

cowo

Kernitem methode 'Van Naerssen'

I N S T R U K T I E.

ben wilbrink 11 juni 1976.

Centrum voor Onderzoek van het Wetenschappelijk Onderwijs

voor: beginselen privaatrecht.

De methode is de eenvoud zelve: u krijgt een aantal toetsvragen voorgelegd in de volgorde van op het tentamen gebleken moeilijkheid (pwaarde). De makkelijkste vraag is nummer 1, de overige zijn doorgenummerd in volgorde van opklimmende moeilijkheid. De p-waarden zélf blijven nog even 'geheim': het is voor de toepassing van de methode essentieel dat u zich niet door de getalsmatige pwaarden kunt laten beïnvloeden.

In de rij vragen wijst u aan wélke vraag volgens u kwa moeilijkheid correspondeert met de grens tussen voldoende en onvoldoende. Of ook: wanneer die grens volgens u tussen twee vragen in ligt: geef aan tussen welke twee vragen in de rij de grens volgens u moet liggen.

Schrijf uw antwoord in het hokje:

bij welke vraag, resp. tussen welke twee vragen, ligt volgens u de grens tussen voldoende en onvoldoende

Probeer ook nu zonder verdere vragen te stellen, zonder te smoezen met collega's, zonder u te laten beïnvloeden door wat collega's opschrijven, en zonder al te lange bedenktijd te gebruiken, de opgave te voltooien. Werk dus snel, maar wel zorgvuldig: bij serieuze toepassing van deze methode hangt van uw antwoord af of bepaalde studenten nog net een voldoende halen, of net als onvoldoende uit de bus gaan komen.

Wacht even met vragen en discussie tot iedereen klaar is. Discussie over deze methode kan onder andere aan de hand van het uit te delen discussiestuk gebeuren.

cowo

ben wilbrink 11 juni 1976.

Centrum voor Onderzoek van het Wetenschappelijk Onderwijs

voor: beginselen privaatrecht.

Bespreking van de resultaten van de methode Van Naerssen.

Vóóordat we naar de verkregen resultaten kijken moet er een afspraak gemaakt worden over de wijze waarop uit de antwoorden van de verschillende docenten de pwaarde gekozen zal worden die in de formules van Van Naerssen ingevuld zal worden. Het ligt erg voor de hand om te zeggen: daarvoor nemen we gewoon het gemiddelde van de pwaarden die door ons aangewezen zijn. Daar zijn echter op voorhand heel weinig redenen voor. Of: het nemen van een gemiddelde is niet iets 'gewoons'.

Stel dat er redenen zouden zijn om inderdaad een of andere

maat voor centrale tendentie te gebruiken. Dan zou je kunnen

kiezen uit het gemiddelde (waarbij nog weer verschil is te maken

tussen rekenkundig, meetkundig, of harmonisch gemiddelde), de modus, of de mediaan. Gezien de onnauwkeurigheid van het soort schattingen waarmee we hier te maken hebben is het af te raden om een gemiddelde te nemen. De modus, dat is de waarde die het vaakst gekozen wordt, kan bij deze kleine aantallen beoordelaars tot rare resultaten leiden, en komt dus ook niet in aanmerking. De mediaan echter is voor dit soort gegevens een heel geschikte maat: de mediaan is de middelste waarde (of het gemiddelde van de twee middelste waarden, wanneer er een even aantal beoordelaars zou zijn). Een voorbeeld:

Stel vijf docenten komen tot de volgende keuze van pwaarden:

.56, .63, .67, .73, .73 (in volgorde gezet).

De mediaan van deze getallen is het middelste getal, dus .67. Het gemiddelde zou .66 geweest zijn, en de modus het getal dat het vaakst gekozen is is in dit geval .73.

