12 Psychometrische benadering binnen de DTT

Daniel van der Palm en Herbert Hoijtink

12.1 Model

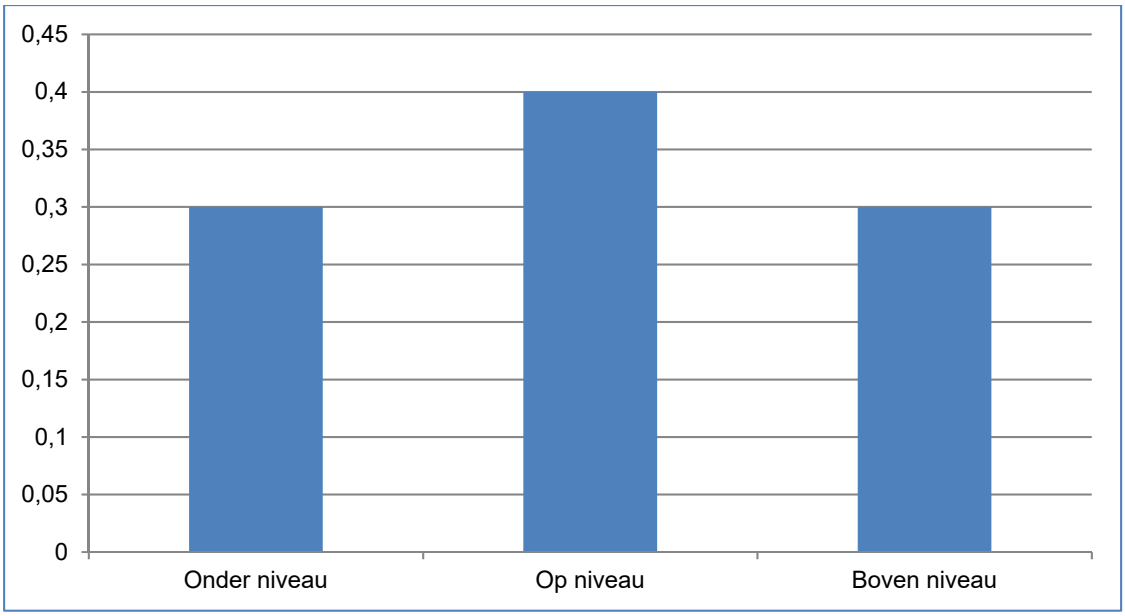
Het psychometrische model voor de DTT is een latenteklassemodel (LK-model) (Goodman, 1974; Lazarsfeld, 1950) dat gecombineerd is met zogenaamde diagnostische hypotheses (Roelofs & Schouwstra, 2012; Hoijtink & Sies, 2013; Sies, 2014). Bij de DTT heeft het LK-model een gefixeerd aantal latente klassen, namelijk drie, en wordt gebruik gemaakt van prior gewichten. Gedurende de afgelopen decennia is het gebruik van het latenteklassemodel gegroeid wat betreft het aantal toepassingen in de praktijk, maar zeker ook wat betreft het aantal wetenschappelijke publicaties. In eerste instantie was het latenteklassemodel bekend als 'Latent Structure Analysis' en binnen de econometrie wordt het ook wel het 'finite mixture'-model genoemd.

Het LK-model wordt gebruikt in situaties waarbij de te meten concept(en) niet direct observeerbaar zijn. Mensen kunnen bijvoorbeeld tot bepaalde latente klassen behoren wat hun persoonlijkheid betreft of politieke overtuiging. Zo kan men bijvoorbeeld niet direct observeren of een persoon introvert/extravert is of bijvoorbeeld boven gemiddeld open staat voor nieuwe ervaringen. Bij een LK-model gebruiken we om die reden indicator variabelen om (indirect) in te schatten tot welke latente klasse elk persoon behoort; dit laatste wordt ook wel classificatie genoemd. In het geval van de DTT worden er drie diagnostische hypotheses (Roelofs & Schouwstra, 2012) gesteld met betrekking tot de beheersing van een leerling: de leerling zit onder niveau (<), op niveau (=) of boven niveau (>). De DTT levert dus geen score op of een zak/slaag resultaat; het doel is om tot een diagnose te komen over de mate waarin een leerling (deel-) aspecten beheerst om, waar nodig, extra tijd en aandacht te kunnen geven aan materie waar een leerling mee worstelt of extra uitdaging te bieden bij materie die de leerling al goed beheerst.

