de typologie	Realisatie
1	Vakgebied/onderwerp	Chemie / zie tabel 6a en 6b
2	Gebruiksdoel	Leermiddel
3	Focus	Kwalitatief/formatief
4	Product	Schriftelijk werk
5	Relatie docentbeoordeling	Vervangend
6	Gewicht van het oordeel	Bijdrage aan het eindcijfer < 5%
7	Oordeelsrichting	Eenzijdig
8	Privacy	Anoniem
9	Contact	Op afstand
10	Jaar	Zelfde opleidingsjaar
11	Bekwaamheid	Gelijk
12	Constellatie beoordelaars	Individueel
13	Constellatie beoordeelden	Individueel
14	Plaats	In contacttijd
15	Tijd	Lestijd
16	Deelnameverplichting	Facultatief
17	Beloning	Marginale toename eindcijfer

Tabel 4: *Typologie van de aard van Peerassessment in de experimenten A en B volgens Topping (1998)*

3.2 Participanten

Uit een totaal van 77 leerlingen uit de Natuur-profielen van 4VWO werden voor experiment A 36 leerlingen geselecteerd. Bij het samenstellen van de groepen werd gebruik gemaakt van de BX, een gerenormaliseerd gemiddelde van alle z-getransformeerde schoolexamen-resultaten uit het voorafgaande semester. Het gemiddelde over de hele jaarlaag van de BX is 100 met een standaarddeviatie van 10. De tweetraps gecomputeriseerde randomisatie verliep als volgt: uit de populatie werd willekeurig een leerling gekozen. Zijn *nearest neighbour* op basis van geslacht en BX werd gezocht. Vervolgens werd de eerste leerling at random óf in groep 1 óf groep 2 geplaatst en de andere leerling in de andere groep. Nadere gegevens staan in tabel 5. De groepen blijken niet van elkaar te verschillen in leeftijd en BX-score gelet op de resultaten van de F-testen en de Fisher-Exact-toets Omdat 3 leerlingen niet het hele experiment mee konden maken, zijn in de uiteindelijke testresultaten de gegevens van 33 leerlingen opgenomen.

Tabel 5 gegevens van participanten in experiment A

groep	1 (O_1 X O_2)	2 (O_1 O_2)	F(1,32)	P
leeftijd ± sd (jr)	15.9 ± 0.32	16.0 ± 0.47	0.57	0.455
BX ± sd	103.4 ± 10.4	104.0 ± 9.21	0.03	0.865
% vrouwelijk	76	69	(Fisher-Exact)	0.50
aantal (N)	17	16		

Voor experiment B waren 44 leerlingen geselecteerd. Bij dit experiment was er geen groepsindeling noodzakelijk. De primaire gegevens staan in tabel 6.

Tabel 6: Gegevens van participanten in experiment B

leeftijd (± sd)	16.2 ± 0.4
% mannelijk	25
aantal (N)	44

3.3 Instrumenten.

Voor experiment A werd een 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 opgaveblad was ruimte voor het opschrijven van het antwoord. Voor het assessment was in het correctiemodel het antwoord op iedere vraag verdeeld in 1-4 essentiële onderdelen. Voor ieder onderdeel mocht een punt worden toegekend. Bij twijfel over het al of niet correct zijn van een vraag mocht een vraagteken worden geplaatst. De posttest bestond uit 15 kort-antwoordvragen, verschillend van de pretest, maar uiteraard wel over dezelfde onderwerpen.
Voor de pretest was ca. 25 minuten nodig, het peer assessment ('X') kostte tussen de 12 en 15 minuten en voor de posttest was ongeveer 15 minuten nodig.
Voor experiment B werd met Wintoets 3.0 een pre/posttest geconstrueerd alsook het assessmentinstrument. Ingescande vragen met bijbehorend leerling-antwoord werden opgeslagen in ca. 400x400 pixel gif-formaat. Voor het dichotoom toekennen van scorepunten aan onderdelen van het antwoord werd het zgn. meer-meerkeuzevraagtype als format gebruikt. Bij ieder getoond antwoord werd een correctievoorschrift voor de desbetreffende vraag gegeven. Het was te merken, dat het aanvankelijk raar was voor een leerling om over te schakelen van een toets, waarbij je zelf het antwoord moet geven naar een setting, 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, het assessment (12 opdrachten) 8.5 ± 3.1 minuten en de posttest (23 vragen) 15.0 ± 3.9 minuten. In tabel 7 staat een overzicht van de leerstof.

