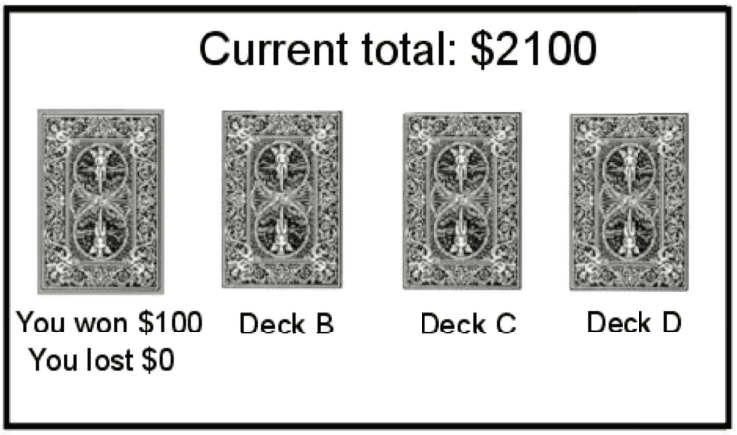
Safe models for risky decisions

Steingröver, H.M.

Nederlandse Samenvatting

Hoe kiezen mensen tussen twee opties die verschillen in hun onmiddellijke en langdurige consequenties? Zou je bijvoorbeeld voor een heerlijk stukje taart gaan of eerder voor een appel? Het stukje taart vind je nu misschien lekkerder, maar als je daar vaker voor kiest, zou dat op lange termijn slecht voor je gezondheid kunnen zijn. In dit proefschrift, getiteld "Risicovolle Beslissingen Veilig Modelleren", heb ik onderzocht hoe mensen zogenoemde risicovolle beslissingen nemen. Hiervoor heb ik gebruik gemaakt van data van de Iowa gambling taak (IGT; Bechara et al., 1994). Deze taak bestaat uit vier speelstapels (Figuur 12.1). Deelnemers worden gevraagd om herhaald kaarten te kiezen van de vier verschillende stapels met als doel een zo hoog mogelijke netto-uitkomst te realiseren. De taak is zodanig ontwikkeld dat dit doel alleen maar bereikt kan worden als je de goede stapels de voorkeur geeft en de slechte stapels vermijdt. Wat karakteriseert de goede en slechte stapels? Om de verschillen tussen de vier stapels te verduidelijken heb ik in Tabel 12.1 de winsten en verliezen van tien voorbeeldkeuzes van elke stapel gegeven. Het wordt duidelijk dat de slechte stapels elke keer een hoge winst geven, maar af en toe een heel hoog verlies; daarom zijn deze stapels op lange termijn ongunstig. De twee andere stapels –de goede stapels– geven elke keer een redelijk lage winst, maar ze zijn op lange termijn gunstig omdat hun onregelmatige verliezen heel laag zijn.

In dit proefschrift stelde ik dat descriptieve analyses van IGT data ons alleen maar kunnen



Figuur 12.1: Voorbeeld van een schermafbeelding van de IGT.

Tabel 12.1: *Winsten en verliezen van 10 voorbeeldkeuzes van de vier verschillende stapels van de IGT. De laatste rij geeft voor elke stapel de netto-uitkomst weer na 10 keuzes.*

	Slechte stapels		Goede stapels	
Keuze	Stapel A	Stapel B	Stapel C	Stapel D
1	100	100	50	50
2	100, –150	100	50, –50	50
3	100	100	50	50
4	100, –350	100	50, –50	50
5	100, –300	100, –1250	50	50
6	100	100	50, –50	50
7	100, –250	100	50, –50	50
8	100	100	50	50
9	100, –200	100	50	50, –250
10	100	100	50, –50	50
Netto-uitkomst	–250	–250	250	250

