

**Het havo examen van 2000;
getallen en meningen**

Stefan Boom

'Onerlijk' en 'gemeen'?

Het havo-eindexamen van het afgelopen jaar is helemaal verkeerd uitgepakt.

Dat is de conclusie die getrokken moet worden uit de reacties op de VGN-site waar docenten hun frustraties over het examen konden spuien.

'Onerlijk, gemeen, de slechtste resultaten uit mijn twintigjarige leservaring, ik zat met een knoop in mijn buik na te kijken, woedend, die lui die examens maken zouden eens een keer....,

Kortom: hier was iets heel erg mis gegaan.

Internet is een mooi medium dat goed past in deze tijd van zappen, snelle reacties en flux handelend optreden. Een prima uitlaatklep ook voor de frustraties van de docent die vernalen wordt tussen de stapels correctie, onredelijk moeilijke opgaven, toch wel lieve leerlingen, eisende schoolleiders en tweede fase perikelen.

Zo heeft de VGN-site gewerkt en zeker zijn nut gehad. Maar kan er op wat langere termijn iets gedaan worden met de geuite kritiek op het afgelopen examen? En zo ja, wat dan? Neem bij lezing van dit artikel het eerste tijdvak van het havo-examen er nog eens bij.

De examenconstructie

Eerst moeten er wat zaken duidelijk worden die te maken hebben met de constructie van de eindexamens. Basis voor de examens zijn de twee onderwerpen die door de minister worden

vastgesteld, op voordracht van de VGN (en voorheen de Intercommissie), afgeleid uit de subdomeinen van het examenprogramma. De precieze inhoud van de stof is vastgelegd in de stofomschrijving. Deze wordt opgesteld door een – weer met inbreng van de VGN – commissie, die bestaat uit docenten, didactici en een vakwetenschappelijke specialist. Op grond van die stofomschrijving wordt een examen geproduceerd, waarvan de productiekant in handen is van de Citogroep, maar de vaststelling en de eindverantwoordelijkheid in handen ligt van de Cevo, een door de minister ingestelde commissie van drie personen.

De Citogroep heeft een wetenschappelijk medewerker voor het vak geschiedenis die de productie van de examen-vragen voor havo en vwo coördineert en organiseert. De vragen zelf worden gemaakt door zes docenten die hun voorstellen indienen en gezamenlijk bespreken. Dit werk wordt aan de Cevo-vaksectie voorgelegd die, na diverse screeningsrondes, het tijdvak vaststelt.

De Citogroep neemt dan verder de coördinatie van het drukwerk, de verspreiding over de scholen en de gegevensverzameling voor de vaststelling van de normen op zich. De Cevo blijft de verantwoordelijke instantie – met de minister van onderwijs natuurlijk. Dit is de formele procedure.

Wat houdt dit voor de praktische gang van zaken in?

In de eerste plaats is dus het uitgangspunt de stofomschrijving, niet de examenkaternen van de diverse uitgevers. Deze katernen zijn ook gebaseerd op de officiële stofomschrijving, maar vertalen deze elk op hun eigen wijze naar de lessituatie.

De productie van de vragen voor het eerste tijdvak van een nieuw onderwerp (nu de Sovjet-Unie) moet al beginnen voordat er katernen geschreven zijn. De antwoorden en de inhoud van het examen zijn dus alleen gebaseerd op de stofomschrijving.

In de Kleio-kritiek op het havo examen betreunde Hans Pols het ontbreken van een vraag over de verzuiling. Maar, in

I feel I should remind
you that the latest
research cautions
against an over-reliance
on testing, Miss Duffy...



