POK Memorandum 2000-1

Algoritmen voor adaptief toetsen

T.J.H.M. Eggen
A.P.J.M. Heuvelmans

Cito

POK Memorandum 2000-1

Algoritmen voor adaptief toetsen

T.J.H.M. Eggen

A.P.J.M. Heuvelmans

Cito
Arnhem, april 2000

Prijs f 15,--

Inhoudsopgave

Voorwoord

Adaptief toetsen, of 'toetsen op maat', mag zich in een toenemende belangstelling verheugen. Kortere toetstijd, betere meetnauwkeurigheid en betere geheimhouding van opgaven zijn vaak gehoorde voordelen. Het basisconcept is van een charmante eenvoud: nadat een kandidaat een of meer opgaven heeft beantwoord, wordt diens vaardigheid geschat. Op basis van die schatting wordt dan uit een opgavenvoorraad een volgend item geselecteerd dat zo goed mogelijk past bij dat vaardigheidsniveau. Deze aanpak impliceert toetsen per computer. De toetsconstructie en -afname geschiedt *in real time.*

De afgelopen jaren heeft het Cito veel ervaring opgedaan met het bouwen van adaptieve toetsen. Dit memorandum behandelt met name de bij het Cito ontwikkelde en beschikbare methoden voor vaardigheidsschatting en itemselectie-algoritmen (en daarbij behorende overwegingen en keuzen), die ingezet kunnen worden bij adaptief toetsen. Het ontwerpen van een adaptieve toets is niet voltooid voordat het itemmateriaal en het schattings- en itemselectiemechanisme zijn beproefd in een of meer simulatiestudies. In het laatste deel van dit memorandum komen kort de voorzieningen (software programma's) aan de orde die zijn ontwikkeld om dergelijke simulaties te kunnen uitvoeren.

Drs. A.P.J.M. Heuvelmans
Projectleider CBT tools

1 Inleiding

In dit memorandum heeft adaptief toetsen de betekenis dat individuele kandidaten met behulp van de computer getoetst worden en dat de toetsconstructie plaatsvindt tijdens de afname: mede op basis van de door een kandidaat getoonde prestatie worden passende opgaven uit een beschikbare geschaalde itembank geselecteerd en afgenomen.
Een algoritme voor een adaptieve toets bevat de voorschriften volgens welke de start, de voortgang, de beëindiging en de rapportage over de toetsafname van een kandidaat plaatsvinden. Voor adaptief toetsen zijn veel (varianten) van algoritmen beschikbaar. In deze notitie zullen we een overzicht geven van die algoritmen die met name op het Cito bestudeerd zijn, en in computerprogrammatuur geïmplementeerd zijn of zullen worden.
Allereerst zullen in paragraaf 2 enkele voorwaarden en beperkingen voor het adaptief toetsen worden besproken, waarna eerst globaal in paragraaf 3 en dan in detail (in paragraaf 4 en 5) de fasen (in algoritmen) van adaptieve toetsen worden behandeld. Daarna worden die men zou kunnen hebben bij het kiezen uit mogelijke varianten aan de orde gesteld. Tenslotte geeft paragraaf 7 enige informatie over de software tools die beschikbaar zijn voor het instellen en beproeven van een gekozen algoritme.

2 Voorwaarden en beperkingen bij adaptief toetsen

De computer

In deze notitie beperken we ons tot computergestuurd adaptief toetsen (CAT). Vormen van adaptief toetsen die we met pen en papier of mondeling zouden kunnen afnemen worden hier buiten beschouwing gelaten. Dit betekent dat voor adaptief toetsen aan de misschien triviaal lijkende voorwaarde moet worden voldaan, dat er een geautomatiseerde omgeving -hardware, software- is, die voldoende kracht heeft om er adaptief mee te kunnen toetsen. Alhoewel de huidige beschikbare apparatuur veel kan en snel is, moet vastgesteld worden dat deze lang niet altijd bij al onze klanten in dezelfde mate aanwezig is. Bij het ontwerpen van de software voor adaptieve toetsen kunnen we er dus niet zonder meer vanuit gaan dat alles kan. Beperkingen zullen er zijn in de aard van de stimuli die kunnen worden aangeboden, maar ook in de toelaatbare responsen van kandidaten op deze stimuli. De responsen moeten automatisch gescoord kunnen worden.

Daardoor hebben de mogelijke responsen vaak een keuzekarakter, of als een kandidaat al een "open" antwoord kan geven zijn de antwoordmogelijkheden (vaak aanvullers, invullers of uitkomsten) beperkt. Daarmee samen gaat een andere beperking van de tot nu toe door het Cito gemaakte adaptieve toetsen. De responsen op een item moeten dichotoom, dat wil zeggen goed of fout, kunnen worden gescoord. Naar de algoritmen voor adaptief toetsen voor items met polytome scoring wordt momenteel wel onderzoek gedaan, maar deze zijn nog niet operationeel. Ten slotte moet worden opgemerkt dat als gevolg van de mogelijkerwijs in de praktijk beperkte rekensnelheid van computers er geen echt rekenintensieve algoritmen in de toetsen kunnen worden gebruikt. Algoritmen die wellicht in sommige omstandigheden evengoed of beter zijn, zullen om die reden in deze notitie niet besproken worden. Bedoeld worden die algoritmen die tijdens de afname de evaluatie van complexe integralen of de oplossing van niet triviale lineaire programmeringproblemen vragen.

De itembank

Adaptieve toetsen veronderstellen de beschikbaarheid van een verzameling items waaruit de toets kan worden samengesteld. Deze items worden gekozen uit een itembank. Een itembank zou men een gestructureerde verzameling items kunnen noemen: behalve de inhoud of fysieke gedaante van een item bevat de itembank een unieke codering (nummer), en bij dat nummer enkele kenmerken van dat item. Voor een idee van de omvang van de itembank gaan we ervan uit dat deze op zijn minst ruim 200 opgaven bevat per vaardigheid die getoetst wordt. Bij de kenmerken van de items die in de bank zijn opgeslagen, kunnen we bijvoorbeeld denken aan een inhoudsclassificatie naar deeldomeinen of aan de geschatte tijdsduur die het een kandidaat kost om het item te maken. Het belangrijkste bij adaptief toetsen zijn echter de psychometrische kenmerken van de items in de itembank. De algoritmen van adaptieve toetsen hebben deze psychometrische kenmerken als basismateriaal.