vertellen *welke keuzes* mensen hebben gemaakt en hoe hun keuzegedrag verandert naarmate de taak voortschrijdt. Echter, om erachter te komen *hoe* mensen de keuzes hebben gemaakt, d.w.z. welke psychologische processen hun keuzes hebben bepaald, moeten we cognitieve modellen gebruiken. Deze modellen maken aannames over de relevante psychologische processen, zoals motivatie (bijv. winsten als net zo belangrijk waarnemen als verliezen), geheugen (bijv. het vermogen eerdere uitkomsten van de stapels te herinneren), en responsconsistentie (bijv. de neiging beslissingen te baseren op verwachtingen van de stapels) en hoe deze processen interacteren om tot uiting te komen in het geobserveerde keuzegedrag. In de context van de IGT zijn zogenoemde *reinforcement-learning* (RL) modellen heel geschikt (Sutton & Barto, 1998). Deze modellen verklaren hoe mensen door *trial and error* hun beslissingen vormen.

De reden waarom de modellen toepasbaar zijn op verschillende patronen in de data is dat de modellen parameters hebben die kunnen variëren. Deze modelparameters staan voor de psychologische processen die verondersteld worden belangrijk te zijn voor risicovolle besluitvorming. Een voorbeeld is een modelparameter die beschrijft hoe belangrijk mensen onmiddellijke winsten achten. Een lage waarde van de parameter betekent dat mensen onmiddellijke winsten niet belangrijk vinden, terwijl een hoge parameterwaarde betekent dat mensen onmiddellijke winsten heel belangrijk vinden. Op deze manier kunnen de modellen gebruikt worden om erachter te komen waarom verschillende groepen verschillende keuzes maken. Hier hebben we echter parameterschattingen nodig voor de verschillende groepen. Om deze te verkrijgen moeten we de modellen passen op de data (zgn. *model fitting*).

In de klinische psychologie worden de IGT samen met RL modellen vaak gebruikt om erachter te komen of een klinische groep suboptimale beslissingen neemt, en zo ja, hoe dit verklaard kan worden. Voor dat soort toepassingen is het dus van groot belang dat de IGT daadwerkelijk risicovolle besluitvorming meet en dat de modellen de relevante processen correct beschrijven. In dit proefschrift beweer ik dat de huidige manier om IGT data te analyseren, de huidige manier om RL modellen toe te passen en de huidige manier om een geschikt RL model te vinden (zgn. modelselectie) gekenmerkt worden door veel problemen. Deze problemen leiden ertoe dat er vaak conclusies worden getrokken die voorbarig zijn. Om dit in de toekomst te voorkomen en om conclusies te trekken die sterker gebaseerd zijn op de data stel ik een aantal maatregels voor. Deze regels hebben betrekking op de volgende gebieden: (1) data-analyse; (2) modelselectie; (3) *model fitting*; en (4) beoordeling van de *model fit*.

In dit proefschrift beweer ik dat elk van de vier zojuist genoemde gebieden enorm kan profiteren

