OVERGANGSBESLISSINGEN IN HET V.H.M.O.

A. D. DE GROOT

Vraagstelling

Aan het einde van het schooljaar wordt in de overgangsvergaderingen op elke school, voor alle leerlingen van alle klassen, beslist of zij al dan niet zullen worden bevorderd. Zoals bekend bepaalt de vergadering hierbij haar eigen beleid. Verschillende scholen gebruiken de cijferlijsten en eventueel daarbij komende gegevens over de leerlingen dus niet noodzakelijkerwijze op dezelfde manier. In grote lijn is er echter vrij veel overeenstemming, voor zover het om hetzelfde schooltype en dezelfde klas gaat: er zijn zekere „belissingsregels" waar iedereen zich in feite aan houdt. Die beslissingsregels staan echter nergens te boek. Rectoren en ervaren leraren kennen ze natuurlijk wel maar zij publiceren ze helaas niet.

De eerste vraag, die in dit artikel aan de orde komt, is of wij op die beslissingsregels door een statistische studie van cijferlijsten vat kunnen krijgen. Daarbij sluit onmiddellijk aan de vraag in hoeverre wij bij verschillende scholen, langs de weg van de statistische analyse, overeenkomsten en verschillen in overgangsbeleid kunnen aantonen.

Een tweede vraag is, of het mogelijk is via de cijfers voor de verschillende vakken en de genomen beslissingen de „invloed" van een vak op de overgangsbeslissing te meten. Bekend is, dat sommige vakken — b.v. lichamelijke oefening — in dit verband niet meetellen, dat sommige andere „zwaar" en weer andere „lichter" worden gewogen; maar een statistische maatstaf ontbreekt tot dusverre. Het leek van belang te proberen er één te ontwikkelen en in zijn samenhang met andere variabelen te analyseren, om daardoor meer inzicht te verkrijgen in de beoordelings- en beslissingsprocessen in de school — die voor de leerlingen van zo groot belang zijn.

Materiaal

Prof. Dr. R. Vuyk was zo vriendelijk ons, na verkregen toestemming van de Rectoren, het volledige materiaal aan eindrapportcijfers en genomen overgangsbeslissingen over twee jaargangen leerlingen der eerste klassen van zeven Amsterdamse openbare Lycea voor een analyse ter beschikking te stellen. In totaal ging het om:

2 jaargangen (1963-64 en 1964-65) van

7 lycea, met tezamen, over 2 jaar gerekend,

57 eerste klassen, met in totaal
1479 leerlingen.
Het aantal leerlingen per klas varieerde van 18 tot 30 met als gemiddelde 26 en mediaan 27; het aantal in de analyse opgenomen klassen per school varieerde van 5 tot 11. De zeven scholen zijn niet in alle opzichten vergelijkbaar — o.m. doordat op twee van de zeven scholen Latijn wordt gegeven, op de andere vijf niet — maar wij kunnen voor de meeste van de volgende bewerkingen wel van deze verschillen afzien. In de bewerking werden opgenomen de rapportcijfers voor acht „meetellende" vakken, namelijk: Nederlands (N), Frans (F), Engels (E), Algebra (A), Meetkunde (M), Biologie (B), Aardrijkskunde (Ar) en Geschiedenis (G), alsmede de door de vergadering genomen beslissing: bevorderd of niet. Uitslagen van herexamens waren bekend, zodat de laatste variabele strikt dichotoom kon worden opgevat — ongeacht het V.H.M.O.-schooltype waartoe de leerling na bevordering werd toegelaten.

Variabelen:
Per leerling werd gewerkt met de volgende variabelen:
• het gemiddelde cijfer, over de acht vakken, c_i;
• het aantal onvoldoende cijfers op de lijst — waarbij de 5 als een halve onvoldoende werd gerekend — n_i;
• de genomen beslissing, bevorderd of niet bevorderd.
Per klas werden de volgende „klasse-variabelen" bepaald:
• het gemiddelde van alle gegeven cijfers, voor de 8 vakken tezamen c_k;
• het aantal onvoldoende cijfers, per vak, n_{vk};
• het gemiddelde aantal onvoldoendes, over alle vakken en alle leerlingen, $\overline{n_k}$.
Naast bovengenoemde variabelen werden in de loop van de analyse nog verscheidene andere ontwikkeld, deels leerling-, deels klasse- en deels vak-variabelen. Deze komen in het volgende nader ter sprake.