25

de stofomschrijving van het Fin-de-Siècle onderwerp komt dat begrip nauwelijks voor, en als het voorkomt is dat in de cursieve gedeeltes, en dat waren nu juist de onderdelen die expliciet van centrale toetsing werden uitgesloten. Een ander voorbeeld: het Duitsland-onderwerp heeft als hoofdvraag de positie van Duitsland in Europa. Binnen dit thema speelt de beeldvorming in andere landen eigenlijk een belangrijker rol dan de Duitse geschiedenis zelf. Het examen bevatte dan ook onder andere passages uit de memoires van mevrouw Thatcher. De havo-kandidaten – en hun docenten – hadden daar niet op gerekend. Blijkbaar was niet duidelijk genoeg geworden welke relatie de stof had met de indeling in het subdomein en leek het onderwerp alleen over de geschiedenis van Duitsland te gaan.

In de tweede plaats is er het veel gehoorde verwijt dat er een grote kloof bestaat tussen de dagelijkse lespraktijk en de examenmakers. Maar is dat wel terecht? Het examen wordt niet gemaakt door professionals in een

laboratoriumsituatie, maar door docenten die gewoon op hun school lesgeven in de examenklassen. Zij vormen gezamenlijk een constructiegroep, die per onderwerp de vragen voor het havo- en het vwo-examen aandraagt. Zij leveren de vragen, zij zoeken de bronnen, zij leveren de ideeën, zij schatten de moeilijkheidsgraad in. In de loop van het productieproces veranderen de vragen wel, maar het blijft de vraag van een constructiegroep-lid. Overigens zet de Citogroep eens per jaar rond de kerstvakantie een advertentie voor gegadigden voor deze functie, solliciteer! Constructiegroep-leden blijven zes jaar in functie. Door deze voortdurende wisseling is gewaarborgd dat er een verschillende benadering plaatsvindt en dat er een breder draagvlak ontstaat van docenten met ervaring in de constructie van examen-vragen.

Dan is het de dag van de examenafname. Als er geen fouten zijn gemaakt hoeft er geen erratum te worden voorgelezen in de warme gymzaal, krijgen

de leerlingen het vragenboekje en het bronnenboekje en gaan ze aan de slag. Drie uur de tijd hebben ze, twee onderwerpen – die totaal verschillend kunnen zijn – worden bevestigd. De stofomschrijvingen hebben als uitgangspunt dat er ongeveer twintig lessen (in de oude stijl) per onderwerp nodig zijn om de stof te bestuderen, veertig (oude) lessen in totaal dus. Voor de tweede fase is dit een overeenkomend aantal studielast-uren geworden. Dat is heel veel. Maar dat mag ook wel: nog altijd bepaalt het examenresultaat voor de helft het eindcijfer voor het vak geschiedenis. Als ze klaar zijn, leveren de kandidaten hun werk in, de docent neemt alles mee naar huis en gaat nakijken aan de hand van het correctievoorschrift.

En daarna beginnen de problemen. De vragen zijn opgesteld in de loop van een langdurig proces, de bijbehorende antwoorden zijn daar zorgvuldig op afgestemd, en toch weten leerlingen volstrekt nieuwe, vaak heel plausibele antwoorden te bedenken die niemand

p' waarde tussen	vraag (oude stijl)	vraag (nieuwe stijl)
10-20		20, 22
21-30	5, 13, 20, 22, 24, 25	5, 13, 24, 25
31-40	4, 15, 17, 18	4, 10, 14, 15, 16, 18, 19
41-50	9, 10, 14, 16, 19, 21	2, 6, 9, 17, 21
51-60	2, 6, 7, 12	7, 12, 23
61-70	1, 3, 23	1, 3
71-80	8, 11	8, 11
gemiddelde p'-waarde	45,2	42,1
gem. p'-waarde Nederland	48,6	46,3
gem. p'-waarde Duitsland	41,6	37,6
gemiddelde score	40,65	37,87
gem. score Nederland	22,34	21,31
gem. score Duitsland	18,31	16,57

Overzicht van de p'-waarden (vraag 1 tot en met 10 vormen het Nederland-deel, 11 tot en met 25 het Duitsland-onderwerp): hoe hoger de p'-waarde hoe makkelijker de vraag.

voorzien heeft. Om hieraan tegemoet te komen is de formule: *Voorbeeld van een juist antwoord is...* boven een groot aantal vragen opgenomen. Daarbij gaat het er niet zozeer om dat historische gebeurtenissen verschillend kunnen worden weergegeven, maar dat bijvoorbeeld de interpretatie daarvan, of de beschrijving van oorzaak-gevolgrelaties op meerdere manieren tot een juist antwoord kunnen leiden. De beoordeling daarvan wordt aan de deskundigheid van de corrector overgelaten. Controle daarop ligt in handen van de tweede corrector en uiteindelijk de inspectie.

