Proefschrift Sylvia P. van Borkulo

The assessment of learning outcomes of computer modeling in secondary science education

Bespreking door:
Wolter Kaper
Universiteit van Amsterdam

Wanneer leerlingen actief met lesmateriaal bezig zijn, steken zij er meer van op. Onder dat brede motto zijn veel onderwijsvernieuwingen te vatten, onder andere ook het onderzoekend leren, het leren natuurwetenschappelijk redeneren en het leren werken met modellen, bijvoorbeeld computermodellen. Vaak wordt geclaimd dat deze onderwijsvernieuwingen ons in staat stellen 'hogere orde' leerdoelen te bereiken. Maar is dat ook zo? Sylvia van Borkulo stelt zich ten doel de vraag te beantwoorden welke leerdoelen er kunnen worden bereikt met het onderwijs in computer modelleren, en ook, hoe deze leeruitkomsten kunnen worden gemeten. Hiervoor vergelijkt zij drie leeractiviteiten: *modelleren*, waar leerlingen een door de leraar gegeven model kunnen inzien, wijzigen en uitbreiden, *simuleren*, waar het door de leraar geprogrammeerde model onzichtbaar blijft, en *traditioneel onderwijs*, bestaande uit een tekst gevolgd door oefening (het schrijven van een werkstuk). De enigszins verrassende uitkomst is, dat leerlingen door het modelleren niet alleen beter leren modelleren, maar dat ook hun antwoorden op reproductievragen verbeteren, mits die reproductievragen niet al te eenvoudig zijn. Omgekeerd geldt voor de eenvoudige reproductievragen (niet onverwacht) dat hier de traditionele aanpak tot betere resultaten leidt. Over 'hogere orde'-redeneervaardigheden wordt het volgende geconcludeerd:

– *toepassen* van een model – hieronder verstaat Van Borkulo het doen van een voorspelling of het verklaren van een gegeven door te redeneren volgens een als juist aangenomen model: hier maakt de onderwijsvorm niet uit als de taak maar eenvoudig genoeg is (het juiste antwoord niet meer dan één redeneerstap omvat). Is de taak ingewikkelder, dan doen leerlingen die getraind zijn in computermodelleren het beter.
– *evalueren* in de context van een model – dat is het beoordelen van het model in het licht van data, of andersom: hier bleek de 'modelleer'-groep licht in het voordeel boven de 'traditioneel onderwezen' groep, weer mits de taak gecompliceerd genoeg is. Omdat echter de resultaten van beide vergeleken groepen laag waren, is de auteur toch niet tevreden hierover.
– *creëren* van een model: hier zijn de resultaten enigszins tegenstrijdig. Bij één van de twee onderwijsexperimenten kwam een verschil naar voren tussen leerlingen die leerden simuleren en zij die leerden modelleren. De laatsten leken in het voordeel bij het zelf creeren

van een simpel model. Maar als ook het traditionele onderwijs in de vergelijking werd betrokken dan bleef dit verschil niet overeind[1].

Bij de twee genoemde negatieve resultaten moet worden vermeld dat het experimentele onderwijs maar kort duurde: netto niet meer dan 5 uur (300 minuten) in twee sessies. Hierbij is nog afgezien van de tijd die nodig is voor het invullen van de tests vóór en na het onderwijs.

De voornaamste trend is dat het leren modelleren van een domein (dat was in alle gevallen 'de warmtebalans van de aarde') een goede invloed heeft op het vermogen van leerlingen om complexe vragen over dat domein te beantwoorden – zowel reproductievragen als model-toepassingsvragen. Als verklaring voor het onverwachte succes bij complexe reproductievragen geeft de auteur, dat complexe reproductievragen wellicht niet uit het geheugen beantwoord worden, maar dat men het antwoord vindt door middel van een redenering. De leerlingen die leerden modelleren hebben in zulke redeneringen meer oefening gehad dan hun collega's uit de 'traditioneel onderwijs'-groep.

Om tot deze resultaten te komen was het nodig eerst te beargumenteren welke verschillende leeruitkomsten er als gevolg van modelleren te verwachen zijn. Hiervoor werd een raamwerk opgezet waarin vier soorten *activiteiten* (reproductie, toepassen van een model, evaluatie en creatie) de belangrijkste dimensie vormden. Een tweede dimensie was de *complexiteit* van de opgaven: hier werd een tweedeling gehanteerd, simpel (niet meer dan één relatie nodig om aan het antwoord te komen) of gecompliceerd (meerdere relaties nodig). De derde dimensie maakte onderscheid tussen domein-specifiek redeneren en domein-generiek redeneren. 'Evenwicht' en 'feedback' zijn bijvoorbeeld concepten die bij modelleren een rol spelen in heel verschillende domeinen. Ongeacht het domein geldt dat bij negatieve feedback het systeem naar een evenwicht tendeert. De derde dimensie onderscheidt zulke domeinonafhankelijke redeneringen van de domeinafhankelijke. Maar deze derde dimensie heeft in dit onderzoek nog niet de rol gekregen die nodig werd gevonden, dus deze dimensie is bij de aanbevelingen voor verder onderzoek terecht gekomen.

