Testtheorie

Testtheorie

Inleiding in de theorie van de psychologische test en zijn toepassingen

prof. dr. P.J.D. Drenth
prof. dr. K. Sijtsma

Vierde, herziene druk

Bohn Stafleu van Loghum
Houten 2006

ISBN 90 313 4747 7
NUR 776

Ontwerp omslag: designwork-bno, Deventer
Ontwerp binnenwerk: Studio Bassa, Culemborg
Automatische opmaak: Pre Press, Zeist

Eerste druk, 1965
Tweede druk, 1975
Derde druk, 1990
Vierde druk, 2006

Bohn Stafleu van Loghum
Het Spoor 2
Postbus 246
3990 GA Houten

www.bsl.nl

Distributeur in België:
Standaard Uitgeverij
Mechelsesteenweg 203
2018 Antwerpen

www.standaarduitgeverij.be

Woord vooraf

Voor u ligt een boek met een lange geschiedenis. De eerste editie van de hand van de eerste auteur verscheen in 1965 en werd in 1975 gevolgd door de tweede editie. In 1990 verscheen met medewerking van de tweede auteur de derde editie. Het voorliggende boek is de vierde, opnieuw ingrijpend gewijzigde en aangepaste editie. Waar liggen de overeenkomsten en verschillen met zijn voorganger?

Eerst maar eens de overeenkomsten. Zo is de indeling van hoofdstukken gehandhaafd; er is alleen een nieuw vierde hoofdstuk bijgekomen dat handelt over de constructie van items en de kwantificering van reacties van personen op items. Daarmee staat het totale aantal hoofdstukken nu op tien. De reden voor het handhaven van deze indeling is dat de hoofdstukken een soort van procedurele volgorde weergeven van het proces van denken over tests en hun constructie. Dus, eerst een historische reflectie (hoofdstuk 1), gevolgd door een uiteenzetting over wat een test is, wat zijn eigenschappen dienen te zijn, wat meten is en hoe het proces van het ontwerpen en maken van een test verloopt (hoofdstuk 2). Vervolgens een overzicht van soorten tests, inclusief een paar concrete voorbeelden (hoofdstuk 3), en daarna een verhandeling over de bouwstenen van tests, de items, en hoe men reacties van personen op items getalsmatig kan bewerken (hoofdstuk 4). De volgende stap (hoofdstuk 5) betreft het voorleggen van een test aan personen en het toekennen van testscores, en ook hoe men die scores op een begrijpelijke manier kan weergeven.

Dan komen de meer technisch-statistische hoofdstukken aan de orde. Eerst (hoofdstuk 6) de bepaling van de betrouwbaarheid volgens de klassieke testtheorie en hiermee samenhangende onderwerpen, vervolgens (hoofdstuk 7) de bepaling van de kenmerken en de kwaliteit

van de test door middel van de moderne testtheorie zoals gevat in de item-responstheorie. Dan komt het bepalen van de validiteit van tests aan bod (hoofdstuk 8) en ten slotte (hoofdstuk 9) wordt de aandacht gericht op het gebruik van tests voor advisering, bijvoorbeeld bij school- en beroepskeuzevragen, in de klinische diagnostiek en bij het nemen van praktische beslissingen over toelating of afwijzing van individuen in onderwijs- en arbeidsorganisaties. Het boek wordt in het laatste hoofdstuk (10) afgesloten met een verhandeling over enkele belangrijke ethische en maatschappelijke problemen met betrekking tot testgebruik.