Op zoek naar een „kritische score"; n_i of $\overline{c_i}$?
Bij het nemen van beslissingen over de bevordering let de vergadering ongetwijfeld op het aantal onvoldoendes, n_i, van een leerling. Ook wordt stellig wel eens — in grote trekken — gekeken naar het gemiddelde cijfer, $\overline{c_i}$. Geen van beide variabelen hebben echter beslissende betekenis; althans niet in „grensgevallen". De eerste vraag is nu, bij hoeveel onvoldoendes (of bij welk gemiddelde) die grensgevallen beginnen. Uit de gegevens blijkt, dat — in ons materiaal — de beslissing van de vergadering *altijd:*

negatief uitvalt bij $n_i \geqq 3$ (drie onvoldoendes is te veel).
positief uitvalt bij $n_i \leqq 1\frac{1}{2}$.
Alle „grensgevallen" hebben dus — op al deze verschillende scholen — of 2 of $2\frac{1}{2}$ onvoldoendes. [1]

Voor verschillende doeleinden zou het prettig zijn om een nauwkeuriger schatting te verkrijgen van de plaats waar de grens tussen slagen en zakken wordt gelegd, zo mogelijk per klas en in ieder geval per school. De vraag is hoe men die grens moet bepalen; en of men daartoe van n_i of van $\bar{c}_i$ moet uitgaan.

Allereerst: aantal onvoldoendes (n_i) en gemiddeld cijfer ($\bar{c}_i$) zijn natuurlijk sterk (negatief) gecorreleerd. Tussen deze twee (leerling-)variabelen mag men, blijkens een paar steekproefberekeningen van de rangcorrelatie binnen een klas, zeker een correlatie van ongeveer: $r(n_i, \bar{c}_i) = -0{,}80$ verwachten. Correleert men de klassegemiddelden, dus $\bar{c}_k$ en $\bar{n}_k$ — twee maten voor het niveau van een klas — dan komt men uiteraard hoger uit. Over alle 57 klassen vonden wij (mintekens weggelaten): $r_r(\bar{n}_k, \bar{c}_k) = 0{,}88$. Berekend per school, liepen de zeven (rang-) correlaties van 0,79 tot 1,00 met als gemiddelde $\bar{r}_r = 0{,}91$. Toch betekent dit niet dat n_i en $\bar{c}_i$ hetzelfde meten; zeker niet voorzover wij in het grensgebied opereren waar de spreiding van beide maten klein is. Waarschijnlijk let de leraarsvergadering meer op het aantal onvoldoendes, n_i; maar daar staat tegenover dat $\bar{c}_i$ minder grof en ongetwijfeld betrouwbaarder is.

Wij hebben getracht voor n_i en $\bar{c}_i$ apart een „kritische score" per klas te definiëren — tweeërlei antwoord dus op de vraag waar bij de beslissingen in deze klas de grens lag tussen bevorderd en gezakt.

Kritische waarden voor $\bar{c}_i$ en n_i: c_{ck} en n_{ck}

De voor een klas kritische waarde van het cijfergemiddelde, c_{ck}, werd gedefinieerd als gemiddelde van de twee laagste $\bar{c}_i$-scores van geslaagden en de twee hoogste $\bar{c}_i$-scores van gezakten. Er werd met tweemaal twee scores gewerkt om de betrouwbaarheid van de grensbepaling althans iets te verhogen.

Op dezelfde wijze werd het „kritische aantal onvoldoendes", n_{ck}, be-

[1] Dit is natuurlijk alleen een empirisch feit. Gaarne zou men een onderwijskundige, pedagogische of psychologische basis geven aan het feit, dat (b.v.) drie onvoldoendes „te veel" blijkt — maar daar zie ik geen kans toe. De beslissingsregel *leidt in de praktijk tot* ca. 25 % doublures, dat wel; maar, hoe „time-honoured" dit laatste getal ook is, een rechtvaardiging kan men er moeilijk aan ontlenen.

rekend — met dien verstande, dat bij deze berekening n_i-scores groter dan 3 en kleiner dan $1\frac{1}{2}$ vervangen werden door 3 respectievelijk $1\frac{1}{2}$.