Het raamwerk van drie dimensies levert in totaal 4 (activiteiten) x 2 (complexiteiten) x 2 (domeinafhankelijk of niet) = 16 cellen op, die elk werden gevuld met drie of vier testvragen. Een belangrijk deel van het onderzoekswerk bestond uit controles op betrouwbaarheid en validiteit van de test. De test bestaat uit open vragen en dus werd gemeten in hoeverre twee beoordelaars tot dezelfde interpretaties komen. Vervolgens werd gecontroleerd of de vier subschalen van de test een goede samenhang hebben, en ook, of de veronderstelling van vier verschillende redeneervaardigheden kon worden bevestigd. Tenslotte werd gekeken of universitaire natuurkundestudenten, die net een cursus Simuleren en Modelleren achter de rug hebben, beter scoren dan psychologiestudenten of leerlingen van een middelbare school met natuur en techniek als profiel. Het bleek dat de universitaire natuurkundestudenten beter scoorden op alle vier de subschalen (reproductie, toepassing, creëren en evalueren), hetgeen de validiteit ondersteunde.

Daarna volgen twee onderwijsexperimenten. Het verschil tussen de experimentele groep en de controlegroep is eerst maximaal groot (universitaire natuurkunde – middelbare school) en wordt dan in twee stappen kleiner gemaakt: eerst modelleeronderwijs vergeleken met traditioneel onderwijs in twee groepen die verder (op basis van een pretest) met een gelijke modelleervaardigheid startten. In een tweede experiment werd een nog kleiner verschil gekozen: onderwijs in modelleren wordt vergeleken met simuleren. Het feit dat er ook in het laatste geval nog duidelijke verschillen werden gevonden, toont aan dat de test voldoende 'gevoelig' is om ook de effecten van relatief kleine verschillen in onderwijsaanpak te laten zien.

Aan de vraag naar de validiteit van een test zitten veel kanten. In hoofdstuk 1 wordt dan ook gesteld dat validiteit niet bewezen kan worden, maar dat men er hoogstens verschillende aanwijzingen voor kan aandragen. Twee vormen van validiteit, namelijk 'validiteit op het eerste gezicht' en 'inhoudsvaliditeit' krijgen in dit onderzoek weinig aandacht. Wel worden bij wijze van voorbeeld enkele testvragen getoond en toegelicht. Juist deze voorbeeldvragen brengen mij aan het twijfelen of er gemeten wordt wat men zegt te meten. Ik geef mijn twijfels hieronder weer.

Bij een 'toepassings'-taak wordt geredeneerd om een voorspelling of verklaring te genereren. Bij een 'simpele toepassing'-taak, is het juiste antwoord gebaseerd op niet meer dan één relatie uit het gegeven model. Als voorbeeld wordt de volgende taak gegeven (pagina 27):

'Choose the correct statement.
A. The higher the albedo, the larger the inflow of energy to the earth.
B. The higher the albedo, the lower the inflow of energy to the earth.
C. The albedo does not influence the inflow of energy to the earth.
Explain your answer.'

Voorafgaand aan de test is het te gebruiken model aan leerlingen gegeven in de vorm van een schema met pijlen. Er is tijd aan besteed om zeker te stellen dat leerlingen deze pijlentaal kunnen lezen, en in de bespreking wordt nergens genoemd dat dit een probleem zou kunnen zijn. In het pijlendiagram zie ik een pijl van 'albedo' naar 'inflow of energy' (naar de aarde) met een minteken erbij, hetgeen betekent: een negatieve invloed.

Mijn conclusie: de taak is geen redeneertaak, maar een reproductietaak. Het antwoord staat immers letterlijk (in pijlen-taal weliswaar) in het aangeboden diagram. Deze taak test niet het modelleren van de toepassing maar uitsluitend het kunnen lezen van het diagram!

Dit roept een principiëlere vervolgvraag op: kan een één-staps-redenering worden getest? Is een dergelijke taak onderscheidbaar van een reproductietaak?