De psychometrische kenmerken van items zijn indices die volgen uit het toepassen van een testtheorie. In een testtheorie wordt een relatie gelegd tussen de te meten eigenschap van personen en de score(s) op het daarvoor gebruikte meetinstrument. Bij adaptief toetsen maken we gebruik van de itemresponstheorie (IRT). Dat heeft veel goede redenen, maar in deze context is de belangrijkste reden eenvoudig aan te wijzen. Bij adaptief toetsen gaan we personen op basis van verschillende opgaven met (op) dezelfde meetlat (schaal) meten. Als we een verzameling opgaven hebben die aan een IRT model

voldoet, dan is deze wijze van meten mogelijk. De psychometrische kenmerken van de items in de itembank worden bepaald in het calibratie-onderzoek. Als hieruit blijkt dat items uit de itembank voldoen aan een gekozen IRT model en de psychometrische kenmerken ofwel itemparameters voldoende nauwkeurig geschat zijn, dan spreken we over een IRT geschaalde itembank.

In onze algoritmen voor adaptief toetsen gaan we er dus van uit dat de items zijn geschaald met een itemresponsmodel. Op het Cito is dit meestal het OPLM-model. In dit model wordt een relatie gespecificeerd tussen de te meten vaardigheid θ en kenmerken of parameters van de items. De huidige algoritmen voor adaptief toetsen gaan er allemaal vanuit dat we met dichotoom scoorbare items te maken hebben, dat wil zeggen een itemscore X_i is 1 (goed) of 0 (fout).

In het OPLM model wordt de kans op het goed maken van opgave i, ook wel de itemrespons-functie genoemd, gegeven door:

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))},$$

waarbij β_i de locatie parameter van de opgave is. Deze parameter is geassocieerd met de moeilijkheid van de opgave. Het is dat punt van de vaardigheidsschaal waarbij er een kans van 50% is op het goed maken van de opgave. Parameter a_i is het discriminerend vermogen van het item, ofwel de raaklijn van de richtingscoëfficiënt van de itemresponsfunctie in het punt $\theta = \beta_i$. In de itembank is van elke opgave een (schatting) van de waarde van a_i en β_i opgeslagen. Deze schattingen zijn bepaald in een calibratie-onderzoek. In de algoritmen die in adaptieve toetsen worden gebruikt, worden deze parameters als bekend en vast verondersteld. Hieruit volgt dat aan de calibratie van items die voor adaptieve toetsen worden gebruikt vrij hoge eisen worden gesteld. Mogelijke calibratie-onzekerheden (fouten) kunnen immers in adaptieve toetsen differentieel, dat wil zeggen per persoon verschillend, doorwerken. Een tweede punt van aandacht in dit verband is dat het calibratie-onderzoek vaak uitgevoerd wordt op basis van met papieren toetsen verzamelde itemantwoorden. Deze gegevens zonder meer van toepassing verklaren op de computerversie van de items is naïef. Op zijn minst moet er, liefst vooraf, evidentie verzameld worden dat dit mogelijk is. Ten slotte dienen we er ons van bewust te zijn dat de in calibratie-onderzoek verzamelde itemgegevens niet lang

(eeuwig) geldig zijn. Mogelijke contexteffecten, het bekend raken van items, curriculumveranderingen en populatieveranderingen zijn enkele zaken die ons nopen om steeds nieuwe afnamegegevens van de items te blijven verzamelen en eventueel tot bijstelling van de calibratie van de itembank te komen.

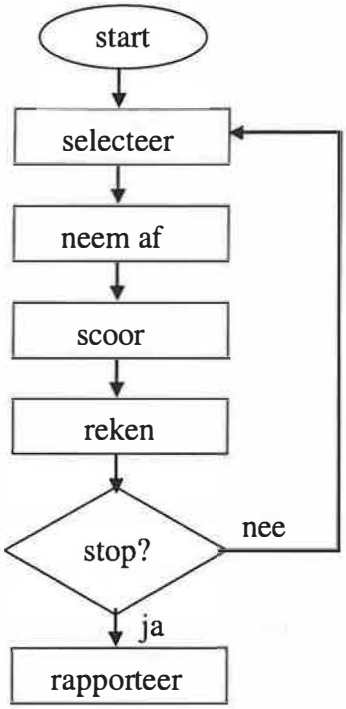
Het toetsdoel

De keuze voor bepaalde (elementen in) algoritmen wordt mede bepaald door het doel van de toets. Toetsen worden voor allerlei doeleinden gebruikt. Vanuit het perspectief van de gebruiker wordt vaak een indeling gemaakt in doelen als afsluiting, selectie, plaatsing, diagnose en sturing. Vanuit het perspectief van de toetsontwikkelaar en de psychometricus kunnen we ons echter beperken tot het onderscheid tussen twee gebruiksdoelen van adaptieve toetsen:

a. **Meten**: het doel is een schatting te verkrijgen voor de vaardigheid van de persoon. Traditioneel zijn bijna alle algoritmen voor adaptieve toetsen er op gericht dit doel zo snel en zo nauwkeurig mogelijk te bereiken.
b. **Classificeren**: het doel is te bepalen in welke van een beperkt aantal vaardigheids-categorieën een persoon thuishoort. Onderzoek heeft uitgewezen dat specifieke algoritmen voor classificatie beter kunnen functioneren dan de onder a bedoelde algoritmen.

3 Beschrijving van de fasen in een adaptieve toetsafname

In de onderhavige beschrijving gaan we uit van de situatie dat we voor een persoon slechts één meting doen. In de praktijk zullen tijdens één afname soms meerdere metingen worden gedaan. De hierna te beschrijven elementen worden dan meerdere malen doorlopen. Schematisch ziet een algoritme voor toetsafname er als volgt uit:



De fasen zijn achtereenvolgens:

Start

In deze fase wordt bepaald welke items als eerst(en) in de adaptieve toets zullen worden afgenomen. Doorgaans beschikt men vóór de afname nog niet over vaardigheids-gegevens van de kandidaat. Als start kan dan worden gekozen voor bijvoorbeeld:

- één of meer aselect getrokken items uit de itembank
- één of meer aselect getrokken items uit een bepaald deeldomein van de itembank (of uit meerdere specifieke deeldomeinen)
- één of meer aselect getrokken items uit een -op grond van psychometrische gegevens ingedeelde- deelverzameling van items uit de itembank (bijvoorbeeld gemakkelijke items).