12.2 Ad-hoc-kalibratie, kalibratie en herkalibratie

Er kunnen drie groepen leerlingen onderscheiden worden: de onder-niveaugroep, de op-niveaugroep en de boven-niveaugroep. Het doel van een kalibratie is om voor elk van de drie niveaugroepen de succeskansen voor items en testlets te bepalen. Een item is de opgave of het opgave-onderdeel dat tot één onafhankelijk beoordeelde respons leidt. Bij een testlet geeft een leerling antwoord op een meerdere opgave-onderdelen die bij elkaar horen en afhankelijk van elkaar zijn. De eerste kalibratie die plaats kan vinden is een ad-hoc-kalibratie en bestaat uit een simpeler proces dan de standaardkalibratie. In deze ad-hoc-kalibratie wordt een relatieve norm toegepast. Een relatieve norm heeft als voordeel dat deze zeer eenvoudig en snel is toe te passen. De ad-hoc-kalibratie stelde ons in staat om na elke pretest de scholen feedback te geven over de prestaties van hun leerlingen op de pretest in vergelijking met andere pretestscholen (beschreven in de pretestverslagen, Cito, 2015; Cito, 2016). De ad-hoc-kalibratie is ook gebruikt bij de toets- en itemanalyses (zie Hoofdstuk 10) om een indicatie te geven van het onderscheidend vermogen.

Bij een ad-hoc-kalibratie wordt er voor elke leerling per aspect een inschatting gemaakt tot welke niveaugroep hij of zij behoort op basis van het percentiel waar de betreffende leerling in valt wat zijn of haar aantal correcte antwoorden (somscore) betreft (zie bijvoorbeeld Figuur 12-1). De leerlingen die in het 20ste percentiel vallen of lager worden ingedeeld in de onder-niveaugroep, leerlingen die boven het 20ste percentiel zitten maar lager dan het 81ste percentiel worden ingedeeld in de op-niveaugroep en de resterende leerlingen in de boven-niveaugroep. Nu de groepen gedefinieerd zijn kunnen de succeskansen voor de items en testlets bepaald worden. Deze laatste stap van de kalibratie gebeurt op dezelfde wijze als bij een standaardkalibratie. Om die reden wordt nu het gehele proces van een standaardkalibratie uiteengezet.



Figuur 12-1. Voorbeeld van proportionele verdeling van leerlingen bij een ad-hoc-kalibratie op basis van aantal correcte antwoorden

Bij een standaardkalibratie wordt een (absolute) criterium-gerelateerde methode gebruikt. Met behulp van experts en de data-driven direct consensus methode worden voor elk vak de standaarden bepaald (zie Hoofdstuk 11). Per aspect zijn er twee standaarden; dit zijn de aantallen correcte responsen die volgens de experts horen bij (a) een leerling wiens beheersing van een aspect nét op niveau is en (b) een leerling wiens beheersing nét boven niveau is. Nadat de standaarden bepaald zijn wordt er voor elke pretestleerling per aspect vastgelegd of hij of zij onder, op of boven niveau zit. Het essentiële verschil tussen een ad-hoc-kalibratie en een standaardkalibratie is dus of de grenswaarden van de drie niveaus gebaseerd zijn op relatieve percentielen of op absolute standaarden. Bij relatieve percentielen (de ad-hoc-kalibratie) hangt de norm af van de prestaties van andere leerlingen. Absolute standaarden die zijn gezet met behulp van vakexperts zijn gebaseerd op de tussendoelen, de toetsinhoud en empirische feedback (zie Hoofdstuk 11).

Na het bepalen van de drie groepen leerlingen moeten de parameters van het LK-model worden geschat. Er zijn drie sets aan parameters: (1) de succeskansen voor losse items, (2) de succeskansen voor testlets en (3) de prior-modelkansen. De succeskansen voor de losse items kunnen simpelweg worden geobserveerd binnen de drie groepen leerlingen (de onder-niveaugroep, de op-niveaugroep en de boven-niveaugroep). De succeskans voor item *j* binnen de boven-niveaugroep, bijvoorbeeld, wordt als volgt berekend,

$$P_j(>) = \frac{N_j(>)}{N_j(>) + M_j(>)}$$

waarbij $N_j(>)$ staat voor het aantal leerlingen in de boven-niveaugroep dat item j correct beantwoord heeft en $M_j(>)$ het aantal boven-niveauleerlingen dat item *j* incorrect beantwoord heeft.