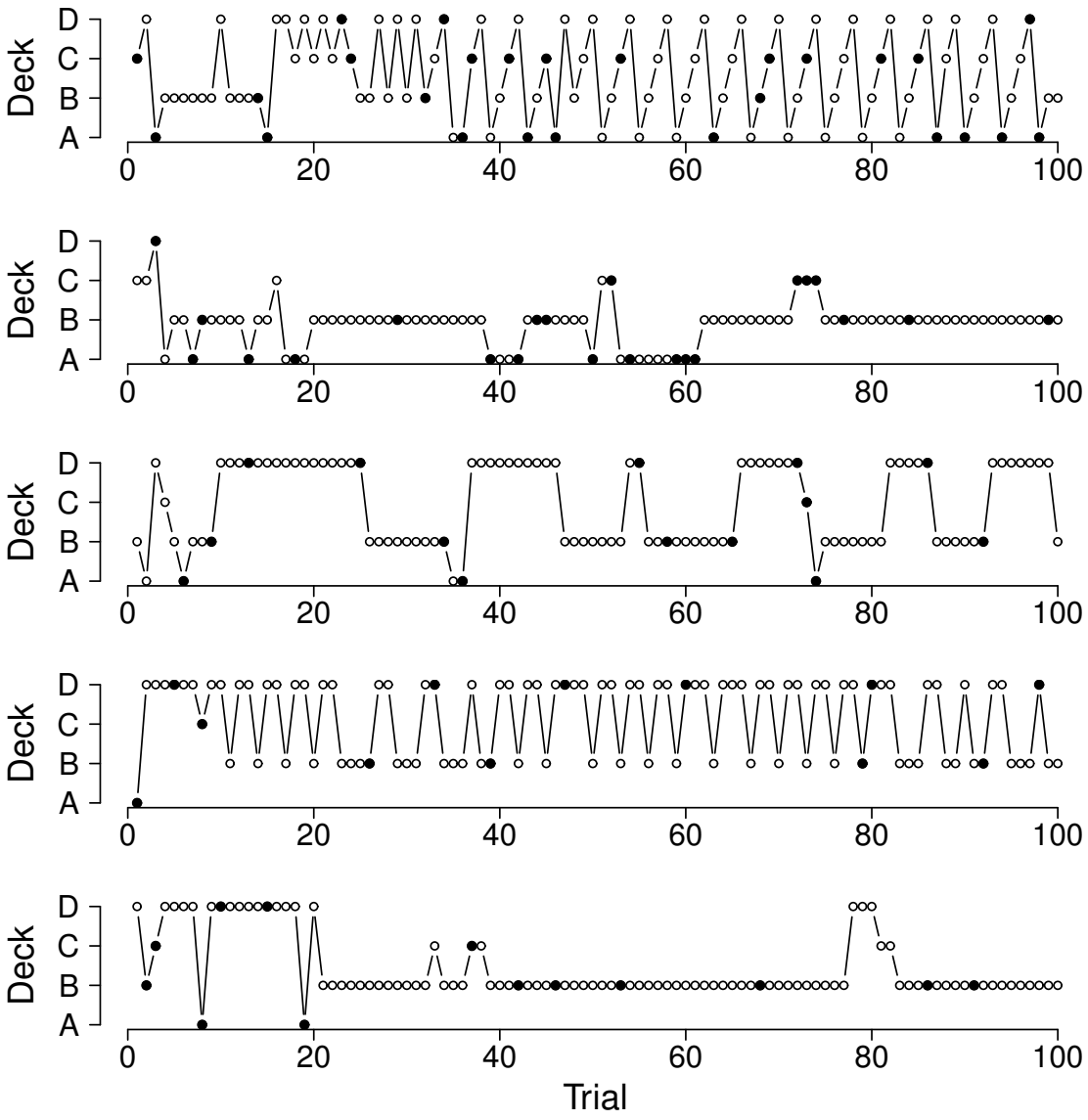
van het gebruik van de Bayesiaanse statistiek in plaats van de klassieke, of de zogenoemde frequentistische statistiek. De Bayesiaanse statistiek heeft de laatste jaren sterk aan populariteit gewonnen en is zeer geschikt voor problemen in de mathematische psychologie (Andrews & Baguley, 2013; Bayarri et al., 2016; Poirier, 2006; Vanpaemel, 2016; Verhagen et al., 2015; Wetzels et al., 2016). In de rest van deze samenvatting zal ik de belangrijkste resultaten van mijn proefschrift samenvatten.

