Het effect van een sensitivering door een pretest op de verwerving van natuurwetenschappelijke begrippen

Floris A.B.H. Bos
Faculteit Gedragswetenschappen, Instituut ELAN, Universiteit Twente

Cees Terlouw
Saxion Hogeschool, Enschede

Albert Pilot
Freudenthal Instituut voor Didactiek van Wiskunde en Natuurwetenschappen, Universiteit Utrecht

Samenvatting
De verminderde beschikbaarheid van vakdocenten in het Nederlandse voortgezet onderwijs maakt het gewenst te zoeken naar effectiever onderwijs voor het leren van begrippen in bètavakken. Vanuit de schematheorie en onderzoek naar voorkennistoetsing lijkt het mogelijk om voorafgaand aan daadwerkelijke begripsverwerving relevante, al bestaande begripsnetwerken te activeren met een test. In een experimentele onderwijsopzet is nagegaan of deze sensitivering door een pretest in combinatie met een interactief digitaal systeem de effectiviteit van het onderwijs vergroot. De inhoud van het experimentele onderwijs betrof een oriëntatie op concepten op het grensvlak van de exacte vakken (inclusief informatica). De effecten van deze experimentele onderwijsopzet werden onderzocht in een uitgebreid Solomon Four onderzoeksdesign. Daarin werd het al dan niet operationeel maken van begrippen al dan niet gecombineerd met pretesten waarin zowel kort-antwoordvragen als meerkeuzevragen voorkwamen. In afzonderlijke experimenten werden betrouwbaarheid en equivalentie van de verschillende testen vastgesteld.

De resultaten laten een hoge onderwijswinst zien, het meest na het toepassen van pretesten, maar ook zonder pretesten werden forse winsten gemeten. Er werd een significante interactie tussen pretest en de hoofdtreatment vastgesteld. Er werd geen verschil in pretesteffect van de twee vraagtypen geconstateerd. Het toepassen van een pretest zonder het operationeel maken van de begrippen had geen effect.

1. Rationale

In het voortgezet onderwijs is rond de millenniumwisseling de Tweede Fase ingevoerd vanuit een specifieke visie op het leerproces. Daarbij is de rol van de docent in kwalitatief opzicht in meer of mindere mate veranderd van kennisoverdrager en certificeerder naar coach bij zelfgestuurd, zelfontdekkend dan wel samenwerkend leren.

Tegelijkertijd is op leerlingniveau deze beschikbaarheid voor wat betreft het aantal klokuren van een vakdocent schei-, wis- en natuurkunde sterk gedaald (Tweede Fase Adviespunt, 2005). Omdat dit laatste mogelijk trendmatig is (Ritzen, 2006; Roes, 2001), is het gewenst te zoeken naar efficiënte middelen die tegelijkertijd ook de effectiviteit verhogen. Het inzetten van ICT in optimale settings is dan voor de hand liggend en veelbelovend (Osborne & Hennessy, 2003). Een illustratie daarvan geeft eerder experimenteel onderzoek naar de effectiviteit van een setting rond discovery learning in een elektronische leeromgeving en hulp door medeleerlingen (*peer support*) waarin – door het gebruik van een Solomon Four-group design – tevens de invloed van een pretest op de treatment kon worden vastgesteld (Bos, Terlouw & Pilot, 2007a). Uit dit onderzoek bleek het inzetten van de elektronische leeromgeving tot zeer hoge leerwinst te leiden. Opvallend genoeg bleek de extra leerwinst door het maken van een pretest groter dan de extra leerwinst ten gevolge van inzetten van peer support. Dit resultaat leidde tot de gedachte dat de door de methodologen als problematisch geziene interactie tussen pretest en treatment (bedreiging van de externe validiteit, Campbell & Stanley, 1963) een interessante component is die effectief zou kunnen zijn in het verhogen van het rendement van een met ICT ondersteund onderwijsarrangement. Dit is de eerste vraag die hier aan de orde komt. Het toepassen van het inzicht, dat pretesten vanuit het perspectief van het onderwijs geen probleem maar juist een extra kans zou kunnen zijn, leidde voorts tot de tweede vraag naar de soort van vragen die dan in een pretest zouden moeten worden gesteld. In het genoemde onderzoek werd bij het pretesten gebruik gemaakt van open vragen, terwijl in een ICT-omgeving het gebruik van gesloten vraagvormen gemakkelijker te implementeren is, zeker als er onmiddellijke feedback aan gekoppeld moet worden.

Om een dergelijk ICT-arrangement daadwerkelijk op effectiviteit te toetsen, is het van belang een onderwijsleersituatie te kiezen, waarin een docent normaliter een cruciale rol vervult. Scott, Asoko, & Leach (2007) noemen hier met name de onderwijsleersituatie waarin de docent de ‘sciencetaal’ laat verwerven aan de hand van het leren van basisbegrippen op beschrijvend niveau. Zij zien deze aandacht voor de ‘sciencetaal' als iets specifieks voor het bèta-domein (Scott, Asoko & Leach, 2007). Er is in het onderzoek dan ook voor gekozen om de twee genoemde vragen te onderzoeken bij een digitale oriëntatie op de inhoud van lessen in de vakken ANW (Algemene Natuur Wetenschappen), scheikunde en informatica aan het begin van 4 vwo. Deze digitale oriëntatie bestaat uit het operationeel maken van een aantal met elkaar samenhangende natuurwetenschappelijke basisbegrippen op beschrijvend niveau als elementaire deeltjes, molecuul, atomen, orde van grootte en logaritmische schaal aan de hand van (digitale) teksten met opdrachten, alsmede het gebruik van het BINAS-tabellenboek hierbij.

2. Theoretisch kader

Bij het leren van nieuwe, met elkaar samenhangende natuurwetenschappelijke basisbegrippen is het van belang dat, voorafgaand aan de daadwerkelijke verwerving van een

nieuw begrippennetwerk, de relevante, al bestaande begrippennetwerken of schemata in het lange termijngeheugen geactiveerd worden ('gevoelig' gemaakt worden) om het leggen van verbindingen met nieuwe kennis gemakkelijker te maken (Ausubel, 1968). Een dergelijk 'sensitiveren' of 'gevoelig maken' wordt gezien als het overbrengen van aanwezige schemata vanuit het lange termijngeheugen naar het sneller toegankelijke korte termijngeheugen. Daardoor kan de aanpassing van het bestaande begrippennetwerk gemakkelijker plaatsvinden (Anderson & Schunn, 2000). Op basis van dit 'cognitief gevoelig zijn' voor de nieuw te leren begrippen kan vervolgens de daadwerkelijke activering en daarmee aanpassing van een bestaand begrippennetwerk effectiever plaatsvinden. Het is derhalve noodzakelijk dat een bestaand relevant begrippennetwerk, de voorkennis, wordt geactiveerd. Hoe kan dit plaatsvinden?

Studenten laten reflecteren en expliciteren

Strangman, Hall en Meyer (2004) beschrijven een aantal strategieën in hun review van onderzoek naar de activering van voorkennis voor het begrijpen van onder andere scienceteksten. Het laten reflecteren gevolgd door een explicitering is de meest simpele aanpak waarin studenten wordt gevraagd na te denken over wat ze al weten van een bepaald onderwerp en dit op te schrijven of mondeling te rapporteren. Ook een vorm als 'reciprocal teaching' waarin studenten aan elkaar uitleggen, kan worden gebruikt. Aan deze strategie kan ook heel goed een interactieve discussie worden gekoppeld ter bevordering van de explicitering van de voorkennis.

Het stellen van vragen

'Pretesting' is een vorm van het stellen van vragen waarin de assessment van voorkennis resulteert in een activering van die voorkennis (Dochy, Segers & Buehl, 1999).

Dochy et al. (1999) concluderen in hun review over de relatie tussen voorkennisassessment en leerresultaten dat er een sterke relatie bestaat tussen voorkennis en prestatie: 92% van de 183 gereviewde studies rapporteren positieve effecten, waarbij de voorkennis 30% tot 60% van de variantie verklaart in de leerprestaties. Hierbij heeft met name het gebruik van objectieve assessmentmethoden een positieve invloed (Dochy & Alexander, 1995). Daarnaast zijn ook leerprocesvariabelen van belang als de leerstrategie, de procedurele metacognitieve kennis, de belangstelling en de 'beliefs'.

Het effect van een assessment van de voorkennis is, zoals hiervoor aangegeven, ook bekend uit de testmethodologie: de veelal als ongewenst geziene effecten van de pretestsensitivering (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002). Twee soorten van effecten worden gerapporteerd:

a. het testing-effect: het effect van de pretest op de posttest als de pretest als posttest de tweede keer wordt afgenomen. Het is één van de bedreigingen van de interne validiteit van een experiment; en

b. het effect van de interactie tussen de pretest en de treatment. Dit is één van de bedreigingen van de externe validiteit (Lana, 1959, 1969).

