EXAMINATOREN GEEXAMINEERD

door Drs. R. Bannink[1] *) en F. Timmers*[2] *)*

Samenvatting. De uitslagen van 504 mondelinge examens in het vak Sociale Economie, in 1969 door het NIvRA afgenomen, zijn onderzocht op systematische invloeden, die bij het examen zouden kunnen zijn voorgekomen. Daarbij blijken duidelijk verschillen tussen examinatoren aanwezig te zijn, zowel in de range der gegeven beoordelingen als in het gemiddelde.

Op het gemiddeld cijfer kan geen invloed worden aangetoond van het uur van de dag, noch van het resultaat van het examen dat het voorgaande uur door dezelfde examinator werd afgenomen.

Wel blijkt het voor de eerste kandidaten van een examendag moeilijker te zijn een voldoende (zes of hoger) te behalen. Deze conclusie leidt in combinatie met de voorgaande ertoe dat zij in het algemeen ook minder kans maken op een extreem lage beoordeling.

1 Inleiding

De methode van examineren is vaak onderwerp van discussie bij alle betrokkenen, de organisatoren, de examinatoren en niet in de laatste plaats de kandidaten zelf.

De gebruikelijke bedenkingen tegen de objectiviteit van een mondeling examen komen veelal neer op de volgende:

a. Er is een verschil in beoordeling tussen de examinatoren.

b. De uitslag wordt beïnvloed door het uur van het examen, b.v. grotere kans op een laag cijfer op een vroeg uur of op een hoog cijfer direct na de lunchpauze van de examinatoren.

c. De uitslag wordt beïnvloed door het resultaat van de voorgaande examenkandidaat, in de zin dat het duidelijk zwaarder zou zijn een hoog cijfer te halen na een goede dan na een slechte voorganger aan de examentafel.

Ten einde deze vermoedens statistisch te controleren zijn de resultaten van de 504 in 1969 door het NIvRA afgenomen examens Sociale Economie geanalyseerd. Alvorens tot de resultaten van deze analyse over te gaan moeten enige opmerkingen worden gemaakt:

a. De interpretatie van een statistisch (vermeend) aangetoonde samenhang impliceert nog geen causaliteit, doch kan wel een hypothetische causaliteit bevestigen c.q. ondersteunen.

b. De statistische methodiek is her en der enig geweld aangedaan, doch naar de overtuiging van de schrijvers en gezien het grondmateriaal niet in die mate dat de conclusies overtrokken zijn.
Waar dat het geval is, zal in de tekst duidelijk een opmerking dienaangaande worden gemaakt.

c. De verwerking van de gegevens gaat in eerste instantie uit van een qua

1) Docent aan de NIvRA-cursus Sociale Economie te Heerlen
2) Statisticus DSM te Heerlen

kwaliteit niet-systematische toewijzing van kandidaten aan uren en/of examinatoren. Daar in de toewijzing van kandidaten aan uren alleen de afstand woonplaats-Amsterdam een rol speelt en in de toewijzing aan examinatoren alleen het criterium wordt gehanteerd dat een docent niet zijn eigen kandidaat examineert is deze veronderstelling acceptabel.

2 Het grondmateriaal

Het Secretariaat van het Bureau der Examens van het NIvRA heeft de examenresultaten verstrekt, gekenmerkt naar:

a. een identificatienummer van de examinator
b. de maand waarin het examen werd afgenomen
c. de dag en het uur van het examen.

Tengevolge van de eis tot anonimiteit der examinatoren kon dit grondmateriaal niet worden gepubliceerd. De consequenties van de verstrekte cijfers zijn: 6 en hoger is geslaagd, 5 betekent gezakt met een mogelijk herexamen na drie maanden, 4 idem na een half jaar en 3 of lager gezakt en herexamen na negen maanden of een jaar.

3 Invloed van het examentijdstip en -uur

In tabel 1 zijn deze cijfers samengevat, gericht op de invloed van het examentijdstip. Uit deze tabel blijkt dat het gemiddelde resultaat van de in juni afgenomen examens niet afwijkt van dat der decemberexamens, doch dat de herexamens (september) tot een hoger gemiddelde leiden.