De variabelen c_{ck} en n_{ck} kunnen beide, door hun afhankelijkheid van de „ligging" der grensgevallen, als klasse-variabelen maar matig betrouwbaar zijn. Weliswaar is hun onderlinge correlatie nog vrij substantieel — per school r_r-waarden over de klassen van -0,16 tot -0,97 met als mediaan r_r = -0,67 — maar dit kan ook betekenen dat zij voor een deel dezelfde toevalsfluctuaties meten. Dat de verschillen in „grens-beleid" tussen de scholen (en de klassen) subtiel moeten zijn — àls ze meetbaar zijn — blijkt uit de verdelingen van c_{ck} en n_{ck}:

Over alle (57) *klassen* gerekend, varieert:

c_{ck} van 5,1 tot 6,1; mediaan 5,66; en

n_{ck} van $1\frac{3}{4}$ tot $2\frac{3}{4}$; mediaan $2\frac{1}{4}$.

Over 7 *scholen* gerekend, variëren de medianen van:

c_{ck} van 5,5 tot 5,8; en

n_{ck} van 2,12 tot 2,44.

De geringe variatie van deze schoolmedianen vooral maakt het onwaarschijnlijk, dat er (op deze wijze betrouwbaar meetbare) systematische verschillen zouden bestaan tussen de scholen in hun „grensbeleid". [1]

Formule in plaats van vergadering?

Interessant is de volgende interpretatie van de algemene medianen:

„Indien men alle vergaderingen over de bevordering van eerste klassers zo goed mogelijk zou willen *vervangen door een* algemene, Amsterdamse *formule*, gebaseerd op het *cijfergemiddelde*, dan zou men als kritische waarde 5,66 moeten nemen; d.w.z. leerlingen met een hoger cijfergemiddelde worden wel, leerlingen met een lager gemiddelde worden niet bevorderd".

Of: Voor een *formule* gebaseerd op het *aantal onvoldoendes:*

„Ieder die niet meer dan 2 onvoldoendes heeft, gaat over; wie $2\frac{1}{2}$ of meer onvoldoendes heeft, doubleert".

Voor de tweede formule hebben wij nagegaan in hoeverre de uitkomsten afwijken van de in werkelijkheid genomen beslissingen, met als resultaat dat in totaal:

16 leerlingen met $n_i = 2$ niet bevorderd werden, en

13 leerlingen met $n_i = 2\frac{1}{2}$ wel bevorderd werden.

[1] Niettemin zijn er wel aanwijzingen: dat één school hogere (gemiddelde-) eisen stelt dan alle andere; dat een andere school van 1964 op 1965 milder is geworden: dat twee andere scholen strenger zijn geworden.

Daarbij moeten wij bedenken, dat deze procedure nog bijzonder grof is; niet alleen houdt de formule geen rekening met mogelijke beleidsverschillen tussen de scholen maar ook werden voor de twee scholen met Latijn als extra vak op het programma de onvoldoendes voor dit vak gewoon meegerekend. Niettemin vallen slechts in 29 van de 1479 gevallen (d.i. 2 %) de formulebeslissingen in werkelijkheid anders uit — slechter? — dan in de vergadering het geval was. Per twee klassen één geval dus. Op één der scholen (zonder Latijn), met acht klassen in het materiaal, vielen zelfs alle beslissingen conform de formule uit. Dit kan dus ook. Op grond hiervan is men geneigd zich af te vragen of het wel zo nodig is om te vergaderen — hierover. [1]

De invloed van een vak op overgangsbeslissingen

Men kan op verschillende manieren trachten de „invloed" die een vak heeft op de overgangsbeslissingen langs statistische weg te achterhalen.

