Algoritmes in de Adaptieve voortgangstoets Geneeskunde iVTG

Jeroen Donkers, Universiteit Maastricht, 16 augustus 2024

Vanaf september 2022 maakt de iVTG gebruik van een computer-adaptieve voortgangstoets, zoals aangeboden in het online toets-systeem Testvision[1] (van Wijk et al., 2024). In een adaptieve toets zijn er 4 plekken waarin computer-algoritmes worden gebruikt. We zullen deze één voor één bespreken. De adaptieve voortgangstoets zoals in Testvision geïmplementeerd is gebaseerd op het 1-parameter IRT model volgens Rasch (Rasch 1960/1980, Wright & Stone, 1990, Howard Wainer, 1990, Linden & Glas, 2000).

Zonder hier diep op IRT en het Rasch model te willen ingaan, is het voor dit stuk goed te weten dat dit model uitgaat van het volgende: iedere student heeft een *ability* die we graag willen meten, het actuele kennisniveau. Iedere vraag heeft een moeilijkheid die we in de studentpopulatie kunnen meten en die vast is en niet afhankelijk van andere vragen. De moeilijkheid (beta) en ability (theta) staan in het model op dezelfde schaal. De kans dat een student een vraag goed beantwoordt, hangt in het model alleen af van het verschil tussen beta en theta. Zijn theta en beta gelijk, dan is deze kans precies 50%.

Algoritme 1: Bepaling van de ability (theta) van een student

Zowel tijdens de adaptieve toets, na het beantwoorden van elke vraag, als aan het eind van de toets wordt de ability (theta) van een student geschat. Deze schatting wordt gemaakt op basis van de vragen die een student vanaf het begin van de toets tot op dat moment in de toets heeft gezien, uitgezonderd de pretest vragen. Hierbij wordt de moeilijkheid van iedere vraag (de beta) zoals die tijdens de kalibratie volgens het Rasch model is bepaald en bij de vraag in Testvision is opgeslagen, gecombineerd met de score (0 of 1) die de student op die vraag haalt. We gebruiken hiervoor de WLE-rekenmethode van Thomas Warm (1989). Deze methode schat de theta-waarde bij vastgestelde beta-waarden en corrigeert daarbij voor statistische bias. De score-berekening is in Testvision geïmplementeerd en voorafgaand aan de pilot-studie in mei 2022 is een controle uitgevoerd waarbij de resultaten van de Testvision-berekening zijn vergeleken met een implementatie van de WLE-rekenroutine in R[2]. De verschillen tussen de beide methoden vielen binnen de afrondingsmarge, dus we kunnen aannemen dat Testvision de WLE-rekenroutine correct heeft geïmplementeerd. De theta-waarde die de WLE-rekenroutine produceert is overigens begrenst tot minus 5 en plus 5, deze range is meer dan ruim genoeg om alle mogelijke scores te bevatten.

Deze rekenroutine kijkt alleen naar de verzameling geziene vragen met hun moeilijkheid en de score van de student op deze vragen. De volgorde van de vragen heeft geen invloed op de uitkomst.

Algoritme 2: Bepaling van de toetsprestatie en uitslag

Aan het einde van de toets wordt dezelfde WLE-rekenroutine gebruikt om de eind-theta van de toets voor de student te berekenen. Deze theta-score wordt vervolgens lineair

[1] www.testvision.nl
[2] www.R-project.org

getransformeerd naar de Toetsprestatie die geldt als toetsresultaat. De formule hiervoor is:

Toetsprestatie = target_mean + target_sd * ((theta – mean_pop) / sd_pop)

Testvision rekent deze transformatie uit en rondt af op 2 cijfers achter de komma. De 4 parameters worden door de iVTG zelf bij de toets in Testvision ingesteld. Momenteel worden de volgende vaste waarden voor de 4 parameters gebruikt:

Target_mean = 35
Target_sd = 15
Mean_pop = 0.0
Sd_pop = 0.35

Deze transformatie is lineair, dat wil zeggen dat de onderlinge volgorde van studenten niet verandert en onderlinge verschillen in gelijke proportie blijven. Ook deze berekening is buiten Tetstvision gecontroleerd en er zijn geen fouten in Testvision gevonden.