In dit onderzoek maken wij, zoals hiervoor aangegeven, vanuit een onderwijskundig perspectief gebruik van het tweede effect, de pretest-treatment interactie. Lana en King (1960) analyseerden de aard van de pretest-treatment interactie en wezen op de relevantie van leerfactoren die vergelijkbaar zijn met de hiervoor genoemde leerprocesvariabelen van Dochy et al. (1999). Willson en Putnam (1982) voerden een meta-analyse aangaande pretest sensitization effects uit over 32 studies met een totaal van 164 resultaten (Willson & Putnam, 1982). Voor de analyse daarvan werden de 134 resultaten gebruikt waarin in het experimentele design gebruik was gemaakt van gerandomiseerde groepen. Er is sprake van een significant en sterk pretest-effect met een gemiddelde effectsize van +0.22 (tussen -0.55 en +4.06), waarbij soort uitkomst, leeftijd en tijd tussen de pre- en posttesten sterk van invloed zijn. Het effect op cognitieve uitkomsten is duidelijk en groot (gemiddelde effect size +0.48).

Computerondersteuning

Biemans (1997) hanteerde in zijn onderzoek een computerondersteuning waarin het stellen van vragen was geïntegreerd. Met het bevorderen van conceptual change als doel testte en verbeterde Biemans een computerondersteund, procesgericht, heuristisch activeringsmodel beschreven door Ali (1990). Preconcepties van basisschoolleerlingen (groep 7/8) op het gebied van de fysische geografie, met andere woorden constructen gevormd vóór de formele instructie door alledaagse ervaring met dergelijke verschijnselen, werden allereerst geactiveerd door middel van een zogenaamde ideevraag (*idea question*): aan de individuele leerling werd een centrale vraag over het betreffende onderwerp (bijvoorbeeld het weer) gesteld en een aantal verklaringsalternatieven geboden, waaronder de algemeen geaccepteerde wetenschappelijke verklaring. De leerlingen moesten een keuze maken en deze vervolgens beargumenteren. Vervolgens kregen ze een informatieve tekst te lezen waarin het betreffende verschijnsel werd uitgelegd. Hun eigen idee dienden de leerlingen vervolgens te vergelijken en te contrasteren met de nieuwe informatie. Als eerder het niet-wetenschappelijk aanvaarde alternatief was gekozen, was het de bedoeling dat inconsistenties werden ontdekt tussen het eigen idee en de aangeboden tekst. Vervolgens werd aan de leerlingen gevraagd om opnieuw hun idee te formuleren, en dit nieuwe idee toe te passen in een andere context en te evalueren. Hierna was ook weer aanpassing van het eigen idee mogelijk. Interessant is dat, vergeleken met de referentiegroep, de hoogste scores werden gehaald voor toetsvragen die direct gerelateerd waren aan concepten uit de geactiveerde kennis, hetgeen enige steun leverde voor ‘selectieve aandacht’ als verklaringshypothese. Vergeleken met de referentiegroep werden lagere scores gehaald voor toetsvragen die minder gerelateerd waren aan de kennis rond de ‘idea question’. In een vervolgstudie werd dit ongewenste effect geëlimineerd door de aandacht ook te richten op belangrijke informatie in de tekst die niet essen-

tieel was voor het beantwoorden van de idea question, strategische informatie te beperken, gemakkelijker toegang te geven tot andere informatie en leerlingen die meteen het wetenschappelijk correcte alternatief kozen, sneller verder te laten werken (Biemans, 1997).

De studie van Biemans wijst op de mogelijkheid van een effectief onderwijsarrangement dat de docent kan ontlasten. Voorts lijkt, gelet op de bevindingen in de review van Dochy et al. (1999), een activering van voorkennis via pretesten met een objectieve assessmentmethode kansrijk, terwijl door inzet van ICT docentbelasting binnen de perken blijft. In een digitale leeromgeving ligt het immers voor de hand over te gaan tot het gebruik van een objectieve assessmentmethode in de vorm van gesloten vragen die de computer zelf kan nakijken en van feedback kan voorzien. Echter, de bevindingen van Strangman et al. (2004) maken ook plausibel dat het zelf formuleren van antwoorden door studenten eveneens bijdraagt aan de effectiviteit. Dit pleit voor vragen waarop leerlingen zelf hun antwoorden formuleren en niet voor gesloten vragen. De vraag rijst dan ook of de effectiviteit van pretestsensitivering mede afhangt van het vraagtype. Aangezien wij zullen werken met een digitaal onderwijssysteem, dat moeite heeft met volledig open vragen, is er gekozen voor twee vraagtypen: gesloten meerkeuzevragen en kort-antwoordvragen waarbij leerlingen zelf kort hun antwoord formuleren.

3. Vraagstellingen

De algemene probleemstelling is of een ondersteunend interactief digitaal onderwijssysteem met (a) een component ‘gevoelig maken’ aan de hand van een pretest (pretestsensitiveren), en (b) een component ‘operationeel maken’ van de te leren natuurwetenschappelijke begrippen aan de hand van opdrachten en daarbij horende terugkoppeling, leidt tot een betere verwerving van nieuwe natuurwetenschappelijke begrippen, zoals gemeten in de toetsing. We vragen ons in het bijzonder af of bij de pretest een toets met kort-antwoordvragen (KAV) of één met meerkeuzevragen (MKV) effectiever is. De onderzoeksvragen zijn derhalve:

1. Is het mogelijk om met een pretest tot betere leerresultaten te komen met een interactief digitaal onderwijssysteem, waarin de leerlingen opdrachten uitvoeren om nieuwe natuurwetenschappelijke begrippen operationeel te maken ?
2. Is er een verschil in leerresultaten tussen een pretest met kort-antwoordvragen (KAV) of meerkeuzevragen (MKV) ?

4. Methode

We gaan achtereenvolgens in op het design, de participanten, de instrumenten, het materiaal, de procedure, de correctieprocedure, de statistische analyse, en de bepaling van de onderwijswinst.

Design

Het Solomon Four-group design (Campbell & Stanley, 1963) is een experimentele onderzoeksopzet waarmee het effect van een pretest, de treatment en de interactie van pretest en treatment kunnen worden geanalyseerd. De generaliseerbaarheid (externe validiteit) van het experiment wordt ermee verhoogd, omdat het genoemde interactie-effect kan worden nagegaan. Scharfenberg, Bogner en Klautke (2006) bevelen dit onderzoeksdesign met nadruk aan voor science education research. In de oorspronkelijke Solomon Four-opzet (Solomon, 1949) is er sprake van vier groepen: (wel/geen pretest) * (wel/geen treatment). Omdat er in onze opzet sprake is van een tweetal typen pretesten (KAV/MKV) is hier sprake van een uitgebreide Solomon 4 opzet (tabel 1).

Tabel 1. uitgebreide Solomon 4 opzet

	geen pretest	pretest A	pretest B
geen treatment	groep 1	groep 2	groep 3
wel treatment	groep 4	groep 5	groep 6

Participanten

Het experiment werd in delen uitgevoerd met in totaal 84 vwo-leerlingen uit het Natuurprofiel. Uit 4 vwo namen 84 leerlingen, gemiddeld 15,5 jaar oud, deel aan het hoofdexperiment, 30 leerlingen uit 4 vwo deden mee aan een hertest van de pretesten. Verder deden nog 70 willekeurig gekozen leerlingen uit 4, 5 en 6 vwo mee aan een kalibratie van pre- en posttesten.

Vanwege het grote belang voor het experiment van de groepen 5 en 6 werden de leerlingen in een tweede (gecomputeriseerde) tweetraps randomisatieprocedure nogmaals verdeeld: at random werd een leerling uit de totale groep gekozen. Vervolgens werd zijn 'nearest neighbour' gezocht op basis van de criteria gemiddelde, paasrapport en geslacht. De eerste leerling werd at random geplaatst of in groep 5 of in groep 6 (en de tweede leerling in de andere groep). Deze procedure werd herhaald, tot alle leerlingen ingedeeld waren.

Instrumenten

Twee digitale pretesten (A en B) werden met het auteurssysteem Wintoets gemaakt, bestaande uit respectievelijk 16 KAV en 15 MKV, waarvan enkele tweekeuzevragen. Deze pretesten meten kennis en begrip van de tot doel gestelde science-basisbegrippen op beschrijvend niveau. Wanneer in de ene test een bepaald leerstofonderdeel in een KAV werd bevraagd, kwam in de andere toets hetzelfde item voor in de vorm van een MKV en omgekeerd. In verband hiermee is het dienstig rekening te houden met *interne diffusie;* wanneer er een bepaald onderwerp via een bepaalde maatregel geactiveerd wordt, ligt het voor de hand, dat daaraan sterk geassocieerde kennis ook beïnvloed wordt (Lawson

& Chinnappan, 2000). Er was dus een onderzoekstechnische noodzaak om effecten bij uiteenlopende onderwerpen te meten, teneinde het gevaar van diffusie te verkleinen.