De primitieve idee achter het hanteren van gemiddelde of mediaan schijnt de gedachte van de verdelende rechtvaardigheid te zijn. Waarbij in dit geval de rechtvaardigheid tegenover collega docenten in het geding is. Bij cesuurbepaling gaat het er niet om aardig te zijn tegenover collega docenten, maar tot een jegens de studenten rechtvaardige bepaling van de zakslaag grens te komen. Uit oogpunt van de rechtvaardigbaarheid van de cesuurbepaling tegenover studenten en eventuele derden is het waarschijnlijk dat een lagere of misschien ook, afhankelijk van soort onderwijs e.d., hogere waarde dan juist de mediaan de voorkeur verdient. De vraag is dan, op grond van welke argumenten kunnen we een redelijke beslissing nemen.

Argument 1). De methode Van Naerssen, noch de oude methode De Groot, houdt rekening met de onbetrouwbaarheid van de toets, wat in gewoon Nederlands betekent dat gezakt of geslaagd zijn bij de kernitem methode afhankelijk is van een forse portie geluk of pech. Onvolkomenheden van de meting van studiestatistiek mogen studenten niet aangerekend worden. Uit dit principe volgt dat een correctie op de met de kernitem methode gevonden cesuur moet worden toegepast die de ongewenste gevolgen van de onnauwkeurigheid van de toets corrigeert. De procedure is u wel bekend van de methode Wijnen, waar een dergelijke correctie gehanteerd wordt. Op basis van dit argument zou een waarde kleiner dan de mediaan gekozen kunnen worden; het is natuurlijk ook mogelijk om de bedoelde correctie pas aan te brengen wanneer de afstestgrens volgens de methode van Naerssen berekend is: in dat geval zou men bijvoorbeeld een aantal maal de standaard meetfout van de berekende afstestgrens (= het referentiepunt, zoals Wijnen zou zeggen) af kunnen trekken.

Argument 2). De methode Van Naerssen houdt op geen enkele wijze rekening met de gevolgen van ten onrechte gegeven onvoldoendes, en ten onrechte gegeven voldoende. Het is echter duidelijk dat al naar gelang de beoordeling van de ernst van die gevolgen voor de betrokken studenten, men de afstestgrens naar beneden of naar boven moet schuiven. Daarbij kan direct opgemerkt worden dat het moeilijk is zich onderwijs situaties voor te stellen waar het 'erger' is iemand ten onrechte een voldoende te geven, dan ten onrechte een onvoldoende. De ongewenste gevolgen van ten onrechte gegeven voldoende zijn moeilijk of niet aantoonbaar. Ten onrechte gegeven onvoldoendes leiden direct tot aanzienlijke verliezen in studietempo, studiemotivatie, en in veel gevallen ook de financiële positie van de student. omdat de kernitemmethode met deze gegevens geen rekening houdt, zal de docent of de staf zélf de nodige correctie op de cesuur moeten aanbrengen.

Ook deze correctie zou achteraf, na berekening van de aftestgrens dus, aangebracht kunnen worden, en zou eventueel ook ingebouwd kunnen worden in de regel die gehanteerd wordt om uit de verschillende door de docenten gekozen pwaarden de voor de verdere berekening te hanteren pwaarde te kiezen.

Vraag die we nu moeten beantwoorden is dan ook op welke wijze uit de verschillende resultaten een keuze gedaan wordt: hanteren we voorlopig de mediaan, en corrigeren we de te berekenen aftestgrens later wel? Of zeggen we op dit moment al: laten we een andere waarde dan de mediaan kiezen, rekening houdend met de genoemde argumenten 1) en 2) (gesteld dat we daar ons in kunnen vinden). Laten we dat nu overleggen, en een, althans voorlopige, beslissing nemen.