bouw en massa van atomen
het Periodiek Systeem
metalen - zouten - moleculaire stoffen
inter- en intramoleculaire wisselwerking
& waterstofbruggen

Tabel 7: *Overzicht van de leerstofonderwerpen in experiment A*

3.4 Materiaal.

De toets van experiment A had betrekking op de inhoud van Chemie Overal deel 1, hoofdstuk 1 (Franken, Kabel-van den Brand, & Korver, 1998). Een aanduiding van de verschillende onderwerpen staat in tabel 6a. Een voorbeeld van een vraag staat in figuur 2.

23. Bereken de verhoudingsformule van de verbinding tussen X en Y:

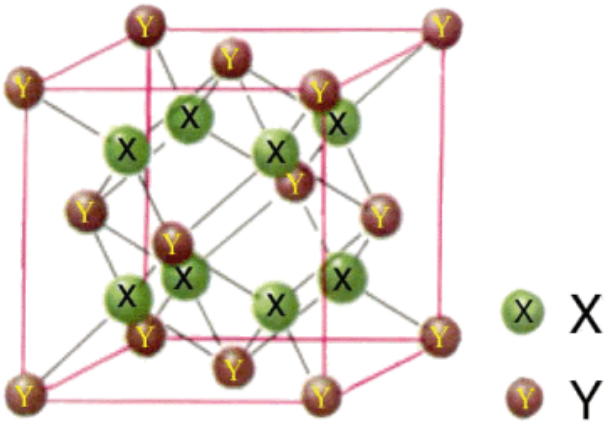


Fig. 2
Voorbeeld van een pré-testvraag, exp. A.

De leerstof in experiment B bestond uit een inleiding op koolstofchemie. In tabel 8 staat een overzicht van de onderwerpen.

typen molecuulmodellen
alkyl-groepen
radicalen, carbokationen, carbanionen
(cyclo-) alkanen, alkenen, alkynen
onverzadigde verbindingen
(cis/trans) isomeren
aldehyden, ketonen, carbonzuren
alcoholen, amines

Tabel 8: *Overzicht van de leerstofonderwerpen in experiment B: de pre- / posttest & assessment taken*

3.5 Correctieprocedure.

Het toekennen van scorepunten aan de open vragen van de testen bij exp. A werd door twee docenten onafhankelijk uitgevoerd aan de hand van een gedetailleerd correctievoorschrift. Op één vraag na (over H-bruggen) bleek er geen systematisch verschil tussen de correctoren. De correlatiecoëfficient tussen de twee beoordelingen bij exp. A was 0.99. Het gemiddelde van de scores van deze twee correctoren werd als de "officiële score" genomen.

3.6 Statistische analyse.

Analyse van testgegevens vond plaats met SPSS 11.0, Vista 6.4, Graphical Analysis 3 en software van de auteurs, geschreven met C++Builder 4 van Borland.

3.7 Bepaling van onderwijswinst.

De gemiddelde genormaliseerde winst <g> is volgens Hake (Hake, 1998) bepaald. De maximale scores voor post en pretest waren op 100 gesteld. De kennisgroeiexponent B (Bos et al., 2006) is, waar mogelijk bepaald via de eerste afgeleide van log(posttest/pretest) tegen log (pretest/100) dan wel uit categoriegemiddelden via de formule

$$B = -\log(\text{posttest}/\text{pretest})/\log(\text{pretest}/100).$$

Een nadere typering van waarden van B staat in tabel 9.

Tabel 9: nominale schaal voor de kennisgroei-exponent (Bos et al., 2006)

exponent	leerwinsttypering
$B \leq 0.40$	laag
$0.40 < B < 0.60$	gemiddeld
$B \geq 0.60$	hoog

Voor de onderwijswinst werden voorts de klassieke effect-grootte E volgens Bloom (Bloom, 1984), 'd' volgens Cohen (Cohen, 1988) en de effectsize d volgens Cooper (Cooper, 1998).

4 Resultaten

Onderzoeksvraag 1a. Wat is het leereffect van 'peerassessment' op de assessor in een conventionele onderwijssituatie?