Het doel van hoofdstuk 2 was te verduidelijken dat de huidige manier waarop IGT data wordt geanalyseerd het zogenoemde *frequency-of-losses* effect verbergt. Het *frequency-of-losses* effect staat voor de bevinding dat het gezonde proefpersonen vaak niet lukt een voorkeur te ontwikkelen voor de goede stapels (d.w.z. stapels C en D). Ze leren uiteindelijk enkel de stapels met regelmatige verliezen (d.w.z. stapels A en C) te vermijden. Dit keuzegedrag zorgt echter niet voor maximalisatie van de netto-uitkomst en duidt volgens de oorspronkelijke interpretatie van de IGT op suboptimale besluitvorming. In hoofdstuk 2 heb ik nog twee andere bevindingen gerapporteerd die aantonen dat belangrijke veronderstellingen over het keuzegedrag van gezonde proefpersonen niet door de data ondersteund kunnen worden. Ik heb namelijk aangetoond dat gezonde mensen (1) idiosyncratisch keuzegedrag vertonen (zie Figuur 12.2); en (2) niet een bepaalde systematiek aanhouden waarbij ze eerst de verschillende stapels verkennen en daarna de beste stapels plunderen. Deze bevindingen stellen de heersende interpretatie van IGT data in twijfel en suggereren dat er bij toekomstige toepassingen van de IGT bijzondere aandacht besteed moet worden aan de belangrijkste veronderstellingen over het keuzegedrag van gezonde mensen. In het bijzonder is het belangrijk dat data van elke stapel apart geanalyseerd worden en opgesplitst in meerdere blokken van trials. In hoofdstuk 9 heb ik bovendien voorgesteld een Bayesiaanse herhaalde metingen variantieanalyse (ANOVA) te gebruiken. Met zo'n ANOVA kan je bijvoorbeeld onderzoeken of proefpersonen meer goede keuzes maken naarmate de taak voortschrijdt, en of er verschillen zijn in de voorkeuren voor de goede stapels tussen twee verschillende groepen. Een groot voordeel van de Bayesiaanse ANOVA is dat deze ook gebruikt kan worden om de bewijskracht van de data te bepalen voor de nulhypothese die ervan uitgaat dat er geen effect is (bijv. geen groepseffect of geen blokeffect).

Op grond van de empirische relevantie van het *frequency-of-losses* effect (d.w.z. een voorkeur hebben voor stapels B en D) in IGT data van gezonde proefpersonen wilde ik vervolgens in hoofdstuk 3 onderzoeken of RL modellen dit datapatroon überhaupt kunnen genereren. Dit is belangrijk omdat we, als we modelparameters interpreteren als representatie van de psychologische processen, zeker moeten weten dat het model goed past op de data. Maar als het model het keuzegedrag van de data überhaupt niet kan genereren, kunnen we ervan uitgaan dat het model niet op de data kan passen. In zo'n situatie kunnen we de parameterschattingen beter niet gebruiken om uitspraken te doen over de onderliggende psychologische processen. De methode die ik hiervoor gebruikt heb heet *parameter space partitioning* (PSP).

De resultaten van hoofdstuk 3 suggereren dat drie bekende RL modellen –het *Expectancy Valence* model (EV; Busemeyer & Stout, 2002), het *Prospect Valence Learning* model (PVL; Ahn et al., 2008), en een combinatie van deze twee modellen, het EV-PU model– geen duidelijke stapelvoorkeuren kunnen genereren, en dat het EV model helemaal geen *frequency-of-losses* effect kan genereren. Samengenomen suggereren deze resultaten dat de drie beschouwde modellen alleen een voldoende data-passend potentieel hebben voor een beperkt aantal keuzepatronen, en dat er nog een model ontbreekt dat universeel gebruikt kan worden.

Daarna heb ik in hoofdstuk 4 laten zien dat het PVL-Delta model –een RL model dat recentelijk veel aandacht heeft gekregen en ook een combinatie is van het EV en PVL model (Ahn et al., 2008; Fridberg et al., 2010; Steingroever et al., 2014)– een beter data-passend potentieel heeft dan de modellen uit hoofdstuk 3. Maar ook dit model kan niet universeel gebruikt worden omdat het geen



Figuur 12.2: Keuzes van vijf proefpersonen uit ons databestand die duiden op de hoge intra-groep variabiliteit in de IGT prestatie van gezonde proefpersonen. De zwarte bolletjes wijzen op het optreden van zowel winsten als verliezen; de witte bolletjes wijzen op het optreden van alleen maar winsten. De data zijn gepubliceerd in Steingroever, Wetzels, Horstmann, et al. (2013).