1. Gegeven het feit dat n_i — met de grens tussen 2 en $2^1/_2$ — een zo goede maatstaf is, kan men stellen dat een vak waarvan de cijfers weinig of niet bijdragen tot de variantie van n_i, niet veel invloed kan hebben. Dit komt in de praktijk neer op iets dat we al weten: wie geen onvoldoendes geeft oefent nauwelijks invloed uit op de beslissingen van de vergadering. Als eerste benadering van de invloed van een vak nemen we derhalve: het *percentage onvoldoendes.*

2. Een tweede mogelijke maatstaf, die althans ons gevoel wel aanspreekt, is het *percentage overeenstemming.* Van „overeenstemming" spreken we in twee gevallen:

(1) als een leerling die voor vak V een voldoende heeft gekregen, slaagt;
(2) als een leerling die een onvoldoende heeft gekregen, zakt.

In de twee-bij-twee-tabel hieronder is het „percentage onvoldoendes"

[1] Natuurlijk moeten wij rekening houden met de mogelijkheid van een vergaderingsbeleid, dat de ingediende cijfers *niet* — zoals vroeger vrij algemeen gebruikelijk was — beschouwt *als vaste* (sacrosancte) *gegevens,* die uitdrukken wat de leerlingen voor de vakken „waard zijn".
Het komt ook voor — tegenwoordig vrij veel, naar het schijnt — dat de gegeven cijfers ter vergadering gewijzigd kunnen worden, d.w.z. dat de cijferlijst enigszins wordt aangepast aan een op „psychologische" gronden gewenste beslissing. Of dit een verbetering is, is echter op zijn minst dubieus. Wij kunnen daarop nu niet ingaan.

gelijk aan c + d, het „percentage overeenstemming" gelijk aan b + c.

	gezakt	geslaagd
vold.	a	b
onvold.	c	d

100 %

A priori mogen we van b + c — al geeft dat psychologisch ongeveer aan, hoe vaak men „gelijk gekregen" heeft — niet veel verwachten. Deze maatstaf kan weinig differentiëren; aangezien (bijvoorbeeld) de „veilige" strategie, die iedere leerling een voldoende geeft, bij een 25 % doubleerders al 75 % overeenstemming (= b) oplevert. Het is moeilijk daar boven te komen; maar wij zullen b + c toch in de analyse opnemen. [1]

3. Wil men de *samenhang* tussen de vak-cijfers (alleen verdeeld in voldoende en onvoldoende) en de overgangsbeslissingen meten, dan kan men het beste op de gegevens van de tabel een tetrachorische correlatiecoëfficient berekenen (r_t; men kan grofweg stellen dat de uitkomst monotoon oploopt met bc/ad). Daarmee meet men echter niet (alleen) de invloed, maar (vooral ook) de mate waarin de vak-cijfers de vergaderingsbeslissingen voorspellen. Ook een vak, dat niet meetelt — dus bij definitie géén invloed uitoefent — kan cijfers opleveren, die hoog met de eind-beslissingen correleren: dus met een hoge r_t.

4. Neemt men in aanmerking dat de negatieve beslissingen (zakken, doubleren) feitelijk de belangrijkste zijn, dan kan men vragen: hoe vaak doet dit vak de gezakten mede de das om? Men verkrijgt dit als *percentage voorspelde doublures* uit de 2x2-tabel door te bepalen: 100 x c/(a +c). Deze maatstaf is waarschijnlijk niet slecht, maar toch wel eenzijdig: een leraar die uitsluitend onvoldoendes zou geven zou namelijk altijd 100 % scoren. Die veronderstelling is niet realistisch; maar een zwakheid van de maatstaf is toch dat men hem doet toenemen door „domweg" het percentage onvoldoendes op te voeren.