Tabel 1. Gemiddelde resultaten naar tijdstip en examenuur.

examenuur	juni		sept.		dec.		totaal	
	aant. ex.	gem. cijf.	aant. ex.	gem. cijf.[3])	aant. ex.	gem. cijf.	aant. ex.	gem. cijf.
9.00- 9.45	24	5,88	4	6,25	39	5,90	67	5,91
9.50-10.35	26	6,12	5	6,40	49	5,73	80	5,90
10.40-11.25	24	5,79	4	6,00	48	5,75	76	5,78
11.30-12.15	24	6,08	5	7,00	49	6,20	78	6,22
14.00-14.45	22	5,55	5	5,40	53	6,17	80	5,95
14.50-15.35	22	6,00	4	6,50	53	5,85	79	5,92
15.40-16.25[4])	8	6,00	6	7,17	30	5,90	44	6,09
Totaal	150	5,91	33	6,42	321	5,93	504	5,96

Aannemend dat de resultaten over de verzameling kandidaten normaal verdeeld zijn, kan met een t-toets worden aangetoond dat het resultaat in

3) Terwille van de leesbaarheid in twee decimalen geschreven, die beide - gezien de aantallen en het afronden in de beoordeling op hele punten - geen of nauwelijks betekenis hebben.

4) Incl. 1 examen van 16.30-17.15 uur in september met een 7.

september significant hoger is (overschrijdingskans 0,5%) dan dat van de beide andere maanden. Weliswaar is de resultatenverdeling iets meer gecentreerd dan de normale (een grotere kurtosis), wat blijkt uit de laatste regel van tabel 3, waardoor de t-toets iets wordt afgezwakt, doch daardoor wordt naar de overtuiging der schrijvers de gevonden overschrijdingskans zeker niet boven de 2,5% gebracht.

Niettemin kan men uit dit significante verschil nog niet concluderen dat het herexamen tot betere resultaten leidt. Daarop komen wij in par. 4 nog terug.

Tussen de gemiddelde resultaten per uur is geen significant verschil te constateren. Daartoe zijn in tabel 2 de juni- en decembergegevens tezamen genomen, teneinde het significante „september-effect", door welke oorzaak dit dan ook ontstaat, uit te schakelen.

Bovendien zijn de examencijfers per examinator gecorrigeerd met het verschil tussen het gemiddeld cijfer per examinator en het totaal-gemiddelde om een eventueel verschil tussen examinatoren als mogelijk storende invloed te elimineren.

Tabel 2. Toetsing van uur-resultaten, juni + december

examenuur (i)	aant. (n_i)	gem.result. ($\bar{x}_i$)	afwijk. t.o.v. gem. ($\bar{x}_i - \bar{x}_.$)	stand. dev.v.h. gem. $= \hat{\sigma}_{\bar{x}_i}$	$z_i = \frac{\bar{x}_i - \bar{x}_.}{\sigma_{\bar{x}_i}}$	z_i^2
9.00- 9.45	63	5,90	-0,02	0,169	-0,14	0,02
9.50-10.35	75	5,85	-0,07	0,155	-0,44	0,19
10.40-11.25	72	5,74	-0,18	0,158	-1,11	1,23
11.30-12.15	73	6,07	-0,15	0,157	0,94	0,88
14.00-14.45	75	6,02	0,10	0,155	0,65	0,42
14.50-15.35	75	5,92	0,00	0,155	0,00	0,00
15.40-16.25	38	5,89	-0,03	0,218	-0,15	0,02
totaal	471	5,93	0,00			2,76

In tabel 2 is de uitvoering van de toets weergegeven. Per examenuur i wordt het gemiddelde bepaald van de in dat uur behaalde resultaten, die zijn gecorrigeerd voor de examinator-invloed. Dit gemiddeld resultaat wordt aangegeven met $\bar{x}_i$. Middelt men de individuele gecorrigeerde resultaten over alle uren, dan verkrijgt men het totaal gemiddelde $\bar{x}_.$. De nulhypothese is nu dat de steekproefgemiddelden $\bar{x}_i$ uit eenzelfde populatie afkomstig zijn (geen verschil tussen uren). Het gemiddelde en de spreiding in deze populatie worden geschat door $\bar{x}_.$ en $\hat{\sigma}$, berekend uit

$$(3.1) \quad \hat{\sigma} = \sqrt{\sum_{i,j} (x_{i,j} - \bar{x}_i)^2 \;/\; (\sum_i n_i - 7)} = 1{,}34 \text{ voor } i = 1 \ldots 7 \text{ en } j = 1 \ldots n_i.$$

$\hat{\sigma}$ is de binnen de steekproeven bestaande spreiding, gepoold over de steekproeven.

m a b blz. 114

De standaarddeviatie van elk steekproefgemiddelde is te schatten door

(3.2.) $\hat{\sigma}_{\bar{x}_i} = \hat{\sigma} / \sqrt{n_i}$

Het kwadraat van de nu te construeren grootheid

(3.3.) $z_i = (\bar{x}_i - \bar{x}.) / \hat{\sigma}_{\bar{x}_i}$

is χ^2 verdeeld met één vrijheidsgraad, de som van de zeven z_i^2-waarden is dus χ^2 verdeeld met zeven vrijheidsgraden.