Bij een testlet geeft een leerling antwoord op een set opgaven of onderdelen van opgaven die bij elkaar horen en afhankelijk van elkaar zijn. Het is om die reden belangrijk dat er rekening gehouden wordt met het feit dat de verschillende antwoorden op deze set opgaven onderling een statistische afhankelijkheid kunnen vertonen. De oplossing voor dit probleem is dat we succeskansen voor testlets op het niveau van responspatronen modelleren. Een testlet die drie responsen oplevert heeft 2^3 = 8 mogelijke responspatronen en voor elk van de drie niveaugroepen moeten we de kans schatten dat elk van de 8 mogelijke responspatronen gegeven wordt door een leerling. Het aantal mogelijke responspatronen neemt snel toe naargelang er meer responsen in een testlet zitten en lang niet alle mogelijke responspatronen zullen daadwerkelijk geobserveerd worden bij een afname. Om die reden gebruiken we voor testlets een

tweede-orde log-lineair model om alsnog voor elk mogelijk responspatroon een kans te kunnen schatten. Zonder deze informatie zou het niet mogelijk zijn om de gehele reeks aan mogelijke responspatronen mee te nemen in de simulatieprocedure. In het geval dat er meer dan vier responsen zijn die eigenlijk tot één testlet behoren, worden deze responsen niet als testlet opgevat maar verwerkt als losse items.

Naast de twee sets aan succeskansen hebben we tot slot ook nog de priors nodig om de niveaugroep van een leerling in te schatten op basis van zijn of haar responspatroon. De priors zijn drie gewichten die ons in staat stellen om tendensen in de diagnoses bij te sturen. Een tendens kan bijvoorbeeld zijn dat minder onder-niveauleerlingen correct gediagnosticeerd worden dan op-niveauleerlingen. In eerste instantie gebruiken we zogenaamde niet-informatieve priors; de drie gewichten zijn dan allen gelijk aan 1/3 en sturen dus nog niet bij. De verdere details omtrent priors komen in sectie 12.4 aan bod.

12.2.1 Herkalibratie

Het resultaat van een kalibratie is een verzameling van geschatte succeskansen voor de items en geschatte kansen voor de responspatronen van de testlets. Wanneer echter nieuwe afnamedata beschikbaar komen voor eerder gekalibreerde items en testlets, dan willen we in principe de kalibratie opnieuw uitvoeren op basis van alle beschikbare data om de invloed van steekproeffluctuaties te verminderen (standaardfouten van modelparameters nemen af naarmate de steekproefgrootte toeneemt). Het is dan van belang om eerst te onderzoeken of de resultaten van een kalibratie op basis van de nieuwe data (item- en testletparameters) niet te veel afwijken van de eerder geschatte parameters. Omdat we te maken hebben met steekproeven ligt het in de lijn der verwachting dat er kleine verschillen zijn in de geschatte kansen tussen verschillende afnames, maar er kan ook sprake zijn van een systematische verschuiving, wat ook wel ‘parameter drift’ wordt genoemd (zie ook Maas, 2017). Indien er sprake is van een verschuiving, dan moet er nader bekeken worden hoe we omgaan met deze discrepantie. Als de resultaten redelijkerwijs voldoende overeenkomen, dan kan er een herkalibratie gedaan worden op basis van alle beschikbare data. Vervolgens wordt de output van deze herkalibratie gebruikt bij het samenstellen van nieuwe toetsen.

12.2.2 Imputatie

Vanwege de adaptieve opzet van de DTT krijgen niet alle leerlingen alle opgaven voorgelegd en is sprake van missende waarden. Binnen de groep onder-niveauleerlingen zullen bepaalde opgaven bijvoorbeeld niet of weinig gemaakt zijn omdat deze opgaven later in de toets voorkomen en alleen worden aangeboden aan leerlingen die waarschijnlijk op-niveau zitten. Een andere oorzaak voor missende waarden die voor komt is dat sommige leerlingen niet alle opgaven voorgelegd hebben gekregen omdat de diagnose al zeker genoeg was (zie Hoofdstuk 13). Als de missende waarden niet zouden worden aangepakt, dan zouden er een aantal opgaven kunnen zijn waarvoor er binnen één of meer van de drie groepen geen of een zeer laag aantal responsen geobserveerd is. Oftewel, de succeskans van die niveaugroep kan dan niet geschat worden of de schatting is zeer instabiel, omdat deze gebaseerd is op een klein aantal observaties.

