KM 6(1982)
nag 2-11

OPSTELBEOORDELING NAAR SOCIAAL MILIEU: EEN DUBBELE GENERALISERING

Ivo W. Molenaar*)

Samenvatting

Van Oudenhoven c.s. (1981) rapporteren een empirisch onderzoek naar de beoordeling van taalprestaties van leerlingen uit verschillende sociale milieus. Daarbij leest een beoordelaar, die de leerlingen niet kent, de door hen geschreven opstellen en moet uit elk opstel het sociaal milieu van de leerling schatten. Het is één van de onderzoekshypothesen dat deze taak lastiger zal zijn, wanneer elk opstel op spelling en grammatica gecorrigeerd is alvorens het aan de beoordelaar wordt voorgelegd. Bij de statistische analyse doet zich de vraag voor of generalisatie naar andere opstellen dan wel naar andere beoordelaars beoogd wordt, dan wel beide. Omdat een soortgelijk probleem zich ook bij andere statistische consultatie kan voordoen, leek het de moeite waard onderstaande analyse van de succeskansen per opstel per beoordelaar in Kwantitatieve Methoden ter discussie te stellen.

1. Probleemstelling

In een onderzoek van Van Oudenhoven c.s. (1981) naar de beoordeling van taalprestaties kwam o.m. de volgende hypothese voor:

> "Beoordelaars zullen op grond van op spelling en grammatica gecorrigeerde opstellen het sociaal milieu van de leerling minder goed kunnen schatten dan op grond van niet gecorrigeerde opstellen".

Hier zal alleen kort worden verteld op grond van welke data de onderzoekers de houdbaarheid van deze hypothese hebben beproefd; in de geciteerde publicatie vindt men desgewenst meer informatie, bijvoorbeeld

*) Vakgroep Statistiek en Meettheorie, FSW, Oude Boteringestr. 23, 9712 GC Groningen.

over het doel van het onderzoek, over de overige hypothesen en over de details van de dataverzameling.

Nadat de leerlingen van twee derde klassen van een basisschool een opstel hadden geschreven, en via de leerkrachten het beroepsniveau van hun vaders was opgevraagd, werden aselect acht opstellen van kinderen van ongeschoolde arbeiders gekozen, en acht opstellen van kinderen waarvan de vader tot een hoog of middelhoog sociaal milieu behoorde. Uit beide groepen werd alsnog een moeilijk leesbaar opstel verwijderd. Een onderzoeker die niet van het milieu van herkomst op de hoogte was, corrigeerde alle opstellen wat betreft spelling, interpunctie, gebruik van hoofdletters en grammaticale conventies. Zowel voor als na correctie werd elk opstel door een derdeklasser van een andere school overgeschreven, zodat van elk opstel zowel een ongecorrigeerde versie als een gecorrigeerde, beide in hetzelfde kinderhandschrift, ontstond.

Als beoordelaars fungeerden 48 leerkrachten van basisscholen (LK) en 48 derdejaars studenten van een Pedagogische Academie (PA). Zowel in de LK als in de PA-groep werden aselect 24 beoordelaars gekozen die de gecorrigeerde opstellen kregen voorgelegd, terwijl de andere 24 de ongecorrigeerde opstellen moesten lezen. De volgorde van de opstellen werd per beoordelaar systematisch gevarieerd om volgorde-effecten tegen te gaan. Nadat de beoordelaar eerst alle opstellen had gelezen en er een waarderingscijfer (1 t/m 10) aan had toegekend, volgde het verzoek op een afzonderlijk formulier per opstel een I toe te kennen indien men meende dat de leerling uit het midden- of hoog milieu afkomstig was, en een II voor een leerling uit het lagere milieu.

De onderzoekers waren eerst van plan de resultaten als volgt te presenteren: er zijn 48 x 14 = 672 oordelen over gecorrigeerde opstellen, waarvan 49.7 procent juist, terwijl van de eveneens 672 oordelen over ongecorrigeerde opstellen 61.2 procent het milieu juist schat. Beide keren (eventueel ook apart voor LK en PA) zou met de tekentoets worden nagegaan of de fractie juiste milieutoeschrijvingen significant groter dan 0.5 was. Misschien wil de lezer eerst over amenderingen op dit voorstel nadenken alvorens paragraaf 2 te lezen.

