Een instrument voor bovenbouw vwo-leerlingen om de kwaliteit van hun natuurwetenschappelijk onderzoek te evalueren

S. A. W. van der Jagt, L. van Rens, H. H. Schalk, A. Pilot en J. J. Beishuizen

Samenvatting

In deze studie is onderzocht in hoeverre een ontworpen zelfevaluatie-instrument bruikbaar is voor het evalueren van de nauwkeurigheid, betrouwbaarheid en validiteit in onderzoek in de bètavakken door leerlingen uit de bovenbouw van het vwo. Het instrument heeft vier ontwerpkarakteristieken, respectievelijk betreffende de inhoud (Concepts of Evidence-model), de mate van complexiteit (SOLO-taxonomie), gedetailleerdheid en algemene toepasbaarheid. In een ontwerpgericht onderzoek is de bruikbaarheid van het instrument voor het onderwijsproces onderzocht in een klas met 27 bovenbouw vwo-leerlingen in drie opeenvolgende onderzoeksmodules. Het door leerlingen en docent ingevulde instrument, lesobservaties, leerlingantwoorden uit interviews en vragenlijsten, en reflectieverslagen van de docent zijn geanalyseerd op vier criteria voor formatieve evaluatie. Geconcludeerd wordt dat de bovenbouw vwo-leerlingen met het instrument de nauwkeurigheid, betrouwbaarheid en validiteit in een onderzoek in voldoende mate zelfstandig kunnen evalueren in verschillende natuurwetenschappelijke onderzoekscontexten.

1 Inleiding

In het Nederlandse curriculum voor de bètavakken is, in navolging van de wereldwijde trend, de afgelopen decennia steeds meer aandacht gekomen voor leren onderzoeken door bovenbouw vwo- leerlingen. Deze verandering richt zich op het ontwikkelen van inzicht bij leerlingen in de *nature of science* en in het denken en werken van een natuurwetenschapper (Aarsen & Van der Valk, 2008; Abd-El-Khalick et al., 2004; Van Rens, Van Muijlwijk, Beishuizen, & Van der Schee, 2011).

Het doen van onderzoek in de bètavakken van het voortgezet onderwijs kan verschillende doelen hebben. De nadruk kan liggen op het vergroten van conceptuele kennis over natuurwetenschappen, het uitbreiden van praktische vaardigheden, of op het vergroten van procedureel inzicht binnen natuurwetenschappelijk onderzoek, bijvoorbeeld het evalueren van nauwkeurigheid, betrouwbaarheid en validiteit (hierna weergegeven als NBV) van een onderzoek (Gott & Duggan, 1995; Millar, 2010).

Het vergroten van procedureel inzicht helpt vwo-leerlingen zich een beeld te vormen van de criteria en werkwijzen die natuurwetenschappers gebruiken wanneer zij een onderzoek bedenken en uitvoeren (Chinn & Malhotra, 2002). Vwo-leerlingen zijn echter beginners in het evalueren van NBV in natuurwetenschappelijk onderzoek. Voor hen is het veelal onduidelijk wat de betekenis is van deze begrippen in zo´n onderzoekscontext. De leerlingen zijn bovendien vooral gewend aan 'kookboek-practica', waarbij de instructies stapsgewijs, zonder eigen inbreng en veelal zonder reflectie worden uitgevoerd. Het gevolg hiervan is dat bovenbouw vwo-leerlingen niet of nauwelijks weten hoe zij NBV bij een onderzoeksopdracht kunnen evalueren (Lunetta, Hofstein, & Clough, 2007; Schalk, Van der Schee, & Boersma, 2011; Van der Jagt, Schalk, & Van Rens, 2011).

Het is voor deze leerlingen bovendien meestal onduidelijk dat er veel overeenkomsten zijn in het doen van natuurwetenschappelijk onderzoek, en het evalueren van NBV, bij de verschillende bètavakken (Roberts & Gott, 2002). Transfer van deze begrippen en leeractiviteiten van de ene onderzoekscontext naar de andere kan worden vergroot als leerlingen hun onderzoek actief controleren en de resultaten op gestructureerde wijze evalueren (Bransford, 2000).

Bransford (ibid.) concludeerde dat beginners in een bepaald domein, zoals de vwo-

leerlingen bij het evalueren van NBV in onderzoek, leeractiviteiten moeten uitvoeren waarin zij domeinspecifieke patronen kunnen herkennen en ondersteund worden bij het structureren van nieuwe kennis en het koppelen hiervan aan voorkennis. Andrade en Valtcheva (2009) toonden aan dat het gebruik van een zelfevaluatie-instrument studenten 'dwingt' tot reflectie op en revisie van hun handelen. Zelfevaluatie richt de studenten op de belangrijkste aspecten van een taak, helpt ze bij het structureren van nieuwe kennis en laat ze de sterke en zwakke kanten van hun verrichtingen herkennen.

Een zelfevaluatie-instrument helpt wellicht ook vwo-leerlingen bij het leren evalueren van NBV van een onderzoek in de bètavakken. Uit onderzoek van Sevian en Gonsalves (2008) met eerstejaarsstudenten is bekend dat hiertoe een samenhangende set *rubrics* ingezet kan worden. Een rubric bevat beschrijvingen van verrichtingen op beginners- en expertniveau, met bij voorkeur één of meer beschrijvingen van tussenliggende niveaus. Iedere niveaubeschrijving zou duidelijk te onderscheiden moeten zijn van de andere beschrijvingen (Burke, 2006). Wanneer de mogelijke verrichtingen van de studenten expliciet en op verschillende beheersingsniveaus zijn beschreven, zijn de rubrics ondersteunend voor het leerproces. Studenten kunnen met rubrics zelfstandig hun werk evalueren en krijgen door de concrete beschrijvingen in de rubrics inzicht in hoe zij hun werk kunnen verbeteren (Jonsson & Svingby, 2007).

Uit de overzichtsstudie van Tierney en Simon (2004) bleek echter dat rubrics die voor het evalueren van leerlingonderzoek zijn ontwikkeld in onvoldoende mate bijdragen aan het leren evalueren van NBV van onderzoek door beginners. Nederlandse docenten ervaren hetzelfde probleem bij het gebruik van rubrics bij het evalueren van een leerlingonderzoek (persoonlijke communicatie). De beschikbare rubrics bevatten bijvoorbeeld vaagheden in de beschrijvingen, zoals: *Je hebt voldoende bronnen gevonden die passen bij de onderzoeksvraag en deelvragen*. Voor de gebruikers wordt niet duidelijk hoeveel bronnen 'voldoende' is en wanneer deze 'passen' bij de onderzoeksvraag en deelvragen. Hierdoor is het voor vwo-leerlingen moeilijk om de kwaliteit van eigen werk te bepalen en verbeteringen te bedenken. Ook zijn rubrics over het algemeen niet getest op interbeoordelaarsbetrouwbaarheid - komen evaluatiescores van verschillende beoordelaars overeen - en validiteit - is het geschikt om te evalueren wat je wilt evalueren (Ledford & Sleeman, 2000). Bovendien blijkt uit de overzichtsstudie van Jonnson en Svingby (2007) dat rubrics veelal gericht zijn op het beoordelen van de kwaliteit van essays en verslagen van studenten en niet op het evalueren van onderzoeksprocessen.

Om dergelijke processen te kunnen evalueren, zijn in een eerdere studie ontwerpkarakteristieken beschreven van een zelfevaluatie-instrument waarmee leerlingen NBV van een onderzoek kunnen evalueren en is de bruikbaarheid van het instrument voor de evaluatie van NBV van een onderzoek in een klassensituatie geëxploreerd (Van der Jagt, Van Rens, Schalk, Pilot, & Beishuizen, ingediend). Reflectie op de resultaten van de voorgaande studie heeft geleid tot een verfijning van de ontwerpkarakteristieken. Het doel van de huidige studie is om inzicht te krijgen in de bruikbaarheid van het herziene zelfevaluatie-instrument bij het evalueren van NBV in verschillende natuurwetenschappelijke onderzoekscontexten door bovenbouw vwo-leerlingen.