De uitslag van een student voor de toets (uitgedrukt in Onvoldoende, Voldoende en Goed) wordt bepaald op basis van door de iVTG van tevoren vastgestelde cesuren per meetmoment. Iedere student wordt bij een toets, op basis van o.a. studiejaar door de eigen opleiding ingedeeld in een meetmoment, van 1 tot 24, met oplopende cesuren. Testvision bepaalt dan op basis van de toetsprestatie het meetmoment van de student en de vooraf ingestelde cesuren de uitslag voor de student.

Algoritme 3: Het adaptief algoritme in Testvision

Het adaptief algoritme maakt gebruik van een bank aan vragen in Testvision waarbij iedere vraag van een moeilijkheid is voorzien. (We noemen dit de “actieve bank”). Voorafgaand aan iedere toetsperiode wordt deze bank geactualiseerd.

Het adaptief algoritme in Testvision maakt gebruik van maximale Fischer-informatie om de beste volgende vraag te kiezen (Wainer, 1990). In het Rasch model betekent dit dat de beste volgende vraag een vraag is die qua moeilijkheid zo dicht mogelijk ligt bij de op dat moment geschatte ability van de student (Wanneer meerdere vragen even dicht bij de ability liggen wordt er één willekeurig gekozen.)

Het adaptief algoritme heeft alleen de lengte van de toets als stopcriterium. Het algoritme stopt pas als een vooraf ingesteld aantal vragen is gesteld, in het geval van de iVTG na 120 vragen.

Om over-exposure van vragen te voorkomen en om studenten een eerlijke toets te geven, krijgen studenten vragen die ze in een toets zien daarna 2 jaar lang niet meer te zien (‘uitsluiten recente items”). Bovendien krijgen studenten in één toets geen combinaties van vragen te zien die als “enemy” van elkaar zijn gemarkeerd in Testvision.

Omdat de voortgangstoets geneeskunde een brede spreiding aan onderwerpen heeft, zorgt het algoritme ervoor dat de vragen die een student krijgt een vooraf ingestelde toetsmatrijs volgt. Alle vragen in de actieve bank zijn voorzien van een categorie en een discipline. De toetsmatrijs die het algoritme gebruikt bevat alleen disciplines. De verdeling over categorieën die een student aangeboden krijgt wordt bepaald door de beschikbare spreiding over categorieën in de actieve bank.

Het algoritme functioneert dan als volgt:

Startvragen: De eerste 6 vragen die een student te zien krijgt worden volledig random uit de bank gekozen (met inachtneming van enemy items, uitsluiten recente items en de toetsmatrijs), waarbij de gemiddelde moeilijkheid van deze 6 items voor *alle* studenten gelijk is. Op dit moment is die vastgesteld op 0. Studenten worden dus niet benadeeld of bevoordeeld door een random keuze van te moeilijke of te makkelijke vragen. Bij de selectie wordt op geen enkele manier gekeken naar het verleden van de student (behalve dan uitsluiting van eerder geziene vragen) of het jaar c.q. meetmoment van de student.

Adaptieve fase: na de 6 startvragen schat het algoritme voor het eerst de theta van de student mbv de WLE-rekenroutine op basis van de antwoorden op en moeilijkheden van de 6 startvragen. Hierbij wordt ook meteen de standaard-meetfout (SEM) bepaald. Beide worden in de database van Testvision opgeslagen voor verdere analyse. Op basis van deze theta en alle hierboven geldende regels wordt de volgende vraag uit de actieve bank geselecteerd en aan de student getoond. Hierna volgt een nieuwe theta-schatting, nu op basis van 7 vragen. Dit gaat door tot het aantal van 120 vragen is gehaald. Dan volgt de laatste theta-schatting op basis van 120 vragen, die de eindscore van de toets vormt.