2. Analyse over opstellen

Laat π de kans aangeven dat een willekeurig opstel door een willekeurige beoordelaar aan het juiste milieu wordt toegeschreven. De onderzoekshypothese is π(gecorrigeerd) < π (ongecorrigeerd), zodat het in elk geval beter lijkt de afzonderlijke toetsingen van $\pi = 0.5$ tegen $\pi > 0.5$ te vervangen door de vergelijking van twee kansen in de 2x2 tabel (juist/onjuist versus gecorrigeerd/ongecorrigeerd). De exacte hypergeometrische toets van Fisher en de bijbehorende benadering met chikwadraat veronderstellen echter dat er sprake is van telkens 672 onafhankelijke experimenten met dezelfde succeskans. Het is de vraag of de data deze veronderstelling ondersteunen, en of de generalisering naar andere beoordelaars en naar andere opstellen niet van elkaar moeten worden onderscheiden.

Tabel 1 laat zien dat er duidelijke verschillen per opstel zijn. De 48 beoordelaars van ongecorrigeerde opstellen leveren per opstel een aantal correcte milieutoeschrijvingen dat uiteenloopt van 4 (voor opstel 7) tot 47 (voor opstellen 3 en 5). Voor de 48 beoordelaars van gecorrigeerde opstellen variëert het aantal successen per opstel van 0 (opstellen 4 en 7) tot 45 (opstel 6). Onderaan de tabel staan de subtotalen over de 7 laag-milieu opstellen (1 t/m 7) en de 7 hoog-milieu opstellen en de totalen over alle 14 opstellen.

Per opstel is in tabel 1 ook het verschil opgenomen tussen de ongecorrigeerde en gecorrigeerde succesaantallen. Omdat elk opstel in beide condities 48 keer is beoordeeld, kan uit deze verschillen een toets worden afgeleid die de generalisatie naar andere opstellen beoogt: de veertien opstellen beschouwend als een steekproef, zou onder de nulhypothese (geen effect van correctie op succeskans bij milieutoeschrijving) het verschil symmetrisch om nul verdeeld moeten zijn. Het geconstateerde verschil in moeilijkheidsgraad tussen opstellen belet ons niet om hier de symmetrietoets van Wilcoxon te gebruiken. Onder de 13 verschilscores in kolom TOT die niet gelijk aan nul zijn, hebben de negatieve verschilscores een rangsom van 9. Exacte bepaling van de linker staart van de verdeling met knopen (Lehmann, 1975, pag.130) levert een eenzijdige overschrijdingskans van 0.003. Afzonderlijk voor

Tabel 1. Aantallen successen per opstel voor elk van de 14 opstellen (= items) voor 96 milieuschatters verdeeld in 48 gecorrigeerd (24 PA, 24 LK) en 48 ongecorrigeerd (24 PA, 24 LK).

		ONGECORR.			GECORR.			VERSCHIL ONG.- GECORR.		
		LK	PA	TOT	LK	PA	TOT	LK	PA	TOT
(laag	1	16	9	25	10	5	15	6	4	10
opstel)	2	3	4	7	4	2	6	-1	2	1
	3	23	24	47	14	17	31	9	7	16
	4	3	4	7	0	0	0	3	4	7
	5	24	23	47	17	18	35	7	5	12
	6	22	23	45	23	22	45	-1	1	0
	7	3	1	4	0	0	0	3	1	4
(hoog)	8	9	8	17	6	9	15	3	-1	2
	9	23	23	46	14	18	32	9	5	14
	10	19	20	39	13	15	28	6	5	11
	11	20	22	42	22	21	43	-2	1	-1
	12	8	14	22	7	10	17	1	4	5
	13	15	13	28	16	15	31	-1	-2	-3
	14	16	19	35	16	20	36	0	-1	-1
subtotaal	1-7	94	88	182	68	64	132	26	24	50
subtotaal	8-14	110	119	129	94	108	202	16	11	27
totaal	1-14	204	207	411	162	172	334	42	35	77

PA resp. LK wordt dit 0.010 resp. 0.005. Gebruik van de tabel voor de verdeling zonder knopen resp. van de normale benadering met voor knopen en nullen gecorrigeerde verwachting en variantie (Lehmann, 1975) had in deze drie gevallen tot een licht resp. matig afwijkend resultaat geleid. Merk op dat de tekentoets leidt tot een eenzijdige overschrijdingskans van 0.05 (3 van 13, TOT) resp. 0.13 (4 van de 13, LK) resp. 0.03 (3 van de 14, PA); omdat ik juist bij de in absolute waarden grote verschillen en plustekens verwacht kies ik voor de meer onderscheidende toets van Wilcoxon. Ik mag dus concluderen dat voor het merendeel van de opstellen correctie inderdaad tot minder juist milieutoeschrijvingen leidt, maar er zijn uitzonderingen.