Als er wel vóórinformatie is, kan eventueel de selectie van het eerste item daarop gebaseerd worden.

Selecteer

In een adaptieve toets vindt itemselectie plaats na elk afgenomen item. Uit de itembank wordt een item gekozen dat in zekere zin goed past bij de tot dan toe gegeven antwoorden: de toetssamenstelling wordt geadapteerd aan de kandidaat. Op de criteria

voor itemselectie zal hierna in meer detail worden ingegaan. Gemeenschappelijk hebben die criteria echter dat ze gebaseerd zijn op het psychometrische begrip **informatie**. Het idee is dat het item dat de hoogste informatie-waarde heeft gerelateerd aan de (tot dan gedemonstreerde vaardigheid van de) kandidaat, als volgende item zal worden aangeboden. Zo'n psychometrisch criterium wordt in de praktijk vaak nog van randvoorwaarden, ten aanzien van deeldomeinen (inhoud) van de itembank (inhoudscontrole = '*content control*'), of ten aanzien van de frequentie van het gebruik van items uit de itembank (afnamecontrole = '*exposure control*'), voorzien.

Neem af en Scoor

Deze twee fasen bij de uitvoering van het algoritmen zullen niet worden besproken. Duidelijk zal zijn dat een item wordt gepresenteerd, waarop de kandidaat een antwoord geeft, waarna de respons wordt gescoord. Het itemantwoord krijgt een numerieke representatie, de itemscore.

Reken

In de rekenfase worden de itemscores van de kandidaat bewerkt: statistische procedures, die hierna meer in detail worden besproken, bepalen op grond van de gescoorde itemantwoorden de vaardigheid van de kandidaat en geven een indicatie van de nauwkeurigheid van die verkregen vaardigheidsschatting. Twee statistisch verschillende aanpakken staan ter beschikking: er vindt statistische schatting plaats of statistische toetsing.

Stop

Na elke afname van een item wordt bepaald of er een nieuw item moet worden geselecteerd of dat de adaptieve toetsafname kan worden gestopt. Het criterium voor stoppen kan onder meer zijn:

a. de nauwkeurigheid waarmee de vaardigheid van de kandidaat wordt geschat;
b. de nauwkeurigheid waarmee een classificatie van een kandidaat kan plaatsvinden;
c. de maximaal (en eventueel ook minimaal) toelaatbare toetstijd of aantal items.

Een combinatie van a. of b. met c. is mogelijk.

Rapporteer

Na de toetsafname wordt de vaardigheid van de kandidaat gerapporteerd. Veel varianten zijn hierbij mogelijk. De belangrijkste zijn:

a. rapportage van de geschatte vaardigheid (met nauwkeurigheidsindicatie) op een lineair getransformeerde schaal die een normgeoriënteerde interpretatie mogelijk maakt;
b. rapportage van de geschatte vaardigheid (met nauwkeurigheidsindicatie) op een getransformeerde schaal die een domein georiënteerde interpretatie mogelijk maakt;
c. eenvoudige (grafische) rapportage van de classificatie van de kandidaat.

4 Het rekendeel van de algoritmen

Op grond van de gegeven antwoorden van een kandidaat wordt in de rekenfase zijn vaardigheid bepaald. In dit onderdeel wordt de basis voor de beslissing over de kandidaat vastgesteld. In onze algoritmen wordt daarbij altijd gebruik gemaakt van de aannemelijkheidfunctie van de vaardigheid θ van de kandidaat. Gegeven de scores $X_1 = x_1, X_2 = x_2, ..., X_k = x_k$ op k items is deze functie:

$$L(\theta; x_1, x_2, ..., x_k) = \prod_{i=1}^{k} p_i(\theta)^{x_i} (1 - p_i(\theta))^{(1-x_i)} = \prod_{i=1}^{k} L(\theta; x_i),$$

waarin voor $p_i(\theta)$ het OPLM model met de uit de calibratie bekende parameters wordt ingevuld. Deze aannemelijkheidfunctie kan op twee statistisch essentieel verschillende manieren gebruikt worden: voor statistische schatting van de vaardigheid en voor statistische toetsing van de vaardigheid. Schatting van de vaardigheid kan voor elke toepassing van adaptief toetsen worden gebruikt, statistische toetsing is een aantrekkelijk alternatief als het toetsdoel classificatie in twee of drie categorieën is. (Zie ook Eggen & Straetmans, 2000).

Schatting

Het doel van adaptief toetsen waarbij de vaardigheid geschat wordt, is dit zo snel mogelijk te doen (met zo weinig mogelijk items) en met vooraf gespecificeerde

nauwkeurigheid. Een veel gebruikte schatting van de vaardigheid van een kandidaat na het maken van k items is de waarde die de aannemelijkheidfunctie maximaliseert. Voor deze (Maximum Likelihood) schatter is een statistisch superieure variant bekend, de gewogen maximale aannemelijkheidschatter van Warm (1989), die in onze algoritmen is ingebouwd. De Warm vaardigheidsschatter na k items, is gegeven door:

$$\hat{\theta}_k = \max_{\theta}(\sum_{i=1}^{k} I_i(\theta))^{1/2}.\prod_{i=1}^{k} L(\theta; x_i)\,.$$

In de uitdrukking voor de Warm schatter is de zogenaamde iteminformatiefunctie $I_i(\theta)$ opgenomen, die een belangrijke rol speelt (of kan spelen) bij de itemselectie (zie paragraaf 5 voor details).

Naast de schatting van de vaardigheid komt na elk item ook een schatting van de standaardfout van de schatting beschikbaar: $se(\hat{\theta}_k)$. De standaardfout is een indicatie van de nauwkeurigheid waarmee de vaardigheid wordt geschat.