In tabel 10 staan de resultaten van experiment A weergegeven, alsook de onderwijswinsten.

groep →	1 (O1 X O2)	2 (O1 O2)	F(1,31)	P
pré-test (max=100) gem ± sd	52.7 ± 15.9	48.9 ± 15.0	0.504	0.483
posttest (max=100) gem ± sd	69.1 ± 20.1	46.6 ± 22.0	9.43	0.004
gain (Hake, 1998) gem ± sem	0.39 ± 0.067	0.00 ± 0.069	ttest	0.0004
kennisgroei-exponent "B" (Bos, 2006)	0.35 ± 0.17	-0.17 ± 0.17	ttest	0.0001

Tabel 10: *resultaten experiment A*

De effect-grootte E volgens Bloom, 'd' volgens Cohen en volgens Cooper staan in tabel 11.

methode	'E', 'd'
Cooper (1998)	1.07
Cohen (1988)	1.07
Bloom (1984)	1.02

Tabel 11: *Effectgroottes*

Cronbachs alfa voor de pretest was 0.665 ($P<10^{-3}$) en voor de posttest 0.783 ($P<10^{-3}$).
Tussen pre- en posttest werd een sterke lineaire correlatie gevonden (groep 1 R=0.812 P>0.9999 groep 2: R=0.875 P>0.9999).
Uit de pré-testwaarden van tabel 9 volgt een initiële equivalentie tussen de twee groepen. Uit de gain volgens Hake valt voor groep 1 een classificatie "average gain" af te leiden, typisch voor het zgn. *interactive engagement* onderwijs (Hake, 1998). Uit de kennisgroeiexponent volgt een *lage leerwinst*. De gemiddelde pré-testscores die door de leerlingen werd gegeven was iets lager dan de officiële score. (48.8 ± 14.6). De effectgrootte berekend volgens Cooper (Cooper, 1998) is -0.13. De gepaarde t-test gaf een verschil van -2.0, t(32)=-2.58 P=0.015. Deze discrepantie van -2.0 is statistisch significant, maar over *educatieve* significantie van een verschil van 2 op de 100-schaal valt te twisten. Ook in de 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, maar deze varieert tussen -0.75 en 1.25. De correlatiecoëfficient tussen de scores gegeven door de peers en de officiële scores was 0.96. Deze correlatie is gezien de literatuur zéér hoog. Analyse van discrepanties naar het geslacht van de peercorrector leverde geen verschil (P=0.87): anders gezegd, het maakt niet uit of de peerassessor een jongen of een meisje is.
De individuele leerwinsten varieerden nogal sterk tussen de peerassessors onderling. Om een verband tussen BX, interventie X (= peerassessment) en gain <g> na te gaan werden de leerlingen ingedeeld in 3 groepen (BX ≤ 100; 100 < BX < 110; BX ≥ 110). De resultaten van een 2-way-ANOVA staan in tabel 12.

Source	Sum-of-Squares	df	Mean-Square	F-Ratio	P-Value
Intervention (group)	1.14	1	1.14	17.98	0.0002
BX (categories 1,2,3)	0.52	2	0.26	4.13	0.0263
All sources	1.67	3	0.56	8.75	0.0003
Error	1.84	29	0.06		
Total	3.51	32			

Tabel 12: *Two-way-ANOVA BX-categorie / Interventie (X) op de variabele gain <g>*

Uit nadere inspectie blijkt geringe winst in BX-groepen 1 en 2 een hoge winst in BX-groep 3. De leerlingen met relatief hoge resultaten voor de schoolexamenvakken behalen dus de meeste leerwinst. Niet alleen verschillen de peerassessoren 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 en zal naar ons vermoeden het leereffect op de peerassessor beperkt zijn. We hebben uit onze onderzoeksresultaten bij dit experiment echter geen onderbouwing kunnen afleiden.