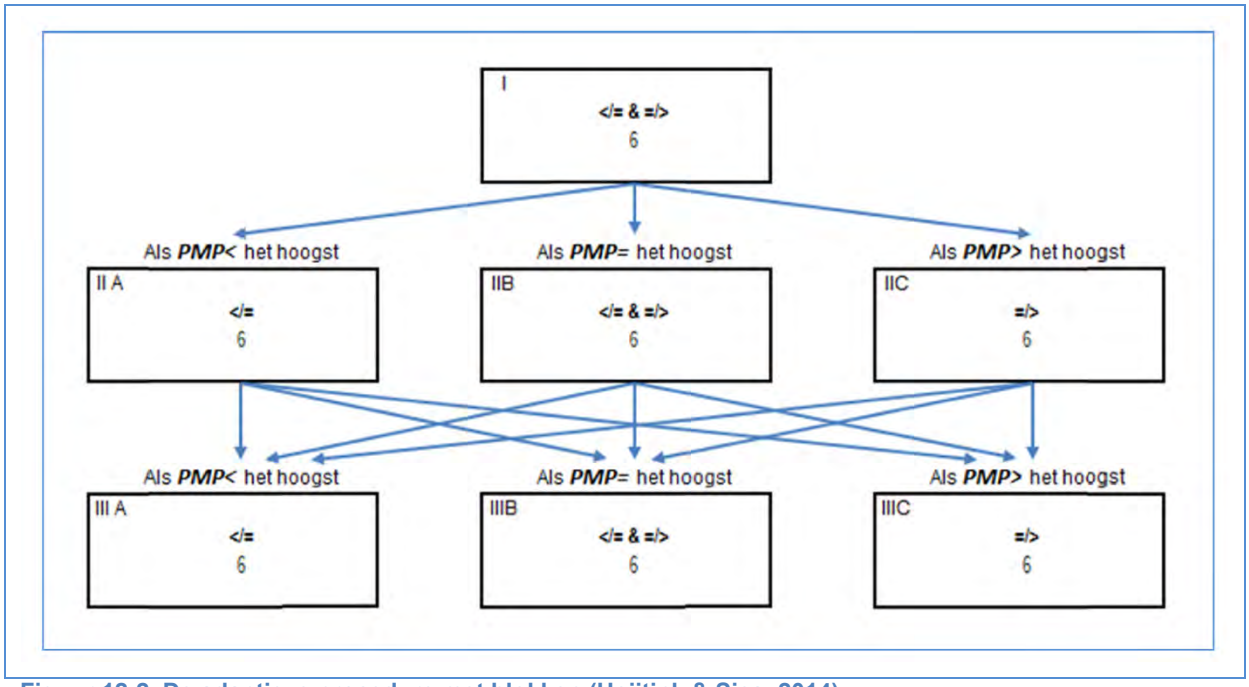
Met behulp van imputatie is het mogelijk om alsnog responsen te schatten voor leerlingen met missende waarden en op deze manier kan de stabiliteit van de geschatte kansen verhoogd worden. In het geval van de DTT wordt er een LK-model gebruikt om de simultane kansverdeling voor alle responsen te schatten. Aan de hand van een dergelijke geschatte kansverdeling kunnen er responsen gegenereerd worden voor leerlingen met missende waarden.

Het aantal latente klassen dat gebruikt wordt voor de imputatie wordt iteratief bepaald aan de hand van een informatiecriterium. Een informatiecriterium maakt het mogelijk om een model te kiezen waarbij modelfit wordt afgewogen tegen het aantal parameters. Voor elke dataset wordt er eerst een LK-model met twee klassen geschat en vervolgens wordt in elke iteratie het aantal klassen opgehoogd, totdat de verbetering in de modelfit niet meer opweegt tegen de toename in het aantal parameters.

Als het LK-model voor imputatie geschat is dan wordt voor iedere leerling met minimaal één missende waarde berekend wat diens kans is om tot elk van de drie groepen te behoren en wordt de leerling willekeurig ingedeeld op basis van die kansen. Vervolgens worden simpelweg de conditionele responskansen gebruikt om voor iedere leerling responsen te genereren voor items met missende waarden.

12.3 Optimalisatie van de blokindeling door middel van simulatie

Met de drie sets aan parameters is het LK-model in principe volledig gedefinieerd. Wat echter nog niet vast staat is welke opgaven we precies voorleggen aan de leerlingen. In theorie is het mogelijk om iedere leerling alle opgaven voor te leggen, ongeacht zijn of haar beheersingsniveau. In de praktijk zou dit echter een zeer tijdrovend proces zijn en een te grote belasting vormen voor de leerling. Daarom wordt gebruik gemaakt van een adaptief design (zie ook Hoofdstuk 13). Een adaptief design bij de DTT bestaat uit 2, 3 of 4 lagen[13] aan opgaven met in elke laag een apart blok aan opgaven voor leerlingen die waarschijnlijk onder niveau zitten, leerlingen die waarschijnlijk op niveau zitten en leerlingen die waarschijnlijk boven niveau zitten.