Omdat het Wintoetspakket uitgebreide digitale presentatievormen kent en een groot deel van de gewenste onderzoeksadministratie kan verzorgen, werd hiermee ook het lesmateriaal (= de treatment) en de posttest gebouwd. Een voordeel hiervan was dat de leerlingen met dezelfde interface werden geconfronteerd.

De posttest, verschillend van, maar equivalent aan de pretesten bestond uit in totaal 32 vragen: 24 vragen, waar één of een paar woorden moest worden ingevuld, zes gatenteksten waarin één woord of formule worden ingevuld, en twee vragen waar een getal moest worden ingevuld. Ook de posttest meet kennis en begrip van de tot doel gestelde science-basisbegrippen op beschrijvend niveau. Dat in dit onderzoek in de posttest uitsluitend gekozen is voor open vragen wordt ingegeven door de overweging, dat het gokelement bij meerkeuzevragen leidt tot grotere foutvarianties en dus tot lagere precisie (Zimmerman, 2003). Met name polytoom gescoorde open vragen blijken betrouwbaarder dan meerkeuzevragen. Eenduidige uitspraken over verschil in validiteit in zijn algemeenheid lijken op grond van de literatuur a priori moeilijk te geven aangezien het kennisdomein en het doel van de toetsing nogal van invloed is (Kuhlemeier, Steentjes & Kleintjes, 2003). Er kan wel verwacht worden, dat open vragen moeilijker zijn dan meerkeuzevragen, omdat het antwoord door de leerlingen zelf geconstrueerd moet worden, 'terugverwijzing' via de keuze-items ontbreekt en verhoging van de kans op een correct antwoord door eliminatie van onwaarschijnlijke keuze-items is niet mogelijk. In het kader van sensitivering zou dit overigens een voordeel kunnen zijn.

Ten slotte zij opgemerkt dat bij géén van de testen tijdens het experiment feedback werd gegeven.

Interventie, lesmateriaal

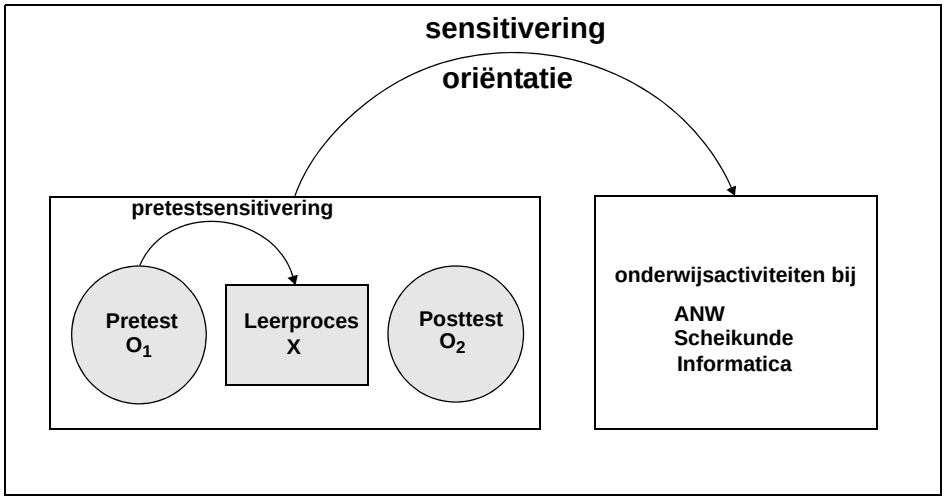
Het onderwijsdoel van de interventie was een eerste oriëntatie op de inhoud van diverse geplande activiteiten aan het begin van 4 vwo:

a. lessen scheikunde over atoom- en molecuulbouw en een stukje koolstofscheikunde speciaal ten behoeve van biologie.
b. een lezing van een hoogleraar theoretische fysica over structuur van de materie gekoppeld aan posterpresentaties daarover door leerlingen (Bais, 2004).
c. een klein project nanotechnologie ANW/Na/Sk enerzijds en Engels, 'Oscillating cantilevers' (Ilic & Craighead, 2004) en Duits, 'Nukleare Mikrobatterien' (Schroeder, 2004) anderzijds.
d. lessen informatica over beeldschermkleuren.

De uitdaging was om deze zeer diverse onderwerpen toch met duidelijke links als één doorlopend verhaal te presenteren.

Het lesmateriaal bestond uit een aantal korte opdrachten voor het leren van nieuwe basis bètabegrippen op beschrijvend niveau die leerlingen naar aanleiding van diverse grafische representaties uit het digitale systeem uitvoerden; het digitale systeem gaf tevens onmiddellijke, beknopte feedback. Meer concreet kwam dit neer op:

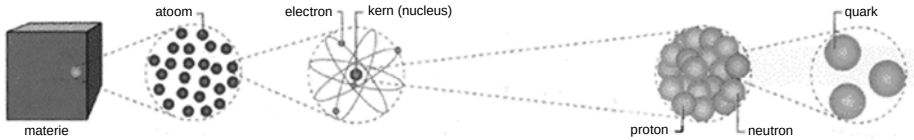
– gebruik/toepassen van het BINAS-tabellenboek
– uitleg en operationeel maken van kennis rond elementaire deeltjes en natuurkrachten
– gebruik van machten van 10 en logaritmische schalen
– het benoemen van de belangrijkste atoomsoorten in het menselijk lichaam (H,O,C,N en sporenelementen)
– het introduceren van kleurconventies van molecuulmodellen
– het genereren van (deze) kleuren op een beeldscherm.



Figuur 1. Plaats en structuur van dit onderzoek in het beoogde onderwijs 4 vwo. Er is sprake van *geneste* sensitivering.

Na een informatiescherm waarin de leerlingen de bedoeling van de interventie werd uitgelegd, verscheen een scherm waarin werd gesteld, dat van een mug geen olifant te maken was, maar op *atomair niveau* de bouwstenen wel dezelfde zijn. Na uitleg van het begrip *orde* (van grootte) werd de leerlingen gevraagd massa's van mug en olifant met elkaar te vergelijken met behulp van BINAS tabel 2 (*vermenigvuldigingsfactoren*). Zonodig was er uitleg over het gebruik hierbij van de grafische rekenmachine.

Vervolgens werd gevraagd via het *register* in BINAS de tabel 26 *Bouw en structuur van de materie* te zoeken (zie ook figuur 2). Omdat in de centrale figuur van deze tabel wordt



uitgegaan van een blok metaal, ontbreekt het sterk verwante begrip *molecuul*. De leerling werd gevraagd naar dit begrip.

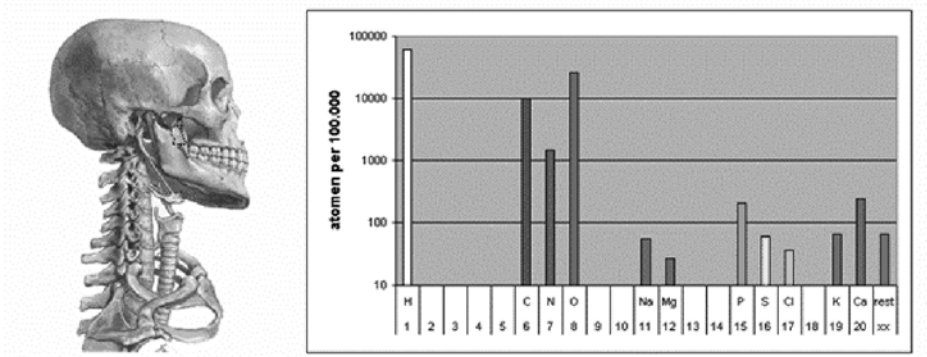
Figuur 2. Voorbeeld van activering van een natuurwetenschappelijk begrip. De opdracht luidt: *De getoonde kubus is een stuk metaal, waarin dus geen aparte [input: goed=moleculen] voorkomen.* (bron: (BINAS, 2004).

Vervolgens werd gevraagd met behulp van BINAS-tabel 1 het Griekse woord ατομοςte vertalen. Als extra feedback kwam er uitleg over de negentiende-eeuwse oorsprong van het woord atoom. Via eenvoudig met BINAS-tabel 26 te beantwoorden vragen werd de aandacht gevestigd op *bekende hadronen, leptonen, wisselwerkingsdeeltjes* en *natuurkrachten*. Bij de behandeling van deze tabel kwam ook een animatie van drie *quarks* en een *gluon* voor. Als afsluiting van dit stuk werd gevraagd tabel 26 uit BINAS te schematiseren op een fotokopie.