Experiment B

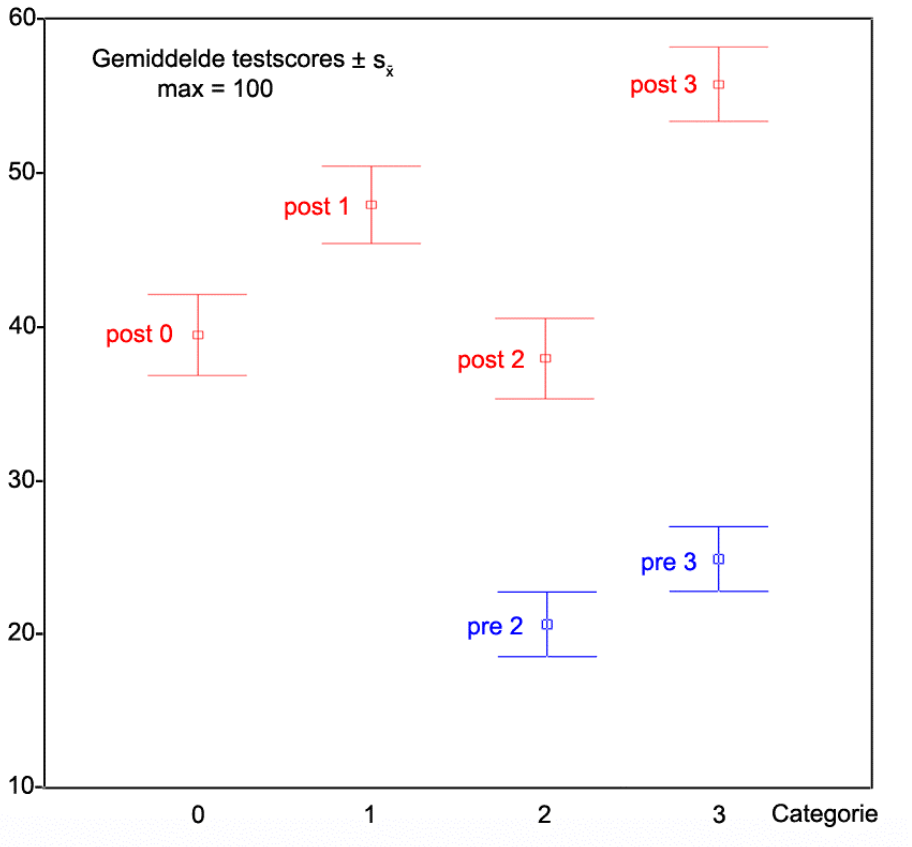
Onderzoeksvraag 1b. Wat is het leereffect van 'peerassessment' op de assessor in een gecomputeriseerde onderwijssituatie en wat is de rol hierbij van (pré-test) sensitisering?

In tabel 13 staan de gemiddelden van de pre- en posttestresultaten voor de diverse categorieën. N.b. de gain en de groei-exponent B voor de categorieën 0 en 1 zijn berekend met behulp van categoriegemiddelden.

categorie→	0	1	2	3
design→	alleen post-test	PA	pre	pre + PA
pré-test ± sd (max=100)	-	-	20.7 ± 14.0	24.9 ± 14.0
posttest ± sd (max=100)	39.5 ± 17.4	47.9 ± 16.6	37.9 ± 17.3	55.7 ± 16.0
gain <g> ± se (Hake, 1998)	0.21 (grp)	0.32 (grp)	0.31 ± 0.053	0.41 ± 0.039
groei-exponent B ± se (Bos, 2006)	0.37 (grp)	0.50 (grp)	0.32 ± 0.13	0.61 ± 0.076

Tabel 13: r*esultaten experiment B*

Cronbach's alfa voor de posttest was 0.85 ($P<10^{-3}$).



Figuur 3. *Gemiddelden per categorie ± se voor pre- en posttest*

Het verschil tussen de pré-testwaarden van categorie 2 en 3 is niet significant F(1,526) = 1.92 (P=0.17). De gemiddelde pré-testwaarde voor categorie 2 en 3 is 23.0 ± 14.

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 14.

	categorie→	0	1	2	3
	design	- -	PA	pre	pre + PA
0	alleen posttest	-	0.14	1	4.6E-05
1	peer assessment (PA)	0.14	-	0.035	0.14
2	pretest (pre)	1	0.035	-	3.4E-06
3	pré-test + peer assessment (pre + PA)	4.6E-05	0.14	3.4E-06	-

Tabel 14: *significantie (P) volgens Bonferroni van verschillen in posttestresultaten*

Met de gemiddelde pré-testwaarde voor categorie 2 en 3 is de gain <g> en groei-exponent B voor categorie 0 en 1 berekend. De leerwinst zowel volgens Hake (Hake, 1998) als volgens onze criteria (Bos et al., 2006) is voor de categorieën 0 en 2 "laag", voor de categorie 1 "gemiddeld". De leerwinst voor categorie 3 valt in de categorie "hoog". Effectgroottes via de klassieke methoden staan in tabel 15.