Een probleem dat in dit stadium pas echt aan het licht komt, is de multi-interpretabiliteit van een vraag. Niet altijd is het helder voor een leerling wat er van hem (of haar) wordt verwacht. Of er zit in de vraag een verwarrend element dat er ongemerkt ingeslopen is. Dat zo'n groot deel van de havo-leerlingen geen idee had dat Abraham Kuyper de voorman was van de orthodox-protestanten, was niet van tevoren ingeschat. Of, een voorbeeld van een aantal jaren geleden, in een briefvraag op het mavo-examen over de Nederlandse Opstand wist een groot aantal leerlingen niet dat een dominee protestant en een pastoor katholiek was.

In een lessituatie lost een leerkracht dit meestal op door een kleine toelichting te geven als de eerste vragende leerling langs is geweest. In een examensituatie is dat uit den boze. Het valt buitengewoon moeilijk te onderkennen welke impliciete voorkennis er van de kandidaten verwacht mag worden.

Het is eigen aan het vak geschiedenis dat er een voortdurende discussie over de interpretatie van het verleden wordt

gevoerd. In de stofomschrijving en in de manier waarop die in de lessen bestudeerd wordt, speelt dit vaak ongemerkt een belangrijke rol. Pas op het examen komt dit werkelijk naar buiten, als blijkt dat er over schijnbaar eenduidige begrippen en noties grote verschillen in opvatting bestaan. Vakken die geen wisselende onderwerpen hebben kunnen deze problemen oplossen door hun vragen vooraf uit te proberen, te pretesten. Bij geschiedenis kan dat niet, de leerlingen van dit schooljaar in de voorexamenklas weten nog niets van de Sovjet-Unie zoals in de stofomschrijving wordt voorgeschreven en kunnen daarover dus geen vragen beantwoorden.

Zonder pretesting zijn de problemen van multi-interpretabiliteit en impliciete voorkennis niet werkelijk oplosbaar. In de komende examens zal dit ongrijpbare element een rol blijven spelen.

Het examen van 2000

Terug naar het examen van dit jaar. Bij welke vragen ging het mis? Iedere leerkracht met examenklassen heeft de versnelde correctie formulieren ingevuld en opgestuurd. Deze formulieren leveren gegevens op voor het vaststellen van de norm, de N-term met ingang van dit examen. Maar er is meer uit te halen. Per vraag is uitgerekend wat de verhouding is tussen de gemiddelde score en de maximale score die op de vraag mogelijk was. Dit levert een indicatie op voor de moeilijkheid van de vraag, een zogenaamde p'-waarde. Hoe hoger de p'-waarde hoe meer leerlingen alle punten voor die vraag gehaald hebben. Een p'-waarde van .90 betekent een erg makkelijke, een p'-waarde van .10 een erg moeilijke vraag.

Een overzicht van de p'-waarden (vraag 1 tot en met 10 vormen het Neder-

land-deel, 11 tot en met 25 het Duitsland-onderwerp) geeft de tabel op de volgende bladzijde.

Voor het Nederland-deel kon een kandidaat maximaal 46, voor het Duitsland-deel maximaal 44 punten halen. Uitschieters bij de oude stijl-kandidaten vormen vraag 22 (de vraag met de leuzen) die de laagste score opleverde ($p' = 23$) en aan de andere kant de vragen 8 en 11, die beide een p' van 76 haalden. Bij de kandidaten nieuwe stijl scoorde vraag 20 ($p' = 17$) het laagst, het hoogst scoorde vraag 11 (75).