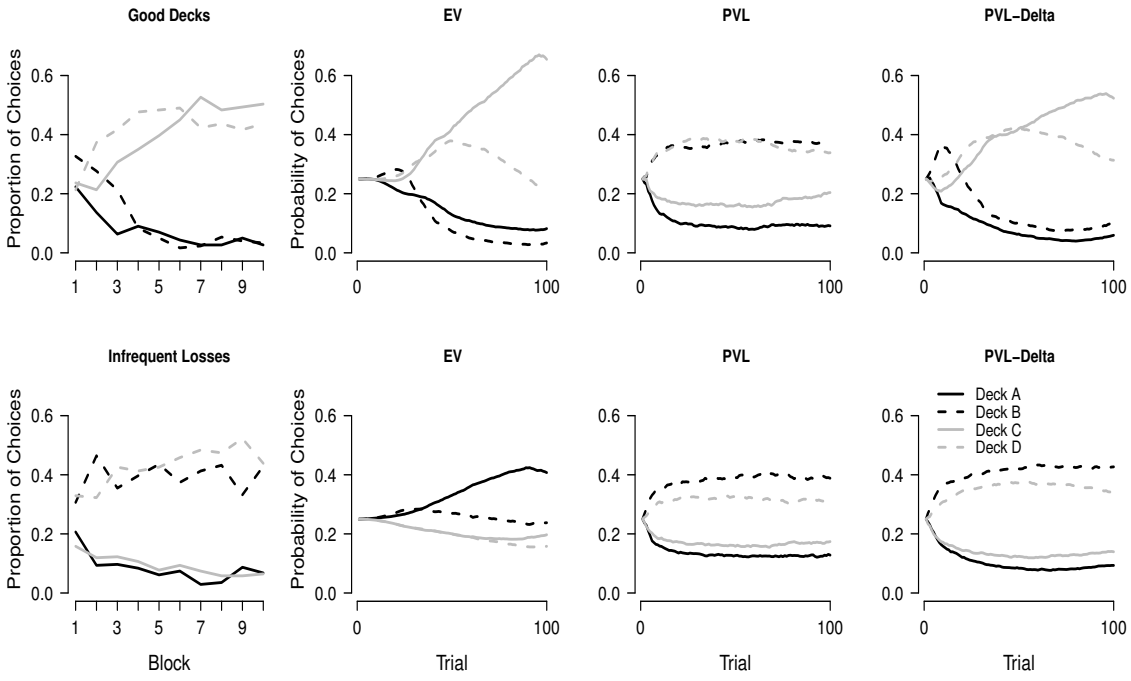
duidelijke voorkeur voor de slechte stapels kan genereren—een keuzepatroon dat voorondersteld wordt voor mensen met suboptimale besluitvorming.

Naast de PSP methode hebben ik een aantal andere methodes besproken die geschikt zijn voor modelselectie. Ten eerste is het belangrijk dat modellen parameters accuraat terugschatten (zie hoofdstukken 4 en 6). Deze eigenschap kan onderzocht worden met zogenoemde *parameter recovery* studies. Voor dit soort studies genereren we data met eigen gekozen parameterwaardes en met het model onder beschouwing. Vervolgens passen we het model op de gegenereerde dataset om parameterschattingen te verkrijgen. Als de parameterschattingen overeenkomen met de waardes die we voor de datageneratie hebben gebruikt, spreken we van een accurate *parameter recovery*. Ten tweede, om te garanderen dat de modelparameters ook daadwerkelijk voor de veronderstelde psychologische processen staan, is het belangrijk dat het model goed scoort op een zogenoemde test van selectieve invloed (zie hoofdstuk 4). Zo'n test onderzoekt of experimentele manipulaties die bedoeld zijn om alleen maar één modelparameter te beïnvloeden zich weerspiegelen in de parameterschattingen. Ten derde heb ik in hoofdstukken 7, 8, en 9 modellen vergeleken met behulp van de Bayes factor. De Bayes factor is in de Bayesiaanse statistiek het standaard middel om modellen te vergelijken. Maar deze was nog niet eerder toegepast in RL modellen voor de IGT omdat de Bayes factor wiskundig gezien moeilijk te verkrijgen is. De relevante code is online beschikbaar waardoor ik hoop dat de Bayes factor in de toekomst vaker gebruikt wordt om modellen te vergelijken.