In adaptieve toetsen waarin het doel de schatting van de vaardigheid is, fungeert de standaardfout $se(\hat{\theta}_k)$ vaak als stopcriterium voor de toets: als de standaardfout onder een bepaalde waarde is gekomen, stoppen we met toetsen. In classificatieproblemen (met schatting) gebruiken we de vaardigheidsschatting en de standaardfout voor de constructie van betrouwbaarheidsintervallen voor de ware vaardigheid van een persoon: $(\hat{\theta}_k - \gamma.se(\hat{\theta}_k), \hat{\theta}_k + \gamma.se(\hat{\theta}_k))$. Daarin is γ een te specificeren constante die afhangt van de gewenste nauwkeurigheid. Met de adaptieve toets stoppen we, zodra er geen grenspunt meer ligt in het betrouwbaarheidsinterval van de vaardigheid. In dat geval nemen we de beslissing die hoort bij dat deel van de vaardigheidsschaal waarin het betrouwbaarheidsinterval ligt. Als bijvoorbeeld in een zak/slaag-probleem met één grenspunt g geldt dat $\hat{\theta}_k + \gamma.se(\hat{\theta}_k) < g$, dan zakt de kandidaat. Bij meerdere grenspunten werkt de procedure analoog.

Als er in het algoritme een maximaal aantal items is vastgesteld, kan het voorkomen dat de hiervoor beschreven procedure nog niet tot het stoppen heeft geleid. In dat geval zal er uiteraard toch gestopt worden. Bij een classificatieprobleem wordt dan de beslissing genomen op grond van de vaardigheidsschatting alleen. Bij een schattingsprobleem kunnen we niets anders doen dan het rapporteren van een grote standaardfout.

Toetsing

Statistische toetsing kan alleen toegepast worden bij adaptief toetsen als het doel is de vaardigheid van de kandidaat toe te wijzen aan één van twee of drie categorieën. De kansen op foute beslissingen dienen daarbij van tevoren worden vastgelegd. Bij twee categorieën toetsen we de volgende twee hypothesen tegen elkaar:

$H0: \theta \leq \theta_0 - \delta = \theta_1$ tegen

$H1: \theta \geq \theta_0 + \delta = \theta_2$.

De maximaal toegelaten beslissingsfouten worden gegeven door:
P(accepteren H0 | H0 is waar) $\geq 1-\alpha$ en P(accepteren H0 | Hl is waar) $\leq \beta$, waarin α en β kleine constanten zijn. Aan deze eisen kan worden voldaan indien als toetsingsgrootheid de ratio van de aannemelijkheidfunctie onder de alternatieve hypothese en de nulhypothese wordt genomen:

$$LR_k(\theta_2;\theta_1) = \frac{L(\theta_2; x_1,...,x_k)}{L(\theta_1; x_1,...,x_k)}$$

en als procedure wordt gebruikt:

als $\beta/(1-\alpha) < LR_k(\theta_2;\theta_1) < (1-\beta)/\alpha$ ga door met waarnemen

als $LR_k(\theta_2;\theta_1) \leq \beta/(1-\alpha)$ accepteer H0

als $LR_k(\theta_2;\theta_1) \geq (1-\beta)/\alpha$ verwerp H0.

De nulhypothese wordt dus verworpen bij grote waarden van de aannemelijkheidsratio omdat in dat geval de door de kandidaat verkregen scores veel waarschijnlijker zijn onder de alternatieve hypothese H1 dan onder de nulhypothese H0. Voor het accepteren van H0 geldt de omgekeerde redenering. Deze statistische toetsing is een toepassing van de SPRT (Sequential Probability Ratio Test) van Wald (1947).

Bij een beslissingsprobleem met drie categorieën, zijn er twee grenspunten θ_1 en θ_2 op de vaardigheidsschaal, waarop we op analoge wijze een tweetal paren hypothesen formuleren, toelaatbare kansen op foute beslissingen specificeren, en toetsings-voorschriften formuleren die in dit geval gecombineerd worden.

$H0\_1: \theta \leq \theta_1 - \delta_{11} = \theta_{11}$ tegen $H1\_\theta \geq \theta_1 + \delta_{12} = \theta_{12}$ en

$H0\_2: \theta \leq \theta_2 - \delta_{21} = \theta_{21}$ tegen $H1\_2: \theta \geq \theta_2 + \delta_{22} = \theta_{22}$.

De kansen op foute beslissingen worden begrensd door:

P(accepteren H0_1 | H0_1 is waar) $\geq 1-\alpha_1$ en P(accepteren H0_1 | Hl_1 is waar) $\leq \beta_1$

P(accepteren H0_2 | H0_2 is waar) $\geq 1-\alpha_2$ en P(accepteren H0_2 | Hl_2 is waar) $\leq \beta_2$

Bij de eerste toets blijven we waarnemen zolang

$$\beta_1/(1-\alpha_1) < LR_k(\theta_{12};\theta_{11}) < (1-\beta_1)/\alpha_1,$$

en bij de tweede zolang

$$\beta_2/(1-\alpha_2) < LR_k(\theta_{22};\theta_{21}) < (1-\beta_2)/\alpha_2.$$

We verwerpen één van de hypothesen, net zoals hiervoor, zodra de onder- of bovengrens van de kritieke ongelijkheid wordt bereikt; hetzelfde geldt voor de andere hypothesenparen. Als beide hypothesenparen tot een besluit zijn gekomen, combineren we de beslissing tot een eenduidige die een indeling in drie categorieën geeft:

Resultaat Toetsing	**Beslissing: vaardigheid ligt**
H0_1 accepteren + H0_2 accepteren	lager dan laagste grenspunt
H0_1 verwerpen + H0_2 accepteren	tussen beide grenspunten
H0_1 verwerpen + H0_2 verwerpen	hoger dan hoogste grenspunt

Als er in het algoritme een maximaal aantal items is vastgesteld, kan het voorkomen dat de hiervoor beschreven toetsingsprocedures nog niet tot een beslissing hebben geleid. In dat geval zal er een beslissing worden geforceerd. In alle gevallen zal de meest voor de hand liggende beslissing worden genomen. Daarbij wordt uitgegaan van de waarde van

de toetsingsgrootheid. Gekozen wordt dan voor het besluit waarvan de kritieke waarde het dichtst bij de waarde van de toetsingsgrootheid ligt.

