50 JAAR C¿TO, EEN HALVE EEUW WISKUNDE-EXAMENS?

DEEL 4

Ruud Stolwijk

In september 2018 bestond Cito 50 jaar. In het vierde artikel over de rol die Cito door de jaren heen heeft gespeeld bij de wiskunde-examens, bespreekt Ruud Stolwijk de psychometrische aspecten van de examens en wat we daar als docent aan kunnen hebben.

Inleiding

Iedere docent die een toets afneemt in zijn of haar klas, kijkt achteraf hoe deze toets gemaakt is, en past daar eventueel de normering op aan. Om dit te kunnen doen, moet je in ieder geval beschikken over iets als de gemiddelde totale score. Maar soms blijken één of meer vragen uit de toets qua resultaat behoorlijk af te wijken (of tegen te vallen), en daarom is een gemiddelde score per opgave of zelfs per vraag zeker ook een handig hulpmiddel bij het vaststellen van de normering van de toets. Bij examens is dat feitelijk niet anders, met daarbij meteen de opmerking dat waar een enkele klas nog wel eens 'bij toeval' afwijkt ('dit is gewoon niet zo'n goede klas', of 'deze klas is wel heel erg getalenteerd'), dit voor examens eigenlijk niet geldt. Het is immers niet aannemelijk dat een hele leerlingpopulatie, een hele jaargang van duizenden kandidaten, qua wiskundige vaardigheid 'zomaar ineens' zal afwijken van het gebruikelijke. Kortom: de resultaten van een examenpopulatie zeggen wel degelijk iets over de moeilijkheid van het betreffende examen in vergelijking met examens van voorgaande jaren. En deze resultaten zijn openbaar! In dit artikel wordt verteld waar die resultaten te vinden zijn, hoe ze te lezen en te interpreteren en wat je als (wiskunde)docent aan deze informatie kunt hebben.
Als voorbeeld in dit artikel gebruiken we steeds hetzelfde examen: havo wiskunde B 2018 1e tijdvak. Het is handig om bij het lezen van dit artikel een laptop, IPad of smartphone paraat te hebben. Vanaf de Cito-homepage is die route als volgt: voortgezet onderwijs - centrale examens - Alles voor de centrale examens - Examenmateriaal om te oefenen - Examens 2016-2018 en dan het gewenste niveau, jaar en tijdvak kiezen, zie figuur 1.

Van elk examen zijn dus beschikbaar de opgaven (Opg.), of er een gesproken versie van is (Spr.), het correctievoorschrift (CV), en de toets- en itemanalyse (TIA). Verder zijn er eventueel nog errata in de opgaven (Errata opg.),

figuur 1

aanvullingen op het correctievoorschrift (Aanv. CV) en door Cito-medewerkers verzorgde publicaties over het examen (Publ.). In het vervolg van dit artikel zullen we ons richten op de toets- en itemanalyse, kortweg de TIA.

Toets- en itemanalyse

Als we de TIA aanklikken, verschijnt (voor havo wiskunde B 2018 1e tijdvak) een Word-document van 55 pagina's. Dat lijkt erg veel, maar voor de meest interessante informatie kunnen we ons beperken tot de eerste bladzijde. De rest wijst zich daarna grotendeels vanzelf en dat laten we dan ook aan de lezer.
De eerste bladzijde ziet er gedeeltelijk uit zoals te zien is in figuur 2.

In deze tabel zijn de gegevens verwerkt van alle in WOLF ingevoerde kandidaten. We bekijken de tabel van links naar rechts en zien dat dit examen bestaat uit 18 vragen, verdeeld over acht opgaven: elke eerste vraag van een opgave is gemarkeerd met een #. Het gewicht van elke vraag (de kolom Gew.) is gelijk - los van het te behalen aantal scorepunten natuurlijk. Daarna zien we twee kolommen (O/D en Missing) waarin te zien is welk percentage van de leerlingen respectievelijk het aantal leerlingen dat de betreffende vraag heeft overgeslagen.

Item Label	Item nr.	Gew.	O/D	Mis-sing	Max	Gem	P	Sd	RSK	Rit	Rir	AR
1#	1	1	0	6	3	2,42	81	1,02	0,34	36	29	78
2	2	1	0	54	6	5,05	84	1,45	0,24	49	39	77
3	3	1	5	630	3	0,53	18	0,84	0,28	35	29	78
4#	4	1	1	163	6	4,74	79	2,02	0,34	56	43	77
5	5	1	2	297	4	3,02	75	1,34	0,34	46	37	78
6#	6	1	2	279	3	2,25	75	1,17	0,39	28	19	79
7	7	1	1	97	3	2,39	80	0,94	0,31	33	26	78
8	8	1	5	595	5	3,55	71	1,89	0,38	52	39	77
9#	9	1	2	191	6	3,04	51	1,81	0,30	60	50	76
10#	10	1	1	128	3	2,21	74	1,02	0,34	41	34	78
11	11	1	4	546	3	1,62	54	1,30	0,43	38	28	78
12	12	1	2	271	5	3,60	72	1,54	0,31	47	36	78
13#	13	1	2	265	5	2,93	59	1,97	0,39	61	50	77
14	14	1	5	641	6	1,70	28	1,90	0,32	45	31	78
15#	15	1	3	406	3	1,93	64	1,28	0,43	44	35	78
16	16	1	12	1484	4	1,86	47	1,56	0,39	51	41	77
17#	17	1	6	701	4	1,78	45	1,49	0,37	53	44	77
18	18	1	7	892	5	2,31	46	1,50	0,30	53	43	77