Na toepassing van de gekozen strategie op de pwaarden die de staf gekozen blijkt te hebben, beschikken we over een fictieve pwaarde die als uitgangspunt voor de verdere berekeningen dient. In plaats van te rekenen, kunnen we ook gebruik maken van het nomogram dat Van Naerssen in zijn artikel geeft (Psychometrische aspecten van de kernitem methode. Nederlands tijdschrift voor de psychologie, 1974, 29, 421-430). Om dat te kunnen doen moeten we ook nog weten welke waarde voor r_{bis} we in moeten vullen. Daarvoor kunnen we nemen de r_{bis} van de toetsvraag die in de voorgaande procedure als 'kernitem' gekozen is (dan wel de toetsvraag die er qua p-waarde het dichtst bij ligt). De aftestgrens valt dan uit de figuur af te lezen. Het 'passen' van deze aftestgrens op de gegevens uit de toetsing verkregen levert dan geen nieuwe problemen meer op (zie eventueel De Groot: De kernitem methode voor de bepaling van de cesuur voldoende/onvoldoende. Paedagogische studiën, 1964, 41, 415-440).

cowo

ben wjlbrink 11 juni 1976.

Centrum voor Onderzoek van het Wetenschappelijk Onderwijs

voor: beginselen privaatrecht.

Proef, bedoeld als voorbereiding op de aanwijzing van kernitems.

INSTRUCTIE.

Kernitem methoden veronderstellen dat de docent belangrijke

eigenschappen van toetsvragen, zoals de moeilijkheid, goed kan schatten. Ook de methode van Van Naerssen die we gaan toepassen, veronderstelt dat de docent de moeilijkheid van de vragen kan beoordelen, en daarom lijkt het van belang een indruk te krijgen in hoeverre deze veronderstelling juist is. Dat kan onder andere door docenten te vragen een aantal toetsvragen te ordenen op moeilijkheid, en de resultaten te vergelijken met de p-waarden verkregen bij het tentamen. We gaan dan als volgt te werk:

U krijgt een aantal vragen die onderdeel uitmaken van het laatste tentamen. Over die vragen zijn een aantal gegevens bekend, die na afloop van de proef bekend gemaakt zullen worden. Probeert u niet om de p-waarden voor bepaalde vragen, die u misschien op de computer output gezien hebt, te herinneren.

Leg de gegeven vragen in volgorde van wat u denkt dat de makkelijkste vraag is, naar wat u denkt dat de moeilijkste vraag is.

Hebt u dat gedaan, nummer dan de vragen in de volgorde van makkelijk naar moeilijk als volgt: 1, 2, 3, enz. Leg dan de vragen op elkaar en lever ze in.

Gebruik niet teveel bedenktijd. Wacht met het stellen van vragen tot iedereen klaar is (tenzij u vragen over de instructie hebt). Laat u niet beïnvloeden door de wijze waarop uw collega's met de vragen schuifelen: het gaat om uw eigen indruk, niet beïnvloed door wat anderen er van schijnen te denken.

COWO

ben wilbrink 11 juni 1976.

Centrum voor Onderzoek van het Wetenschappelijk Onderwijs

voor: beginselen privaatrecht.

Bespreking van de 'volgorde' proef.

De proef is bedoeld om een voorlopige indruk te krijgen in hoeverre verschillende docenten het met elkaar en met de werkelijkheid eens zijn over de volgorde in moeilijkheid (p-waarden) van een aantal al afgenomen toetsvragen.

De resultaten zijn van belang voor het vertrouwen dat we kunnen stellen in de kernitem methode, ook de methode in de variant van Van Naerssen.

Immers, wanneer er belangrijke verschillen geconstateerd worden, betekent dat dat de student in forse mate aan het toeval is overgeleverd waar het gaat om de beslissing of zijn score tot de voldoende, dan wel tot de onvoldoenden gerekend zal worden. Op een aantal andere belangrijke veronderstellingen van kernitem methoden zullen we in het discussie stuk nog ingaan, voor wat deze proef betreft gaat het alleen om de vraag of docenten een goede indruk van de verschillen in moeilijkheid van de tentamen vragen hebben.