5 Het itemselectie deel van de algoritmen

Na elk afgenomen item wordt een volgend af te nemen item uit de itembank gekozen. Bij de keuze van het item wordt met inachtneming van het toetsdoel het beste item bij de tot dan toe gedemonstreerde vaardigheid van de kandidaat gekozen. De volgende itemselectie methoden zijn geïmplementeerd:

a. aselect;
b. maximale Fisher informatie bij de lopende vaardigheidsschatting;
c. maximale Fisher informatie bij het (dichtstbijzijnde) grenspunt;
d. maximale Kullback-Leibler informatie bij de getoetste hypothesen.

Aselect

Het aselect trekken van items betekent, dat na elke afname elk nog niet gebruikt item een even grote kans heeft om geselecteerd en afgenomen te worden. Bij deze methode van items kiezen wordt de toetsinhoud niet aangepast aan de vaardigheid van de kandidaat en wordt er niet efficiënt getoetst. Uit simulatiestudies is gebleken dat door aselect trekken de mogelijke winst van adaptief toetsen verdwijnt.

Maximale Fisher informatie bij de lopende vaardigheidsschatting

Deze methode van itemselectie is optimaal bij adaptief toetsen met schatting van de vaardigheid van de kandidaat als toetsdoel. Bovendien functioneert deze methode ook goed als op grond van de toets personen in één van twee of drie categorieën moeten worden ingedeeld. De methode werkt dan zowel in combinatie met een rekenprocedure die gebaseerd is op statistische schatting als met een procedure die gebaseerd is op statistische toetsing. Formeel is de Fisher iteminformatie functie gedefinieerd als:

$$\mathrm{I}_i(\theta) = -E\left(\frac{\partial^2 \ln L(\theta; x_i)}{\partial \theta^2}\right) = E\left(\frac{\partial L(\theta; x_i)/\partial \theta}{L(\theta; x_i)}\right)^2$$

Deze iteminformatie functie drukt de bijdrage uit die een item kan leveren aan de nauwkeurigheid van de meting van een persoon, als functie van de vaardigheid. Dit is goed in te zien als we ons realiseren dat de schattingsfout van de vaardigheidsschatting uit te drukken is in de som van de iteminformaties van de afgenomen items:

$$se(\hat{\theta}_k) = \sqrt{\sum_{i=1}^{k} I_i(\hat{\theta}_k)} .$$

In het algemeen is bij dichotome items de Fisher iteminformatie een ééntoppige functie van de vaardigheid. Bijvoorbeeld in het OPLM is de functie gegeven door:

$$I_i(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta)) = \frac{a_i^2 \exp(a_i(\theta - \beta_i))}{(1+\exp(a_i(\theta - \beta_i)))^2}$$

In te zien is dat voor elk item de informatie zijn maximum bereikt bij de waarde van de locatieparameter (moeilijkheid) van het item ($\theta = \beta_i$) en dat de discriminatieparameter een grote invloed heeft op de informatie.

De itemselectie vindt met behulp van de informatiefunctie als volgt plaats. Nadat de vaardigheidsschatting $\hat{\theta}_k$ is bepaald, wordt voor elk nog niet afgenomen item de informatie uitgerekend in dat punt. Het item waarvoor de informatiewaarde het hoogst is wordt vervolgens gekozen en afgenomen.
In de geïmplementeerde itemselectieprocedures vindt de feitelijke selectie enigszins anders plaats. In principe is het mogelijk om na elke itemafname een maximaliseringalgoritme toe te passen op de uit te rekenen iteminformaties van alle nog beschikbare items uit de itembank. Dit is echter rekenintensief en niet erg efficiënt omdat steeds voor dezelfde en voor verschillende personen steeds dezelfde berekeningen gemaakt moeten worden. Dit wordt in de praktijk vermeden door te werken met zogenaamde informatietabellen. De tabel wordt vóór de adaptieve afnames gemaakt en bevat voor een (groot) aantal vaardigheidspunten een reeks itemnummers uit de itembank in volgorde van de grootte van de iteminformatie op dat punt. Tijdens de afname wordt de rij in de tabel met de vaardigheid die het dichtst bij de vaardigheidsschatting ligt gekozen en vervolgens wordt uit die rij itemnummers het meest informatieve, nog niet afgenomen item geselecteerd.

Maximale Fisher informatie bij het (dichtstbijzijnde) grenspunt

Bij deze methode worden items geordend op de informatie bij de waarde van de vaardigheid die hoort bij het (dichtstbijzijnde) grenspunt. Deze methode van itemselectie werkt bij classificatieproblemen met één grenspunt beter dan selectie op basis van de maximale Fisher informatie bij de lopende vaardigheidsschatting. Bij problemen met twee grenspunten is geen verbetering van deze methode te verwachten (Eggen & Straetmans, 2000). Men dient zich te realiseren dat bij deze methode van itemselectie (bij één grenspunt) elke persoon in dezelfde volgorde dezelfde items krijgt aangeboden.

Maximale Kullback-Leibler informatie bij de getoetste hypothesen

Zoals Fisher informatie hoort bij schatting van de vaardigheid, hoort de Kullback-Leibler (KL) informatie bij het toetsen van hypothesen. Eggen (1999) heeft itemselectie procedures op grond van dit informatiebegrip voorgesteld en aangetoond dat bij gebruik van een toetsingsalgoritme hiermee een verbetering van de prestaties van adaptieve toetsen kan worden bereikt. Bij toetsing tussen twee hypothesen $\theta = \theta_1$ tegen $\theta = \theta_2$ is de KL iteminformatie gedefinieerd als:

$$K_i(\theta_2 \langle\rangle \theta_1) = E \log\left(\frac{L(\theta_2 ; x_i)}{L(\theta_1 ; x_i)}\right).$$

Het is een uitdrukking van het vermogen van een item om te discrimineren tussen twee hypothesen. Naarmate de KL informatie hoger is, is een item daar geschikter voor. In het OPLM wordt de uitdrukking gegeven door:

$$K_i(\theta_2 \langle\rangle \theta_1) = a_i.(\theta_2 - \theta_1) p_i(\theta_2) + \ln[(1 - p_i(\theta_2))/(1 - p_i(\theta_1))]$$

Evenals het kiezen van items met maximale Fisher informatie bij het grenspunt, dient men zich te realiseren dat ook bij deze itemselectiemethode (bij één grenspunt) elke persoon in dezelfde volgorde dezelfde items krijgt aangeboden.