2 Theoretisch kader:

De ontwerpkarakteristieken voor een zelfevaluatie-instrument waarmee NBV van een onderzoek geëvalueerd kan worden, zijn gebaseerd op de literatuur en de resultaten van de eerste ontwerpcyclus (Van der Jagt et al., ingediend). Verloop (2003) stelt dat docenten veelvuldig onbewust handelen op basis van hun praktijkkennis. Op grond hiervan wordt verwacht dat bij het uitwerken van de ontwerpkarakteristieken ook onbewust aspecten zijn toegevoegd aan het zelfevaluatie-instrument, doordat de ontwerpers reeds ervaring hebben met het gebruik van zelfevaluatie-instrumenten in het voortgezet onderwijs. De ontwerpkarakteristieken, die verderop nader uitgewerkt worden, hebben betrekking op:

– *inhoud* voor het evalueren van NBV van een natuurwetenschappelijk onderzoek,
– *mate van complexiteit* van de beschrijvingen in het instrument
– *mate van gedetailleerdheid* met het oog op het gebruik door beginners,
– *mate van algemene toepasbaarheid* in verschillende natuurwetenschappelijke onderzoekscontexten.

Betreffende de eerste ontwerpkarakteristiek, inhoud van het zelfevaluatie-instrument, laten recente studies naar leren onderzoeken bij de schoolvakken biologie en scheikunde (Schalk, Van der Schee, & Boersma, 2009; Van Rens, Pilot, & Van der Schee, 2010) zien dat het gebruik van het *Concepts of Evidence-model* (Gott, Duggan, Roberts, & Hussain, z.j.) geschikt is voor het vergroten van het procedureel inzicht bij leerlingen, waaronder het (leren) evalueren van NBV van een onderzoek bij de natuurwetenschappelijke schoolvakken. Gott e.a. (z.j.) beschrijven 82 *concepts of evidence (CoE)* die van belang zijn bij de opbouw van bewijs in een natuurwetenschappelijk onderzoek. Gott e.a. maken hierbij onderscheid tussen concepten van bewijsvoering die betrekking hebben op:
– afzonderlijke metingen,
– alle metingen gezamenlijk (resultaten),
– verbanden tussen resultaten,
– patronen in de resultaten,
– vergelijken van resultaten met resultaten uit ander onderzoek,
– de sociaal-maatschappelijke betekenis van de bewijsvoering uit een onderzoek.

In een vorige studie zijn 19 CoE als relevant beoordeeld voor het evalueren van NBV van een natuurwetenschappelijk onderzoek door bovenbouw vwo-leerlingen (Van der Jagt et al., ingediend). Deze 19 algemeen geformuleerde CoE zijn omgezet naar het niveau van Nederlandse leerlingen uit bovenbouw vwo en worden hierna ´items´ genoemd. Van der Jagt e.a. (ingediend) concludeerden echter dat slechts 10 van deze 19 items veelvuldig voorkwamen in het leerlingonderzoek. Het gebruik van rubrics over deze 10 items ondersteunt de leerlingen (beginners) bij het evalueren van NBV van een onderzoek. De resterende negen items zijn niettemin van belang voor het evalueren van NBV van een onderzoek en zouden moeten voorkomen in een zelfevaluatie-instrument voor beginners. Deze negen moeten zodanig uitgewerkt zijn dat leerlingen er controlehandelingen mee kunnen uitvoeren, zoals nagaan of meetapparatuur voorafgaand aan een experiment steeds op nul is gezet. Hiertoe wordt een

Tabel 1

Items uit CoE-model en uitwerking in zelfevaluatie-instrument

Items in volgorde van toepassing in onderzoek	Opgenomen in rubrics	Opgenomen in checklist
1. Steeds dezelfde onafhankelijke en afhankelijke variabele		X
2. Specifieke en afgeperkte onderzoeksvraag	X	
3. Hypothese is toetsbaar door middel van beschreven onderzoek	X	
4. Onderzoeksopzet geschikt om onderzoeksvraag te beantwoorden en/of hypothese te toetsen	X	
5. Trekken van een steekproef	X	
6. Meetapparatuur voldoende nauwkeurig voor onderzoek		X
7. Meetapparatuur ijken of op nul zetten		X
8. Beperken van invloed van omgevingsvariabelen		X
9. Controle-experiment of blanco		X
10. Herhalen van metingen		X
11. Metingen en waarnemingen objectief uitvoeren		X
12. Metingen en waarnemingen systematisch uitvoeren		X
13. Gemiddelde meetwaarden en spreiding	X	
14. Trekken van een conclusie	X	
15. Evaluatie van de nauwkeurigheid	X	
16. Evaluatie van de betrouwbaarheid	X	
17. Evaluatie van de validiteit	X	
18. Vergelijkbare resultaten en/of conclusies uit ander onderzoek		X
19. Valide vervolgonderzoek	X	

Noot. Alle items zijn opgenomen in de oriënteringskaart.

deelinstrument, een checklist, toegevoegd aan het zelfevaluatie-instrument. De 19 items en de wijze van uitwerking in het zelfevaluatie-instrument zijn weergegeven in Tabel 1.

De tweede ontwerpkarakteristiek, de mate van complexiteit in de beschrijvingen in het zelfevaluatie-instrument, is van toepassing op de 10 items die in rubrics worden uitgewerkt. De negen items die in een afzonderlijk deelinstrument worden uitgewerkt, kunnen als controlemomenten op zo laag mogelijk complexiteitsniveau worden uitgewerkt. De beschrijvingen in de rubrics zouden een hiërarchische rangschikking moeten hebben waarin de verschillende complexiteitsniveaus van een uitgewerkt item zichtbaar worden (Jonsson & Svingby, 2007). Hierdoor kunnen leerlingen en docenten bepalen aan welk complexiteitsniveau het betreffende onderdeel van hun onderzoek voldoet en wat gedaan zou moeten worden om aan een complexer niveau te voldoen.

Als basis voor de hiërarchische beschrijvingen in de rubrics is gekozen voor de *Structure of Observed Learning Outcomes (SOLO)-taxonomie* (Biggs & Tang, 2007). Deze taxonomie bleek in de studie van Chan, Tsui, Chan en Hong (2002) het meest geschikt voor het vaststellen van verschillende soorten leeruitkomsten van leerlingen. Ook bleek de SOLO-taxonomie ondersteuning te bieden aan universitaire studenten bij het evalueren van hun verrichtingen op verschillende momenten tijdens de uitvoering van een leertaak (o.a. Hodges & Harvey, 2003; Levins & Pegg, 1993; Minogue & Jones, 2009).

In de SOLO-taxonomie worden vijf niveaus van toenemende complexiteit onderscheiden: prestructureel, unistructureel, multistructureel, relationeel en uitgebreid-abstract. Voor het evalueren van NBV is sprake van prestructureel niveau als leerlingen gebruik maken van alledaagse taal om NBV te beschrijven, bijvoorbeeld: “We hebben het zo exact mogelijk gedaan”. Er is sprake van het unistructureel niveau als een leerling één aspect noemt, meestal door imitatie van taalgebruik uit het lesmateriaal of van de docent. Op unistructureel niveau kunnen controlehandelingen worden uitgewerkt, zoals de negen items die geschikt zijn voor verwerking in een checklist. Op multistructureel niveau beschrijven leerlingen meerdere aspecten die van belang zijn, zonder in te gaan op inconsistenties of mogelijke verbanden tussen deze aspecten. Op relationeel niveau gaan leerlingen wel in op deze inconsistenties en verbanden. Van beheersing van het uitgebreid-abstract niveau is sprake wanneer een leerling kan aangeven hoe zijn onderzoek past in het betreffende domein. Wanneer de SOLO-taxonomie op juiste wijze is uitgewerkt in een instrument mag verwacht worden dat een bepaald niveau alleen beheerst wordt als alle voorgaande niveaus ook worden beheerst. Een verrichting kan dus alleen voldoen aan de beschrijving op het relationeel niveau wanneer deze ook volledig aan het multistructureel niveau voldoet (Biggs & Tang, 2007; Van der Jagt et al., 2011).