Steeds is geprobeerd om voldoende diepgang te koppelen aan helderheid. Overigens is de moderne testtheorie al heel wat verder dan wat in dit boek aan de orde wordt gesteld. Moderne testtheorie zoals die in het zevende hoofdstuk wordt behandeld, wordt tegenwoordig vaak ingebed in een omvattende behandeling van statistische methoden, zodat de vergelijking van groepen (bijv. samengesteld op basis van schooltypen, onderwijsmethoden, ontwikkelingsniveau), de afhankelijkheid van testgegevens binnen deze groepen, en de relatie met andere variabelen in een theoretisch netwerk, in één onderzoeksopzet kunnen worden meegenomen. Dit zijn zeer belangrijke ontwikkelingen, maar ze veronderstellen een uitvoerige kennis, niet alleen van de toetsende statistiek maar ook van allerlei nogal gevorderde onderzoeksmethoden, die in bacheloropleidingen doorgaans niet of onvoldoende aan de orde komen. Om de geïnteresseerde student niet in de kou te laten staan, bevat dit boek een groot aantal referenties. Met behulp daarvan kan men zich nader op de hoogte stellen.

Ter ontnuchtering kan worden vastgesteld dat het gros van de tests en vragenlijsten vandaag de dag nog steeds met behulp van de klassieke testtheorie en de factoranalyse wordt geconstrueerd. Vandaar dat wij hier nog steeds veel aandacht aan besteden. Overigens moet het aantal onderzoekers dat toch steeds meer gebruik maakt van de item-responstheorie om op basis daarvan, vaak nog in combinatie met klassieke methoden, hun tests te construeren, niet worden onderschat. In een belangrijk instituut voor toetsontwikkeling als het CITO geldt de item-responstheorie al als een standaard. De verwachting is dat deze 'revolutie' zich ook verder doorzet, ook al vertoont de praktijk van de testconstructie in Nederland vooralsnog een opvallende hang naar het gebruik van de klassieke methoden.

Er is in dit boek gekozen voor een behandeling van in hoofdzaak reeds gevestigde methoden en procedures, met inbegrip van de item-responstheorie. De reden hiervoor is dat dit een lesboek is dat in het algemeen aan het begin van een studie zal worden gebruikt. Dan dient men zich, naar onze mening, te beperken tot zaken die algemeen aanvaard zijn. Ook al biedt de moderne testtheorie talloze interessante ontwikkelingen en is de verleiding groot hierover eens uit te pakken, men moet zich ook realiseren dat nogal wat van deze ontwikkelingen 'ver voor de troepen uitlopen', en dat het, los van hun vaak opvallende vernuftigheid, toch meestal nog onduidelijk is in hoeverre zij zullen stand houden, laat staan doorbreken, in de test-, toets- en vragenlijstconstructie. Een lesboek voor de beginnende student dient de lezer juist op de hoogte te stellen van de gangbare en geaccepteerde beginselen van het vak en nieuwere ontwikkelingen aan te stippen. Overigens bieden diverse universitaire opleidingen cursussen over de meer geavanceerde onderwerpen aan, en kunnen wij geïnteresseerde studenten aanraden zich vooral bij hun docenten te melden. Die kunnen hen dan zonder moeite verder helpen.

Vervolgens de verschillen ten opzichte van de vorige editie. Het eerste wat zal opvallen in vergelijking met de editie uit 1990 is het taalgebruik. Misschien is dit verschil niet zo spectaculair, maar het trof ons dat, na een aanvankelijk grote waardering van de kant van de lezers, zich in de loop van de jaren negentig een kentering in deze mening leek voor te doen. Vond men het boek aanvankelijk helder geschreven, latere generaties vonden de stijl en het taalgebruik nogal ouderwets en nodeloos ingewikkeld. In deze editie is daarom getracht om hierin verbetering te brengen door de dingen zakelijker en bondiger te formuleren, zonder overigens te vervallen in overdreven taalkundige eenvoud.

Een andere wijziging heeft betrekking op de uitleg van veelal ingewikkelde begrippen en procedures. Tot in de jaren negentig van de vorige eeuw was het aan de universiteiten nog gebruikelijk om een redenering of een bewijs niet helemaal uit te leggen en de student zelf te laten zoeken naar ontbrekende schakels en oplossingen. Een mooi voorbeeld hiervan vormen de vaak gevorderde statistiekboeken, waarin sommige afleidingen of oplossingen niet in de hoofdtekst worden behandeld maar door de lezer zelf via het maken van opgaven moeten worden gevonden. In de huidige tijd is het onderwijs veel meer gericht op het aanbieden van kant-en-klare modules, inclusief alle oplossingen, en lijkt men ernaar te streven alle onzekerheid over hoe

de vork nu precies in de steel zit te vermijden door alles precies uit te leggen. Hoe men daar ook over denkt, een modern lesboek moet hiermee rekening houden. Deze editie biedt daarom meer uitleg en concrete voorbeelden dan de vorige, hoewel we niet hebben willen uitsluiten dat het maken van opgaven tot aanvullende inzichten kan leiden.