6 Randvoorwaarden op de itemselectie

Bij het selecteren van items op basis van maximale informatie maken we gebruik van een psychometrisch criterium. Door de praktijk worden doorgaans ook andere eisen gesteld of aanvullende wensen ten aanzien van de toetssamenstelling geuit. Deze wensen kunnen worden gerealiseerd door restricties op te leggen aan het algoritme dat selecteert op maximale informatie. Gemeenschappelijk hebben deze maatregelen dat ze enige meetefficiëntie kosten, doorgaans is dat evenwel zo gering of is het praktisch belang zo groot dat dat geen probleem is. (Zie bijvoorbeeld Eggen & Straetmans, 2000.) Twee verschillende soorten randvoorwaarden zijn in de algoritmen opgenomen. Het betreft inhoudscontrole (Engels: content control) en afnamecontrole (Engels: exposure control).

Inhoudscontrole

Als bij adaptief toetsen items volgens een psychometrisch criterium worden geselecteerd dan wordt steeds het item gekozen dat de meeste informatie geeft bij de tot dan gedemonstreerde vaardigheid. Er vindt een optimale afstemming plaats tussen de vaardigheid van de persoon en de kenmerken van de items. Dit kan echter in strijd zijn met gewenste inhoudelijke toetsspecificaties. Een gebruiker zou kunnen wensen dat bepaalde deelonderwerpen van de te meten vaardigheid in een bepaalde verhouding in een toets voorkomen. Bijvoorbeeld: een adaptieve toets Rekenen moet bestaan uit evenveel items over percentages als over breuken. Overigens worden de toets-specificaties ook vaak ontleend aan een papieren toets die men in gedachten heeft, of die door de Computergestuurde Adaptieve Toets (CAT) wordt vervangen. Er zijn verschillende manieren om zo'n specificatie te realiseren. Twee ervan worden veel toegepast.

1. Verdeel de itembank in zoveel deel-itembanken als er onderscheiden moeten worden en laat de kandidaat in feite meerdere adaptieve toetsen maken. De toetsen gaan dan achtereenvolgens over de verschillende onderwerpen. Als we adaptief toetsen met variabele lengte, dan hebben we op deze manier geen volledige controle over de verhoudingen van het aantal items van de deelonderwerpen die door elke kandidaat worden gemaakt. Desalniettemin wordt deze variant vaak toegepast. Een belangrijke reden hiervoor is dat soms items van een bepaald deelonderwerp dezelfde soort stimulus hebben (bijvoorbeeld

vraagtype), die anders is dan van andere deelonderwerpen. Het afwisselen van de items zou dan problemen kunnen opleveren. Een taaltoets zou bijvoorbeeld items met luisterfragmenten en met leesfragmenten kunnen bevatten. Het is dan praktisch om die items niet door elkaar af te nemen. Daarnaast zijn redenen voor het partitioneren van de itembank in deelbanken, dat de vaste verhouding tussen deelonderwerpen niet zo erg belangrijk is, en of dat men wel met vaste toetslengte wil werken.

2. Soms is het gewenst om de items over deelonderwerpen in gefixeerde proporties in elke adaptieve toets te krijgen. In het algoritme wordt er dan voor gezorgd dat gedurende de afname in de itemselectie de gewenste verhoudingen tussen het aantal items over deelonderwerpen zo goed mogelijk benaderd wordt. Daardoor zal op ieder moment van stoppen met de toets de gewenste specificatie zo goed mogelijk worden bereikt. Deze inhoudscontrole, voorgesteld door Kingsbury & Zara (1991), werkt als volgt. Na elke itemafname worden van de percentages afgenomen items over de deelonderwerpen de gewenste percentages afgetrokken. Van het deelonderwerp waarbij dit verschil het grootst negatief is wordt het volgende item (volgens maximale informatie) afgenomen. In het algoritme wordt dus eerst het deelonderwerp bepaald en daarna pas daarbinnen een item op grond van informatie gekozen.

Afnamecontrole

De itemselectiemethoden die per persoon de optimale toets samenstellen, hebben in de praktijk een aantal minder wenselijke eigenschappen. Alhoewel elke kandidaat bij een adaptieve toets doorgaans een andere toets krijgt, komt het vaak voor dat een aantal items uit de itembank heel veel afgenomen wordt terwijl een ander deel van de itembank niet of nauwelijks gebruikt wordt. Twee problemen kunnen geïdentificeerd worden:

1. overbenutting: sommige items worden zo vaak geselecteerd, dat de geheimhouding zeer snel en direct in gevaar komt.
2. onderbenutting: sommige items worden zo weinig gebruikt dat men zich kan afvragen waarom de kosten zijn gemaakt om ze te construeren.

Door nadere restricties aan de itemselectiemethoden op te leggen, d.w.z. toepassing van zogenaamde afnamecontrole, kan een oplossing voor genoemde problemen worden verkregen. Dit zal enig verlies van efficiëntie van de toets tot gevolg hebben.

Hierna zullen de geïmplementeerde methoden tegen overbenutting en onderbenutting worden beschreven. Voor de overbenutting wordt (een variant van) de Sympson-Hetter methode gebruikt (Sympson & Hetter, 1985). Voor de onderbenutting is gebruik gemaakt van een generalisatie van de progressieve methode van Revuelta en Ponsoda (1998).

Remedie tegen overbenutting

Deze vorm van afname controle probeert de exposure of afname ratio van items kleiner te maken. Deze ratio is voor elk item gedefinieerd als de kans dat een item i uit de itembank wordt afgenomen. Deze kans laat zich schatten door na een groot aantal toetsafnames het aantal keren dat het item in de toets zit te delen door het totaal aantal toetsafnames. Het effect van toepassing van deze afname controle moet zijn dat items die zonder controle (te) vaak worden afgenomen, minder vaak worden afgenomen.