Gevraagd werd om de volgorde van geschatte moeilijkheid aan te geven, en niet om de p waarden zélf te schatten. De reden is dat op deze manier de proef sneller en eenvoudiger uit te voeren is, en ook dat de rangordening vrijwel evenveel informatie bevat als de rijtjes met geschatte pwaarden zouden opleveren. Dit soort schattingen zijn immers dermate ruw, dat we niet mogen verwachten dat veel meer dan alleen de volgorde nog redelijk te schatten is. Bovendien bevrijden we ons van een bij voorbaat misschien wat vruchteloze discussie over de mate waarin de ene collega de vragen gemiddeld een stuk moeilijker beoordeelt dan een andere collega dat doet. Verschillen in dit opzicht zullen overigens bij de uitvoering van de kernitem methode in de variant Van Naerssen nog aan de orde komen.

zie omme zijde:

Een aanduiding voor de mate van overeenstemming tussen geschatte volgorde van moeilijkheid en de werkelijk verkregen volgorde van moeilijkheid is bijvoorbeeld het aantal keren dat naastliggende vragen van plaats zouden moeten wisselen om tot de waargenomen volgorde van pwaarden te komen. Een voorbeeld:

De volgorde van vijf toetsvragen in op het tentamen gebleken moeilijkheid (pwaarden) is 1 - 2 3 4 - 5.

Een docent schatte de volgorde: 3 1 4 5 2.

Er zijn vier paarsgewijze verwisselingen nodig om van de geschatte volorde tot de werkelijke volgorde te komen, als volgt:

(0) 3 1 4 5 2 (geschatte volorde)

(1) 1 3 4 5 2 (3 en 1 verwisseld)

(2) 1 3 4 2 5 (5 en 2

(3) 1 3 2 4 5 (4 en 2

(4) 1 2 3 4 5 (3 en 2

Je zou kunnen zeggen dat de afstand of het verschil tussen de door deze docent geschatte volgorde, en de werkelijke volgorde, vier is.

Wanneer een verschil groot of klein is, valt moeilijk te zeggen, omdat vergelijkingsmateriaal ontbreekt. Het grootste verschil dat met vijf vragen mogelijk is, bedraagt 10 (wanneer de geschatte volgorde precies de omgekeerde van de waargenomen volgorde is). Een verschilscore van twee, misschien drie, lijkt voor vijf vragen redelijk, en zou althans wat de schatting van moeilijkheid van items betreft, aan de methode-Van-Naerssen niets in de weg hoeven leggen.

Hopelijk heeft Ben ondertussen de scores voor de gedane proef berekend.

ben wilbrink 11 juni 1976.

cowo Centrum voor Onderzoek van het Wetenschappelijk Onderwijs voor:
beginselen privaatrecht.

Item analyse toets 12 mei 1976.

Toepassing van de kernitem methode Van Naerssen vraagt de beschikbaarheid van de biseriële correlaties van items met test score. Omdat het computer programma slechts de puntbiseriële correlaties berekent (de 'r-at' waarden voor de goede alternatieven), zullen we met de hand de biseriële correlaties moeten uitrekenen. Daartoe maken we gebruik van een tabel gegeven door Guilford (Fundamental statistics in psychology and education, 1965, tabel G, kolom D), waarin de factor gegeven wordt waarmee de puntbiseriële correlatie vermenigvuldigd moet worden om de biseriële correlatie te verkrijgen. Die vermenigvuldigingsfactor is afhankelijk van de p-waarde van het item, en wel de voor raden gecorrigeerde:

$$P_c = (A_p - 1) / (A - 1) \quad A = \text{aantal alternatieven}$$

(In principe mogen geen biseriële correlaties berekend worden voor vragen die door raden goed gemaakt kunnen worden, zie Lord & Novick Statistical theories of mental test scores, 1968, blz. 353. Wat dit betreft lijkt Van Naerssen's methode slechts toepasbaar voor open-eind- en niet voor meerkeuzevragen).