Ten aanzien van de derde ontwerpkarakteristiek, de mate van gedetailleerdheid in het instrument ten behoeve van de evaluatie van NBV binnen een onderzoekscontext, bleek uit eerdere studies dat een analytisch instrument het meest geschikt is voor zelfevaluatie door studenten (Arter & McTighe, 2001; Mertler, 2001). Van een instrument met analytisch karakter verwacht Mertler (ibid.) dat leerlingen handvatten hebben om gedetailleerd(er) naar hun product te kijken en informatie krijgen over het verbeteren van hun prestatie. Vooral leerlingen die weinig ervaring hebben met zelfevaluatie van hun producten leren meer van het gebruik van analytische dan van holistische rubrics.

Van der Jagt e.a. (ingediend) namen echter waar dat leerlingen met het zelfevaluatie-instrument wel de NBV van de verschillende onderdelen van een onderzoek evalueerden, maar geen passende uitspraak deden over de NBV betreffende het gehele onderzoek. De vwo-leerlingen zagen de rubrics als losse onderdelen. Om hun meer inzicht te geven in de samenhang tussen de 19 items en om het geheel van items in één oogopslag te kunnen overzien, kan een holistisch overzichtsinstrument worden toegevoegd aan het zelfevaluatie-instrument.

Voor de vierde ontwerpkarakteristiek, de mate van algemene toepasbaarheid in verschillende natuurwetenschappelijke onder-

zoekscontexten, is gekozen voor het ontwerpen van een generiek zelfevaluatie-instrument (Arter & McTighe, 2001; Jonsson & Svingby, 2007). Daarmee kunnen leerlingen NBV van een onderzoek evalueren in verschillende natuurwetenschappelijke onderzoekscontexten. Een knelpunt bij het gebruik van een generiek instrument, in het bijzonder door beginners, is de transfer van de door leerlingen opgedane kennis van de ene naar een andere context (Bransford, 2000). Jonsson en Svingby (2007) stellen dat leerlingen de algemene beschrijvingen in rubrics beter kunnen begrijpen als deze vergezeld worden van een normstellend voorbeeld. Zij suggereren om de diversiteit aan voorbeelden zo breed te maken als het aantal verschillende contexten waarin het instrument gebruikt wordt. Van der Jagt e.a. (ingediend) concludeerden echter dat variatie in de voorbeelden verwarrend is voor leerlingen, omdat zij daardoor bijvoorbeeld de hiërarchie in de rubrics en samenhang tussen rubrics niet doorzien. De voorbeelden in het herontwerp van de rubrics uit het zelfevaluatie-instrument zouden daarom allemaal moeten aansluiten bij hetzelfde onderzoeksthema. Hierbij moet gezocht worden naar een thema dat voldoende algemeen is om voor leerlingen transfer naar verschillende vakspecifieke onderzoekscontexten inzichtelijk te maken en breed genoeg om uitgewerkt te kunnen worden in alle niveaus van alle rubrics. Tabel 2 geeft een overzicht van de ontwerpkarakteristieken voor een zelfevaluatie-instrument waarmee de NBV van een onderzoek geëvalueerd kan worden.

Alvorens onderzocht kan worden wat de leeropbrengst is voor leerlingen als zij het zelfevaluatie-instrument gebruiken, is het van belang om eerst opnieuw na te gaan of de bijgestelde ontwerpkarakteristieken bijdragen aan de bruikbaarheid van het zelfevaluatie-instrument voor de evaluatie van NBV in onderzoek van bovenbouw vwo-leerlingen. Immers, wanneer het instrument onvoldoende bruikbaar blijkt te zijn voor dit doel, dan is het niet zinvol de leeropbrengst te onderzoeken.

Hierbij wordt uitgegaan van drie functies die het instrument binnen het onderwijsleerproces zou moeten vervullen 1) zelfevaluatie van NBV van een onderzoek door de leerlingen, 2) informatie leveren aan de docent over de ondersteuning die leerlingen hierbij nodig hebben, en 3) het ontwikkelen van een 'onderzoekstaal' waarin leerlingen en docenten effectief met elkaar kunnen communiceren over NBV van een onderzoek. Op grond van deze functies zijn vier criteria afgeleid waarmee geëvalueerd kan worden of de ontwerpkarakteristieken leiden tot een bruikbaar zelfevaluatie-instrument voor de ondersteuning van het beoogde onderwijsleerproces (Nieveen, 2009). Deze criteria zijn:

a. Het zelfevaluatie-instrument wordt door leerlingen toegepast zoals beoogd.
b Leerlingen (beginners) en docent kunnen het zelfevaluatie-instrument hanteren in de verschillende onderzoeksfasen en onderzoekscontexten.
c. Gebruik van het zelfevaluatie-instrument levert de docent informatie op over de ondersteuning die leerlingen nodig hebben bij het evalueren van NBV van hun onderzoek.

Tabel 2
Overzicht ontwerpkarakteristieken

Ontwerpkarakteristieken voor het bijgestelde zelfevaluatie-instrument	
1.	De inhoud van het instrument is gebaseerd op 19 relevante items uit het CoE-model. Tien daarvan hebben een gedetailleerde uitwerking, de andere negen minder gedetailleerd.
2.	Een deel van het instrument bestaat uit rubrics. Iedere rubric bevat beschrijvingen op vijf hiërarchische niveaus van complexiteit die gebaseerd zijn op de SOLO-taxonomie en geschikt zijn voor leerlingen uit bovenbouw-vwo.
3.	De rubrics en checklisten hebben een analytisch karakter. Ze worden aangevuld met een overzichtsinstrument.
4.	Het instrument is bruikbaar voor leerlingen uit bovenbouw-vwo in verschillende onderzoekscontexten in de bètavakken. Iedere rubric bevat normatieve voorbeelden om de algemene beschrijvingen te verduidelijken. Deze voorbeelden zijn opgesteld bij hetzelfde onderzoeksonderwerp dat globaal genoeg is voor transfer naar verschillende vakspecifieke onderzoekscontexten.

Tabel 3

Koppeling criteria en ontwerpkarakteristieken

Criteria	Sluit aan bij ontwerpkarakteristiek(en)			
	1	2	3	4
Het zelfevaluatie-instrument:				
a. Wordt toegepast zoals beoogd.		X	X	
b. Is praktisch hanteerbaar en sluit aan bij het beginnersniveau van leerlingen.	X	X	X	
c. Levert informatie over de ondersteuning die leerlingen nodig hebben.		X	X	X
d. Leidt tot toename van passende 'onderzoekstaal' bij leerlingen	X		X	X

d. Gebruik van het zelfevaluatie-instrument leidt tot een toename in het gebruik van passende 'onderzoekstaal' bij de evaluatie van NBV in een onderzoek.

Tabel 3 toont de criteria voor het onderzoek naar de bruikbaarheid van en de reflectie op de ontwerpkarakteristieken.

De onderzoeksvraag van deze studie is: In welke mate voldoet het zelfevaluatie-instrument aan de criteria voor bruikbaarheid voor de zelfevaluatie door bovenbouw vwo-leerlingen van de nauwkeurigheid, betrouwbaarheid en validiteit van een leerlingonderzoek in de bètavakken?

3 Methode

De bruikbaarheid van het zelfevaluatie-instrument is onderzocht door middel van een formatieve evaluatie als onderdeel van een ontwerpgericht onderzoek (Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006). Hierbij is een kwalitatieve onderzoeksmethode (Cohen & Manion, 1994) gebruikt met triangulatie van data (Yin, 2003).

Voorafgaand aan het onderzoek in een klassensituatie werd het zelfevaluatie-instrument door twee onderzoekers, onafhankelijk van elkaar, geanalyseerd op geschiktheid en volledigheid voor evaluatie van NBV van een onderzoek, bruikbaarheid in het lesmateriaal voor de onderzoekslessen, bruikbaarheid in verschillende fasen van een onderzoek en op volledige aansluiting bij alle leerling-producten. De analyses vertoonden een grote mate van overeenkomst. Het zelfevaluatie-instrument werd geschikt bevonden.