Enige interpretaties

Deze getallen kunnen gebruikt worden bij de verklaring van de problemen rondom dit examen. Welke conclusies kunnen op grond van deze cijfers getrokken worden?

Er is een duidelijk verschil tussen de kandidaten oude stijl en de kandidaten nieuwe stijl. In het geheel lijkt de oude stijl beter te scoren: één vakje hoger in de indeling van de tabel. In de gemiddelde score en in de gemiddelde p'-waarde komt dit verschil ook naar voren. De extremen lijken bij de nieuwe stijl kandidaten wat groter: op de moeilijke vragen is lager gescoord. Hierbij moet worden aangetekend dat de omvang van de populatie zeer verschilt. Voor de nieuwe stijl gaat het om 557, voor de oude stijl om een steekproef van 1778 kandidaten.

Er is een opvallend verschil in resultaten tussen het Duitsland- en het Nederland-deel. Het Duitsland-onderwerp was het nieuwe thema, in het examen het tweede gedeelte.

In de gegevensverzameling is geen onderscheid tussen een kandidaat die een vraag volledig fout heeft en daarom 0 punten scoort en een kandidaat

die een vraag overslaat wegens tijdgebrek en daarom 0 punten haalt. Het ligt voor de hand te veronderstellen dat de kandidaten vooral bij het Duitsland-onderwerp last hebben gehad van tijdgebrek. De laatste vragen (22 tot en met 25) zijn bijna allemaal slecht gemaakt.

Maar er is meer. In het algemeen heeft de stelopdracht redelijk goed gescoord in vorige examens. Je kon er, zeg maar, je cijfer mee ophalen. Dit jaar was dat anders. Bij de oude stijl kandidaten treftefen we een gemiddelde score van 6,3 en een p'-waarde van 42 aan. Bij de nieuwe stijl kandidaten was dat respectievelijk 5,91 en 39. Voor beide groepen kandidaten heeft de stelopdracht dus niet bijgedragen aan een verhoging van de totaalscore. Het inleven in een Friese landarbeider is blijkbaar voor veel leerlingen toch een stap te ver geweest.

Het is overigens in de havo-examens niet gebruikelijk een briefopdracht te geven; dat is wel zo bij het mavo-examen en het kan op het havo-examen voorkomen.

Andere conclusies zijn gecompliceerder. Zijn de vragen zonder bronnen moeilijker geweest? Als we vraag 10, de stelopdracht, uitsluiten, waren de vragen 1, 3, 5, 15 tot en met 20, 22 en 25 zonder bron. Op vraag 1 en 3 na, die scoorden beter, zitten al deze vragen onder een p'-waarde van 51. Bij de samenstelling van het examen wordt gestreefd naar een p'-waarde tussen 55 en 65, dit zijn dus moeilijke vragen. Als we kijken naar "zuiver" reproductieve vragen – en over welke vragen dat zijn is altijd discussie mogelijk – dan is in de vragen 5, 6, 7, 8, 11, 15, 16, en 15 op zijn minst een sterk reproductief element aanwezig. Deze vragen zijn redelijk gespreid over de moeilijkheids-categorieën, hoewel het opvalt dat de twee vragen waarop het hoogst gescoord werd beide een sterk reproductief karakter hebben. Blijkbaar hebben de leerlingen toch wel de stof geleerd voor dit examen.