Daarmee is een derde vernieuwing genoemd. Elk hoofdstuk wordt nu afgesloten met een serie vragen en opdrachten. Het oefenen van de stof door middel van deze vragen en opdrachten is naar onze mening een goede manier om na te gaan of men de stof beheerst, maar ook om nieuwe inzichten te ontwikkelen. Waar het berekeningen betreft zijn de antwoorden te vinden op de website van de uitgever, uiteraard bij voorkeur pas te raadplegen ná het uitvoeren van de opdrachten.

Een vierde vernieuwing is de toevoeging van een verklarende appendix waarin de statistische begrippen zijn opgenomen die men in dit boek nodig heeft. Opnieuw hebben wij gemeend ons te moeten beperken tot een descriptieve behandeling van de testtheorie, ook al worden begrippen als steekproef en populatie en steekproevenverdeling niet gemeden. Onze keuze is echter ingegeven door de overtuiging dat het van het grootste belang is in dit boek studenten vertrouwd te maken met de logica, de procedures en de belangrijkste begrippen van de testtheorie. Ook al is het formele deel van de testtheorie een specialisatie van de statistiek, toch menen wij dat in een inleiding niet onnodig veel lastige zaken tegelijk de revue moeten passeren. Een vollediger begrip van de veelzijdige testtheorie kan in een later stadium van de studie zonder veel bezwaar worden verkregen in combinatie met de inmiddels opgedane kennis van de toetsende statistiek.

Een vijfde vernieuwing betreft het nieuwe hoofdstuk 4 over de constructie van items en de kwantificering van reacties op die items. In voorgaande edities vormde dit onderdeel in sterk verkorte vorm een onderdeel van hoofdstuk 3, maar het werd door ons toch te belangrijk gevonden om zo onopvallend te blijven. Wij hopen uiteraard met de uitvoerige behandeling van de vraag hoe een echte test of vragenlijst er uitziet de stof minder abstract te hebben gemaakt. Hiertoe draagt wellicht ook bij dat in hoofdstuk 3 niet alleen een overzicht van soorten tests wordt gegeven, maar dat nu ook drie echte tests bij wijze van voorbeeld meer in detail worden behandeld.

Tot slot willen wij de volgende collega's danken voor hun bijdragen aan de totstandkoming van de huidige editie van dit boek: Andries van der Ark, Luc van Baest, Samantha Bouwmeester, Hans Landsheer en Rob Meijer gaven commentaar op voorversies van diverse hoofdstukken, Wilco Emons leverde de figuren, Jos ten Berge en Frits Zegers stelden enkele opgaven bij hoofdstuk 6 en de appendix ter beschikking, en Arne Evers verschafte informatie over een aantal tests. Vele, niet genoemde collega's inspireerden ons in de afgelopen decennia tot onze huidige inzichten en standpunten. Wij blijven uiteraard zelf verantwoordelijk voor de inhoud van dit boek, inclusief eventuele onjuistheden.

Pieter J.D. Drenth
Klaas Sijtsma
Amsterdam/Bussum, voorjaar 2006

Inhoud

Woord vooraf 5

1 Historische ontwikkeling van het testen 15
1.1 Periode tot het verschijnen van de Binet-Simon-test 16
1.2 Periode tussen het verschijnen van de Binet-Simon-test en de Eerste Wereldoorlog 20
1.3 Van het begin van de Eerste tot de Tweede Wereldoorlog 22
1.4 Van het begin van de Tweede Wereldoorlog tot heden 28
1.4.1 *Ontwikkelingen in de Verenigde Staten* 28
1.4.2 *Ontwikkelingen in Europa, vooral in Nederland* 32
Opdrachten 35