3.1 Beschrijving zelfevaluatie-instrument

Het zelfevaluatie-instrument bestaat uit rubrics, een checklist (beide analytisch) en een oriënteringskaart (holistisch). Figuur 1, 2 en 3 bevatten fragmenten uit deze drie deelinstrumenten (zie Van der Jagt (2012) voor het volledige zelfevaluatie-instrument).

De rubrics zijn bedoeld voor de evaluatie van NBV van een onderzoeksplan en voor de reflectie op de NBV van een volledig onderzoek. Tien items uit het CoE-model zijn uitgewerkt tot 11 rubrics. Iedere rubric bevat vijf niveaubeschrijvingen in oplopende mate van complexiteit. De voorbeelden in alle rubrics zijn uitgewerkt bij het onderzoeksthema 'Meten aan het menselijk lichaam bij inspanning'. De gebruikers kunnen in iedere rubric omcirkelen op welk niveau zij vinden dat zij het betreffende item hebben uitgewerkt.

Met de checklist kunnen gebruikers controleren of zij alle noodzakelijke handelingen hebben gedaan. De checklist bevat 16 vragen over de voorbereiding van het onderzoek en 15 vragen over de uitvoering, zoals: "Geef je in het onderzoeksplan aan hoe vaak je iedere meting wilt herhalen?" De vragen zijn afgeleid van de 19 items, waarbij het zwaartepunt ligt bij de negen items die als controlehandelingen zijn uitgewerkt in het zelfevaluatie-instrument. De gebruikers kruisen steeds één van de vier categorieën aan die achter de vraag vermeld zijn: 'gedaan', 'deels gedaan', 'niet gedaan', 'niet van toepassing op dit onderzoek'. Wanneer gebruikers aankruisen dat een handeling deels of niet is gedaan, dan krijgen zij via een aanwijzing op de checklist de opdracht om deze handeling te verbeteren.

Hoe ver ben je? Omcirkel het behaalde niveau ↓	HET THEORETISCH KADER:	VOORBEELDEN
1	is gebaseerd op informatiebronnen uit het dagelijks leven.	*Ik zag in Studio Sport een interview met een wielrenner over de invloed van sporten op zijn hartslag.*
2	is gebaseerd op één wetenschappelijke bron	*In de wetenschapsbijlage van de krant las ik dat Jansen in 2008 het verband tussen sporten en hartslagfrequentie heeft onderzocht.*
3	is gebaseerd op verschillende wetenschappelijke bronnen	*Jansen (2008) gaat in op de relatie tussen sporten en hartslagfrequentie. Hij heeft onderzocht dat…. Ook Owen (2004) heeft onderzoek gedaan naar dit verband en haar conclusies zijn…..*
4	is gebaseerd op informatie uit verschillende wetenschappelijke bronnen die verwerkt zijn tot een samenhangende tekst, waarin het centrale onderzoeksonderwerp duidelijk naar voren komt.	*Wanneer je de conclusies van Jansen (2008) vergelijkt met die van Owen (2004) dan blijkt dat uit het ene onderzoek komt dat door sporten de hartslagfrequentie verlaagd wordt en uit het andere onderzoek komt dat er geen verband is tussen sporten en de verandering van hartslagfrequentie.*
5	is gebaseerd op informatie uit verschillende wetenschappelijke bronnen die verwerkt zijn tot een samenhangende tekst, waardoor het onderzoeksonderwerp vanuit verschillende onderzoeksrichtingen bekeken wordt.	*Uit het onderzoek van Contador (2005) hebben we afgeleid dat de meetapparatuur uit het onderzoek van Jansen (2008) en Owen (2004) waarschijnlijk niet tot vergelijkbare resultaten leidt.*

Figuur 1. Voorbeeld van een rubric uit het zelfevaluatie-instrument

HEB IK ALLES GEDAAN? Controleer of je alle handelingen hebt uitgevoerd

In de eerste kolom staan de omschrijvingen van handelingen met betrekking tot je onderzoek.
Zet achter iedere omschrijving een kruisje in de kolom die van toepassing is op jouw onderzoek.

VOORBEREIDING ONDERZOEK	Geheel beschreven in onderzoeksplan	Deels beschreven in onderzoeksplan	Niet beschreven in onderzoeksplan	Niet van toepassing op onderzoek
VALIDITEIT **BEVAT HET ONDERZOEKSPLAN…**				
…bij ieder onderdeel steeds dezelfde onafhankelijke en afhankelijke variabelen?	☐	☐	☐	☐
…een theoretisch kader?	☐	☐	☐	☐
…een specifieke en afgeperkte onderzoeksvraag?	☐	☐	☐	☐
… een toetsbare hypothese?	☐	☐	☐	☐
…een onderzoeksmethode waarmee de onderzoeksvraag beantwoord kan worden?	☐	☐	☐	☐
… een onderzoeksmethode waarmee de hypothese getoetst kan worden?	☐	☐	☐	☐
BETROUWBAARHEID				
Noem je omgevingsvariabelen die de meetwaarden	☐	☐	☐	☐

Figuur 2. Fragment checklist uit het zelfevaluatie-instrument

De oriënteringskaart bevat een beknopte beschrijving van alle 19 uitgewerkte items. Deze kaart kan als overzichtsdocument worden gebruikt, onder andere bij het geven van een algemeen oordeel over NBV van een onderzoek.

3.2 Procedure beoogd gebruik zelfevaluatie-instrument

Het zelfevaluatie-instrument is ingezet in een lessenserie van drie onderzoeksmodules. De eerste onderzoeksmodule ('startmodule') was algemeen-natuurwetenschappelijk van aard, daarna volgde een biologiemodule en tenslotte een natuurkundemodule. In de startmodule kregen de leerlingen een onderzoeksopdracht over de afkoelingssnelheid van koffie, in de biologiemodule over het proeven van smaak door mensen en in de natuurkundemodule over de verbetering van de verkeerssituatie op een gevaarlijk kruispunt. In een vorige onderzoekscyclus gaven

Figuur 3. Fragment oriënteringskaart

de leerlingen aan dat zij bovenstaande onderzoeksopdrachten voldoende open en stimulerend vonden voor het bedenken en uitvoeren van een eigen onderzoek.

De leerlingen schreven in iedere module een onderzoeksplan met een onderzoeksvraag, hypothese en onderzoeksopzet bij een gegeven theoretisch exposé over het onderzoeksonderwerp. In de biologiemodule bestond dit theoretisch exposé bijvoorbeeld uit een beschrijving van twee experimenten over het proeven van smaak door mensen waarbij de resultaten en conclusies van beide experimenten elkaar tegenspraken. De leerlingen kregen de opdracht om te onderzoeken voor welk van beide theorieën zij het sterkste bewijs konden vinden.

Hierna controleerden en verbeterden de leerlingen hun onderzoeksplan met het checklistdeel 'Voorbereiding'. Het verbeterde onderzoeksplan werd gebruikt door de leerlingen om met de relevante rubrics de NBV van hun onderzoek te evalueren. Vervolgens maakten de leerlingen de laatste verbeteringen in hun onderzoeksplan. De volgende stap was om de definitieve onderzoeksopzet uit te voeren en vervolgens te controleren met het checklistdeel 'Uitvoering'. Hierna verwerkten zij de onderzoeksdata tot overzichtelijke resultaten, werkten de bewijsvoering uit en trokken conclusies. Tenslotte evalueerden de leerlingen de NBV van hun gehele onderzoek met behulp van de relevante rubrics.

De docent controleerde en evalueerde na afloop van elke les met dezelfde checklist en rubrics alle onderzoeksproducten die ook door de leerlingen waren geëvalueerd. Voorafgaand aan de volgende les kregen de leerlingen de checklist en rubrics met de docentevaluatie uitgereikt en vergeleken deze met hun eigen evaluatie. Op grond daarvan konden leerlingen na eventuele discussie met de docent hun onderzoek aanpassen.