De adaptieve fase maakt op geen enkele manier onderscheid tussen studenten. Alleen de antwoorden van de student *tijdens deze toets* worden in het algoritme gebruikt. Er wordt niet gekeken naar hoelang iemand over een vraag nadenkt, of iemand in jaar 1 of jaar 6 zit, of iemand al 3 keer een onvoldoende heeft gehaald, of met welk meetmoment de student deelneemt, etc.

Pretestvragen: Tussen de 120 meetellende vragen krijgt de student 15 vragen te zien en te beantwoorden die niet meetellen in de score en geen invloed hebben op het adaptief algoritme. Dit zijn pretest vragen die we hieronder verder zullen bespreken.

Algoritme 4: Bepaling van de moeilijkheid van vragen (kalibratie)

Omdat het adaptief algoritme wordt gestuurd door de moeilijkheid van vragen, is het belangrijk te beschrijven hoe deze bepaald wordt.

Bij de start van de adaptieve toets in 2022 is gebruik gemaakt van vragen die vanaf september 2007 in de voortgangstoets geneeskunde aan bod zijn gekomen. Al deze vragen zijn uitgebreid gereviewed door de iVTG voordat ze in de actieve bank zijn opgenomen. Alleen vragen die door deze review komen en dus nog relevant en van goede inhoudelijke en psychometrische kwaliteit waren zijn opgenomen. De moeilijkheden van deze vragen zijn geschat op basis van de antwoorden die studenten tijdens deze papieren toetsen hebben gegeven. Hierbij is eerst elke papieren toets apart gekalibreerd volgens een aangepast Rasch model dat rekening houdt met het feit dat de papieren toets een vraagtekenoptie kende. Daarna zijn de moeilijkheden van alle 40 papieren toetsen aan elkaar gelijkgesteld, waarbij voor schommelingen tussen de toetsen is gecorrigeerd. Van deze kalibratie is een uitgebreide rapportage geschreven waarin alle kwaliteitskenmerken zijn beschreven.

Nieuwe vragen voor de adaptieve toets worden, net als voorheen, geschreven door auteurs binnen de deelnemende UMCs. De nieuwe vragen ondergaan eerst een lokale en peer-review voordat ze opgenomen worden als kandidaat voor de adaptieve toets. Het is natuurlijk belangrijk om ook van deze nieuwe vragen de moeilijkheid te bepalen. Men kan de moeilijkheid meten door de vragen aan een testpopulatie voor te leggen. Meestal

is dat een ingewikkelde en kostbare operatie. Het feit dat de adaptieve toets op het Rasch model is gebaseerd geeft ons de mogelijkheid om dit in de toets zelf te integreren. Voor het Rasch model geldt dat je de moeilijkheid het beste kunt schatten als je een vraag aan een zo breed mogelijk publiek voorlegt (Stocking, 1988, Ban, 2001, Verschoor, 2019). Door de populaties volledig random over alle studenten te kiezen krijgen we een representatieve sample over alle UMC's, meetmomenten en niveaus van studenten.

We bereiken dit als volgt. Iedere toets worden 210 vragen ge-pretest. We delen deze vragen in Testvision in 14 "bakjes" in en voorzien de vragen van het label "pretest vraag". Het adaptief algoritme van Testvison kiest bij de start van de toets voor iedere student volledig willekeurig een bakje en zorgt ervoor dat de student de 15 vragen uit dit bakje te zien krijgt, random tussen de andere vragen in.

Ieder bakje wordt dus door 1/14 van de populatie gezien, in de praktijk zo'n 800 studenten. Na de toets analyseren we de populaties om te zien of ze niet significant verschillen in samenstelling. Dit blijkt tot nu toe steeds zo te zijn. Mocht er wel een significante afwijking zijn, dan moet hiervoor statistisch gecorrigeerd worden.