5. Om een (nog) betere maatstaf voor „uitgeoefende invloed" te verkrijgen kan men niet met de gegevens van een 2x2-tabel volstaan. Men zou moeten weten, in hoeveel gevallen men kan volhouden, dat het

[1] Stellwag's „selecterende waarde" van een vak lijkt hier veel op; uit de tekst (*STELLWAG* 1955, p. 332-334) is echter niet duidelijk of met deze grootheid bedoeld wordt b +c, of 100 b/(b+d) of 100 c/(a+c).

cijfer voor een bepaald vak (mede) *de doorslag heeft gegeven.* Leerlingen zeggen vaak: „Ik ben op die ene 4 gezakt", of: „Als ik voor dat vak een (twee) punt(en) meer had gehad, zou ik geslaagd zijn". Hierbij aanknopend kan men als maatstaf voor de invloed van een leraar (of een vak) nemen: het percentage van alle gezakten, dat met (b.v.) 2 punten meer voor het vak in kwestie zou zijn geslaagd. Deze maatstaf is ook wat eenzijdig — hij gaat evenals de vorige van de gezakten uit — maar die eenzijdigheid is tamelijk reëel. Deze grensgevallen zijn het onaangenaamst — en in principe, met bijlessen of andere hulpmiddelen, het beste te voorkomen. Wij noemen dit het *percentage (v + 2)-redbaren;* d.w.z. het percentage niet bevorderde leerlingen, dat met een cijfer v + 2 (in plaats van v) voor vak V wèl zou zijn bevorderd.

De vraag of die 2 punten meer, voor één van de 8 vakken, zouden hebben geholpen hebben wij iets subtieler trachten te beantwoorden dan door alleen op het na de verhoging resulterende aantal onvoldoendes, n_i', te letten (als het oorspronkelijke aantal n_i was). Wij hebben de volgende beslissingsregels aangehouden. Als *na* de verhoging voor een vak:

a) $n_i' \geq 3$, dan: „niet gered";
b) $n_i \leq 2$, dan: door deze (V-)verhoging „gered" — wat o.a. betekent, dat een leerling met $n_i = 2$, die niet bevorderd werd, geacht wordt door 2 punten winst in onverschillig welk vak te zijn gered;
c) $n_i' = 2^1/_2$, dan: alléén „gered", indien in de vijf vakken N, F, E, A, M compensatie aanwezig is voor de in die vakken voorkomende onvoldoendes; daarbij wordt onder „compensatie" verstaan: minstens één 7 tegenover een 4 als laagste onvoldoende, of een 8 of twee 7's tegenover een 3; enz.

Deze laatste regel is waarschijnlijk gecompliceerder dan nodig is — maar zo is het berekend. [1]

Vergelijking van invloedsmaatstaven: resultaten

De uitkomsten van deze berekeningen zijn te vinden in onderstaande

[1] De talloze berekeningen waarvan de resultaten hier worden weergegeven, waren voor een klein deel reeds door de medewerkers van Prof. Vuyk verricht; voor het grootste deel werden zij uitgevoerd door mevr. D. A. van Naerssen-Court, voor een ander deel door de schrijver zelf.

Tabel: Invloed op overgangsbeslissingen; diverse maatstaven en uitkomsten voor 8 schoolvakken in de eerste klas V.H.M.O.

	N	F	E	A	M	B	Ar	G
1. % onvoldoendes: c + d	21	31	30	34	26	14	15	15
2. % overeenstemming: b + c	84	85	83	81	79	81	83	84
2a. Mdn (b + c), 57 klassen	85	85	84	80	80	82	82	83
3. correlatie: $r_t = f(bc/ad)$	.82	.87	.82	.81	.79	.78	.84	.86
3a. % r_t (per kl.) $< 0{,}50$	11	2	4	2	7	16	18	14
3b. % r_t (per kl.) $\geq 0{,}70$	86	89	88	80	79	79	72	77
4. % vsp. dbl.: 100 c/(a+c)	59	79	74	78	79	41	45	48
4a. Mdn. 100 c/(a+c), 57 kl.'en	64	81	77	81	78	51	49	44
5. % (v + 2)-redb., v. alle dbl.	14	19	19	18	18	8	9	10
5a. % (v + 2)-redb., v. alle redb.	42	59	58	59	56	25	27	31

Het is verleidelijk deze uitkomsten nader te analyseren — b.v. door correlaties tussen de rijen te berekenen — of ze, geheel of gedeeltelijk, in een grafiek weer te geven. Wij laten dat echter aan de lezer over; de voornaamste conclusies kunnen ook direct uit de tabel worden getrokken.