figuur 2

Zo is te zien dat vraag 16 door maar liefst 1484 kandidaten (12%) is overgeslagen. Deze informatie is van belang bij het bepalen van de normering, want met name als er aan het eind veel kandidaten vragen hebben overgeslagen, dan zou dit een teken van tijdnood kunnen zijn. Als we verder kijken, zien we kolommen met de maximumscore die voor de vraag behaald kan worden (Max), de behaalde gemiddelde score door de kandidaten (Gem), en de zogenoemde p-waarde (P): het percentage van de scorepunten die er gemiddeld voor de vraag werden behaald. Uiteraard geldt hierbij dat P = 100 × Gem / Max. Dan volgen twee kolommen die te maken hebben met de spreiding van de scores binnen de betreffende vraag (Sd en RSK). De eerste geeft de standaardafwijking van de behaalde scores, de relatieve standaardafwijking (RSK) maakt de standaardafwijkingen van de verschillende vragen onderling vergelijkbaar door deze steeds te delen door de maximumscore (dus RSK = SD / Max). De naam RSK is overigens ontleend aan oud-Cito-medewerker Wiel Knops (Relatieve Standaardafwijking Knops).

Rit en Rir

De kolommen Rit en Rir verdienen wat extra aandacht. De eerste, Rit, is de correlatie (weergegeven in een getal tussen –100 en 100) tussen de score van de vraag en de totaalscore van de toets inclusief de score op de vraag zelf (het hele examen dus). De Rit van een vraag geeft dus aan in hoeverre de betreffende vraag representatief is voor de toets als geheel. Een hoge Rit-waarde betekent dat veelal de goede kandidaten hoge scores op deze vraag wisten te behalen, en minder goede kandidaten minder hoge scores. Maar wat is nu een hoge Rit-waarde? Gangbaar is dat een Rit-waarde onder de 20 als 'laag' wordt gezien: de betreffende vraag levert dan statistisch gezien eigenlijk niet of nauwelijks een bijdrage aan het examen als geheel. En vragen met negatieve Rit-waarden zijn zelfs uit den boze, want die zouden aangeven dat een kandidaat die een goed examen maakt op een dergelijke vraag juist slecht zou scoren. En dan meet je met zo'n vraag wel iets heel merkwaardigs - bijvoorbeeld iets wat in de rest van de toets niet wordt gemeten. Verder zijn Rit-waarden vanaf 30 'goed' te noemen, en vanaf 40 zelfs 'zeer goed'.

De Rir-waarde is vrijwel hetzelfde als de Rit-waarde, zij het dat de Rir-waarde de correlatie weergeeft tussen de score van de vraag en de totaalscore van de toets minus de score op de vraag zelf - de correlatie met de rest van de toets dus. De Rit zal dus nooit lager kunnen zijn dan de Rir. In de tabel is te zien dat wat betreft de Rit- en Rir-waarden het examen havo wiskunde B 2018 1e tijdvak louter vragen bevatte die voor een duidelijk onderscheid zorgden tussen vaardige en minder vaardige kandidaten. En dat is uiteraard de bedoeling van een examen.

Betrouwbaarheid

Het boven de kolommen vermelde 'Gewogen' geeft overigens aan dat er bij de berekening rekening is gehouden met alleen die kandidaten die de vraag ook echt gemaakt hebben - dus de Missing-kandidaten zijn daarbij weggelaten. De laatste kolom in de tabel (AR) betreft de betrouwbaarheid. AR staat hierbij voor *alpha-rest* en dit is de betrouwbaarheid (vermenigvuldigd met 100) van het examen minus het betreffende item. Maar wat is dan precies betrouwbaarheid? Betrouwbaarheid is in het geval van toetsen en examens de mate waarin de scores van de kandidaten consistent, nauwkeurig en reproduceerbaar zijn. Dat betekent dat het zo zou moeten zijn dat als dezelfde kandidaat met dezelfde voorkennis onder dezelfde omstandigheden dezelfde toets tweemaal zou maken, de kandidaat ook tweemaal dezelfde score zou moeten halen. Dat dit echter louter theorie is moge duidelijk zijn... Daarom zijn er statistische methoden ontwikkeld om deze theorie zo dicht mogelijk te benaderen en de betrouwbaarheid zo goed mogelijk te schatten. Een van de meest gebruikte schattingen voor de betrouwbaarheid van een toets is de volgende:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\left(S_i^2\right)}{\left(S_x\right)^2}\right)$$