In het verlengde van het begrip *orde van grootte* ligt het werken met machten van 10 en daarmee samenhangend de *logaritmische schaal*. Na uitleg van het principe van zo'n schaal werd de leerling gevraagd uit drie assen (een lineaire, een logaritmische en een fantasieschaal) de logaritmische te kiezen. Het nut van een logaritmische schaal werd getoond door de afmetingen van proton, atoom, bacterie, mug en mens op één as te plaatsen. Hiermee was het gemakkelijk een idee te geven van afmeting van structuren in de nanotechnologie. Via de *oscillating cantilever*, waarmee de gezamenlijke massa van een paar duizend moleculen kan worden gemeten, werd ingezoomd op extreem grote aantallen en kleine afmetingen. Voor de belangrijkste atoomsoorten in het menselijk lichaam werd naar BINAS-tabel 34 (*Samenstelling*) verwezen (zie ook figuur 3). Ook in deze tabel wordt een logaritmische schaal (voor de y-as) gebruikt. Dezelfde gegevens werden ook op een 'gewone' schaal getoond met behulp van een taartdiagram.

Gevraagd werd nu via de CPK-kleuren de vier meest frequente elementen (H, O, C en N) te benoemen. In figuur 3 is te zien hoe bijvoorbeeld ingezoomd werd op het element P (fosfor) in tabel 34. Dit onderdeel werd afgesloten met uitleg van het begrip *sporenelementen*, een leesstukje over anemie en tekorten aan de elementen ijzer en cobalt in het lichaam (en een simpele vraag hierover).

Tabel 40 uit BINAS (*Elementen*) werd toegepast om afmetingen van atomen (in *pm*) om te zetten in representaties op het beeldscherm (gekleurde cirkels met een bepaalde



Het skelet bestaat uit water, eiwit en een vorm van calciumfosfaat. Dit wijst naast calcium, C, H, O en N op het element (geef het symbool :) ☐
Tip: kijk ook in het histogram.

doorsnede in pixels). Er werd uitgelegd hoe je een en ander in een grafische editor kunt verwezenlijken.

Figuur 3. Voorbeeld van een invulvraag. De y-as van het histogram is logaritmisch en de staafjes hebben CPK-kleuren. Bronnen : Schedel – met dank aan Ciba-Geigy. Diagram – een sterk verwant staafdiagram staat in (BINAS, 2004).

Vervolgens werd duidelijk gemaakt hoe gewenste *websafe CPK-kleuren* van atoommodellen via een *RGB-code* kunnen worden gemaakt. Net als in de hele interventie werd hier dus telkens na het geven van enige informatie of aanwijzingen gevraagd het gepresenteerde meteen toe te passen. Vervolgens werd onmiddellijk feedback gegeven.

In deze interventie werden dus nogal uiteenlopende onderwerpen via duidelijke *bruggetjes* tot één aaneensluitend, doorlopend verhaal gecombineerd. Een centrale rol was weggelegd voor het BINAS-tabellenboek als belangrijke, blijvende informatiebron. Het lesmateriaal was dus duidelijk verbonden met de onderwerpen die in andere settings in de weken erna aan de orde zouden komen.

De gebruikte vragen in de interventie waren van het type gatentekst (13), woord (4), meerkeuze (6), en tweekeuze (1). Daarnaast waren er 12 infoschermen.

In bijlage 1 staat een schematisch overzicht van de natuurwetenschappelijke basisbegrippen dat als bron diende voor de pretestvragen en de opdrachten in het interactieve digitale systeem.

Ten slotte zij opgemerkt dat de werkwijze die in de interventie, het onderwijs, wordt gevolgd componenten bevat die Treagust (2007) als specifiek ziet voor het bèta-onderwijs (Treagust, 2007):

– sciencebegrippen worden visueel gedemonstreerd aan de hand van met computersoftware gemaakte animaties (Linn, 2003);
– sciencebegrippen worden nadrukkelijk expliciet uitgelegd aan de hand van kernbegrippen in science (bijvoorbeeld molecuul, atoom) (Ogborn, Kress, Martin & McGillycuddy, 1996). Dit gebeurt in zorgvuldig ontworpen informatieschermen;
– het operationeel maken van de sciencebegrippen gebeurt aan de hand van opdrachten en vragen waarbij op de antwoorden direct terugkoppeling wordt gegeven (Lemke, 1990). Voorts maken de leerlingen daarbij kennis met een 'tool' (Vygotsky, 1978) dat specifiek is voor het bètadomein, het BINAS-tabellenboek; en
– verschillende representatievormen worden gebruikt om de sciencebegrippen te laten begrijpen: staafdiagrammen, taartdiagrammen, tabellen, plaatjes, animaties, etc. (Kozma, 2000). Hierbij wordt er ook rekening gehouden met het niveau van representatie (Johnstone, 1991): macro-niveau (de zichtbare, concrete wereld van verschijnselen, bijvoorbeeld de vergelijking van de massa van de mug met de olifant), het (onzichtbare) micro-niveau (bijvoorbeeld molecuul, atomen), en het symbolische niveau (bijvoorbeeld formules). De interventie beperkt zich tot de eerste twee niveaus.

Procedure

De leerlingen werd vooraf meegedeeld dat ze meewerkten aan een onderzoek, waarvan de inhoud te maken had met lessen ANW, scheikunde en informatica, daarmee samenhangende activiteiten in de komende periode, en dat er geen consequenties waren verbonden aan deelname. Het maken van een pretest A of B nam ca. 10 minuten in beslag, het doorwerken van het lesmateriaal ca. 40 minuten. Tussen de diverse onderdelen waren geen pauzes. Bij alle groepen werd de posttest afgenomen bestaande uit 32 KAV. De afname van de posttest nam 10 à 15 minuten in beslag.

Correctieprocedure

Het scoren van de open vragen verliep als volgt: de sleutel (het correcte antwoord) was in een aantal gevallen één woord (quark, positron etc.), een symbool (P voor fosfor), een uitdrukking (4+C) of een getal. Alle antwoorden werden daarom in de vorm (*vraag ID, student ID, antwoord*) opgeslagen in een relationele database en in deze vorm tezamen met een strak correctieprotocol aan twee onafhankelijke correctoren voorgelegd. In uiteindelijk 1% van alle antwoorden was er discrepantie tussen de twee correctoren en werd voor de berekeningen het gemiddelde genomen.

Statistische analyse

Met SPSS werd een variantieanalyse en een meervoudige vergelijking volgens Bonferroni uitgevoerd (significantiecriterium 5%), alsook Cronbachs alfa berekend. Met het VISTA 6-pakket werd een 2-way-ANOVA uitgevoerd.

Bepaling van onderwijswinst

Wanneer er alleen naar posttestresultaten wordt gekeken, kunnen alleen verschillen in uiteindelijke effecten van diverse behandelingen in beeld komen. Voor een berekening van winst is het nodig het aanvangsniveau te kennen. Om het effect van een treatment bij leerlingen of groepen van leerlingen met verschillend pretestniveau met elkaar te kunnen vergelijken, is het noodzakelijk voor het pretestniveau te corrigeren.

Bij diverse test-retest-experimenten werd in de vakken Frans, informatica en scheikunde empirisch een sterk verband vastgesteld tussen posttest en pretest (Bos, Terlouw & Pilot, 2007b). Dit verband kan worden gebruikt om de variabele *pretest* te elimineren en aldus onderwijswinst te berekenen.

Wanneer de pretestscores worden gedeeld door de maximaal te behalen pretestscore en we deze variabele x noemen (x = pretestscore/maximum_pretestscore $0 \leq x \leq 1$) en dezelfde bewerking wordt toegepast op de posttestscores en deze variabele y noemen, kan de verhouding $f = y/x$ met de machtsfunctie $f = x^{-B}$ worden beschreven. De exponent B is een robuuste maat voor de kennisgroei in een pretest-interventie-posttest-design (OXO'-design). Normaliter is het posttestresultaat groter dan het pretestresultaat (anders is er niets geleerd) en ligt B tussen 0 en 1. Met de schattingen van de fout in parameter B kunnen uitspraken over de significantie van verschillen worden ondersteund. Via ijking met behulp van data uit een review van Hake (1998) is een nadere nominale karakterisering van B vastgesteld. Als $B \leq 0.40$ wordt de leerwinst gekarakteriseerd als *laag*, met waarden voor $B \geq 0.60$ is de leerwinst *hoog* te noemen. Tussenliggende waarden $0.40 < B < 0.60$ krijgen de leerwinsttypering *gemiddeld*.