3. Analyse over beoordelaars

In tabel 2 staat een frekwentietabel van het aantal successen Y per beoordelaar (theoretisch variërend van 0, als elk opstel fout, tot 14, als elk opstel goed was toegeschreven). De slechtste beoordelaar scoort Y = 4 en de beste Y = 11, maar de overgrote meerderheid heeft Y tussen 7 en 10 (ongecorrigeerd) resp. Y = 6, 7 of 8 (gecorrigeerd). De vaardigheid per beoordelaar loopt dus veel minder uiteen dan de moeilijkheidsgraad per opstel.

Tabel 2. Aantallen successen Y per beoordelaar (frekwentietabel) gesommeerd over de 14 opstellen.

Y =	4	5	6	7	8	9	10	11	totaal
LK, ongecorr.	0	0	1	6	4	7	5	1	24
PA, ongecorr.	0	0	1	3	4	12	4	0	24
Totaal ongecorr.	0	0	2	9	8	19	9	1	48
LK, gecorr.	1	3	7	6	5	1	1	0	24
PA, gecorr.	0	2	4	10	5	2	1	0	24
totaal gecorr.	1	5	11	16	10	3	2	0	48

We kunnen nagaan of tabel 2 verenigbaar is met de hypothese dat de succeskans constant is over personen maar variabel over opstellen: stel dat de tabel het resultaat is van 48 onafhankelijke realiseringen van een grootheid Y die zelf de som is van 14 onafhankelijke experimenten met verschillende succeskansen $\pi_j (j = 1, 2, \ldots, 14)$. In deze samengesteld binomiale verdeling heeft Y verwachting $\Sigma\, \pi_j$ en variantie $\Sigma\, \pi_j(1-\pi_j)$.

We kunnen deze variantie schatten door voor elke π_j de steekproeffractie uit tabel 1 in te vullen (dus ongecorrigeerd $\hat{\pi}_1 = 25/48$, gecorrigeerd $\hat{\pi}_1 = 15/48$, enzovoorts). Zo ontstaat een schatting $\hat{\sigma}$. Bij een verschil in vaardigheid tussen beoordelaars verwacht ik dat de geobserveerde standaardafwijking 5 groter is dan $\hat{\sigma}$ is. Het resultaat is:

	$\Sigma\, \hat{\pi}_j$	$\hat{\sigma}_2 = \Sigma\, \hat{\pi}_j(1-\hat{\pi}_j)$	$\hat{\sigma}$	s
ongecorr.	8.56	1.68	1.29	1.18
gecorr.	6.96	2.23	1.49	1.30

waarbij in de laatste kolom de standaardafwijking s staat zoals die uit de frekwenties in tabel 2 voor telkens 48 waarnemingen volgen. De empirische s is telkens zelfs iets kleiner dan de geschatte σ, zodat de gelijkwaardigheid van de personen plausibel is.

In deze analyse is impliciet aangenomen dat er onafhankelijkheid is tussen personen (bij deze onderzoeksopzet aannemelijk) en tussen opstellen (dat komt, evenals het ontbreken van interactie, terug in paragraaf 5). Uitsluitend op grond van de onafhankelijkheid tussen personen kan ik in tabel 2 al toetsen of de aantallen successen gecorrigeerd lager zijn dan ongecorrigeerd. Dat levert met de Student t-toets voor twee onafhankelijke steekproeven voor de LK-groep $t_{46} = 4.47$, voor de PA-groep $t_{46} = 4.46$ en voor alle beoordelaars $t_{94} = 6.31$; voorzover de beoordelaars als een aselecte steekproef mogen gelden zouden dus ook andere soortgelijke beoordelaars gemiddeld meer juiste milieuschattingen maken bij ongecorrigeerde opstellen.