De uit tabel 2 verkregen waarde van deze som (2,76) heeft volgens de genoemde verdeling een overschrijdingskans van 90%, dus er is géén reden om een verschil in beoordeling tussen de examenuren te veronderstellen.

4 Verschil tussen examinatoren?

Tabel 3 geeft per examinator de frequentieverdeling van de gegeven beoordelingen in juni en december, alsmede de daaruit volgende gemiddelde beoordeling per examinator.

Tabel 3. Frequentieverdeling van cijfers, gemiddelde beoordeling en z_i^2-waarde, per examinator, juni + december:

examinator	frequentie van beoordeling behaalde cijfers							Gemid. beoord.	Aant. exam.	z_i^2
	⩽3	4	5	6	7	8	⩾9			
1	-	4	10	10	2	1	1	5,61	28	1,4
2	-	2	2	5	8	6	1	6,71	24	7,2
3	1	2	-	3	-	-	-	4,83	6	3,6
4	1	7	7	4	6	5	2	5,94	32	0,0
5	-	-	4	9	4	2	-	6,21	19	0,7
6	1	5	7	6	8	4	3	6,15	34	0,8
7	-	1	4	7	8	7	2	6,76	29	9,9
8	-	2	17	8	11	4	-	5,95	42	0,0
9	4	6	4	9	8	1	-	5,44	32	3,8
10	1	5	6	5	2	-	1	5,30	20	3,9
11	1	-	1	1	2	-	-	5,60	5	0,2
12	3	1	10	4	2	2	-	5,18	22	6,2
13	1	2	12	8	9	3	-	5,89	35	0,0
14	-	2	1	3	5	3	4	7,06	18	11,4
15	1	2	3	-	-	-	1	5,00	7	3,0
16	1	-	1	6	5	-	1	6,29	14	0,9
17	-	1	2	4	4	-	-	6,00	11	0,0
18	-	1	7	1	2	2	-	5,77	13	0,1
19	-	5	6	7	5	-	1	5,52	23	1,9
20	-	-	2	2	1	1	-	6,17	6	0,1
21	-	5	8	5	4	3	1	5,81	26	0,1
22	-	3	5	10	5	2	-	5,92	25	0,0
totaal	15	56	119	117	101	46	17	5,93	471	55,9

Uit een eerste beschouwing blijken reeds duidelijk verschillen te bestaan tussen de verdelingen van cijfers per examinator, zowel in range (het bestreken beoordelingsgebied) als in gemiddelde en de spreiding daaromheen.

Uit de laatste kolom - de z_i^2-waarden, berekend als in tabel 2, maar hier op basis van een over examinatoren gepoolde spreiding - blijken significante verschillen tussen examinatoren. Significant hoge gemiddelde beoordelingen worden gegeven door de examinatoren 2, 7 en 14, significant lage door de examinatoren 3, 9, 10 en 12.

Bij nadere beschouwing blijkt dat de september-examens zijn afgenomen door 3 „neutrale" en 2 „hoge" beoordelaars. Het geconstateerde gemiddeld hogere septembercijfer zou dus èn door een betere kwaliteit der geëxamineerden èn door een systematische vertekening t.g.v. beoordelaars kunnen zijn ontstaan. Corrigeren we weer als in tabel 2 voor de examinator-invloed, dan wordt het gemiddeld septembercijfer 6,08, wat voor de groep septemberexamens leidt tot een z_i^2-waarde van 0,36 welke niet significant is. Herexamens geven dus geen statistisch duidelijke verbetering in resultaat.

5 Invloed van de voorgaande kandidaat?

Voor de examinatoren die in hun gemiddeld cijfer niet significant uitspringen is nagegaan in hoeverre een goede of slechte beoordeling het resultaat van het daaropvolgend examen beïnvloedt.