In de eerste helft van de tabel de gegevens voor de 'kernitems oude stijl', op de volgende bladzijde de gegevens voor de overige items.

('kern')item	p	p - e	rp-bis	r-bis
--------------	---	-------	--------	-------

2	.739	.652	.128	.165
10	.525	.367	.292	.373
12	.521	.361	.474	.608
18	.500	.333	.458	.594
24	.676	.568	.492	.620
26	.327	.103	.304	.520
27	.454	.272	.498	.669
29	.792	.723	.191	.257
33	.504	.339	.378	.489
43	.412	.216	.329	.460
46	.560	.413	.244	.309

overige items p p - e r-pbis r-bis

1	.852	.803	.422	.603
3	.919	.892	.256	.426
4	.542	.389	.290	.369
5	.746	.661	.293	.379
6	.722	.629	.331	.423
8	.588	.451	.303	.381
9	.518	.357	.108	.139
11	.567	.423	.294	.371
13	.694	.592	.251	.318
14	.493	.324	.349	.455
16	.560	.413	.347	.439
17	.644	.525	.374	.469
19	.810	.747	.319	.435
20	.813	.751	.456	.622
22	.665	.553	.358	.450
25	.518	.357	.249	.319
28	.824	.765	.396	.546
30	.447	.2e3	.265	.358
31	.750	.667	.256	.332
32	.746	.661	.285	.369
34	.394	.192	.049	.071
35	.320	.093	.358	.631
36	.806	.741	.144	.195
37	.880	.840	.320	.482

38	.937	.916	.164	.299
39	.673	.564	.380	.478
40	.549	.399	.312	.396
41	.785	.713	.496	.6.58
42	.560	.413	.351	.444
44	.299	.065	.211	.401
47	.246	negatief	.349	--
48	.556	.408	.417	.528
49	.239	negatief	.237	--
50	.489	.319	.341	.445

De items met de hoogste biseriële correlaties in volgorde:

item	r bis	pc	p	'kern'?	r-pbis
27	.669	.272	.454	ja	.498
41	.658	.713	.785	nee	.496
35	.631	.093	.320	nee	.358
20	.622	.751	.813	nee	.456
24	.620	.568	.676	ja	.492
12	.608	.361	.521	ja	.474
1	.603	.803	.853	nee	.422
18	.594	.333	.500	ja	.458
28	.546	.765	.824	nee	.396
48	.528	.408	.556	nee	.417
26	.520	.272	.454	ja	.498
33	.489	.339	.504	ja	.378
37	.482	.840	.880	nee	.320
39	.478	.564	.673	nee	.380

Opmerkingen: item 35 lijkt, ondanks zijn hoge biseriële correlatie, niet geschikt om in de kernitem methode procedure te laten meespelen, gezien het heel erg kleine aantal studenten dat dit item zonder er bij te hoeven raden, heeft kunnen goed maken.

Voor items 47 en 49 zijn geen biseriële correlaties berekend, omdat de gecorrigeerde p-waarde op een negatief getal uitkwam (een mogelijkheid die bij dit soort correctieformules inderdaad bestaat).

De gemiddelde pwaarde (ongecorrigeerd) van de kernitems zoals op blz. 1 getabelleerd, is $6.010/11 = .546$. De gemiddelde gecorrigeerde pwaarde

voor kernitems is $4.347/11 = .395$. De conclusie is heel erg duidelijk: de p-waarden voor deze kernitems zijn te laag om de items als kernitems te kunnen handhaven. De vragen zijn te moeilijk.

De gemiddelde p-waarde (ongecorrigeerd) voor alle getabelleerde vragen, dus inclusief de 'kernitems', is $27.161/45 = .604$.

De gemiddelde gecorrigeerde p-waarde voor alle items bij elkaar is

(items 47 en 49 niet daarin meegeteld): $21.233/43 = .494$.