Figuur 12-2. De adaptieve procedure met blokken (Hoijtink & Sies, 2014)

Zoals te zien is in Figuur 12-2 maken alle leerlingen voor één aspect eerst allemaal hetzelfde blok aan opgaven. Deze eerste set opgaven zou dus goed onderscheid moeten kunnen maken tussen leerlingen onder en leerlingen op niveau, evenals tussen leerlingen op en leerlingen boven niveau. Na het doorlopen van het eerste blok wordt er een inschatting worden gemaakt van de niveaugroep. Deze berekening is als volgt:

> De priors en de succeskansen voor elk van de drie niveaugroepen zijn geschat en dus is het mogelijk om voor elke niveaugroep te berekenen wat de kans is dat een bepaald responspatroon wordt geobserveerd (Hoijtink, Béland, & Vermeulen, 2012); namelijk, de prior behorende bij een niveaugroep vermenigvuldigd met de kans dat voor item 1 de gegeven respons zich voordoet en zo verder de productsom over alle items die beantwoord zijn door een leerling. Als we deze drie kansen elk delen door de som van deze drie kansen dan komen we tot de 'posterior membership probabilities' (PMP). Deze PMP's geven aan hoe waarschijnlijk het is dat een leerling gegeven zijn of haar responspatroon tot een specifieke niveaugroep behoort.

Na het eerste blok wordt elke leerling voorlopig geplaatst in de niveaugroep waarvoor hij of zij de grootste PMP heeft (die het meest waarschijnlijk is). Vervolgens krijgt de leerling in laag 2 een blok opgaven dat hoort bij deze niveaugroep. Bijvoorbeeld, als de leerling waarschijnlijk onder niveau zit krijgt de leerling opgaven die een goed onderscheid maken tussen onder en op niveau (linker blok in Figuur 12-2). Als de

[13] In 2016 was sprake van 2 lagen en in 2017 sprake van 3 lagen zonder doortoetsen (wiskunde) of met doortoetsen (talen)

leerling waarschijnlijk boven niveau zit, krijgt de leerling juist opgaven die een goed onderscheid maken tussen op- en boven-niveau (rechter blok in Figuur 12-2). Aan het eind van laag 2 (en laag 3 als er 4 lagen zijn) wordt dit proces herhaald. Aan het eind van de toets zijn de PMP's hetgeen gerapporteerd wordt: de toetsuitkomst. Aan het eind van de toets weten we hoe waarschijnlijk het is dat de leerling gegeven zijn of haar responspatroon bij de onder-niveaugroep hoort (PMP<), bij de op-niveaugroep hoort (PMP=) en bij de boven-niveaugroep (PMP>) hoort. De leerling wordt geplaatst in de niveaugroep die het meest waarschijnlijk is (de hoogste PMP heeft). Bijvoorbeeld, als de leerling het meest waarschijnlijk bij de op-niveaugroep hoort, krijgt de leerling de diagnose op niveau.

Hiermee is in de technische zin een toetsafname voltooid voor één hoofdaspect. Het is echter dan nog niet duidelijk op welke wijze we opgaven zouden moeten verdelen over de verschillende blokken. Om die vraag te kunnen beantwoorden maken we gebruik van een simulatieprocedure.

12.3.1 Simulatie

Door middel van simulatie is het mogelijk om te zoeken naar een zo goed mogelijk verdeling van de opgaven en responsen over de blokken. Allereerst is het nodig om te definiëren welke criteria we hanteren om de kwaliteit van een indeling te beoordelen. Er zijn twee criteria waarmee rekening is gehouden: (1) de grootte van de kansen op correcte diagnoses gegeven het ware niveau van iedere leerling en (2) de balans tussen de drie kansen op correcte diagnoses.

In de Tabel 12-1 worden de 9 mogelijke uitkomsten weergegeven van de toets, waarbij we onderscheid maken tussen de werkelijke niveaugroep van een leerling en de ingeschatte niveaugroep aan het eind van de toetsafname. In het hypothetische voorbeeld krijgt 88% van de onder-niveauleerlingen de juiste diagnose onder niveau, de rest krijgt een verkeerde diagnose: 9% de verkeerde diagnose op niveau en 3% de diagnose boven niveau. Van de op-niveauleerlingen wordt 87% correct gediagnosticeerd en van de boven-niveauleerlingen 91%.

Tabel 12-1. Voorbeeld van classificatietabel

	Geschat niveau		
Ware niveau	Onder	Op	boven
Onder	0,88	0,09	0,03
Op	0,07	0,87	0,06
Boven	0,01	0,08	0,91

Het eerste criterium, de kans op correcte diagnoses, houdt dus in dat we de som van de diagonaal van deze 3-bij-3 tabel proberen te maximaliseren, aangezien dit de uitkomsten zijn waarbij de ware en geschatte niveaugroep overeenkomen. De simulatie is ontwikkeld om zo goed mogelijk aan de twee criteria te voldoen en werkt als volgt:

1) De opgaven worden op willekeurige wijze verdeeld over de blokken.
    a. In het geval van bijvoorbeeld 3 lagen (zonder doortoetsen) zijn er 7 blokken, zoals weergegeven in Figuur 12-2. Als er 7 blokken moeten komen, worden de opgaven over 5 blokken verdeeld, omdat het middelste op-niveaublok in laag 2 en 3 wordt gevormd door opgaven uit onder-niveaublok en uit het boven-niveaublok. Hiervoor wordt per deelaspect willekeurig de helft van de opgaven genomen uit het onder-niveaublok en de helft uit het 'boven niveau' blok.
    b. Het aantal opgaven dat aan een blok toegewezen mag worden wordt beperkt door de beschikbare toetsduur. Bij aanvang is bekend wat de maximale toetsduur is die beschikbaar is per hoofdaspect. Bijvoorbeeld, als er drie uur beschikbaar is voor de hele afname en er zijn vier hoofdaspecten en zaai-opgaven, dan is er per hoofdaspect 2160 seconden beschikbaar (10800 secondes/5). De beschikbare toetsduur per hoofdaspect wordt gedeeld door het aantal blokken dat een leerling moet maken (lagen) om tot de beschikbare tijd per blok te komen. Vervolgens delen we de beschikbare tijd per te maken blok door de beschikbare antwoordtijd per opgave om tot een maximaal aantal opgaven per blok te komen. Bijvoorbeeld, bij een toetsduur per hoofdaspect van 2100 seconden, en

een design met 3 lagen met doortoetsen krijgt een leerling vier blokken: 2100/(4 blokken) = 525 seconden per blok, 525/(50 seconden antwoordtijd benodigd per opgave) = 10 opgaven, dus mag een leerling per blok maximaal 10 opgaven aangereikt krijgen. Het kan dus betekenen dat niet alle beschikbare opgaven daadwerkelijk in de toets voorkomen. De benodigde antwoordtijd per opgave werd uitgerekend op grond van de waargenomen responstijden tijdens de voorgaande afname.

2) Voor de blokindeling worden 10.000 gesimuleerde leerlingen aangemaakt die onder niveau zitten, 10.000 leerlingen die op niveau zitten en 10.000 leerlingen die boven niveau zitten.
   Deze gesimuleerde leerlingen doorlopen de toets met de gecreëerde blokindeling. Er worden antwoorden op de opgaven gegenereerd die de gesimuleerde leerlingen volgens de blokindeling krijgen, op basis van de succeskansen conditioneel op de niveaugroep. Aan het einde van elke laag wordt de niveaugroep van de gesimuleerde leerling geschat. Aan het eind van de toets is de diagnose bekend en blijkt in welke van de negen mogelijke uitkomstgroepen de gesimuleerde leerling is terecht gekomen (in de classificatietabel, zie Tabel 12-1); de frequentie van de betreffende uitkomst in de 3-bij-3 tabel wordt met één opgehoogd.
3) Deze procedure van blokindelingen maken en simuleren loopt door totdat er 100 blokindelingen gevonden zijn die voldoen aan de eisen. Er zijn drie restricties:
   a. De tijdsrestrictie
   b. De eis dat ieder blok minimaal één respons moet hebben voor elk deelaspect
   c. De eis dat de aantallen responsen voldoende evenredig verdeeld zijn over de blokken
4) Van deze 100 indelingen wordt uitgerekend wat de som is van de drie kansen op correcte diagnoses en worden de 10 indelingen aangehouden die de hoogste som van de diagonaal hebben.
5) Tot slot worden de priors geoptimaliseerd om de verschillen in grootte van de kansen op correcte diagnoses te minimaliseren (zie Sectie 12.4).

Deze tien beste oplossingen worden teruggekoppeld en toetsdeskundigen bepalen in samenspraak met de afdeling psychometrie welke oplossing de voorkeur heeft (zie verder Hoofdstuk 13).

Opgavegroepen in de blokindeling

Soms moet bij de optimalisaties rekening gehouden worden met de aanwezigheid van opgavegroepen. Dit zijn opgaven die op verschillende schermen gepresenteerd worden en die bij elkaar horen, bijvoorbeeld drie opgaven die over dezelfde tekst gaan. Dit betekent dat dat deze opgaven verplicht tegelijkertijd aangeboden moeten worden in een specifieke volgorde. Het zoekalgoritme om een optimale verdeling van opgaven te vinden houdt rekening met opgavegroepen door van tevoren dergelijke opgaven samen te bundelen in een tijdelijke structuur. Als we twee opgaven hebben met ieder één respons en deze opgaven behoren tot dezelfde opgavegroep, dan maakt de simulatie er tijdelijk één opgave van met twee responsen. Deze kwestie speelt overigens alleen in het deel van de simulatie waar bekeken wordt of willekeurige blokindelingen voldoen aan de verschillende restricties. Als er eenmaal een blokindeling is gevonden die voldoet aan de restricties dan worden deze bundels weer teruggebracht naar hun originele vorm, maar dan hebben we wel de garantie dat de betreffende opgaven direct achter elkaar zijn geplaatst in hetzelfde blok.

