KALEIDOSCOOP

Optimale toetsconstructie

betrouwbaar meten = zweten

In de eerste aflevering van de nieuwe rubriek Kaleidoscoop aandacht voor het project van Bernard Veldkamp, AIO bij prof. dr. W.J. van der Linden. Hij voert zijn project uit bij de afdeling Onderwijskundige Meetmethoden en Data-analyse van de faculteit Toegepaste Onderwijskunde, Universiteit Twente.

Bernard Veldkamp

De resultaten van je eindexamen, een beroepskeuzetest, of een persoonlijkheidstest bij een sollicitatie kunnen veel invloed hebben op je carrière. Je mag er dan ook van uitgaan dat deze resultaten goede voorspellers zijn van je toekomstige prestaties. Vroeger werd dit afdoende gegarandeerd doordat een autoriteit zijn of haar naam aan een test verbond. Tegenwoordig moeten er echter heel wat meer inspanningen geleverd worden om dit te kunnen garanderen. Van deze inspanningen maakt ook mijn AiO-project deel uit. Het doel van het project was het ontwerpen van modellen en algoritmes voor automatische toetsconstructie.

Rol van de computer

Vanaf de tachtiger jaren worden computers meer en meer ingezet voor het op grote schaal meten van mentale vaardigheden. In grote toetsinstituten in de Verenigde Staten zoals *Educational Testing Service* (ETS), *American College Testing* (ACT) en de *Law School Admission Council* (LSAC), maar ook bij het Centraal Instituut voor Toets Ontwikkeling (CITO) in Nederland worden computers niet alleen ingezet om kandidaten te beoordelen, maar ook voor het construeren van toetsen.

Automatische toetsconstructie

In het proces van automatische toetsconstructie kunnen vier fasen onderscheiden worden. Allereerst worden de vragen geschreven. Dit gebeurt door inhoudelijke experts. Vervolgens worden de statistische eigenschappen van deze vragen bepaald. De kans dat een kandidaat een vraag goed beantwoordt, wordt gemodelleerd met een logis-

tische functie van de latente vaardigheid van de kandidaat en enkele parameters die de vraag karakteriseren. Deze parameters worden geschat met een Expectation-Maximization (EM) algoritme of een Markov Chain Monte Carlo methode. De vragen, waar het model uit de Item Response Theorie goed op fit, worden opgeslagen in een vragenbank. De grootte van een vragenbank varieert van een paar honderd tot een paar duizend vragen. In de derde fase worden de eigenschappen van de toets gespecificeerd. Bij deze eigenschappen kan gedacht worden aan inhoudelijke onderwerpen, het soort vragen, aan achtergrondvariabelen zoals geslacht en etniciteit, maar ook aan de betrouwbaarheid van de schattingen van het niveau van de kandidaten.

Optimalisatiemodel

Tijdens de laatste fase wordt een optimalisatiemodel geformuleerd voor de selectie van vragen uit de vragenbank. Dit optimalisatiemodel heeft een meervoudige doelfunctie. Fisher's informatiefunctie, een maat voor de betrouwbaarheid, moet gemaximaliseerd worden voor verschillende waarden van de latente vaardigheid. De randvoorwaarden worden gevormd door restricties op de in fase 3 geformuleerde eigenschappen. Voor een gemiddelde toets resulteert dit in een paar honderd randvoorwaarden. De 0-1 beslissingsvariabelen geven aan of een vraag wel of niet geselecteerd is voor de toets. Om het probleem op te lossen worden methodes uit de Multiple Objective Optimization gebruikt.

Adaptieve toetsen

Naast het klassikale toetsen worden ook nieuwe vormen ontwikkeld. Een opkomend fenomeen is Computer Adaptive Testing (CAT). Dit adaptieve toetsen is te vergelijken met een mondeling examen, waarbij een computerprogramma de rol van examinator overneemt. Na elke vraag wordt de vaardigheid van de kandidaat geschat en op basis van deze schatting wordt de volgende vraag geselecteerd die de betrouwbaarheid maximaliseert. Net als bij een mondeling examen wordt het niveau van de vragen dus aangepast aan het niveau van de kandidaat. Dit heeft een aantal voordelen ten opzichte van klassikale toetsen.

Voordelen CAT

Een groot voordeel is de toegenomen flexibiliteit. Omdat de moeilijkheid van de vragen kan worden aangepast aan het niveau van de individuele kandidaat, worden er geen vragen gesteld die te moeilijk of te makkelijk zijn. Daardoor raakt de kandidaat minder snel gefrustreerd en verslapt zijn of haar aandacht minder snel. Een tweede voordeel is dat er veel minder vragen hoeven te worden beantwoord. De vragen zijn afgestemd op het niveau van de kandidaat en geven daardoor veel meer informatie, zodat er minder vragen nodig zijn voor een betrouwbare schatting van het niveau. Het construeren van deze toetsen brengt twee nieuwe problemen met zich mee. De restricties zijn gedefinieerd voor de hele toets terwijl er elke keer maar één vraag geselecteerd hoeft te worden. Daarnaast moet de optimalisatie binnen beperkte tijd gebeuren omdat de volgende vraag binnen enkele seconden op het scherm moet verschijnen.

Resultaten onderzoek

Binnen het project wordt uitgebreid ingegaan op Computer Adaptive Testing en op toetsen die meerdere vaardigheden meten. Voor het construeren van deze toetsen zijn modellen, heuristieken en algoritmen ontwikkeld, die op data van Law School Admission Council en American College Testing zijn toegepast. De resultaten zijn veelbelovend en op het moment wordt onderzocht hoe de modellen kunnen worden toegepast in de praktijk.

Bernard Veldkamp is AIO bij prof. dr. W.J. van der Linden, afdeling Onderwijskundige Meetmethoden en Data-analyse, faculteit Toegepaste Onderwijskunde, Universiteit Twente, <Veldkamp@edte.utwente.nl>.