Een 'evaluatie'-taak is een taak waarbij model en data met elkaar worden geconfronteerd. Volgens de auteur moet hierbij ófwel het model, ófwel de data ter discussie worden

gesteld, niet beide (pagina 40). Als voorbeeld van een complexe evaluatietaak wordt gegeven (pagina 44):

> ‘Andre performed two experiments with the simulation to investigate the relation between inflowing radiation and temperature.’
> [Er volgen twee tabellen en twee grafieken met gegevens uit beide computerexperimenten. De grafieken tonen de temperatuur tegen de tijd: een stijgende curve bij experiment 1, dalend bij experiment 2. Zowel de ‘inflow of energy’ als de ‘albedo’ hebben verschillende waarden in de beide experimenten]
> ‘From these data, Andre concludes that, the higher the inflowing radiation, the higher the temperature on earth. Is it correct for Andre to draw this conclusion? Explain your answer.’

Uit de formulering blijkt dat het model hier als waar moet worden aangenomen, terwijl Andre's uitspraak (als ‘data’) ter discussie staat. Om de juistheid van Andre’s uitspraak te beoordelen zou ik (1) het model zelf toepassen (een meerstaps-toepassingstaak) en dan (2) mijn eigen resultaat met dat van Andre vergelijken. Bij een verschillende uitkomst concludeer ik dat Andre ongelijk heeft. In dit geval heeft hij ongelijk omdat bij hogere instraling de temperatuur best kan dalen. Dat gebeurt namelijk – volgens het gegeven model – als de albedo sterker toeneemt dan de instraling.

Is dit nu een ‘hoog niveau’-redeneervaardigheid zoals bedoeld door de auteur bij haar verwijzing naar Bloom’s herziene taxonomie (Anderson & Krathwohl, 2001, geciteerd op pagina 37)?

Mijn redenering hierboven is equivalent aan het beantwoorden van een toepassingstaak, gevolgd door een ja/nee beslissing. Wat maakt ‘evaluatie’ volgens Bloom tot een hoger niveau dan ‘toepassing’? Is dat die ja/nee beslissing, of moet het meer zijn?

Wanneer evaluatie optreedt in een authentieke onderzoekssituatie dan staat er nooit zwart op wit gegeven welke van de twee voor waar moet worden aangenomen: het model of de data. In plaats daarvan kent de onderzoeker de achtergrond van de data (hoe kwam Andre aan zijn uitspraak, welke aanwijzingen voerde hij aan) en de onderzoeker kent de status van het model: successen en twijfels. Als model en data conflicteren, is de hamvraag: aan welke van de twee moeten we nu de schuld geven, aan het model of aan de data?

Mijn conclusie luidt: als die keuze al door de maker van het test-item voor ons gemaakt is, dan wordt het hoge niveau van een evaluatietaak aanzienlijk verlaagd.

Zijn er testmethoden voor hogere orde leerdoelen? Voorstanders van competentiegestuurd leren argumenteren dat niet elk leerdoel kan worden getoetst in een examensituatie. Aan de universiteit is het eindwerkstuk een bachelor- of master-scriptie, geen tentamen. Dat is niet voor niets. De reden is dat niet elke authentieke taak (zoals onderzoek doen, of modelleren) kan worden opgesplitst in kleine stukjes, die dan elk apart in onaf-

hankelijk te beantwoorden opgaven[2] worden getoetst. Iets dergelijks zou ook kunnen gelden voor sommige van de leerdoelen die Van Borkulo heeft uitgekozen. Werkstukken en projecten zijn minder intersubjectief betrouwbaar te beoordelen dan tentamens, en dat is jammer. Maar toch is er een reden waarom ze onlangs in de mode zijn gekomen, en dat heeft veel met hogere orde vaardigheden te maken. Enige discussie daarover zou een academisch proefschrift niet misstaan in deze tijden van snel wisselende (nu moet het evidence-based, maar kort geleden moest het vooral authentiek) modegrillen.

Onze kennis over hoe de drie eerste Bloom-niveaus (reproductie, begrijpen, toepassen) kunnen worden bereikt en getoetst in onderwijs dat modelleren inzet als middel voor begripsvorming, is door dit proefschrift toegenomen. De conclusie dat modelleren vooral de vaardigheid beïnvloedt om complexere problemen te beantwoorden is grondig onderbouwd, belangrijk, en toepasbaar.

Noten

1. Het lijkt nu een vreemde anomalie, dat de ‘simulatie’-leerlingen uit het tweede experiment bij het creëren van een simpel model onder de maat presteerden (zie tabel 6-2, pagina 95)! Een verklaring voor dit merkwaardige verschijnsel wordt niet gegeven.
2. Meerdere onafhankelijk te beantwoorden opgaven zijn nodig om statistiek te kunnen bedrijven. Een toets met maar één opgave (extreem voorbeeld: schrijf een proefschrift) heeft geen Cronbach alfa.