12.3.2 Uitstapkansen

In het adaptieve ontwerp is er ook rekening mee gehouden dat een leerling niet alle blokken opgaven hoeft te maken als de diagnose al heel zeker is, bijvoorbeeld als de leerling de opgaven consistent heel goed of heel slecht maakt (zie verder Hoofdstuk 13). De leerling mag als het ware uit de toets stappen.
Een uitstapkans is een criterium waarmee per niveau aan het einde van elke laag bekeken kan worden of een leerling kan uitstappen omdat er voldoende zekerheid is omtrent diens niveau. Deze uitstapkans zou berekend kunnen worden voor elke laag apart, maar bij de technische opzet van de adaptieve module is gekozen voor één set aan uitstapkansen per hoofdaspect (en één set voor elk deelaspect). De set uitstapkansen wordt berekend op basis van de opgaven in de eerste laag, omdat dit de meest conservatieve waarden oplevert. Deze uitstapkansen worden vervolgens na elke laag toegepast.

Bij het bepalen van de uitstapkansen maken we allereerst de aanname dat items en testlets in een diagnostische tussentijdse toets -toets onderscheidend vermogen hebben. De mate van onderscheidend vermogen varieert van item tot item en van testlet tot testlet. Desalniettemin geldt dat elke additionele

opgave die we de leerlingen voorleggen, de zekerheid omtrent het ingeschatte niveau verhoogt (tenzij de succeskansen dus gelijk zijn voor de verschillende niveaus, maar dit kunnen we bij de kalibratie detecteren en komt hooguit voor een zeer beperkt aantal items voor).

Onder het DTT-model geldt dat hoe groter het aantal opgaven is dat een leerling maakt, des te groter de kans is dat er een correcte diagnose plaatsvindt (i.e., het geschatte niveau is gelijk aan het ware niveau van een leerling). Theoretisch gezien is het zo dat de kans op correcte diagnose naar één gaat als het aantal opgaven naar oneindig gaat, aangezien de invloed van toeval onder die omstandigheden meer en meer wegvalt. In de praktijk hebben we natuurlijk een beperkt aantal opgaven en is de invloed van toeval groter. We komen dus niet zo snel tot een toets die perfect onderscheid maakt.

Ook al maakt de toets geen perfect onderscheid, het observeren van een lichtelijk afwijkend responspatroon (gegeven het niveau van een leerling) is waarschijnlijker dan het observeren van een zeer afwijkend responspatroon. Oftewel, de kans dat een leerling uit de onder-niveaugroep één heel moeilijke opgave toevallig toch correct beantwoord (door gokken of door een moment van extra inzicht) is groter dan de kans dat diezelfde leerling tien heel moeilijke opgaven correct weet te beantwoorden. Dit betekent ook in het algemeen, als er een verkeerde diagnose gemaakt wordt, dat het verschil tussen de waarschijnlijkheid (PMP) van de ingeschatte niveaugroep en de waarschijnlijkheid (PMP) van de ware niveaugroep doorgaans relatief klein is. Bijvoorbeeld, als een onder-niveauleerling de verkeerde diagnose op niveau krijgt is het waarschijnlijker dat de toetsuitkomst van die leerling [$PMP_{<}$ = 0,3; $PMP_{=}$ = 0,5; $PMP_{>}$ = 0,2] is, dan dat de toetsuitkomst [$PMP_{<}$ = 0,1; $PMP_{=}$ = 0,2; $PMP_{>}$ = 0,7] is. We kunnen dan ook concluderen dat er gemiddeld genomen een hogere proportie aan incorrecte diagnoses voor komt in de verzameling van PMP-vectoren die weinig afwijken van een uniforme verdeling [1/3, 1/3, 1/3] en dat we minder fouten zullen aantreffen onder de leerlingen met PMP-vectoren die sterk neigen naar één niveau, bijvoorbeeld [0,7, 0,2, 0,1].