Bij iedere module was een werkboek met opdrachten beschikbaar voor de leerlingen (Van der Jagt, 2012). Alle leerlingen hadden steeds de beschikking over het volledige zelfevaluatie-instrument. Het beoogde gebruik van de onderdelen van het zelfevaluatie-instrument werd klassikaal besproken voorafgaand aan het eerste gebruik ervan. Terminologie uit het zelfevaluatie-instrument werd in iedere module toegelicht in het werkboek en via klassikale uitleg. De werkboeken bevatten aanwijzingen om op diverse momenten de oriënteringskaart, checklist of rubrics te gebruiken. De leerlingen werden steeds gestimuleerd om het onderzoek te verbeteren naar aanleiding van hun controle of evaluatie.

Bij alle modules hoort een docentenhandleiding waarin het beoogde gebruik van het zelfevaluatie-instrument en handreikingen voor instructie door de docent tijdens de uitvoering van de module zijn beschreven (Van der Jagt, 2012). Bovenstaand onderwijsleerproces is elders in uitgebreidere vorm beschreven (Van der Jagt, Van Rens, Schalk, Pilot, & Beishuizen, in voorbereiding)

3.3 Participanten

Aan de studie hebben 27 leerlingen uit vwo-4 en vwo-5 deelgenomen, die zich allemaal vrijwillig hebben aangemeld. Er heeft geen selectie van leerlingen plaatsgevonden. De studie is uitgevoerd buiten de reguliere lestijd om invloed van storende factoren uit een reguliere les, zoals lesonderbrekingen na 50 minuten, te voorkomen. De leerlingen ontvingen een financiële beloning wanneer zij bij alle onderzoekbijeenkomsten aanwezig waren geweest.

De leerlingen werkten in 12 groepen aan drie opeenvolgende onderzoeksmodules gedurende zeven onderzoeksmiddagen verdeeld over een periode van drie maanden. Alle leerlingen volgen de vakken biologie, natuurkunde en scheikunde op vwo-niveau. Zij hadden ervaring met het uitvoeren van praktische opdrachten bij de bètavakken, maar waren onbekend met het systematisch evalueren van de NBV in een natuurwetenschappelijk onderzoek.

Een eerstegraads biologiedocent met acht jaar ervaring heeft de onderzoeksmodules onderwezen. Dit was tevens één van de ontwerpers van het zelfevaluatie-instrument en één van de onderzoekers. Door deze betrokkenheid bij het ontwerp was de docent goed op de hoogte van het beoogde onderwijsleerproces. Uit de eerste onderzoekscyclus met het zelfevaluatie-instrument is gebleken dat de docent goed op de hoogte moet zijn van de inhoud en beoogde toepassing van het zelfevaluatie-instrument om het evalueren van NBV te kunnen onderwijzen. Hiernaast was het door de 'dubbelrol' van de docent mogelijk om het onderwijsleerproces, indien nodig, snel bij te stellen. Er is gekozen voor één docent die alle onderzoeksmodules onderwees om verstoringen door wisseling van docenten te voorkomen.

Tijdens de onderzoeksmiddagen waren steeds twee observanten aanwezig om het verloop van de les schriftelijk vast te leggen en video-opnames te maken. Beide observanten zijn docent in de bètavakken in het voortgezet onderwijs, maar waren niet als onderzoeker of als ontwerper betrokken bij deze studie.

3.4 Dataverzameling en -analyse

De twee observanten legden iedere onderzoeksmiddag het verloop van de les vast in een lesverslag en video-opnames. Gesprekken tussen leerlingen onderling en met de docent werden vastgelegd via audio-opnames. Na iedere les hebben de docent en de observanten het gebruik van het zelfevaluatie-instrument door de 12 groepen besproken. Deze gesprekken werden vastgelegd via audio-opnames. De docent schreef een reflectieverslag over het gebruik van het zelfevaluatie-instrument in vergelijking met het beoogde gebruik en over de nabespreking met de observanten. Het reflectieverslag werd ter aanvulling en verbetering voorgelegd aan de observanten.

De leerlingen vulden na afloop van alle onderzoeksmiddagen een vragenlijst in over hun ervaringen met de bruikbaarheid van het zelfevaluatie-instrument. In totaal zijn 187 vragenlijsten ingevuld. Na de laatste onderzoeksmiddag zijn de leerlingen groepsgewijs geïnterviewd over hun ervaringen met de bruikbaarheid van het zelfevaluatie-instrument door een onderzoeker die niet bij de onderzoeksmiddagen aanwezig was geweest. Hiernaast werden alle ingevulde rubrics en checklisten van leerlingen en de docent verzameld samen met alle onderzoeksproducten van de leerlingen uit hun werkboeken. Ook aantekeningen van de leerlingen en de docent in de werkboeken en op de oriënteringskaart, checklist en rubrics werden verzameld.

Vervolgens is door twee onderzoekers onafhankelijk van elkaar geanalyseerd in hoeverre aan de vier evaluatiecriteria is voldaan. Bij verschillen in de analyse vond discussie plaats tussen deze onderzoekers tot overeenstemming werd bereikt. De uitkomst van de analyse is vervolgens ter controle voorgelegd aan de andere onderzoekers.

Om vast te stellen of het zelfevaluatie-instrument door leerlingen gebruikt is zoals beoogd (criterium a), is het gebruik van de oriënteringskaart, de checklist en rubrics door de leerlinggroepen in alle modules geanalyseerd aan de hand van de ingevulde rubrics, de lesverslagen, notities van leerlingen en de video-opnames. De norm hierbij is dat minimaal 10 leerlinggroepen (>80%) het zelfevaluatie-instrument gebruiken zoals beoogd (Juran, Gryna & Bingham, 1974).

Om te bepalen of de leerlingen het zelfevaluatie-instrument kunnen hanteren in

verschillende onderzoeksfasen en onderzoekscontexten (criterium b) zijn alle leerlingantwoorden uit de vragenlijsten en interviews geanalyseerd. Minstens 22 leerlingen (>80%) moeten hierover positieve uitspraken doen om te kunnen spreken van een hanteerbaar instrument (Juran et al., 1974). De reflectieverslagen van de docent en de lesverslagen zijn geanalyseerd op informatie waaruit blijkt of het zelfevaluatie-instrument ook hanteerbaar is voor de docent.

Om vast te stellen of het zelfevaluatie-instrument informatie oplevert over de ondersteuning die leerlingen nodig hebben van de docent bij de evaluatie van NBV van een onderzoek (criterium c) werden leerlingscores en docentscores in rubrics en checklisten vergeleken. Hierbij is als norm gehanteerd dat klassikale uitleg bij een item nodig is als minstens zes groepen (50%) een andere rubric- of checklistscore hebben dan de docent (Hafner & Hafner, 2003).

Om te bepalen of gebruik van het zelfevaluatie-instrument leidt tot een toename in het passend gebruik van 'onderzoekstaal' bij de evaluatie van NBV van een onderzoek (criterium d) zijn de audio-opnames van gesprekken tussen leerlingen en docent en de onderzoeksproducten van leerlingen geanalyseerd op veranderingen in het gebruik van items uit het CoE-model door leerlingen. Bij deze analyse is onderscheid gemaakt tussen gebruik op ´passende´ of 'niet-passende' wijze (Van der Jagt et al., 2011). Aan criterium d is voldaan als minstens 10 leerlinggroepen (>80%) minimaal één item over nauwkeurigheid, één over betrouwbaarheid en één over validiteit noemden. Ook moet van de genoemde items minimaal 80% op passende wijze gebruikt worden door de leerlingen om te kunnen spreken van voldoende toename (Juran et al., 1974).

4 Resultaten

4.1 Toepassing zelfevaluatie-instrument (criterium a)

De analyse van de lesobservaties laat zien dat alle 12 groepen de oriënteringskaart gebruikten bij het formuleren van een algemene uitspraak over de NBV van een onderzoek. Hiernaast gebruikten alle 12 groepen de kaart op één of meer andere momenten tijdens het onderzoek. Acht groepen keken regelmatig terug naar de opdracht uit de startmodule waarin de items uit de oriënteringskaart omgezet zijn naar een concrete onderzoekssituatie. Drie leerlingen schreven ook opmerkingen op de oriënteringskaart uit hun persoonlijke werkmap om items over nauwkeurigheid voor zichzelf te verduidelijken.