Een apart probleem treedt op bij het berekenen van leerwinst bij groep 4 (zie tabel 1). Formeel gesproken is het berekenen van leerwinst onmogelijk, omdat deze groep géén pretest maakt. Wanneer echter op goede gronden wordt uitgegaan van equivalentie van de zes groepen kan met behulp van de formule $B = -\log(y_{gem}/x_{totgem})/\log(x_{totgem}$ een schatting van de leerwinst worden gemaakt. Hierbij is y_{gem} de gemiddelde posttestwaarde van de desbetreffende groep 4 en x_{totgem} is het gemiddelde over de pretestwaarden van de groepen 2, 3, 5, en 6. Er kleven twee bezwaren aan deze benadering. Deze methode levert vaak te lage B-waarden (Bos et al., 2007b). Een tweede probleem is het ontbreken van gegevens over de standaardfout in de B-waarde.

Als extra service voor de lezers die gewend zijn aan klassieke effectmetingen is ook de effectgrootte volgens Cohen (1988) bepaald. Effectgroottes van meer dan drie standaarddeviaties berekend volgens de Cohen-methode, mogen als zeer groot beschouwd worden.

Bewerking en analyse van de testen

Op de meerkeuzevragen van de pretesten werd $-1/(k-1)$ als correctie voor gokken toegepast, waarbij $k = 4$ voor vierkeuzevragen en $k = 2$ voor tweekeuzevragen. De gemiddelden en standaarddeviaties na correctie voor gokken staan in tabel 2.

Tabel 2. Resultaten pretesten A en B op een schaal van 0-100.

		na correctie voor gokken		leerlingen
groep	pretest	gemiddelde	std.dev.	N
2	A	16.9	12.1	8
3	B	15.8	10.0	8
5	A	21.8	7.31	13
6	B	16.7	10.7	14
alle		18.1	9.9	43

Variantieanalyse van pretestresultaten gaf geen significant verschil tussen de vier groepen voor correctie voor gokken $F(3, 39) = 0.580$ $(p > 0.050)$ en na correctie voor gokken $F(3, 39) = 0.883$ $(p > 0.05)$. De groepen kunnen dus vóór de interventie als equivalent worden beschouwd.

Voor het berekenen van de leerwinst werden gecorrigeerde waarden gebruikt.

Voor de posttest was Cronbachs alfa 0.92. Zeer lage waarden van alfa werden gevonden voor de pretesten tijdens het experiment, vandaar dat de pretesten en de posttest in aparte experimenten werden gecalibreerd. Toetsen over dezelfde leerstof zijn equivalent als (Drenth & Sijtsma, 2006):

a. er geen verschil tussen gemiddelde scores is,
b. er een hoge lineaire correlatie tussen de uitkomsten is, en
c. de standaarddeviaties gelijk zijn.

In de calibratie-experimenten werden (1) géén significante verschillen tussen scores uit de verschillende testen gevonden, (2) een hoge correlatiecoëfficiënt tussen de drie testen gemeten, en (3) waren de standaarddeviaties vrijwel gelijk. Hieruit kon worden geconcludeerd, dat de pretesten A en B en de posttest equivalent waren.

Bij deze experimenten was Cronbachs alfa voor de pretesten 0.904.

De begripsvaliditeit van de instrumenten

Zijn de gebruikte pre- en posttest een instrumentele realisering van het begrip dat wij willen meten, met andere woorden is er sprake van begripsvaliditeit (ook wel aangeduid met 'construct validity': De Groot, 1971; Swanborn, 1987; Den Hertog & Van Sluijs, 2000)? De instrumenten beogen de meting van kennis en begrip van natuurwetenschappelijke basisbegrippen op beschrijvend niveau op het grensvlak van de vakken scheikunde, natuurkunde, biologie, en informatica als een eerste oriëntatie op activiteiten in deze vakken in 4 vwo. Dit betekent vooral het verwerven van de 'natuurwetenschappelijke taal'. Het gaat hier om kennis en begrip in termen van het cognitieve domein van de leerdoelentaxonomie van Bloom, respectievelijk aangeduid met 'knowledge' en 'comprehension' (Bloom, Hastings & Madaus, 1971). In de herziene taxonomie van Bloom door Anderson en Krathwohl (2001) wordt *knowledge* vervangen door *remembering* en *comprehension* door *understanding* met de volgende definities:

Remembering: Retrieving, recognizing, and recalling relevant knowledge from long-term memory.
Understanding: Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining. (Anderson & Krathwohl, 2001).

Er zij opgemerkt dat beide taxonomieën een cumulatieve hiërarchische opbouw hebben: leerlingen kunnen pas verder met elementen van *understanding*, wanneer de relevante elementen uit het voorgaande niveau voldoende beheerst worden. De hogere niveaus uit de taxonomie komen later, gaandeweg in de loop van het vierde en vijfde leerjaar meer en meer aan de orde.

In de pretest wordt 'remembering' gemeten met twee- en vierkeuzevragen, en 'understanding' met kortantwoordvragen. In de posttest wordt vooral 'understanding' gemeten aan de hand van kortantwoordvragen waarin één of meer woorden, een formule, of een getal moesten worden ingevuld.

De testen waren ontworpen in het kader van onderwijs, en de daarbij horende leerdoelen, dat in de weken na het experiment werd aangeboden. Er werd zorg gedragen dat alle hiervoor genoemde relevante vakken waren vertegenwoordigd in de testvragen.

Tabel 3. Correlatiecoëfficiënt voor het lineair verband tussen posttestscore en officiële schoolexamencijfers voor de groepen 5 en 6 (n=27).

vak	R	*significantie*
ANW	0.441	*
NA	0.427	*
SK	0.441	*
WB	0.413	*

* = significant op 0.05 niveau

Met deze aanpak aan de hand van Bloom et al. (1971) c.q. Anderson en Krathwohl (2001) wordt gewaarborgd dat er sprake is van inhoudsvaliditeit, omdat het instrument qua inhoud en qua dekking van het bedoelde, overeenstemt met het te meten begrip. Daarmee wordt een bijdrage gegeven aan de begripsvaliditeit van de instrumenten (De Groot, 1971). Naast deze bijdrage aan de begripsvaliditeit is er ook een empirische bijdrage in de vorm van 'congruent validity'. De Groot (1971) vertaalt dit met 'soortgenoot validiteit'. Dit is een speciaal soort predictieve validiteit die voornamelijk betekenis heeft als bijdrage tot de begripsvaliditeit. Hierboven gaven wij aan dat de instrumenten waren ontworpen in het kader van leerdoelen van het onderwijs. Het is te verwachten dat de posttestscores significant correleren met het eerstvolgende, op dezelfde leerdoelen gebaseerde schoolexamen, de 'soortgenoot' van de posttest. Om de 'soortgenoot validiteit' van de posttest

met empirische informatie te onderbouwen is voor de groepen 5 en 6 de correlatie-coëfficiënt tussen de posttestscore en de eerstvolgende officiële schoolexamens voor een aantal betrokken vakken na zes weken berekend. In alle gevallen is er sprake van een behoorlijke correlatie die tevens in alle gevallen significant is. Zie tabel 3.

Onze conclusie is dat met de aangegeven inhoudsvaliditeit en soortgenoot validiteit een bevredigende inhoudelijke en empirische bijdrage wordt gegeven aan de begripsvaliditeit van de instrumenten.

5. Resultaten

Resultaten onderzoeksvraag 1: Is het mogelijk om met een pretest tot betere leerresultaten te komen met een interactief digitaal onderwijssysteem, waarin de leerlingen opdrachten uitvoeren om nieuwe natuurwetenschappelijke begrippen operationeel te maken?
De gemiddelde scores op de posttest voor de leerlingen in de zes groepen staan in tabel 4.

Tabel 4. Resultaten posttest (maximum = 100)

groep	pretest/treatment	gemiddelde	std. dev.	N
1	geen pretest geen treatment	20.5	8.4	27
2	pretest A geen treatment	20.8	5.8	8
3	pretest B geen treatment	22.2	9.9	8
4	geen pretest treatment	51.8	15.8	15
5	pretest A treatment	66.8	14.1	12
6	pretest B treatment	67.6	9.2	14
totaal		40.7	23.7	84

ANOVA van de resultaten van posttest gaf een significant verschil tussen de zes groepen $F\,(5\,,\,78) = 60.64\ (p < 0.05)$.

Meervoudige vergelijking volgens Bonferroni (zie tabel 5) geeft de volgende resultaten:

– De groepen 2 (alléén pretest A) en 3 (alléén pretest B) scoren hoger dan groep 1 (geen pretest, geen treatment), maar dit verschil is niet significant.
– De groepen 5 (pretest A + treatment) en 6 (pretest B + treatment) scoren significant hoger dan groepen 2 en 3 (beide alléén respectievelijk pretest A of B). Er is een fors effect van de interventie ‘operationaliseren met feedback’ (= ‘X’) gecombineerd met de pretest.