4. Analyse per opstel

Het ziet er tot nu toe naar uit dat er meer verschil tussen opstellen dan tussen beoordelaars is. We willen daarom voor elk opstel afzonderlijk een 2x2 tabel analyseren met als classificaties goede/foute milieuschatting en ongecorrigeerd/gecorrigeerd. Dit is gedaan in tabel 3: voor opstel 1 bijvoorbeeld 25 resp. 15 goede schattingen bij telkens 48 beoordelaars. We toetsen nu de gelijkheid van de kans op een goed oordeel eenzijdig tegen het alternatief dat deze kans voor gecorrigeerde opstellen lager is. Per 2x2 tabel kan dat met chikwadraat, maar omdat de verwachte aantallen dan soms wat klein worden en omdat eenzijdig toetsen dan een extra kunstgreep vereist heb ik een normale benadering van de hypergeometrische verdeling volgens Molenaar (1970,1973) gebruikt: in de geobserveerde tabel

r-k	N-n-r+k	N-n
k	n-k	n
r	N-r	N

is onder de nulhypothese van gelijke succeskans per regel de kans op k of minder in de cel links onder vrij goed te benaderen met de kans dat een standaard normaal verdeelde grootheid hoogstens gelijk is aan

$$z = 2(N+1)^{-\frac{1}{2}}\{(k+1)^{\frac{1}{2}}(N-n-r+k+1)^{\frac{1}{2}} - (n-k)^{\frac{1}{2}}(r-k)^{\frac{1}{2}}\}.$$

Deze z-waarden zijn per 2x2 tabel in tabel 3 berekend en onderaan gecombineerd onder veronderstelling van onafhankelijkheid tussen opstellen. Er lijkt wel aanleiding te zijn de nulhypothese te verwerpen en de onderzoekshypothese bevestigd te achten, maar de gedetailleerde analyse maakt duidelijk dat dit in sterke mate is veroorzaakt doordat opstellen 1, 3, 4, 5, 9 en 10 na correctie lastiger werden, bij de overige opstellen heeft correctie geen significant effect ($\alpha = 0.05$ eenzijdig).

Dit maakt generalisering naar alle denkbare andere opstellen meer dubieus dan generalisering naar alle denkbare andere beoordelaars: niet alleen is het ene opstel blijkbaar moeilijker toe te schrijven dan het andere, maar het effect van correctie is ook verschillend. Opstel 6 en 11 bijvoorbeeld zijn en blijven eenvoudig, en opstel 2 wordt zowel voor als na correctie door bijna alle beoordelaars fout ingedeeld.

Dat ook hoog-milieu opstellen soms lastiger in te delen zijn na correctie is geenszins in strijd met de onderzoekshypothese: misschien sprongen opstellen 9 en 10 er ongecorrigeerd juist uit door opvallend goede spelling en grammatica. Het zou interessant zijn te weten op grond van welke aspecten de beoordelaars zelf zeggen dat ze de toeschrijving uitvoeren. Een speciaal punt van aandacht daarbij vormt de "base rate": als ik de instructie krijg 14 opstellen elk aan hoog of laag milieu toe te schrijven zal ik mijn verwachtingen hebben over de aantallen die ik ongeveer in elke groep moet krijgen, bijvoorbeeld elk even veel of naar rato van de Nederlandse bevolking. In elk geval zal ik vermoeden dat de proefleider niet alle opstellen uit dezelfde milieugroep heeft gerecruteerd. De beoordelaars hebben kortom hun eigen gedachten over de taak, en dat kan van invloed zijn op het resultaat. Uit tabel 1 valt af te leiden dat de LK-groep 152 ongecorrigeerde en 142 gecorrigeerde opstellen aan het lage milieu toeschrijft; voor de PA-groep is dit 137 resp. 124 (suggestie van Dr. Ch. Lewis).

Tabel 3. Analyse per opstel via 2x2 tabel met eenzijdige z-score volgens Molenaar (1970).

	goed (ongec.)	fout (ongec.)	goed (gecorr.)	fout (gecorr.)	z-score
opstel 1	25	23	15	33	-1.87
opstel 2	7	41	6	42	0.00
opstel 3	47	1	31	17	-4.16
opstel 4	7	41	0	48	-2.43
opstel 5	47	1	35	13	-3.33
opstel 6	45	3	45	3	+0.40
opstel 7	4	44	0	48	-1.47
opstel 8	17	31	15	33	-0.22
opstel 9	46	2	32	16	-3.53
opstel 10	39	9	28	20	-2.24
opstel 11	42	6	43	5	+0.63
opstel 12	22	26	17	31	-0.83
opstel 13	28	20	31	17	+0.84
opstel 14	35	13	36	12	+0.46
Gecombineerde toets					$z = -17.94/\sqrt{14} = -4.79$

5. Het Rasch-model

Voor elk opstel wordt door elke beoordelaar een dichotome score verworven, waarbij de succeskans van de moeilijkheid van het opstel en de vaardigheid van de beoordelaar kan afhangen. Als er geen interactie tussen beoordelaar en opstel is zouden de data beschouwd kunnen worden met het logistische schaalmodel van Rasch (zie b.v. Fischer, 1974; Molenaar, 1982).