Hierbij is α de betrouwbaarheid, k het aantal vragen in het examen, S_i^2 de variantie (ofwel het kwadraat van de standaardafwijking) van de scores bij vraag i en S_x^2 de variantie van de totaalscores. Met behulp van deze formule is het (zeker in Excel) overigens niet zo heel moeilijk om (een schatting voor) de betrouwbaarheid van je eigen toets te berekenen.

Je kunt deze formule ook anders schrijven:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\left(S_i^2\right)}{\sum_{i=1}^{k}\left(Rit \cdot S_i\right)}\right)$$

Hierbij is Rit de correlatie tussen de scores van vraag i en de scores op de totale toets. Aan deze tweede formule voor α is te zien dat de eerder besproken Rit-waarden inderdaad verband houden met de betrouwbaarheid: hoe hoger de Rit, hoe hoger de betrouwbaarheid. Verder is ook in te zien dat, als het aantal vragen groter wordt (mits natuurlijk vragen die toetsen wat je ook echt in het examen wil toetsen), de betrouwbaarheid dan ook groter zal zijn.

Wat nu precies de minimale betrouwbaarheid van een toets of examen moet zijn, hangt onder andere af van het aspect of de toets of het examen het enige middel is om de vaardigheid van de kandidaat te bepalen. Bij examens is het zo dat het examen per vak voor 50% deze vaardigheid bepaalt (er is immers ook nog het schoolexamen); in een dergelijke situatie worden in de literatuur betrouwbaarheden van minstens 0,65 genoemd. De α van het examen dat we hier steeds bekijken, havo B 2018-1, is met 0,79 dan ook 'heel behoorlijk' te noemen.
De theorie gaat ervan uit dat een toetsscore altijd is opgebouwd uit twee delen: een ware score en een meetfout. De ware score wordt daarbij bepaald door de vaardigheid van de kandidaat in combinatie met de moeilijkheid van de toets. De toetsscore hangt echter ook af van 'de vorm van de dag', de toevallige verdeling van de vragen over de stof (en net dat éne onderwerp waar de kandidaat niet zo goed in is zat er wat vaker in), hoofdpijn, of de net iets te hoge temperatuur in de examenzaal. Deze laatste aspecten vertroebelen in feite het zicht op de werkelijke vaardigheid van de kandidaat en vormen gezamenlijk de meetfout. Als je nu een kandidaat meermalen dezelfde of een vergelijkbare toets zou laten maken, dan zou het effect van die meetfout wel uitgefilterd kunnen worden. Maar dat is nu eenmaal in de praktijk niet mogelijk. Daarom is het van belang om de meetfout zo klein mogelijk te krijgen, met als probleem dat de meetfout moeilijk te meten valt... we werken immers louter met toetsscores. Statistisch kan echter uit de betrouwbaarheid (waarvan hierboven is aangegeven hoe deze uit de toetsresultaten berekend kan worden) worden bepaald hoe groot de bijdrage van de meetfout aan de variantie van de toetsscores is... maar dat voert in het kader van dit artikel wel wat te ver.

Tot slot

Wat heb je hier als docent nu allemaal aan - los van mogelijk wat meer begrip voor de analyse van examens op zich? Om te beginnen kun je natuurlijk terugkijken op de resultaten van je eigen examenklas met behulp van de door Cito geleverde groepsrapportage. Maar dat is achteraf... Misschien is het grootste voordeel wel dat je, als je oude examenopgaven en -vragen gebruikt voor eigen toetsen, je van tevoren te weten kunt komen hoe moeilijk deze opgaven en vragen zijn door in de betreffende TIA te kijken. En dat vragen met hoge Rit- en Rir-waarden leerlingen met goede en minder goede vaardigheden goed weten te onderscheiden. En dat kan - mits de leerlingen de (immers openbare) opgaven niet al geoefend hebben natuurlijk - een hulpmiddel zijn om te zien of jouw klas wel of niet op het gewenste niveau zit.

Noot

[1] *https://www.cito.nl/onderwijs/voortgezet-onderwijs/centrale-examens-voortgezet-onderwijs/examenmateriaal-om-te-oefenen/havo-2018/havo-2018-tv1*

Over de auteur

Ruud Stolwijk is toetsdeskundige wiskunde bij Cito.
E-mailadres: ruud.stolwijk@cito.nl