– De groepen 5 (pretest A + treatment) en 6 (pretest B + treatment) scoren significant hoger dan groep 4 (geen pretest + wèl treatment). Er is een sterk effect, als er een pretest in combinatie met oefening met feedback heeft plaatsgevonden.
– Groep 2 (alléén pretest A) scoort niet significant verschillend van groep 3 (alléén pretest B). Groep 5 (pretest A + treatment) verschilt niet significant van groep 6 (pretest B + treatment). De invloed van pretest A is gelijk aan die van B in deze twee gevallen.
– Groep 4 (geen pretest + wèl treatment) scoort significant hoger dan groep 1 (geen pretest, geen treatment). Er is een treatmenteffect van operationaliseren met feedback.

Tabel 5. Significantie van verschillen in posttest scores

groep	pretest/treatment	1	2	3	4	5
1	geen pretest geen treatment					
2	pretest A geen treatment	ns				
3	pretest B geen treatment	ns	ns			
4	geen pretest treatment	***	***	***		
5	pretest A treatment	***	***	***	*	
6	pretest B treatment	***	***	***	**	ns

* = significant op 0.05 niveau
** = significant op 0.01 niveau
*** = significant op 0.001 niveau

Uit de posttestwaarden werd de onderwijswinst berekend. Met behulp van pre- en posttestwaarden kon de leerwinstexponent *B* worden bepaald. Tussen de groepen 2 en 3 bleek geen significant verschil in leerwinst te bestaan $(p > 0.05)$. Voor de gezamenlijke groepen 2 en 3 was *B* gelijk aan 0.10 ± 0.074. Deze waarde verschilt niet significant van 0.

Via berekening van de leerwinstexponent *B* bleek tussen de groepen 5 en 6 geen significant verschil in leerwinst te bestaan $(p > 0.05)$. Voor de gezamenlijke groepen 5 en 6 was *B* gelijk aan $B = 0.79 \pm 0.021$.

Via het gemiddelde van de pretest van de groepen 2, 3, 5, en 6 werd een geschatte waarde van de leerwinst voor groep 4 van $B = 0.62$ berekend.

De effectgrootte ‘d’ volgens Cohen (1988) van de gecombineerde groepen 5 en 6 ten opzichte van de gecombineerde groepen 2 en 3 was 3.5.

Om na te gaan of er een significante statistische interactie tussen de pretest en treatment was, werd een univariate lineaire regressie uitgevoerd van de afhankelijke variabele *posttestscore* met *'wel of geen pretest'* en *'wel of geen treatment'* als factoren. De belangrijkste gegevens staan in tabel 6. Uit deze analyse volgt een significant effect van het afnemen van een pretest, $(p < 0.05)$ alsook voor de treatment $(p < 0.05)$. De interactie tussen pretest en treatment is eveneens significant $(p < 0.05)$. De waargenomen statistische power, berekend met $\alpha < 0.05$, is voor alle drie de bronnen groter dan 0.80. Zie ook tabel 6.

Hoewel de statistische power voldoende is, zij toch opgemerkt dat kleine groepsaantallen uiteraard tot enige voorzichtigheid nopen voor de conclusies en de punten van discussie.

Tabel 6. Enige gegevens uit de univariate regressieanalyse van de afhankelijke variabele *posttestscore* met *'wel of geen pretest'* en *'wel of geen treatment'* als factoren.

Bron	$F\ (1\ ,\ 83)$	significantie	power
wel of geen pretest	11.03	***	0.91
wel of geen treatment	242.8	***	1.00
interactie pretest * treatment	8.638	***	0.83

*** = significant op 0.001 niveau

Resultaten onderzoeksvraag 2: Is er een verschil in leerresultaten tussen een pretest met kortantwoordvragen (KAV) of meerkeuzevragen (MKV)?

Door verschillende typen vragen (KAV of MKV) in pretest A of B was het mogelijk na te gaan of er verschil in pretesteffect was voor de verschillende vraagtypen. In de analyse werd in eerste instantie alleen gekeken naar KAV versus 4-KV van de groepen 5 en 6. Met een ANOVA werd geen significant verschil geconstateerd tussen de posttestscores $(p > 0.05)$. Ook met een two-way ANOVA pretest A of B versus pretestvraagtype werd geen verschil geconstateerd met betrekking tot de variabele *pretestvraagtype* $(p > 0.05)$ of de variabele *test A of B* $(p > 0.05)$.

Tabel 7. Score op posttest na sensitivering in pretest

type vraag in pretest	gemiddelde	std. dev.	N
kortantwoordvraag	69.6	45.5	340
vierkeuzevraag	71.7	43.8	338
totaal	70.6	44.7	678

De resultaten van de posttestresultaten gegroepeerd naar pretestvraagtype staan in tabel 7.

Nadere analyse van de leertaaktijd
Teneinde aandacht te geven aan een alternatieve hypothese waarin taaktijd een rol speelt, werden ook nog diverse relevante tijdanalyses uitgevoerd.

De tijd die leerlingen uit de groepen 5 en 6 nodig hadden voor pretest, treatment en posttest werden nader onderzocht. Er was een enorme variatie onder de deelnemers: de snelste leerlingen waren drie tot vijf keer sneller dan de langzaamste.

Er werd geen correlatie gevonden tussen de *posttestscore* en de tijd die aan de leertaken werd besteed (*time-on-task*). De covariaat tijd was ook niet significant $(p > 0.05)$ in een univariate variantie-analyse. Hierin was de posttestscore de afhankelijke variabele. De vaste factor bij deze analyse was *pretest ja/nee*.

6. Conclusies

We zullen per onderzoeksvraag de conclusies weergeven en de taaktijdanalyses samenvatten.

Onderzoeksvraag 1: Is het mogelijk om met een pretest tot betere leerresultaten te komen met een interactief digitaal onderwijssysteem, waarin de leerlingen opdrachten uitvoeren om nieuwe natuurwetenschappelijke begrippen operationeel te maken?
Zoals aangegeven bestond het interactief digitaal systeem uit twee componenten, (a) een component ‘gevoelig maken’ aan de hand van een pretest (pretestsensitiveren), en (b) een component 'operationeel maken' van de te leren natuurwetenschappelijke begrippen aan de hand van opdrachten en daarbij behorende terugkoppeling. Het interactieve systeem bestaande uit de twee componenten, leidde tot significant hogere scores. De berekende onderwijswinst aan de hand van de waarde $(B = 0.79)$ kan zéér hoog worden genoemd. Ook zonder de pretest bleken significant hogere scores en een hoge leerwinst $(B = 0.62)$ te worden bereikt, maar pretest plus operationeel maken leverde significant hogere scores dan operationeel maken alleen.

Alléén pretesten, dat wil zeggen zonder enig vervolg in de vorm van opdrachten en terugkoppeling met het oog op het operationeel maken van de begrippen, heeft geen effect op de leerresultaten.

Onderzoeksvraag 2: Is er een verschil in leerresultaten tussen een pretest met kortantwoordvragen (KAV) of meerkeuzevragen (MKV) ?
Er is geen verschil in leerresultaten tussen een pretest met kort antwoordvragen of met meerkeuzevragen.

Nadere analyse van de leertaaktijd
Een alternatieve hypothese, dat de *leertaaktijd* een significante variabele is, wordt in dit onderzoek op een onmiddellijke tijdschaal (in de orde van minuten vóór de treatment) niet ondersteund.

7. Discussie

Uit het experiment blijkt de hoogste leerwinst ($B = 0.79$) geboekt te worden, als er eerst een pretest wordt afgenomen en vervolgens een toegesneden multimediaal interactief systeem wordt ingezet, maar ook zonder pretesten is er een hoge leerwinst ($B = 0.62$). Het heeft dus zin om pretesten te koppelen aan een interventie van goed doordachte uitleg, gekoppeld met vragen en onmiddellijke feedback. Bij het pretesten treedt er een vorm van gevoelig maken op. Mogelijk is hier sprake van mentale kiemvorming (Vos, 1990), maar het inzetten van een pretest sec heeft geen of wellicht weinig effect. Om *leren* te bewerkstelligen moet er na een pretest een onmiddellijk hierop aansluitend leerproces volgen. Gelet op de review van Strangman et al. (2004; zie het theoretisch kader) zijn de gevonden goede resultaten te verklaren uit een combinatie van door hen als effectief geziene onderwijsstrategieën: activeren van voorkennis door het vooraf stellen van vragen (pretest), het stellen van vragen (interventie), en het geven van feedback (interventie). Volgens dit onderzoek maakt het niet uit of in de pretest meerkeuzevragen of kort-antwoordvragen worden gebruikt. Overigens sluit dat niet uit dat, als er met echte 'open vragen' was gewerkt (kort-antwoordvragen staan tenslotte nog vrij dicht bij gesloten vragen), er dan wellicht wel verschillen waren gevonden. Vanuit het oogpunt van onderwijsefficiëntie is het gebruik van open vragen echter een probleem. Bedrijfsmatig gezien is de geautomatiseerde correctie van toetsen met vragen, waarin de leerling zelf een (open) antwoord formuleert, met de huidige 'off-the-shelf-software' niet rendabel, omdat de docent nog een nacorrectie moet uitvoeren.