categorie→	1	2	3
Cooper (1998)	0.49	-0.09	0.97
Cohen (1988)	0.49	-0.09	0.97
Bloom (1984)	0.51	-0.09	1.01

Tabel 15: *Effectgroottes (t.o.v. cat. 0)*

Het hoogste resultaat op de posttest wordt bereikt als eerst een pretest wordt afgenomen, de leerling een aanverwante vraag van een andere leerling nakijkt ("pre +PA)". Het resultaat is ook significant hoger vergeleken met het afnemen van een pretest en vergeleken met de blanco setting (geen ingreep, alleen posttest). Het resultaat van PA+pretest is echter niet significant hoger vergeleken met een PA zonder pretest. Ook is het verschil tussen groep 2 en groep 0 niet significant.

5 Conclusies en discussie

Door toepassen van Peer Assessment wordt een significant hogere leerwinst behaalt in beide settings. De gevonden negatieve leerwinst in experiment A heeft te maken met het feit, dat de twee toetsen niet gelijkwaardig zijn. De posttest was moeilijker dan de pretoets. Wanneer dit verdisconteerd wordt, blijkt bij experiment A een *gemiddelde leerwinst* behaald te worden. Opmerkelijk is de bevinding, dat de betere leerlingen meer opsteken dan de zwakkere.
Opmerkelijk bij experiment B is de leerwinst als er géén peerassessment plaatsvindt. Wij zijn ons ervan bewust, dat de design van de orthogonale randomisatie een inherente bedreiging van de interne validiteit herbergt. Omdat iedere deelnemer in alle vier de groepen voorkomt kan er diffusie (Cook & Campbell, 1979) optreden. Weliswaar worden in het peerassessment vragen beoordeeld waarin één bepaald concept aan de orde komt of een bepaalde competentie vereist is, maar het beoordelen van een vraag kan ook een uitstraling hebben naar andere onderwerpen. De gevonden verschillen tussen de diverse categorieën kunnen dan ook als ondergrens beschouwd worden van effecten die bij een "echt" Solomon Four design zullen worden gevonden. Een praktisch bezwaar van dit alternatief is de dreiging van een type II -fout, omdat de groepsgroottes dan nog maar een kwart zouden zijn. Als gevolg van het diffusie-effect is naar onze mening de invloed van de pretest bij experiment B niet duidelijk. Weliswaar wordt bij de combinatie pretest + peerassessment het hoogste resultaat behaalt. Er wordt wel een lager resultaat behaald bij een peerassessment zonder pretest, maar het verschil is niet significant. Het resultaat