Allereerst blijkt, dat de variabelen 2, 2a, 3 en 3b weinig differentiëren tussen de vakken. Van 2 en 2a (percentage overeenstemming) was dit te verwachten. Wat de correlatie-maten betreft (3 en 3b) is het tamelijk verrassend. Het betekent, dat *alle acht vakken in hun voldoende/onvoldoende-verdeling de eindbeslissing zakken/slagen ongeveer even goed „voorspellen"*. De correlaties zien er zeer hoog uit; maar daarbij moeten wij in het oog houden dat de tetrachorische correlatiecoëfficient vaak nogal geflatteerd is, vooral bij sterk van 50-50 afwijkende tweedelingen (dus bij weinig onvoldoendes en weinig gezakten). [1]

Als *invloedsvariabele* leent zich variabele 5 het beste tot een directe, „absolute" interpretatie: het percentage van alle (381) zittenblijvers, dat met twee punten méér voor het betreffende vak gered zou zijn. Voor *Frans, Engels, Algebra en Meetkunde* is dit een kleine 20 %; voor *Biologie, Aardrijkskunde en Geschiedenis* een kleine 10 %. De eerste groep — die van de echte selectievakken — heeft, zo bekeken, *minstens twee-*

[1] Berekent men *phi-coëfficienten* voor dezelfde overall twee-bij-twee-tabellen, dan zijn de uitkomsten respectievelijk: phi (N) = 0,55; phi (F) = 0,62; phi (E) = 0,58; phi (A) = 0,54; phi (M) = 0,53; phi (B) = 0,45; phi (Ar) = 0,51 en phi (G) = 0,53. Ook zo is de differentiatie gering. Frans is weer de beste en Biologie de slechtste voorspeller. Verder zijn de vakken B, Ar en G (met lage onvoldoende-percentages) nu tot de onderste drie plaatsen afgezakt.

maal zo veel invloed op grensbeslissingen als de tweede. Het vak *Nederlands* ligt daar tussenin.

Vraagt men, hoeveel van de leerlingen die zijn blijven zitten, in totaal gered hadden kunnen worden door twee punten meer — op de juiste plaats, d.w.z. voor het vak waarbij dit het meeste zou hebben geholpen — dan komt men op 123 van de 381. *Bijna één derde van alle doubleerders zou met twee punten meer* — het effect van een goede bijles? — *zijn overgegaan.* Van deze groep, de „redbaren", zou, blijkens de laatste rij van de tabel, telkens een kleine 60 % gered zijn met 2 punten meer òf voor Frans, òf voor Engels, òf voor Algebra òf voor Meetkunde. Enz.

Nemen wij variabele 5 of 5a als uitgangspunt voor de bepaling van de bruikbaarheid van de *andere, gemakkelijker te berekenen invloedsvariabelen,* dan blijkt het volgende:

Variabele 4, het percentage voorspelde doublures, vertoont *vrijwel hetzelfde profiel* als de beste invloedsmaat, namelijk variabele 5. Het lijkt dus een goede vervanger. Een bezwaar is echter dat bij kleinere aantallen, bijvoorbeeld bij gebruik in één klas, een breuk als $c/(a + c)$ een onbetrouwbare grootheid is.

Variabele 1, het percentage onvoldoendes, vertoont echter óók vrijwel hetzelfde profiel als 5. De overeenstemming is iets minder mooi dan bij variabele 4; maar toch nog groot genoeg om te stellen, dat *de invloed van een vak op overgangsbeslissingen vrijwel evenredig is met het percentage onvoldoendes,* die in dat vak vallen.

Een verdere analyse van de tabel — b.v. van de relatieve uitzonderingspositie van Meetkunde onder de selectievakken — laten wij nu achterwege. Alleen nog één vraag. Mag men uit deze laatste bevinding afleiden: Hoe meer onvoldoendes (*een leraar* geeft), des te meer invloed (kan *hij* krijgen)? Wij gaan dit niet meer analyseren; wij stellen alleen, dat het antwoord binnen zekere grenzen en in bepaalde vakken (b.v. Nederlands) helaas zonder meer bevestigend luidt.