Ten slotte is in een nevenexperiment ook het pretesteffect van tweekeuzevragen nagegaan. Er is geen significant pretesteffect van tweekeuzevragen geconstateerd, en de kans lijkt groot, dat dit bij een omvangrijk experiment ook niet wordt gevonden.

Vanuit de noodzaak van onderwijsefficiëntie lijkt het voor de hand te liggen om in een digitale omgeving meerkeuzevragen als pretestmateriaal te gebruiken, want dit geeft mogelijkheden voor onmiddellijke feedback. Deze mogelijkheid werd in dit experiment niet gebruikt in verband met de onderzoeksopzet. De combinatie van een test en feedback is een invloedrijke interventie in onderwijskundig en methodologisch opzicht, en is vergelijkbaar met groep 4 in onze opzet (geen pretest, wèl treatment). Het is te verwachten dat bij het gebruik van een toets met feedback ook significant hogere leerwinsten verkregen worden, in vergelijking met het afnemen van een toets zonder feedback. Met meerkeuzevragen is het technisch goed uitvoerbaar onmiddellijk adequate feedback te geven.

De conclusie dat de tijd die aan de tests en de treatment wordt besteed als covariaat niet significant is, schijnt in tegenspraak met onderzoek waaruit blijkt dat de tijd die besteed wordt aan een taak, een goede voorspeller is van uiteindelijk gemeten prestaties (Admiraal, Wubbels & Pilot, 1999; Cotton, 2001). Het meeste onderzoek betreft echter een distale tijdschaal (enige maanden). Bovendien gaat het om de *relevante* taaktijd, die van belang is (Wellman & Marcinkiewicz, 2004). De analyses suggereren, dat een marginale, extra tien minuten op een *onmiddellijke* tijdschaal (in de orde van minuten vóór de

hoofdinterventie) die besteed worden aan een pretest véél effectiever is, dan een extra tien minuten die aan de rest van de interventie wordt besteed. Pretesttijd is van een hogere *kwaliteit*, omdat het maken van een pretest de treatment op een specifieke manier beïnvloedt.

De gevonden resultaten hebben op een beschrijvend (micro-)niveau betrekking op sciencebasisbegrippen waarbij een directe relatie bestaat met het macro-niveau, de wereld van de waarneembare verschijnselen. Aanvullend onderzoek is nodig naar de verwerving van begrippen op theoretisch (micro-)niveau in een uiteindelijk symbolische representatie: Omdat in natuurwetenschappelijke settings diepe verwerking en transfer van groot belang zijn, dient de posttest anders te zijn dan de pretest, omdat het louter reproduceren van declaratieve kennis meestal geen hoofddoelstelling is. Bovendien moet ook om methodologische redenen de posttest anders dan maar wel equivalent, zijn aan de pretest. Immers, het ging in dit onderzoek om het gebruikmaken van de pretest-treatmentinteractie, die statistisch significant in de resultaten is aangetoond. Als de pretest geheel gelijk is aan de posttest bevordert men juist het ongewenste 'testingeffect', een interactie van de pretest met de posttest, en niet de gewenste pretest-treatmentinteractie (Campbell & Stanley, 1963). Nu is desalniettemin een testingeffect niet geheel uit te sluiten, maar dat kon in dit onderzoek niet worden gecontroleerd, omdat voor een andere posttest is gekozen. In een verder uitgebreide variant van de Solomon Four-group design zou wel apart het (methodologische en onderwijskundig ongewenste) 'testing effect' en het (onderwijskundig gewenste) 'pretest-treatment' interactie-effect kunnen worden gemeten. Scharfenberg et al. (2006) zien een Solomon Four-group design als een krachtige onderzoeksopzet met vele voordelen aangezien zowel de interne als externe validiteit wordt versterkt, omdat een potentieel 'confounding effect' van de pretest onder controle wordt gehouden. Dit is met name van belang voor 'science education research' aangezien de voorkennis veelal wordt getoetst, zeker in de veel voorkomende 'conceptual change' benadering (Anderson, 2007; Scott et al., 2007). Scharfenberg et al. (2006) constateren eveneens dat om een aantal redenen dit design toch niet zoveel wordt gebruikt: er zijn veel meer proefpersonen nodig; de onderzoeker gelooft dat het pretesteffect in zijn onderzoek niet aan de orde is; het is moeilijker om conclusies te trekken omdat het design complexer is; en de statistische verwerking is eveneens complexer. De slotconclusie van Scharfenberg et al. (2006) is toch dat het aanbeveling verdient een Solomon Four-group design, of varianten daarvan, in science education research te gebruiken; hun eigen onderzoek geeft daarvan een mooi voorbeeld (Scharfenberg et al., 2006).

Als eerste handreiking voor de onderwijspraktijk zou de opzet van het experiment als ontwerpstructuur kunnen functioneren voor een introductiemodule. Het is technisch haalbaar om een meerkeuzetest in een elektronische leeromgeving te zetten. Het is dan aan te bevelen om onmiddellijk feedback op de vragen te geven. De test zou gevolgd kunnen worden door een aantal informatieschermen gekoppeld aan digitaal te controleren opdrachten met onmiddellijke feedback. De leerlingen zouden een dergelijke module

asynchroon voor de start van een nieuwe onderwijsactiviteit door kunnen werken. Procesgegevens zouden voor de leraar of coach bij de start van de onderwijsactiviteit beschikbaar moeten zijn. Een dergelijke opzet is overigens niet geheel nieuw: een vergelijkbare opzet is aanwezig in het COO-pakket SCOOR (Paulides & Pilot, 1996) dat is bedoeld om deficiënties van instromende hbo-studenten op te sporen en weg te werken. De pretest in deze software heeft daarin wel primair een *allocatieve functie* – afhankelijk van de pretest wordt doorverwezen naar één of meer specifieke modules – en geen sensitiverende functie, maar deze pretest zou wel eens zo kunnen werken. De combinatie van pretestsensitivering, gevolgd door computerondersteund onderwijs (COO) bestaande uit een gerichte instructie, oefening en toegesneden feedback is mogelijk een belangrijke verklaring voor het grote leereffect van het pakket dat wij hebben berekend. Uit historische gemiddelde pre/posttestgegevens kon voor de groep studenten die de scheikundemodule volgden een leerwinst van $B = 0.73$ worden berekend. Voor de niet-COO-groep is B gelijk aan 0.12 (SCOOR, 1986). Inmiddels zijn echter zowel de multimediamogelijkheden als asynchrone toegankelijkheid van het SCOOR-pakket sterk vergroot waarmee wellicht de leerwinst nog verder vergroot kan worden.

Als tweede handreiking voor de onderwijspraktijk zou pretestsensitivering – wellicht in combinatie met andere vormen van voorkennisactivering – relevant kunnen zijn voor begripsvorming bij de context-conceptbenadering (Bulte et al., 2005). Daarin is het sensitiveren van het begrippennetwerk dat relevant is voor het uitvoeren van taken binnen de gekozen context en voor het faciliteren van transfer een lastig probleem (Pilot & Bulte, 2006). Ook uit de review van Strangman et al. (2004) bleek dat het gebruik van authentieke situaties niet zonder meer tot activering van voorkennis leidde. Wellicht moet een strategie van activering en ontwikkeling worden gevolgd, die geïnspireerd wordt door de in dit experiment gevonden resultaten.

De vraag rijst, of dit experiment specifiek is voor het bètadomein. In vrijwel alle disciplines is pretestsensitivering vastgesteld en het gericht gebruiken van dit effect om onderwijs effectiever te maken is geen exclusieve zaak voor de bètawetenschappen. De bètainhoud leent zich echter uitermate goed voor dit type aanpak, zeker waar het de eerste twee categorieën in de taxonomie van Bloom (1971) c.q. Anderson & Krathwohl (2001) betreft.

Correspondentie over dit artikel aan A.B.H.Bos, abh.bos@home.nl

English summary

The design of effective instruction for concept learning in science subjects is necessary in Dutch secondary education, because of the decrease in the amount of face-to-face instruction and the relatively high student/teacher ratio. Schema theory and research about the assessment of prerequisite knowledge show that pre-test sensitising – deliber-

ately activating relevant conceptual schemata by a pre-test – appears to be an important learning variable. Moreover, the pre-test was embedded in an instructional design in an interactive digital system through which all relevant instructional functions could be realised. The instructional content concerned an orientation on complex contexts in the joint area of physics, chemistry, biology, applied mathematics, and computer sciences. An extended Solomon Four-Group design was applied with experimental and control groups in which the pre-test and assignments with science concepts were present or lacking. The pre-test consisted of both short-answer questions and multiple-choice questions. In separate experiments the reliability and equivalence of the different tests were confirmed. The results show a strong educational gain, especially in the combination of pre-test and assignments. A significant interaction between pre-test and treatment is found. No significant differences were found in the use of two different question types. Only applying the pre-test also did not result in a significant learning effect.