voor posttestwaarden van categorie 2 in exp. B, in combinatie met het strenge karakter van de vergelijkingsmethode volgens Bonferroni beïnvloedt de opmerkelijke significanties in de kolom 1, waar de invloed van pretest naar voren zou moeten komen. In een metastudie vonden Willson en Putnam (Willson & Putnam, 1982), dat de wisselwerking tussen pretest en treatment niet groot was met *effectsizes* in de orde van 0.30-0.50. Zij concludeerden dat in onderwijskundige, psychologische en sociologische research "there is a general pretest effect which cannot be safely ignored". Ook wij vonden eerder duidelijke pré-testeffecten en gebruikten deze om de onderwijswinst te verhogen (Bos & Terlouw, 2005). Uit de resultaten van dit onderzoek willen we dan ook geen negatieve conclusies ten aanzien van positieve effecten van pretesten trekken, zeker niet als de pretest onmiddellijke feedback geeft. Wij vermoeden, dat diffusie hier het oplossend vermogen verminderd heeft en dat in een echt Solomon Four-design met voldoende groepsgrootte wel degelijk een significant verschil constateerbaar zal zijn.
Het nadeel van de klassieke effectmetingen komt in deze case sterk naar voren. Er wordt vergeleken met een referentie (categorie 0) en niet met pretesten. Omdat in categorie 0 er door diffusie wel degelijk een leereffect optreedt schijnen de effecten kleiner dan de vergelijkbare winstmetingen via de Hake-gain en groei-exponent. Het leereffect in de referentiecategorie zou zonder pretest niet waarneembaar zijn. We treffen hier een sterk argument aan vóór het O_1XO_2-design. Anders gezegd: in een goed onderzoeksdesign zou er -zo mogelijk- altijd een pretest moeten plaatsvinden, in weerwil van de gesignaleerde pre/post paranoia (Hake, 2001).
Er vond tijdens of onmiddellijk na de pretest geen feedback plaats. Uit de vergelijking van categorie 0 en 2 blijkt dat het dus geen zin lijkt te hebben om een pretest sec, dus zonder feedback of follow-up te geven. Het enigszins lagere resultaat voor de posttest in groep 2 zou verklaard kunnen worden uit 1) vermoeidheid 2) de weerzin om voor de tweede keer op een vraag in te gaan, waarop geen correct antwoord lijkt te kunnen worden gegeven.
De leerlingen scoren in het algemeen wat lager dan de professionals. Uit observatie van het assessmentproces blijkt een mogelijke verklaring hiervoor. Wanneer een antwoord iets afwijkt van het correctiemodel zijn leerlingen geneigd het antwoord fout te rekenen, terwijl een professional snel de merites van een alternatieve oplossing doorziet én waardeert. De verschillen tussen de officiële scores en de peerassessments zijn weliswaar statistisch significant, maar op zich zeer gering (in de orde van 2% van de totaalscore). Deze verschillen worden bepaald door het onderwerp, type vraag, correctiemodel en door de kwaliteit van de peerassessoren. Het eerste experiment was ook bedoeld om de leerlingen enige ervaring te geven met nakijken van andermans werk. Met de gebruikte correctiemodellen voor toetsen in dit deel van de Scheikunde met dit type leerlingen kan dus een redelijk precies resultaat bereikt worden. De echte winst zit echter naar onze mening in forse leerwinst voor de leerling met als bonus een lastenverlichting voor de docent.

Correspondentie over dit artikel aan A.B.H.Bos, abh.bos(at)home.nl

6 Referenties

Annis, L. F. (1983). The Processes and Effects of Peer Tutoring. *Human Learning, 2*, 39-47.

Benware, C. A., & Deci, E. L. (1984). Quality of Learning With an Active Versus Passive Motivational Set. *American Educational Research Journal, 21*(4), 755-765.

Berg, B. A. M. v. d. (2003). *Peer Assessment in universitair onderwijs.* Universiteit Utrecht, Utrecht.

Berg, B. A. M. v. d., Admiraal, W., & Pilot, A. (2006). Designing student peer assessment in higher education: analysis of written and oral peer feedback. *Teaching in Higher Education, 11*(2), 135-147.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Bos, A. B. H., & Terlouw, C. (2005). *Met ICT gevoelig maken voor het leren van bètabegrippen.* Paper presented at the Meten en Onderwijskundig Onderzoek, Proceedings van de 32e Onderwijs Research Dagen 2005, Gent.

Bos, A. B. H., Terlouw, C., & Pilot, A. (2006). Het leereffect van pré-testsensitivering. *in preparation*.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*(5), 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. (1979). *Quasi-Experimentation. Designs & Analysis Issues for Field Settings.* Boston: Houghton Mifflin Company.

Cohen, J. (1988). Statistical power analysis for the behavioral science. In (2nd ed., pp. 40). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & Analysis Issues Designs for Field Settings.* Boston: Houghton Mifflin Company.

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

Dochy, F. J. R. C., & McDowell, L. (1997). Assessment as a Tool for Learning. *Studies in Educational Evaluation, 23*(4), 279-298.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. M. A. (1999). The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education, 24*(3), 331-350.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis. *Review of Educational Research, 70*(3), 287-322.

Franken, P. W., Kabel-van den Brand, M. A. W., & Korver, E. J. (1998). *Chemie Overal* (Vol. vwo NG/NT 1). Houten: 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. *Am. J. Phys., 66*, 64-74.

Hake, R. R. (2001). *Pre/Post Paranoia*, from <http://lists.asu.edu/cgi-bin/wa?A2=ind0105&L=aera-d&P=R19884>.

Saddler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Terlouw, C., Kramers-Pals, H. & Pilot, A. (2004). Over het leren aanpakken van eindexamenopgaven bij scheikunde in het voortgezet onderwijs. *TDß, 21*(2), 107-144.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pré-test sensitization effects in experimental design. *American Educational Research Journal, 19*, 249-258.