Alle 12 groepen vulden in iedere onderzoeksmodule de checklist en rubrics volledig in. Na de controle met behulp van de checklist stelden in alle drie de onderzoeksmodules minstens 10 groepen hun onderzoeksplan bij. Analyse van de lesverslagen en de notities in de werkboeken laat zien dat in iedere onderzoeksmodule minimaal zes groepen de beschrijvingen uit de rubrics gebruikten om het onderzoeksplan en de reflectie op het onderzoek bij te stellen. Uit de analyse van de lesverslagen blijkt dat geen van de groepen na het invullen van de checklist delen van het onderzoek opnieuw uitvoerde.

Samengevat, alle deelinstrumenten zijn in alle onderzoeksmodules door minstens 80% van de leerlinggroepen gebruikt bij de controle en evaluatie van hun onderzoek.

4.2 Hanteerbaarheid zelfevaluatie-instrument voor leerlingen (criterium b)

De analyse van de interviews met de 27 leerlingen geeft aan dat alle deelinstrumenten – oriënteringskaart, checklist en rubrics – door 25 leerlingen (93%) als hanteerbaar zijn ervaren.

Nadere analyse van de leerlingantwoorden in de vragenlijsten leverde 156 antwoorden van 27 leerlingen op die betrekking hadden op de hanteerbaarheid van het zelfevaluatie-instrument. De antwoorden zijn gelijk verdeeld over de vragenlijst van de drie onderzoeksmodules. 129 antwoorden (83%) gingen over positieve aspecten van het zelfevaluatie-instrument, zoals: *"Je zag nu pas hoe slecht je onderzoeksvraag is."* en *"Ik kon snel zien wat er verkeerd ging en wat ik vergeten was."* Dertien antwoorden (8%) gingen zowel over de positieve aspecten als beperkingen van het werken met het zelfevaluatie-instrument, bijvoorbeeld: *"Het is goed om zo naar je eigen onderzoek te kij-*

ken, maar hulp van leraren heb ik ook nodig, want om alles alleen uit te zoeken, is moeilijk." Tien antwoorden (6%) betroffen beperkingen van het instrument: *"Het was bij de smaakproef onmogelijk om genoeg meetwaarden te verzamelen om een betrouwbaar onderzoek te doen en dat te beoordelen met de* [rubric]*tabellen."* Vier antwoorden (3%) konden niet gecategoriseerd worden.

De analyse van de reflectieverslagen laat zien dat de docent heeft ervaren dat de checklist en rubrics vlot in te vullen zijn, dat alle items aansloten op minstens twee van de drie onderzoeksmodules en alle onderdelen van het onderzoek van de leerlingen geëvalueerd konden worden. De voorbeelden uit de rubrics gebruikte de docent wanneer zij twijfelde tussen twee niveaus in een rubric. Voor alle rubrics gold volgens de docent dat bij een score op het ene niveau aan de omschrijvingen op alle voorgaande niveaus was voldaan. Wel laat de analyse van de reflectieverslagen zien dat de docent in vier rubrics - Onderzoeksvraag, Evaluatie nauwkeurigheid, Evaluatie betrouwbaarheid en Evaluatie validiteit - tussenliggende niveaus nodig had om nauwkeuriger aan te kunnen geven 'hoe ver' de leerlingen zijn. De analyse van lesverslagen toont dat de docent voor zeven leerlinggroepen tijdens de uitvoering de relevante checklistvragen kon invullen. Zij schreef daarover in een reflectieverslag: *"Door tijdgebrek kon ik pas na de uitvoering voor sommige groepen de checklist invullen. Ik heb wel alle groepen mondelinge feedback gegeven op hun uitvoering, waarna de leerlingen direct veranderingen doorvoerden."*

Analyse van de lesverslagen toont dat tijdens de uitvoering van een experiment acht groepen een gemiddelde meetwaarde berekenden met de meetwaarden van de verschillende thermometers. Bij de constatering van grote verschillen in meetwaarden legden zij uit zichzelf geen verband met de verschillende meetafwijking van meetinstrumenten. Uit de lesverslagen is af te leiden dat de docent alle acht groepen hierover mondelinge feedback gaf. Onder een ingevulde checklist noteerde de docent dat een vraag ontbreekt over het gebruik van verschillende meetapparatuur en de invloed daarvan op de nauwkeurigheid.

Samengevat, meer dan 80% van de leerlingen en de docent gaven aan het zelfevaluatie-instrument als hanteerbaar te hebben ervaren. Wel werden kanttekeningen gemaakt bij de hanteerbaarheid van de checklist tijdens de uitvoeringsfase en de evaluatie van de betrouwbaarheid in de biologische onderzoekscontext.

4.3 Benodigde ondersteuning leerlingen door docent (criterium c)

Analyse van de vergelijking tussen leerling- en docentenscores in de rubrics laat zien dat in de startmodule de leerlingscore en docentscore in vier van de 11 rubrics bij 50% of meer van de groepen verschilde. In de biologiemodule gold dit voor vijf van de 10 rubrics en in de natuurkundemodule voor

Tabel 4

Overeenkomst tussen leerling- en docentscores in de rubrics

Titel rubric	Startmodule (%)	Biologiemodule (%)	Natuurkunde-module (%)
De onderzoeksvraag	50	50	33
De hypothese	42	42	50
De onderzoeksopzet	50	25	25
Trekken van een steekproef	58	17	33
Gemiddelde meetwaarde en spreiding	58	-	-
Het antwoord op de onderzoeksvraag	42	50	-
De bewijsvoering	25	50	-
Evaluatie van de nauwkeurigheid	67	25	-
Evaluatie van de betrouwbaarheid	50	58	-
Evaluatie van de validiteit	50	75	-
Ideeën voor vervolgonderzoek	42	17	-

Noot 1. De rubric Gemiddelde meetwaarde was niet van toepassing op het biologieonderzoek.
Noot 2. Door slechte weersomstandigheden werd het natuurkundeonderzoek gedeeltelijk uitgevoerd. Hierdoor kon slechts een deel van de rubrics worden ingevuld.

drie van de vier rubrics. In Tabel 4 is aangegeven welke rubrics dit betreft.

Nadere analyse van de reflectieverslagen laat zien dat de docent bij het bespreken met leerlingen van verschillen in de checklistscores en rubricscores in iedere module steeds bij drie (25%) of vier (33%) groepen merkte dat de mondelinge uiteenzetting van leerlingen meer informatie bevatte dan hun onderzoeksnotities in het werkboek. De docent noteerde hierover: *"Ik kan de rubrics beter invullen nadat ik de leerlingen hun onderzoeksplan heb laten toelichten*. […] *Misschien is eerst een korte presentatie per onderzoeksplan nog niet eens zo'n gek idee."*

Uit de analyse van de lesverslagen komt naar voren dat verschillende begrippen uit het instrument, zoals 'representatieve steekproef' en 'controleproef/blanco' in eerste instantie onvoldoende bekend waren bij de leerlingen. Wanneer de docent een dergelijk knelpunt signaleerde, gaf zij een klassikale toelichting. Analyse van de lesverslagen toont dat leerlingen hierna beter uit de voeten konden met deze begrippen.

Analyse van de reflectieverslagen van de docent toont dat leerlingen sommige items flexibeler hanteerden dan in het CoE-model wordt bedoeld. Een voorbeeld is het item *Meetapparatuur ijken of op nul zetten*. In de biologiemodule noteerden zeven groepen op de checklist dat zij bij de uitvoering van het smaakonderzoek de apparatuur bij iedere meting op nul hebben gezet, ondanks dat zij geen meetapparatuur hebben gebruikt. Navraag door de docent leerde dat deze leerlingen het spoelen van de mond door de proefpersonen beschouwden als 'het op nul zetten'.

Samengevat, bij verschillende items hadden 50% of meer van de leerlinggroepen een andere score dan de docent en leek in eerste instantie klassikale uitleg benodigd. Nadere inventarisatie door de docent liet zien dat de verschillen regelmatig veroorzaakt werden door onvolledige informatie in de werkboeken van de leerlingen of door onbekendheid met terminologie uit het zelfevaluatie-instrument.