De afnamecontrole vindt in de praktijk als volgt plaats. Per item uit de itembank hanteren we een Exposure Control Parameter (ECP) die we noteren als k_i. De ECP (of k_i) heeft een waarde tussen 0 en 1 en geeft de kans weer dat bij het toegepaste selectiealgoritme een geselecteerd item i daadwerkelijk wordt afgenomen.

Twee uitvoeringen van afnamecontrole die hetzelfde effect hebben:

- Als tijdens een toetsafname een item i geselecteerd is, trek dan een aselect getal R_i uit de uniforme verdeling op (0, 1), als $R_i \leq k_i$ neem dan het item af. Is hier niet aan voldaan, kies dan het volgend item en herhaal de procedure.
- Voer hetzelfde kansexperiment uit voor alle items voordat een persoon aan de toetsafname begint en verwijder de niet af te nemen items voor de betreffende persoon uit de itembank.

Methodes voor afnamecontrole kunnen verschillen in de manier waarop de k_i's worden bepaald.

- De globale methode k_i, met $i = 1, ..., I$ is een constante, bijvoorbeeld 0,5. Het effect hiervan is dat alle vaak geselecteerde items minder vaak zullen worden afgenomen.

- Deze simpele methode (toegepast in Eggen & Straetmans, 2000) kan verbeterd worden door gebruik te maken van simulatieresultaten van de adaptieve toets. Twee vormen zijn de C-methode en de Sympson & Hetter (SH) methode (1985)

Gemeenschappelijk hebben deze twee vormen dat voor de vaststelling van de k_i's grote aantallen (per run van bijvoorbeeld 1000) adaptieve toetsafnames, volgens de gewenste specificaties van het algoritme (bijvoorbeeld itemselectiealgoritme, beslisalgoritme, stopcriterium), worden gesimuleerd. Bij de C-methode vindt deze simulatierun één keer plaats, bij de SH methode moet deze simulatie een aantal keren herhaald worden.

De C-methode

Van elk item $i = 1,...,I$ uit de itembank wordt bij een gegeven algoritme de kans geschat dat een item geselecteerd wordt $P_i(S)$ door

$$f_i(S) = \frac{\text{aantal afnames waarin item } i \text{ geselecteerd is}}{\text{aantal afnames}}$$

En $k_i = 1 - f_i(S)$. Deze parameters zijn na één volledige simulatierun gevonden.

De SH-methode

De SH methode begint met de vaststelling van de maximaal toelaatbare afname ratio, r, bijvoorbeeld 0,2. Dat wil zeggen dat in 2 van de 10 gevallen dat het item is geselecteerd het item ook wordt aangeboden. Vervolgens worden de k_i's in een aantal simulatieruns bepaald. In elke simulatierun schatten we voor elk item de selectieratio $P_i(S)$ en de afnameratio $P_i(A)$. Dus bijgehouden wordt het aantal keren dat een item geselecteerd wordt en het aantal keren dat het item afgenomen wordt. In de eerste simulatierun zijn deze ratio's gelijk aan elkaar.

Stapsgewijs werken dan de simulaties als volgt:

1. Stel alle k_i's voor item $i = 1,...,I$ op 1,
2. Voer de simulatie uit en bepaal $P_i(S)$ en het maximum van $P_i(A)$,

3. Stel de k_i's voor item $i = 1,...,I$ bij:
    a. als $P_i(S) > r$, dan $k_i = r / P_i(S)$,
    b. als $P_i(S) \leq r$, dan $k_i = 1.0$,
    c. zorg ervoor (door de hoogste op 1 te zetten) dat er minstens zoveel k_i's gelijk aan 1.0 zijn als de maximale toetslengte is,
4. Herhaal de stappen 2 en 3 totdat het maximum van $P_i(A)$ ongeveer gelijk is aan r.

Recentelijk is een aantal studies gepubliceerd waarin exposure control methoden worden vergeleken. Hierin werden ook steeds nieuwe varianten voorgesteld. De methodes die gericht zijn op het opheffen van overbenutting zijn dan meestal varianten of uitbreidingen van de SH methode. Bijvoorbeeld een SH methode die varieert met de toetslengte, of een conditioneel op vaardigheid werkend SH methode. Deze methoden zijn over het algemeen ingewikkelder en de resultaten van vergelijkingen met de gewone SH methode geven geen directe aanleiding om ze nu al uit te proberen.

Remedie tegen onderbenutting

Van de hiervoor beschreven methoden is aangetoond dat ze effectief zijn tegen overbenutting van de itembank. Voor de bestrijding van onderbenutting is recent door Revuelta en Ponsoda (1998) een effectieve methode voorgesteld: de zogenaamde progressieve methode voor afnamecontrole. Aan dit voorstel werd een eigen uitbreiding gegeven welke vervolgens is geïmplementeerd. Het basisidee is simpel: de itemselectie vind plaats op een mengsel van twee criteria, volgens toeval (R) en op basis van maximale informatie bij de lopende vaardigheidsschatting (MI). In het begin van een toetsafname is het gewicht van het R criterium groot en van het MI-criterium klein, terwijl naarmate de toetsafname vordert het gewicht van R steeds kleiner en van MI groter wordt. De methode wordt hierna kort beschreven.

Definieer:

h: het aantal items dat al is afgenomen bij een kandidaat,

m: het maximaal aantal items dat een kandidaat krijgt voorgelegd,

$s = \min(a.h / m, 1)$: de relatieve positie van een item in de toets,

I_i^h: de informatie van item i bij $\hat{\theta}_h$, de vaardigheidsschatter na h items,

1. Bepaal het maximum van de informatie voor alle ongebruikte items na afname van h items en noem dit $H := \max_i I_i^h$,
2. Trek voor elk item i een aselect getal R_i uit de uniforme verdeling op het interval $(0, H)$
3. Bepaal voor elk ongebruikt item een gewicht, een lineaire combinatie van een aselecte en een informatie-component: $w_i = (1-s).R_i + s.I_i^h$
4. Neem het item met maximale w_i af.