Daartoe is de volgende tabel samengesteld:

Tabel 4. Relatieve frequenties van resultaten per categorie van voorgaand resultaat:

voorgaand resultaat	relatieve freq. % volgend resultaat					aantal
	⩽4	5	6	⩾7	totaal	
⩽5	12,1	29,7	22,0	36,2	100,0	91
6	18,0	21,3	26,2	34,5	100,0	61
⩾7	11,3	30,0	26,2	32,5	100,0	80

De hypothese die onderzocht wordt is: beoordelingen worden in een relatieve schaal gegeven, dus na een goede beoordeling is de kans op een slechte groter (en omgekeerd) dan na een „neutrale" beoordeling.

Gezien de consequenties (par. 2) is de neutrale beoordeling een 6. De kans op een zeven of meer is vrijwel onafhankelijk van het voorgaand cijfer, al is er misschien sprake van een lichte verschuiving van 6 naar 7 waar te nemen als de voorganger 5 of minder had. Hieruit kan worden geconcludeerd dat de goede kandidaat niet van een voorgaand examen te „lijden" heeft. De kans op „voldoende" is in alle drie de regels nagenoeg gelijk, dus ook de kans op „onvoldoende". In de mate van onvoldaanheid der examinatoren zit echter een rem, zowel na een slecht als na een goed examen. Men zou dit kunnen interpreteren als een hantering van de relatieve norm indien deze ten gunste van de kandidaat zou werken (dus na een zwakke beoordeling), doch een

bewuste terughoudendheid indien dit mogelijk ten ongunste van de kandidaat zou werken.

6 Onzekerheid t.a.v. het nulpunt?

De voorgaande paragraaf roept een zekere twijfel op ten aanzien van de nulpuntsbepaling van de beoordelingsschaal. Deze twijfel zou misschien niet geheel worden weggenomen, maar toch zeker worden afgezwakt indien geen verschil kan worden aangetoond tussen „eerste" beoordelingen op een examendag en „volgende" beoordelingen, eventueel tussen „eersten" en „op een 6 volgenden".

De betreffende relatieve frequenties zijn gegeven in tabel 5 voor de in de vorige paragraaf geselecteerde groep examinatoren.

Tabel 5. Relatieve frequenties van eerste en volgende examens per examendag:

categorie examen	relatieve frequenties in % met beoordeling								aanta
	3	4	5	6	7	8	9	totaal	
eersten	1,3	13,9	35,4	30,4	8,9	7,6	2,5	100,0	79
volgenden:									
-totaal	2,2	11,2	27,6	24,5	22,4	9,1	3,0	100,0	232
- na 6	3,3	14,8	21,3	26,2	24,6	8,2	1,6	100,0	61

Een eerste beschouwing geeft het vermoeden dat er ondanks de in par. 3 bereikte conclusie (gelijke uurgemiddelden) toch sprake is van een zekere concentratie op 5- en 6-beoordelingen in het eerste examen, m.a.w. dat examen wordt als „nulpunt" genomen.

Uit de volgende tabel kan deze veronderstelling worden getoetst, tegen de bovengenoemde nulhypothesen.

De daarin vermelde hypothetische frequenties zijn verkregen door de bovengenoemde percentages voor „volgenden" toe te passen op het totaal aantal eerste-uurs examens. (Een eerste-uur examen is niet noodzakelijk van 9.00-9.45 uur, enkele examinatoren beginnen later; de eerste-uurs examens zijn per examinator gedefinieerd)

Tabel 6. Waargenomen en hypothetische frequenties van resultaten der eerste-uurs examens:

	$\leqslant 4$	5	6	7	$\geqslant 8$	totaal
waargenomen	12	28	24	7	8	79
hypoth. o.g.v.						
- tot. volgenden	10,6	21,8	19,4	17,7	9,5	79
- op 6 volgenden	14,3	16,8	20,7	19,4	7,8	79

De hieruit te berekenen χ^2 (met 4 vrijheidsgraden), toetsingsgrootheid voor

de overeenstemming tussen waargenomen en hypothetische verdeling, bedraagt 9,74 (t.o.v. totaal volgenden) en 16,29 (t.o.v. op 6 volgenden). Beide waarden zijn significant, m.a.w. de verdeling van de eerste-uurs resultaten wijkt duidelijk van de hypothetische verdelingen af.

Uit de eerste regel van tabel 5 blijkt dat het gemiddeld cijfer van de eerste-uurs kandidaten ook lager ligt: de verdeling over de cijfers 5, 6 en 7 is bij eerste-uurs examens meer naar de vijf gecentreerd.