Veel klachten gingen over de casus, vraag 21. Inderdaad een moeilijke vraag, maar toch niet extreem. Deze vraag zit netjes in de middencategorie van de gemeten p'-waarden. Vraag 22, de vraag met de leuzen, kende een streng correctie-voorschrift. Een docent die dit voorschrift juist interpreteerde kon alleen of 0 of 3 punten toekennen. Bij de versnelde correc-

tieprocedure hebben de docenten die regel toegepast. Na de normbesprekingen in den lande waarschijnlijk niet meer. Daarmee komen we aan een gevoelig punt: hoe betrouwbaar zijn deze gegevens? Het is waarschijnlijk dat deze getallen – gebaseerd op een snelle eerste nakijkronde, nog geen overleg met collega's, nog geen centrale normbespreking, streng nagekeken – een neerwaartse bias vertonen. Toch is op grond van deze gegevens de N-term vastgesteld. Voor beide examens is die op N=2 vastgesteld. Dat betekent dat er, een bijstelling met tien scorepunten heeft plaatsgevonden (bij een 'normaal' examen is N gelijk aan 1). Of dit redelijk was, is alleen na te gaan als de uiteindelijke resultaten op de eindlijsten vergeleken worden met de uitkomst van de versnelde correctie.

Lessen voor het komend examen

De verwijten aan het adres van Cito en Cevo zijn luid en duidelijk overgekomen. Er is duidelijk geworden dat dit havo-examen te moeilijk was en te lang. Het kwam in ieder geval niet overeen met het niveau waarop kandidaten en docenten gerekend hadden. Voor het komende examen heeft de Cevo daarom een aantal veranderingen doorgevoerd (informatie daarover staat in de 'novemberbrief' die aan alle scholen is gestuurd).

Het meest ingrijpend is de verandering in de scoreschaal. Het havo-examen zal tussen de 80 en 85 punten krijgen. Daarmee wordt bereikt dat het aantal vragen, of 'toetsmomenten' wordt teruggebracht. Aan de klacht over het tijdgebrek wordt hiermee tegemoetgekomen. In dezelfde lijn ligt het terugschroeven van aantal en lengte van de tekstuele bronnen. Speciale aandacht zal de toegankelijkheid van het taalgebruik in de bronnen krijgen. Bij de bronvermeldingen zal dit leiden tot een groter aantal vermeldingen 'ontleend aan' – als de bron bewerkt is – in plaats van 'bron' – als de bron in oorspronkelijke vorm is gehouden.

In het verleden heeft de Cevo al gewezen op de tijdsinvestering die de stelopdracht kost. Door het oefenen op dit soort opdrachten kunnen de kandidaten hiermee meer ervaring opdoen en daardoor de benodigde tijd beter inschatten. Verder is hier en daar gewezen op het voorkomen van een dubbele instructie in een vraag. In het vervolg zal dat zoveel mogelijk vermeden wor-

den. Uitzondering daarop vormen vragen in de vorm van: Noem en leg uit. Als dit soort vragen in tweeën worden gesplitst kan dat weer moeilijkheden opleveren bij de kandidaten.

Tenslotte

Dit artikel exploreert slechts het topje van een ijsberg. Net als het ijs van zo'n spreekwoordelijke berg vormt de examenproblematiek glibberige materie, waarop moeilijk houvast is te krijgen. Net als bij een ijsberg zit een groot deel onder de oppervlakte verstopt. Het eindexamen is voor alle betrokkenen belangrijk en de daarbij horende emoties kunnen hoog oplopen. Om meer grip te krijgen op de processen die een rol spelen in dit geheel is meer onderzoek nodig. Naar de stelopdracht bijvoorbeeld. Of naar de mate van impliciete voorkennis die verondersteld wordt. Of naar de veranderingen die de tweede fase brengt in de inzet en de aanpak van de leerlingen die geschiedenis kiezen. Of naar de wijze waarop stofomschrijvings-commissies de domeinen vertalen in een stofomschrijving. Of naar de manier waarop de uitgevers dit weer in hun katernen omzetten. Of naar zo zijn er nog veel meer te noemen.

We moeten blijven discussiëren over wat uit dit soort onderzoek naar voren zal komen. Die discussies worden al gevoerd. In het kader van de commissie-De Rooy, in *Kleio*, in VGN-commissies, in de geschiedenissecties op de scholen. Zowel de Cevo als de Citogroep leggen hun oor daar graag en regelmatig bij te luisteren en willen hun bijdrage leveren. Zoals in dit artikel.