2 Definitie, kenmerken en toepassingen van de test 38
2.1 Wat is een test? 38
2.1.1 *Onderdelen van een test* 38
2.1.2 *Eerste omschrijving* 40
2.1.3 *Kenmerken van een test* 41
2.2 Meten van eigenschappen door middel van tests 53
2.2.1 *Meetniveaus en toegestane operaties* 53
2.2.2 *Opvattingen over meten* 57
2.2.3 *De gangbare procedure voor het meten van psychologische eigenschappen* 61
2.3 Definitie van een test 67
2.4 Toepassingsmogelijkheden 68
2.4.1 *Beoordeling van individuen* 68
2.4.2 *Beoordeling van groepen* 70

2.4.3 *Beoordeling van invloed van situaties en methoden* 71
Opdrachten 72

3 Indelingen, onderscheidingen en begrippen 76
3.1 Indeling naar testgedrag 76
3.1.1 *Tests voor prestatieniveau* 78
3.1.2 *Tests voor gedragswijze* 86
3.1.3 *Drie voorbeelden van tests* 96
3.2 Indeling naar instructie en afneming 106
3.2.1 *Individuele test en groepstest* 106
3.2.2 *Snelheidstest en niveautest* 107
3.3 Onderscheid op basis van testvragen 109
3.3.1 *Cultuurvrije en niet-cultuurvrije tests* 110
3.3.2 *Directe tests en indirecte tests* 112
3.3.3 *Vrije-antwoordentests en keuze-antwoordentests* 113
Opdrachten 113

4 Constructie van items en kwantificering van reacties 116
4.1 Van de respondent gevraagde activiteit 117
4.2 Vorm waarin het antwoord wordt gegeven 119
4.3 Itemvormen: het speciale geval van geprecodeerde items 125
4.3.1 *Items voor prestatieniveautests* 125
4.3.2 *Items voor tests voor gedragswijze* 129
4.4 Kwantificering van antwoorden 131
4.4.1 *Kwantificering, diverse informatiebronnen* 131
4.4.2 *Itemscores* 133
4.5 Beoordeling van de kwaliteit van items in vooronderzoek 136
4.5.1 *Dichotome items* 138
4.5.2 *Polytome items* 142
Opdrachten 143

5 Afneming van tests en verwerking van testgegevens 146
5.1 Tests afnemen 146
5.2 Scoring van antwoorden 151
5.2.1 *Scoring van reacties op items met open-vraagvorm* 152
5.2.2 *Scoring van reacties op geprecodeerde items* 153
5.2.3 *Toevalscorrectie* 156
5.2.4 *Weging van itemscores* 160
5.3 Testen per computer 161
5.3.1 *Technologische bijdragen en veranderingen* 162

5.3.2 *Wetenschappelijke bijdragen en veranderingen* 166
5.3.3 *Adaptief testen* 169
5.4 Bewerkte scores en normen 172
5.4.1 *Vergelijking met een absolute standaard* 175
5.4.2 *Verhoudingsnormen* 176
5.4.3 *Vergelijking en normen gebaseerd op een rangorde* 179
5.4.4 *Vergelijking en normen gebaseerd op gemiddelde en spreiding* 182
Opdrachten 187

6 Betrouwbaarheid 190
6.1 Herhaalbaarheid van metingen 190
6.2 De klassieke testtheorie 194
6.2.1 *Betrouwbare score en meetfout* 194
6.2.2 *Betrouwbaarheid van testscores en de standaardmeetfout* 202
6.2.3 *Belangrijke onderscheidingen* 204
6.3 Bepaling van de betrouwbaarheid 205
6.3.1 *Parallelvormmethode* 206
6.3.2 *Test-hertestmethode* 210
6.3.3 *Splitsingsmethode* 212
6.3.4 *Interne-consistentiemethode* 215
6.4 Speciale onderwerpen 226
6.4.1 *Nauwkeurigheid van metingen* 226
6.4.2 *Betrouwbaarheid en testlengte* 235
6.4.3 *Betrouwbaarheid en validiteit* 238
6.4.4 *Betrouwbaarheid van verschilscores* 241
6.4.5 *Betrouwbaarheid en spreiding van scores* 243
6.4.6 *Betrouwbaarheid van heterogene tests* 244
6.4.7 *Generaliseerbaarheid van metingen* 245
6.5 Tot besluit 247
Opdrachten 248