Eenvoudig in te zien is dat s een getal tussen 0 en 1 is en stijgt bij toenemende lengte van de toets. Verder is aan de definitie van w_i te zien dat de bijdrage van de aselecte component in het begin van de toets groot is en dat verderop de iteminformatie zwaarder weegt.
Een bijzonder geval hebben we als we a = 1 kiezen omdat gedurende de gehele toetslengte de selectie dan zal plaatsvinden op een combinatie van toeval en informatie. Kiezen we a groter (bijvoorbeeld bij 2) dan werkt het toeval in de itemselectie korter (bijvoorbeeld alleen in de eerste helft van de toets).

Revuelta & Ponsoda rapporteren dat deze methode, met beperkt verlies aan nauwkeurigheid, er zorg voor draagt dat het itemgebruik beter over de bank verdeeld is. Het is met name een oplossing van het onderbenuttingsprobleem. Deze methode kan een goede aanvulling zijn op de Sympson-Hetter methode voor afnamecontrole die een oplossing geeft voor het overbenuttingsprobleem

7 Configuratie van een adaptieve toets

Ingrediënten

Voor adaptief toetsen is tenminste nodig:

- een itemverzameling, ook wel genoemd 'itembank'. Dit is de reeks van items waaruit het mechanisme dat de itemselectie verzorgt kan putten tijdens een adaptieve toetsafname. Merk op dat hier het begrip 'itembank' een iets andere inhoud en vorm

heeft dan bij het Cito itembankprogramma THPro. De voor adaptief toetsen benodigde itembank kan echter worden afgeleid uit een THPro itembank.
- een toetsspecificatie. Dit is een (geformaliseerde) beschrijving van verschillende eigenschappen waaraan elke adaptieve toetsafname zal moeten voldoen. Denk aan iets wat we kennen als een toetsmatrijs.

Een itembank zoals hier bedoeld bevat behalve de items zelf (tekstcomponenten) en de itemgegevens (bijvoorbeeld minimum- en maximum itemscore, sleutel, aantal afleiders, etc.) ook de calibratiegegevens. Dat zijn kwantitatieve, psychometrische itemkenmerken (zoals moeilijkheidsgraad, discriminatiewaarde en een schatting van de verdeling van de vaardigheid van de doelgroep). Ten slotte bevat de itembank waar nodig informatie over zogenaamde *toetssegmenten*. Dat is een inhoudelijke groepering van items zodanig dat een adaptieve toetsafname later kan bestaan uit elkaar opvolgende deeltoetsingen, bijvoorbeeld eerst Leesvaardigheid, daarna Luistervaardigheid.
In het navolgende worden enkele gereedschappen voor het manipuleren van al dit materiaal kort aangeduid.

Software voor het maken van adaptieve toetsen

Voor het opzetten van en het werken met een itembank is het programma IT.EXE (**Items & Tests**) beschikbaar, ontwikkeld door Verschoor (Cito, afdeling POK).
Toetsdefinities kunnen worden gemaakt met het programma CASDEF.EXE (**Casper Definitie),** ontwikkeld door Verschoor (Cito, afdeling POK).
Met de twee programma's samen kan men bepalen hoe tijdens de afname een adaptieve toetsafname zich zal gedragen. We noemen dat het opzetten en 'configureren' van een adaptieve toets.

Stappenschema

De volgende activiteiten zijn nodig om een adaptieve toets te kunnen opzetten en configureren:

1. Maak met het programma IT.EXE een **nieuwe (lege) itembank** aan. Daarbinnen:
   - Definieer een **classificatiestructuur**. Items zijn dan toe te wijzen aan de categorieën die later een rol spelen bij het gebruik van een toetsmatrijs. Denk bijvoorbeeld aan de classificatie in Kennis, Inzicht, Toepassing, of de

classificatie op basis van de leerstofstructuur (hoofdstuk, Periode) of vaardigheid (bijvoorbeeld Spreken, Lezen, Luisteren)

- Definieer **hoeveelheden**. Bepaal of het programma werkt met aantallen items, of met hoeveelheden scorepunten, of met percentages.
- **Maak** in het programma de **items aan**. Dan gaat het over nummers en labels van items, de tekstcomponenten.
- **Classificeer** de items. Wijs elk item aan de categorie(en) toe tot welk het behoort.
- Definieer **inclusie- en exclusiegroepen** en wijs in aanmerking komende items daaraan toe.
- Definieer de **toetsmatrijs** (of: toetsmatrijzen).
- Definieer de **toetssegmenten** en geef aan hoe elk segment wordt behandeld qua itemselectie procedure (lineair, adaptief).

2. Maak met het programma CASDEF.EXE de **toetsspecificatie** aan. Daarbinnen:
   - Plaats de in de itembank gedefinieerde **toetssegmenten** in de gewenste **afname-volgorde.**
   - Specificeer onder welke voorwaarden getoetste kandidaten van het ene naar het (en welk) andere segment gaan.

Software voor simulatieonderzoek

Met de in de itembank beschikbare items en de gekozen instellingen is het zaak te controleren of alles naar behoren werkt en zal werken op de toetslocatie waar de adaptieve toets later beschikbaar komt. Die controle gebeurt door de testdriver software op geautomatiseerde wijze te voeden met itemantwoorden van grote aantallen (duizenden) gefingeerde kandidaten. Welke virtuele kandidaat bij welk item welk antwoord geeft wordt daarbij bepaald op basis van door de onderzoeker zelf ingestelde waarden, bijvoorbeeld de kenmerken van een verwachte vaardigheidsverdeling. Indien het algoritme goed werkt, moeten aan het einde van dit type simulaties de toetsuitkomsten dicht in de buurt liggen van de specificaties die de aard van de antwoorden hebben bepaald. Voor dit simulatie onderzoek, zonder welk een adaptieve toets redelijkerwijs niet kan worden uitgebracht, is bij de afdeling POK specifieke software ontwikkeld: CASSIM.EXE (Verschoor). Per simulatie-run rapporteert dit

programma zodanig dat inzichtelijk wordt op welke punten de instellingen voor het algoritme kunnen worden aangepast. Documentatie en een handleiding bij de genoemde POK software komt separaat beschikbaar.

Literatuur

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Eggen, T.J.H.M., & Straetmans, G.J.J.M. (To appear in 2000) Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement.*

Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Revuelta, J., & Ponsada, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 311-327.

Sympson, J.B., & Hetter, R.D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the annual conference of the Military Testing Association, San Diego.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Warm, T. A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.