4.4 Gebruik van 'onderzoekstaal' (criterium d)

Analyse van de onderzoeksproducten toont dat in de startmodule 11 van de 12 leerlinggroepen minimaal één item vermeldden bij de evaluatie van NBV van hun onderzoek. Door de 11 groepen samen zijn in totaal 13 van de 19 items opgeschreven. Zes leerlinggroepen (50%) noteerden ieder minimaal één item over nauwkeurigheid, één over betrouwbaarheid en één over validiteit. Eén leerlingengroep noemde items uit twee categorieën en vier leerlinggroepen uit één categorie. Hierbij werden 38% van de items op passende wijze gebruikt door de leerlinggroepen.

In de natuurkundemodule vermeldden alle 12 leerlinggroepen minimaal één item over nauwkeurigheid, één over betrouwbaarheid en één over validiteit. Alle 19 items zijn door minstens één leerlinggroep genoteerd. Alle leerlinggroepen noteerden in de natuurkundemodule 79% van de items op passende wijze.

Samengevat, gedurende de lessenserie was een toename te zien van 50% naar 100% van de leerlinggroepen die bij de evaluatie van een onderzoek minimaal één item over nauwkeurigheid, één over betrouwbaarheid en één over validiteit gebruiken. Het gebruik op passende wijze nam toe van 38% naar 79%.

5 Discussie en conclusie

Voorafgaand aan het onderzoek naar het zelfevaluatie-instrument zijn vier criteria geformuleerd om de bruikbaarheid van het zelfevaluatie-instrument voor het evalueren van de NBV in de bètavakken door vwo-leerlingen in de bovenbouw vast te stellen.

Aan criterium a. *Het zelfevaluatie-instrument wordt door leerlingen toegepast zoals beoogd* is voldaan. De onderdelen van het instrument – rubrics, checklist en oriënteringskaart – zijn door minimaal 80% van de leerlinggroepen in alle modules op de beoogde momenten en beoogde wijze tijdens het onderzoeksproces ingezet.

Ook criterium b. *Het zelfevaluatie-instrument is praktisch hanteerbaar en sluit aan bij het voorkennisniveau van leerlingen* is behaald. Uit de rapportages van leerlingen kan worden afgeleid dat meer dan 80% van de leerlingen (de gestelde norm) het

zelfevaluatie-instrument hanteerbaar vonden. De ervaringen van de docent sloten hierbij aan.

Het eerste ontwerp van het instrument werd door leerlingen en docenten ervaren als te omvangrijk (Van der Jagt et al., ingediend). Hierover zijn in deze onderzoekscyclus door leerlingen geen opmerkingen gemaakt. Wel bleek het voor de docent onmogelijk om tijdens de uitvoering een checklist met 15 vragen in te vullen voor alle leerlinggroepen. Het is wellicht niet nodig om de checklist daadwerkelijk in te vullen. Het geven van directe feedback in de uitvoeringsfase aan de hand van de checklistvragen leidt tot bijstelling van de uitvoering. Het is hierbij wel een voorwaarde dat de docent een expertrol kan vervullen door voldoende inhoudelijke kennis van de items en dit adequaat kan toepassen op de leerlingonderzoeken (Bransford, 2000).

De beschrijvingen in het instrument sluiten voldoende aan bij het niveau van beginners: de leerlingen kunnen de meeste items en deelinstrumenten hanteren in de verschillende onderzoekscontexten en -fasen. Wanneer de leerlingen een item of omschrijving niet begrepen, konden zij er na toelichting wel mee uit de voeten. Uit de lesverslagen bleek dat in de checklist een vraag over het onderling vergelijken van meetapparatuur ontbrak. Dit betreft een nadere uitwerking van het CoE *Beperken invloed van omgevingsvariabelen* (Gott et al., z.j.). Ook is voor leerlingen een nadere uitwerking nodig over de evaluatie van de betrouwbaarheid van een onderzoek als sprake is van weinig proefpersonen.

Criterium c. *Het zelfevaluatie-instrument levert de docent informatie op over de ondersteuning die leerlingen nodig hebben* is ook in voldoende mate behaald. Op basis van de ingevulde checklisten en rubrics kon de docent voor ieder item bepalen of 50% of meer van de leerlinggroepen ondersteuning nodig hadden en welke items dus klassikaal toegelicht moesten worden. Een kanttekening kan gemaakt worden bij de voorspellende waarde van verschillen tussen de leerling- en docentscores. Anders dan verwacht, gaven de onderzoeksnotities in het werkboek bij 25 à 33% van de leerlingengroepen geen goede indicatie van de beheersing van een item door leerlingen, omdat de leerlingen bepaalde elementen in het onderzoek wel hadden bedacht, maar niet genoteerd. Deze elementen zijn daardoor niet door de docent, maar wel door de leerlingen gebruikt bij de controle en evaluatie van hun onderzoek. Het verschijnsel dat leerlingen minder noteren dan zij weten, kan duiden op een verschuiving van instrumenteel naar mentaal handelen, wat juist een toename van begrip door leerlingen en minder benodigde ondersteuning door de docent impliceert (Bransford, 2000). Hiernaast lijken verschillen in leerling- en docentscores, en daarmee de indicatie van de benodigde ondersteuning, veroorzaakt doordat leerlingen beschrijvingen uit de checklist en rubrics ruimer interpreteren dan de docent. Gesprekken met leerlingen waren hierbij verhelderend en dit voorkwam dat de docent klassikaal uitleg gaf over een item dat voldoende leerlingen reeds beheersten.

Tenslotte criterium d. *Het gebruik van het zelfevaluatie-instrument leidt tot toename van gebruik van passende 'onderzoekstaal'*. Uit de analyse van de leerlingproducten bleek dat alle leerlinggroepen in de laatste module minimaal uit iedere categorie één item noemden, een toename van 50% ten opzichte van de eerste module. Hierdoor is sprake van een toename van het gebruik van 'onderzoekstaal' en is voldaan aan de norm van 80%.

Het gebruik van items op passende wijze nam toe van 38% in de startmodule naar 79% in de laatste module. Opvallend is dat alle 19 items in deze module genoemd zijn. De norm van 80% is vrijwel bereikt. Geconcludeerd kan worden dat er een forse toename is in het gebruik van passende 'onderzoekstaal'. De leerlingen zijn hiermee op weg naar een expertrol in het communiceren via 'onderzoekstaal' uit het zelfevaluatie-instrument (Bransford, 2000).

Op basis van de evaluatiegegevens over de criteria wordt geconcludeerd dat het zelfevaluatie-instrument het beoogde onderwijsleerproces in voldoende mate ondersteunt. Gezien de eerdere argumentatie is het aannemelijk dat de ontwerpkarakteristieken hebben bijgedragen aan de ontwikkeling van dit zelfevaluatie-instrument voor de evaluatie van NBV van een onderzoek door beginners.

De items uit het CoE-model (*ontwerpkarakteristiek 1*) blijken passend voor de evaluatie van de NBV met behulp van het instrument. De uitwerking van 10 items in rubrics en negen items in een checklist heeft geleid tot een instrument dat door beginners in alle onderzoeksfasen en onderzoekscontexten gebruikt kon worden.

De 10 items in de rubrics zijn wat betreft complexiteit adequaat uitgewerkt in de vijf niveaus van de SOLO-taxonomie (*ontwerpkarakteristiek 2*). De docent en de leerlingen ondervonden geen problemen met de hiërarchische opbouw in de rubrics. Het toevoegen van extra tussenliggende niveaus in de rubrics kan leiden tot een gedetailleerdere beschrijving van de mate van complexiteit waarin een item is uitgewerkt. Hierdoor zouden docent en leerlingen beter kunnen begrijpen wat met een beschrijving wordt bedoeld en kan de overeenkomst in score tussen docent en leerlingen verder worden vergroot (Chan et al., 2002). Echter, te gedetailleerdere beschrijvingen op tussenliggende niveaus kunnen de overdraagbaarheid van het instrument naar verschillende onderzoeksdomeinen teniet doen. Dit laatste is niet wenselijk wanneer rubrics geschikt moeten zijn voor gebruik in verschillende vakspecifieke onderzoekscontexten zoals de rubrics uit deze studie.