Binnen ieder bakje wordt nu de moeilijkheid van alle 15 vragen geschat op basis van de antwoorden van de 800 studenten en hun score op de adaptieve toets (Stocking, 1988). Dit gaat in twee rondes: in de eerste ronde worden de kwaliteitsmaten van de Rasch-schatting bepaald, met name voor de studenten. Studenten die te ver buiten de kwaliteitsmaten vallen (bijvoorbeeld omdat ze vragen niet serieus hebben geantwoord) worden uit de populatie verwijderd. Dit gaat meestal om zo'n 2-4 procent van de studenten. Daarna worden de moeilijkheden opnieuw geschat. De outlier-studenten worden in een lijst verzameld om te zien of bepaalde studenten vaker in deze lijst voorkomen (en er dus een bias kan voorkomen). Dit blijkt echter zeer zeldzaam te zijn (29 studenten komen over de afgelopen 2 jaar meer dan 2 keer voor als outlier) en er dus geen reden hier een bias te vermoeden.

De resultaten van deze kalibratie worden ter review aangeboden aan de WiV – pas bij positief besluit worden deze nieuwe vragen opgenomen in de actieve bank.
Bij de Rasch-schatting van de moeilijkheid op basis van bekende theta-scores van studenten wordt eveneens een WLE-rekenroutine gebruikt. De kalibratie van items vindt niet in Testvision plaats maar wordt door de psychometrici van iVTG uitgevoerd in R. Deze procedure is getest en in een publicatie (nog onder review) beschreven. Van elke kalibratieronde wordt een rapport opgesteld.

Pilotstudie

Voordat de adaptieve voortgangstoets in september 2022 in gebruik werd genomen is er in mei 2022 een uitgebreide pilotstudie gehouden waarbij studenten zowel de papieren voortgangstoets als de adaptieve voortgangstoets hebben gemaakt (van Wijk et al., 2024).

Zomerstudies

In de zomerperiode wordt de hele actieve bank aan kwaliteitscontroles blootgesteld. Hierbij worden slecht-functionerende vragen of vragen die in psychometrisch gedrag zijn veranderd sinds de kalibratie uit de bank verwijderd. Dit is ook de periode waarin onderzoek wordt gedaan om te zien op welke wijze de kwaliteit van de toets en het verdere proces verder kan worden verbeterd, op basis van nieuwe wetenschappelijke ontwikkelingen.

Tenslotte

Uit bovenstaande blijkt dat de uitslag van de studenten alleen wordt bepaald door hun eigen prestatie: de antwoorden op de vragen die ze te zien krijgen tijdens de adaptieve voortgangstoets. De algoritmen veroorzaken daarbij geen bias of verschil tussen studenten. De algoritmen die we gebruiken zijn openbaar en al decennialang bekend.

Literatuur

Ban, Jae Chun & Hanson, Bradley & Wang, Tianyou & Yi, Qing & Harris, Deborah. (2001). A Comparative Study of On-line Pretest Item—Calibration/Scaling Methods in Computerized Adaptive Testing. Journal of Educational Measurement. 38. 191 - 212. 10.1111/j.1745-3984.2001.tb01123.x.

Linden, W.J van den & Glas, C.A.W. (Eds) (2000) Computerized Adaptive Testing: Theory and Practice. Kluwer Academic Publishers. ISBN 0-7923-6425-2.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Stocking, M. (1988). Scale drift in on-line calibration. Research Report (pp. 88–28). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x

Verschoor, Angela & Berger, Stéphanie & Moser, Urs & Kleintjes, Frans. (2019). On-the-Fly Calibration in Computerized Adaptive Testing. 10.1007/978-3-030-18480-3_16

Wainer H (ed) (1990). Computer adaptive testing: a primer. LEA publishers, ISBN 0 8058 0636 9.

Warm T.A. (1989) Weighted Likelihood Estimation of Ability in Item Response Theory. Psychometrika, 54, 427-450.

Van Wijk EV, Donkers J, De Laat PCJ, Meiboom AA, Jacobs B, Ravesloot JH, Tio RA, Van Der Vleuten CPM, Langers AMJ, Bremers AJA. (2024). Computer Adaptive vs. Non-adaptive Medical Progress Testing: Feasibility, Test Performance, and Student Experiences. *Perspect. Med. Educ.* 2024 Jul 26; 13(1):406-416. doi: 10.5334/pme.1345. PMID: 39071727; PMCID: PMC11276406.

Wright, B. D., & Stone, M. H. (1999). Measurement Essentials. Wilmington: Wide Range, Inc.