Om modellen te passen hebben ik gebruik gemaakt van het Bayesiaanse hierarchische raamwerk (hoofdstukken 4, 5, 8, en 10). Dit raamwerk leidt tot nauwkeurigere en informatievere schattingen dan alternatieve methoden omdat het raamwerk gebruik maakt van zowel de verschillen als gemeenschappelijkheden van proefpersonen van een groep (Ahn et al., 2011; Horn et al., 2015; Lejarraga et al., 2016; Navarro et al., 2006; Rouder & Lu, 2005; Rouder et al., 2005, 2008; Scheibehenne & Pachur, 2015; Shiffrin et al., 2008; Wetzels, Vandekerckhove, et al., 2010).

Om tenslotte te beoordelen hoe goed een model op de data past, heb ik laten zien dat het belangrijk is de absolute *model fit* te beoordelen (hoofdstukken 4 – 6, en 9). Dit is in strijd met de huidige manier waarin alleen maar de relatieve *model fit* wordt beschouwd—een methode die bepaalt hoeveel beter het RL model onder beschouwing op de data past dan een eenvoudig referentiemodel. Dit geeft een getal dat helaas moeilijk te interpreteren is. Dit getal zegt alleen of het RL model beter past, maar niet hoeveel beter en ook niet of het model überhaupt voldoende op de data past (bijv. of het model kwalitatief hetzelfde datapatroon voorspelt als door de data vertoond wordt). Het zou dus kunnen dat het RL model onder beschouwing volgens de relatieve index beter past op de data dan het referentiemodel, maar eigenlijk passen beide modellen redelijk slecht. Om te bepalen of het RL model daadwerkelijk op de data past, is het daarom beter om te kijken of data gegenereerd met de parameterschattingen hetzelfde keuzegedrag opleveren als de data zelf. Alleen in deze situatie kunnen we ervan uitgaan dat de parameterschattingen voor de relevante psychologische processen staan.

Dat het zo belangrijk is de absolute *model fit* te bekijken heb ik in hoofdstuk 5 verduidelijkt met behulp van de EV, PVL, en PVL-Delta modellen. Hiervoor hebben ik gebruik gemaakt van twee gestileerde datasets en vijf gepubliceerde datasets. Ik heb eerst de modellen op de data gepast, en vervolgens heb ik met de parameterschattingen data gegenereerd. De geobserveerde data van de twee gestileerde datasets staan in de eerste kolom van Figuur 12.3 en de voorspellingen van de drie modellen in kolommen 2 – 4. Dit figuur verduidelijkt dat alleen het PVL-Delta model beide datapatronen kwalitatief correct voorspelt. In het geval van het EV model en de gestileerde dataset met een voorkeur voor de stapels met incidentele verliezen, zien we bijvoorbeeld dat het EV model een voorkeur voor de slechte stapels voorspelt, dus een kwalitatief verschillend datapatroon. Dit suggereert dat we de parameterschattingen beter niet kunnen gebruiken om conclusies te



Figuur 12.3: Voorspellingen van de drie RL modellen met betrekking tot de twee gestileerde datasets. De eerste kolom laat de geobserveerde gemiddelde keuzeproporties zien van elke stapel in 10 blokken. Elk blok bevat 10 trials. De tweede, derde, en vierde kolom tonen de gemiddelde keuzeproporties van elke stapel gegenereerd door respectievelijk het EV, PVL, en PVL-Delta model.

trekken over de psychologische processen. Om voorbarige conclusies te voorkomen in toekomstige toepassingen van RL modellen op IGT data is het daarom van groot belang dat de absolute fit van de modellen zorgvuldig wordt bekeken.

Als we de hierboven genoemde belangrijke stappen nastreven komen we hopelijk dichter bij het ambitieuze doel de IGT en cognitieve modellen te gebruiken om suboptimale besluitvorming te diagnosticeren en beter te begrijpen.