Bij de analyse ten behoeve van criterium a en b werd duidelijk dat de leerlingen de oriënteringskaart inderdaad gebruiken als hulpmiddel bij het doen van een algemene uitspraak over NBV van een onderzoek en als naslagwerk. Hierbij zijn geen problemen gesignaleerd (*ontwerpkarakteristiek 3*).

Tenslotte zijn alle deelinstrumenten en de meeste items geschikt voor evaluatie van NBV in drie verschillende onderzoekscontexten (*ontwerpkarakteristiek 4*). Alle items sloten aan bij minimaal twee van de drie onderzoeksmodules en met alle items konden de leerlingen de NBV evalueren zoals beoogd was. In de analyse zijn geen moeilijkheden met het gebruik van de set samenhangende voorbeelden gevonden, dus deze bieden waarschijnlijk de beoogde ondersteuning.

Samengevat, op basis van bovenstaande resultaten en discussie kan geconcludeerd worden dat het zelfevaluatie-instrument bruikbaar is voor bovenbouw vwo-leerlingen om NBV in een onderzoek te evalueren in verschillende natuurwetenschappelijke, vakspecifieke onderzoekscontexten. Echter, hoe het 'ideale' verloop van het onderwijsleerproces eruit kan zien en wat dan de leeropbrengst voor bovenbouw vwo-leerlingen is als zij het zelfevaluatie-instrument gebruiken, dient nader onderzocht te worden.

Literatuur

Aarsen, M., & Valk, T. van der (2008). Onderzoekende houding, een leerlijn. *NVOX, 33*(8), 354-356.

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., & Hofstein, A., (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397-419.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice, 48*(1), 12-19.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks: Corwin Press.

Biggs, J., & Tang, C. (Eds.). (2007). *Teaching for quality learning at university* (3rd ed.). Buckingham: Open University Press.

Bransford, J. D. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington: National Academy Press.

Burke, K. (2006). *From standards to rubrics in 6 steps. Tools for assessing student learning* (Revised ed.). Thousand Oaks: Corwin Press.

Chan, C. C., Tsui, M. S., Chan, M. Y. C., & Hong, J. H. (2002). Applying the Structure of the Observed Learning Outcomes (SOLO) taxonomy on student's learning outcomes. *Assessment & Evaluation in Higher Education, 27*(6), 511 - 527.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools. *Science Education, 86*(2), 175-218.

Cohen, L., & Manion, L. (1994). *Research methods in education* (4th ed.). London: Routledge.

Gott, R., & Duggan, S. (1995). *Investigative work in the science curriculum*. Buckingham: Open University Press.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (z.j.). *Research into understanding scientific evidence*. Opgehaald op 23 maart 2012, van http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool. *International Journal of Science Education, 25*(12), 1509-1528.

Hodges, L. C., & Harvey, L. C. (2003). Evaluation of student learning in organic chemistry using the SOLO taxonomy. *Journal of Chemical Education, 80*, 785-787.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144.

Juran, J. M., Gryna, F. M., & Bingham, R. S. (Eds.). (1974). *Quality control handbook* (3rd ed.) (pp. 2-16-2-19). New York: McGraw-Hill.

Ledford, B. R., & Sleeman, P. J. (2000). *Instructional design*. Greenwich: Information Age Publishing.

Levins, L., & Pegg, J. (1993). Students' understanding of concepts related to plant growth. *Research in Science Education, 23*, 165-173.

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and teaching in the school science laboratory. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 393-442). Mahwah: Lawrence Erlbaum.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, *7*(25). Opgehaald op 30 maart 2009, van http://PAREonline.net/getvn.asp?v=7&n=25.

Millar, R. (2010). *Analysing practical science activities to assess and improve their effectiveness*. Hatfield: The Association for Science Education.

Minogue, J., & Jones, G. (2009). Measuring the impact of haptic feedback using the SOLO taxonomy. *International Journal of Science Education, 31*(10), 1359-1378.

Nieveen, N. (2009). Formative evaluation in educational design research. In T. Plomp & N. Nieveen (Eds), *An introduction to educational design research.* Enschede: SLO.

Roberts, R., & Gott, R. (2002). Investigations: Collecting and using evidence. In D. Sang & V. Wood-Robinson (Eds.), *Teaching secondary scientific enquiry*. London: Association for Science Education.

Schalk, H. H., Schee, J. A. van der, & Boersma, K. T. (2009). The use of concepts of evidence by students in biology investigations. In M. Hammann, K. Boersma & A. J. Waarlo (Eds.), *The nature of research in biological education: old and new perspectives on theoretical and methodological issues (ERIDOB).* Utrecht: Bèta Press.

Sevian, H., & Gonsalves, L. (2008). Analysing how scientists explain their research: A rubric for measuring the effectiveness of scientific explanations. *International Journal of Science Education, 30*(11), 1441-1467.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, *9*(2). Opgehaald op 30 maart 2009, van http://PAREonline.net/getvn.asp?v=9&n=2.

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). *Educational design research*. London: Routledge.

Van der Jagt, S., Schalk, H., & van Rens, L. (2011). Teachers' and students' use of Concepts of Evidence in judging the quality of an inquiry. In A. Yarden & G. S. Carvalho (Eds.), *Authenticity in biology education: Benefits and challenges (ERIDOB)* (pp. 41-52). Braga: Universidade do Minho.

Van der Jagt, S. (2012). *Materialen voor (leren) evalueren van nauwkeurigheid, betrouwbaarheid en validiteit bij een natuurwetenschappelijk onderzoek door bovenbouw vwo-leerlingen.* Vrije Universiteit Amsterdam. http://hdl.handle.net/1871/38422.

Van der Jagt, S.A.W., van Rens, L., Schalk, H.H., Pilot, A., & Beishuizen, J.J. (ingediend). *A self-evaluation instrument in inquiry learning for pre-university science students: Design and exploration.*

Van der Jagt, S.A.W., van Rens, L., Schalk, H.H., Pilot, A., & Beishuizen, J.J. (in voorbereiding). *Educational process for learning to evaluate the quality of inquiries at pre-university level by using a self-evaluation instrument.*

Van Rens, L., Pilot, A., & van der Schee, J. (2010). A framework for teaching scientific inquiry in upper secondary school chemistry. *Journal of Research in Science Teaching, 47*(7), 788-806.

Van Rens, L., van Muijlwijk, J., Beishuizen, J., & van der Schee, J. (2011). Upper secondary chemistry students in a pharmacochemistry research community. *International Journal of Science Education*. DOI/10.1080/09500693.2011.591845.

Verloop, N. (2003). De leraar. In N. Verloop & J. Lowyck (Eds.), *Onderwijskunde* (pp.195-248). Wolters Noordhoff, Groningen.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks: Sage.

Auteurs

Saskia van der Jagt is verbonden aan de Faculteit Psychologie en Pedagogiek, Vrije Universiteit Amsterdam en het Coornhert Gymnasium te Gouda. **Lisette van Rens, Herman Schalk** en **Jos Beishuizen** zijn verbonden aan de Faculteit Psychologie en Pedagogiek, Vrije Universiteit Amsterdam. **Albert Pilot** is verbonden aan de Faculteit Bètawetenschappen, Universiteit Utrecht.

Correspondentieadres: Saskia van der Jagt, Coornhert Gymnasium, Jan van Renesseplein 1, 2805GT, Gouda.
E-mail: cgjt@coornhert-gymnasium.nl

Abstract

An instrument for pre-university students to evaluate the quality of their scientific inquiries

This article considers the feasibility of the design of a self-evaluation instrument for pre-university science students with which they can evaluate the accuracy, reliability and validity of their inquiries. The design is based on four design characteristics: content (CoE-model), complexity (SOLO-taxonomy), extent of details and extent of generality. Design-based research is used to test the feasibility of the instrument in a class with 27 pre-university science students in three successive inquiry units. The analysed data are: the filled-out instrument, classroom observations, students' responses in questionnaires and interviews, and teacher's reflection reports. From this study it can be concluded that the four characteristics lead to a design of a self-evaluation instrument that is feasible for pre-university students to evaluate the accuracy, reliability and validity in different inquiry contexts.