Literatuur

Admiraal, W., Wubbels, T. & Pilot, A. (1999). College teaching in legal education: Teaching Method, Students' Time-on-Task, and Achievement. *Research in Higher Education*, *40*(6), 687-704.

Ali, K. (1990). *Instructiestrategieën voor het activeren van preconcepties*. Tilburg: Universiteit van Tilburg.

Anderson, C. W. (2007). Perspectives on science learning. In S.K. Abell & N.G. Lederman (ed.), *Handbook of research on science education* (pp. 3-31). Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J.R. & Schunn, C.D. (2000). Implications of the ACT-R Learning Theory: No Magic Bullets. In R. Glaser (ed.), *Advances in instructional psychology* (Vol. 5). Mahwah, NJ: Erlbaum.

Anderson, L.W. & Krathwohl, D.R. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives (pp. 67-68). *New York*: Longman.

Ausubel, D.P. (1968). *Educational psychology. A cognitive view*. New York: Holt, Rinehart & Winston, Inc.

Bais, S. (2004). *Nooit meer rechtdoor. Over het wankele evenwicht tussen waarneming en verbeelding*. Deventer: Thieme.

Biemans, H.J.A. (1997). *Fostering activation of prior knowledge and conceptual change*. Radbout Universiteit, Nijmegen.

BINAS (2004). *Informatieboek havo/vwo voor het onderwijs in de natuurwetenschappen*. Groningen, The Netherlands: Wolters-Noordhoff.

Bloom, B.S., Hastings, J.T. & Madaus, G.F. (1971). Handbook on formative and summative evaluation of student learning (pp. 271-272). New York: McGraw-Hill.

Bos, A.B.H., Terlouw, C. & Pilot, A. (2007a). *Het effect van pretest-sensitisatie bij ontdekkend leren met behulp van een simulatie in het voortgezet onderwijs*. Paper presented at the ORD 2007, Zorgvuldig en Veelbelovend Onderwijs, Proceedings van de 34e Onderwijs Research Dagen, Groningen.

Bos, A.B.H., Terlouw, C. & Pilot, A. (2007b). *A Pre-test-Corrected Learning Gain. F*rom http://www.utwente.nl/elan/onderzoek/publicaties/elandoc/2007/2007-004.pdf

Bulte, A., Klaassen, K., Westbroek, H., Stolk, M., Prins, G., Genseberger, R. et al. (2005). Modules for a new chemistry curriculum. Research on the meaningful relation between contexts and concepts. In P. Nentwig & D. Waddington (ed.), *Making it relevant. Context based learning of science* (pp. 273-299). Munster: Waxmann.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

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

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation. Design & Analysis Issues Designs for Field Settings*. Chicago: Rand McNally College Publishing Company.

Cotton, K. (2001). *Educational Time Factors*. Retrieved Nov. 17, 2007. From http://www.nwrel.org/scpd/sirs/4/cu8.html

Dochy, F. & Alexander, P.A. (1995). Mapping prior knowledge: a framework for discussion among researchers. *European Journal of Psychology of Education*, *10*(3), 225-242.

Dochy, F., Segers, M. & Buehl, M.M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, *69*(2), 145-186.

Drenth, P.J.D. & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4 ed.). Houten: Bohn Stafleu van Loghum.

De Groot, A.D. (1971). *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen*. Den Haag: Mouton.

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.

Hertog, F. den & Sluijs, E. van (2000). *Onderzoek in organisaties: een methodologische reisgids*. Assen: Van Gorcum.

Ilic, B. & Craighead, H.G. (2004). Attogram detection using nanoelectromechanical oscillators. *J. Applied Physics*, *95*(7), 3694-3703.

Johnstone, A.H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of computer assisted learning*, *7*, 75-83.

Kozma, R.B. (2000). The use of multiple representations and the social construction of understanding in chemistry. In M.J. Jacobson & R.B. Kozma (ed.), *Adcanced design for technologies of learning* (pp. 11-46). Mahwah, NJ: Lawrence Erlbaum Associates.

Kuhlemeier, H., Steentjes, M. & Kleintjes, F. (2003). *De gelijkwaardigheid van open en meerkeuzevragen bij wiskunde. Effecten van vraagtype en scoringswijze op gemeten vaardigheden, betrouwbaarheid, moeilijkheid en afnametijd*. Arnhem.

Lana, R.E. (1959). Pretest-treatment interaction effects in attitudinal studies. *Psychological Bulletin*, *56*, 293-300.

Lana, R.E. & King, D.J. Learning factors as determiners of pretest sensitization. *Journal of Applied Psychology*, *44*(3), 189-191.

Lana, R.E. (1969). Pretest Sensitization. In Rosenthal & Rosnow (ed.), *Artifact in Behavioral Research*. New York: Academic Press.

Lawson, M.J. & Chinnappan, M. (2000). Knowledge Connectedness in Geometry Problem Solving. *Journal for Research in Mathematics Education*, *31*(1), 26-43.

Lemke, J.L. (1990). *Talking science: language, learning and values*. Norwood, NJ: Ablex.

Linn, M.C. (2003). Technology and science education: starting points, research programs and trends. *International journal of science education*, *25*, 727-758.

Ogborn, J., Kress, G., Martin, I. & McGillycuddy, K. (1996). *Explaining science in the classroom*. Buckingham, UK: Open University Press.

Osborne, J. & Hennessy, S. (2003). *Literature review in Science Education and the role of ICT: promise, problems and future directions* (No. 6). Bristol: NESTA Future Lab.

Paulides, J.P. & Pilot, A. (1996). SCOOR for Windows. In M. van Geloven & A. Pilot (ed.), *Multimedia in het hoger onderwijs*. Groningen: Wolters-Noordhoff.

Pilot, A. & Bulte, A.M.W. (2006). The use of 'contexts' as a challenge for the chemistry curriculum: its successes & the need for further development and understanding. *International Journal of Science Education*, *28*(9), 1087-1112.

Ritzen, J. (2006). Hoger onderwijs tussen kennis en koopje. *TH&MA*(1).

Roes, T. (2001). *De sociale staat van Nederland*. (Vol. 2001-14). Den Haag: SCP.

Scharfenberg, F.J., Bogner, F.X. & Klautke, S. (2006). The suitability of external control-groups for empirical control purposes: a cautionary story in science education research. *Electronic Journal of Science Education*, *11*(1).

Schroeder, B. (2004). *Nukleare Mikrobatterien*. From http://www.heise.de/bin/tp/issue/r4/dl-artikel2.cgi?artikelnr=18351&zeilenlaenge=72&mode=html

SCOOR. (1986). *Verslag Symposium 'COO in het HBO'*. Enschede: Werkgroep COO-HBO en TH Twente, CDO.

Scott, P., Asoko, H. & Leach, J. (2007). Student Conceptions and Conceptual Learning in Science. In S.K. Abell & N.G. Lederman (ed.), *Handbook of research on science education* (pp. 31-56). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Shadish, W.R., Cook, T.D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Solomon, R.L. (1949). An extension of control group design. *Psychological Bulletin*, *46*, 137-150.

Strangman, N., Hall, T. & Meyer, A. (2004). *Background knowledge instruction and the implications for UDL implementation*. From http://www.cast.org/publications/ncac/ncac_backknowledgeudl.html

Swanborn, P.G. (1987). *Methoden van sociaal-wetenschappelijk onderzoek.* Amsterdam, Meppel: Boom.

Treagust, D.F. (2007). General instructional methods and strategies. In S.K. Abell & N.G. Lederman (ed.), *Handbook of research on science education* (pp. 373-393). Mahwah, NJ: Lawrence Erlbaum Associates.

Tweede Fase Adviespunt. (2005). *Zeven jaar Tweede Fase, een balans*. From www.tweedefase-loket.nl

Vos, H. (1990). *Transfer of learning: het gebruik van een leerkiem. Intern rapport OC-doc 90-43*. Enschede: Universiteit Twente, OC / IBTE.

Vygotsky, L.S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wellman, G.S. & Marcinkiewicz, H. (2004). Online learning and time-on-task: impact of proctored vs. un-proctored testing. *Journal of Asynchronous Learning Networks*, *8*(4), 93-104.

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

Zimmerman, D.W. & Williams, R.H. (2003). A New Look at the Infuence of Guessing on the Reliability of Multiple-Choice Tests. *Applied Psychological Measurement*, *27*(5), 357-371.

Bijlage 1. Begrippennetwerk