Voor de gehele datamatrix van 14 opstellen en 96 beoordelaars past dat model heel slecht, maar dat model verwaarloost dan ook dat de opstellen voor de helft van de beoordelaars extra moeilijk zijn gemaakt door correctie. De vraag is of het model past voor de twee afzonderlijke groepen van 48 personen die de gecorrigeerde resp. de ongecorrigeerde versie voorgelegd kregen.

De nullen in tabel 1 spelen ons nu parten: van de opstellen 4 en 7 die door geen van de 48 beoordelaars juist zijn beoordeeld kan in een Rasch-analyse geen moeilijkheidsschatting worden bepaald. Deze twee opstellen zijn dus uit de analyse verwijderd, ook voor de ongecorrigeerde groep waar ze nog 7 resp. 4 keer goed waren ingedeeld.

Ik vind dan, met het computerprogramma PML van Gustafsson (1979, zie ook Molenaar 1981), in elk van beide groepen dat het Rasch-model goed lijkt te passen. Wegens het kleine aantal personen (48) is het onderscheidingsvermogen van de aanpassingstoetsen van Andersen en Martin-Löf maar matig, en in één groep is de Andersen-toets zelfs niet mogelijk, maar ook de gedetailleerde analyse met de BINO-optie van PML wijst uit dat de 12 overblijvende items voor elke groep van 48 beoordelaars aan het Rasch-model lijken te voldoen. Het Rasch-model houdt geen rekening met door puur gissen verkregen juiste resultaten, en vooral bij de gecorrigeerde opstellen verwacht ik dat er nogal eens gegist is. Desalniettemin lopen de opstellen duidelijk uiteen in moeilijkheid, en de beoordelaars veel minder, zoals we al eerder hebben gezien.

Wanneer de hoog-milieu opstellen bij correctie even moeilijk bleven en de laag-milieu opstellen na correctie moeilijker in te delen zouden zijn, zou dit binnen een Rasch-analyse fraai aantoonbaar zijn door de item-moeilijkheden in de twee groepen van 48 personen te

vergelijken. Maar we hebben al in tabellen 1 en 3 gezien dat het niet zo fraai ligt, en in de discussie bij tabel 3 dat die verbijzondering van de onderzoekshypothese ook inhoudelijk niet juist behoeft te zijn. Ik denk dat mijn voorstel om de data met het Rasch-model te analyseren dus weinig nieuws oplevert, en dat de analyse volgens tabel 3 de meest voor de hand liggende manier is om aan te tonen dat de opstellen na correctie in doorsnee inderdaad lastiger aan een milieugroep toe te schrijven zijn, maar dat dit effect maar bij een deel van de opstellen duidelijk optreedt. Dit neemt niet weg dat er ongetwijfeld ook andere mogelijkheden zijn om deze data te analyseren, en ik houd mij aanbevolen voor suggesties. Mijn dank gaat uit naar diverse student-assistenten van onze Methodologiewinkel, die bij het rekenwerk hebben geholpen.

Referenties

Gustafsson,J.E. (1979) PML: A computer program for conditional estimation and testing in the Rasch-model for dichotomous items, Reports from the Institute of Education, University of Göteborg, no. 85.

Lehmann, E.L. (1975) Nonparametrics: Statistical Methods Based on Ranks, San Francisco, Holden-Day.

Molenaar, W. (1970) Approximations to the Poisson, binomial and hypergeometric distribution functions, MC Tract no. 31, Mathematisch Centrum, Amsterdam.

Molenaar, W, (1973) Simple Approximations to the Poisson, binomial and hypergeometric distributions Biometrics, 29, 403-407.

Molenaar, I.W. (1981) Programmabeschrijving van PML (versie 3.1) voor het Rasch model, Heymans Bulletin HB-81-538-RP, Oude Boteringestr. 23 Groningen.

Molenaar, I.W. (1982) Mensen die het beter meten, Kwantitatieve Methoden, 5, 3-39.

Van Oudenhoven, J.P., Withag, J. & Siero, F. (1981) De invloed van spellings en grammaticale fouten op de beoordeling van taalprestaties van leerlingen uit verschillende sociale milieus, Vakgroep Sociale Psychologie, R.U. Groningen, ter publicatie aangeboden.

Fischer, G.H. (1974) Einführung in die Theorie Psychologischer Tests, Bern, Huber.