De regel is geïmplementeerd dat er alleen vroegtijdig uitgestapt mag worden als het percentage correcte diagnoses in de groep uitstappers 95% of hoger is (zie Hoofdstuk 13) en uitstappen mag op zijn vroegst na het volledig doorlopen van de eerste laag. Om een optimale blokindeling te vinden in de simulatieprocedure, wordt voor elk van de drie niveaugroepen uitgerekend wanneer voldaan kan worden aan deze 95% eis. De gesimuleerde leerlingen—die bijvoorbeeld als onder niveau worden ingeschat—hebben in ieder geval een PMP voor dit niveau groter dan 1/3 en een PMP kan oplopen tot (praktisch) 1,0. Zoals eerder genoemd is het de verwachting dat er meer foute diagnoses voorkomen onder de leerlingen wiens PMP voor ‘onder niveau’ maar nét de grootste is. Om die reden worden de leerlingen die ingeschat worden op ‘onder niveau’ geordend op basis van hun PMP(<) van klein naar groot. Het is in principe mogelijk dat het aantal foute diagnoses direct al lager is dan 5%, als bijvoorbeeld de eerste laag opgaven bevat die zeer goed onderscheid kunnen maken tussen onder- en op-niveau en tussen op-niveau en boven-niveau. In dat geval wordt de uitstapkans gelijkgesteld aan ,01 en mag dus vrijwel iedereen uitstappen na de eerste laag als ze gediagnosticeerd zijn als onder niveau. Het is ook mogelijk dat de eis van 95% helemaal niet gehaald kan worden door minder of minder goede opgaven; de uitstapkans wordt dan op ,99 gezet. Voor de gevallen tussen deze twee extremen in wordt in de simulatieprocedure de minimale PMP steeds wat opgehoogd totdat de groep met PMP’s boven een specifieke PMP-waarde minimaal voor 95% bestaat uit correct gediagnosticeerde leerlingen. Wanneer er voldaan is aan de 95% eis wordt de uitstapkans gelijkgesteld aan deze PMP.

12.4 Optimalisatie van de priors

Zoals eerder al kort genoemd zijn priors de gewichten die gebruikt kunnen worden om tendensen in de kansen op correcte diagnoses bij te sturen. We hebben bij de DTT drie niveaus en dus drie prior gewichten. Als we eenmaal een bepaalde blokindeling gevonden hebben in de simulatieprocedure dan weten we in ieder geval dat de som van de kansen op correcte diagnoses het hoogst was voor deze indeling (binnen de set van onderzochte blokindelingen). Dit garandeert echter niet dat de kans op correcte diagnoses voor de drie niveaus even groot zijn. De verschillen tussen deze kansen zijn meestal niet bijzonder groot omdat er doorgaans vele alternatieve indelingen zijn die deze afwijking niet vertonen en in het zoekalgoritme verdringen deze alternatieve indelingen eventuele indelingen die mankementen vertonen.

Doordat we echter te maken hebben met discrete eenheden die we moeten verdelen (items en testlets) lukt het nooit om de drie kansen op correcte diagnoses exact gelijk te krijgen. Om dit alsnog zo goed mogelijk te bewerkstelligen gebruiken we priors. Voor de 10 beste oplossingen worden de volgende stappen gezet om de priors te optimaliseren en de kansen op correcte classificatie zo gelijk mogelijk te krijgen:

(1) neem van een indeling de kansen op correcte classificatie, $pr_<, pr_= \text{ en } pr_>$, en bereken de vector **m** als volgt: $\mathbf{m} = \left(pr_>, \frac{pr_>}{pr_=/pr_<}, pr_<\right)$
(2) Normaliseer vector **m** door elk van de drie elementen te delen door de som van **m**
(3) Vermenigvuldig de genormaliseerde vector **m** met de oude priors en normaliseer het resultaat opnieuw om tot een nieuwe set aan priors te komen
(4) herbekeken $pr_<, pr_= \text{ en } pr_>$, gegeven de nieuwe priors en ga weer naar stap 1 zolang de aanpassing van de priors in een iteratie groter is dan de ingestelde ondergrens.

12.5 Literatuur

Cito (2015). *Diagnostische tussentijdse toets: Verslag pretest 2015*. Arnhem: Cito.

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016*. Arnhem: Cito.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Hoijtink, H., & Sies, A. (2013). De psychometrie van de diagnostisch tussentijdse toets. In S. Schouwstra (Red.), *De diagnostisch tussentijdse toets: onderzoek 2013* (pp. 13-62). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2014). Adaptieve procedure met itemblokken. In S. Schouwstra (Red.), *De diagnostisch tussentijdse toets: onderzoek 2014* (pp. 33-40). Arnhem: Cito.

Hoijtink, H., Béland, S., & Vermeulen, J. A. (2014). Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychological Methods*, *19*(1), 21-38.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 361-412). Princeton: Princeton University Press.

Maas, L. (2017). *Potential impact of item parameter drift on diagnostic accuracy in educational testing*. [ongepubliceerde master thesis]. Utrecht: Utrecht University.

Roelofs, E., & Schouwstra, S. (Red.). (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie*. Arnhem: Cito.

Sies, A.M.A, (2014). *Cognitive diagnostic testing using calibrated hypotheses* [ongepubliceerde master thesis]. Utrecht: Utrecht University.