7 Nieuwe ontwikkelingen in testtheorie en testconstructie 253
7.1 Principes en begrippen van de item-respons-theorie 256
7.2 Enkele modellen uit de item-responstheorie 262
7.2.1 *Het Rasch-model* 263
7.2.2 *Modellen met respectievelijk twee en drie itemparameters* 273
7.2.3 *De modellen volgens Mokken* 278
7.2.4 *De onderlinge relaties van de item-responsmodellen* 285
7.3 Meten met item-responsmodellen 287
7.3.1 *Betekenis en gebruik van metrische schalen* 288

7.3.2 *Nauwkeurigheid van de meting* 291
7.4 Praktisch gebruik van de item-responstheorie 294
7.4.1 *De itembank en equivalering van scores en kenmerken van items* 294
7.4.2 *Testconstructie op basis van een itembank* 299
7.4.3 *Adaptieve tests* 302
7.4.4 *Vraagonzuiverheid* 306
7.4.5 *Afwijkende patronen van itemscores* 312
7.5 Tot besluit enkele speciale onderwerpen 317
7.5.1 *Item-responstheorie voor polytoom gescoorde items* 318
7.5.2 *Vergelijking klassieke testtheorie en item-responstheorie* 320
7.5.3 *Rol van item-responstheorie in psychologische theorievorming* 322
Opdrachten 323

8 Validiteit en betekenis 328
8.1 Het begrip validiteit 329
8.2 Enkele andere onderscheidingen in validiteit 334
8.2.1 *Vier belangrijke soorten validiteit* 334
8.2.2 *Andere onderscheidingen in het begrip validiteit* 338
8.3 Predictieve validiteit 341
8.3.1 *Nadere bepaling van het criteriumbegrip* 343
8.3.2 *Opzet van een test of testbatterij met predictieve validiteit* 346
8.3.3 *Differentiatie in het criteriumonderzoek* 359
8.3.4 *Validiteitsgeneralisatie* 363
8.3.5 *Beperkingen van predictieve validiteit* 368
8.4 Betekenis en begripsvaliditeit 370
8.4.1 *Begripsvalidering* 370
8.4.2 *Betekenisanalyse: op zoek naar de betekenis* 376
8.4.3 *Alternatieve verklaringen* 383
8.5 Nogmaals betrouwbaarheid en validiteit 388
Opdrachten 392

9 De bijdrage van de test in het beslissingsproces 396
9.1 Taxonomie van beslissingen 398
9.2 Enkelvoudig selectie- c.q. afwijzingsmodel 402
9.2.1 *Het gebruik van een enkele test* 403
9.2.2 *Het gelijktijdig gebruik van diverse tests* 413
9.2.3 *Selectie in een of meer fasen* 415
9.3 Plaatsingsbeslissingen 419
9.3.1 *Plaatsing en niveauverschillen* 419
9.3.2 *Plaatsing en kwalitatieve verschillen* 423

9.4 Individuele beslissingen 426
9.5 Open vraag 429
9.6 Tot besluit 433
Opdrachten 433

10 Ethiek van het testen 437
10.1 Levensbeschouwelijke en menselijke bezwaren 440
10.2 Technische en methodologische bezwaren 445
10.3 Misbruik 449
10.3.1 *Schending van vertrouwen* 449
10.3.2 *Misleiding* 450
10.3.3 *Binnendringen in het privéleven* 451
10.3.4 *Discriminatie* 453
10.4 Tot besluit 459
Opdrachten 459

Appendix 462
Eenvoudige statistische begrippen 462
Opdrachten 469

Literatuur 474

Register 497