ARTIKELEN

Studiesucces of -falen van eerstejaarsstudenten voorspellen: een nieuwe aanpak

Peter G.M. van der Heijden, David J. Hessen & Theo Wubbels[*]

In dit methodologische artikel wordt een nieuwe aanpak gepresenteerd om te kijken naar studiesucces. De responsvariabele 'studiesucces' wordt gecodeerd als een ordinale variabele met de categorieën (i) vertrek in jaar 1, (ii) vertrek na jaar 1 dan wel langer studeren dan 4 jaar, (iii) diploma in jaar 4 en (iv) diploma in jaar 3. Studiesucces wordt voorspeld met ordinale regressie, waarbij gebruik wordt gemaakt van gegevens beschikbaar in het studievoortgangsysteem van de Faculteit Sociale Wetenschappen (FSW) van de Universiteit Utrecht. In de presentatie van de resultaten wordt de nadruk gelegd op het interpreteren van de kansen dat een student in een van de vier categorieën van de variabele 'Studiesucces' terechtkomt. Door de tijd heen worden analyses verricht, waarbij de tentamenresultaten een steeds belangrijkere rol blijken te krijgen. Er wordt een Excel-programma gepresenteerd waarmee de kansen van een individuele student eenvoudig kunnen worden bepaald door zijn gegevens in te voeren. De gepresenteerde methodiek is eenvoudig door andere instellingen in het hoger onderwijs te implementeren.

Inleiding

Op vele plekken in Nederland en in de wereld wordt gekeken naar de voorspelling van studiesucces in het wetenschappelijk onderwijs. Dit gebeurt zowel in de wereld van het onderzoek van het hoger onderwijs (zie bijvoorbeeld Annema & Ooijvaar, 2011; Bruinsma & Jansen, 2009; Cassidy, 2011; Jansen & Bruinsma, 2005; Meeuwissen, Van Wensveen & Severiens, 2011; De Koning & Loyens, 2011; Torenbeek, Suhre, Jansen & Bruinsma, 2011; Visser, 2011; Willcoxson, 2010) als in de wereld van de *institutional research* (zie bijvoorbeeld Brogt, Sampson, Comer, Turnbull & McIntosh, 2011; Campbell, 2007; Whitmer, Fernandes & Allen, 2012; Yu, DiGangi, Jannasch-Pennell & Kaprolet, 2010). In deze bijdrage wordt in dit door velen betreden onderzoeksgebied een nieuwe methodische aanpak uiteengezet, die relatief eenvoudig is te implementeren. De nieuwe aanpak wordt toegelicht met statistische analyses voor de Faculteit Sociale Wetenschappen van de Universiteit Utrecht.

[*] Prof. dr. P.G.M. van der Heijden (p.g.m.vanderheijden@uu.nl) is werkzaam bij de afdeling Methoden en Technieken, Faculteit Sociale Wetenschappen, Universiteit Utrecht en bij de onderzoeksafdeling S3RI van de University of Southampton. Dr. D.J. Hessen is werkzaam bij de afdeling Methoden en Technieken, FSW, Universiteit Utrecht. Prof. dr. T. Wubbels is werkzaam bij het Facultair Management Team van de FSW, Universiteit Utrecht.

Het oorspronkelijke doel was om studiesucces in de universitaire bacheloropleiding te voorspellen op basis van eenvoudig beschikbare gegevens uit de studentenadministratie. Vanuit de literatuur komen daarvoor data over geslacht, leeftijd, en eerder verworven kennis en vaardigheid in aanmerking (zie bijvoorbeeld Bruinsma & Jansen, 2009; Onzenoort, 2009). Daarbij werd in een eerste fase gekeken naar het 'langstuderen', dat wil zeggen naar studenten die nog in het vijfde studiejaar staan ingeschreven omdat zij in de voorgaande jaren geen diploma hebben gehaald. Dit 'langstuderen' werd op verschillende momenten in de studie voorspeld op basis van achtergrondkenmerken van studenten, zoals geslacht, eindexamencijfers en de tot dan toe behaalde studieresultaten. Gaandeweg werden ook data over studenten die niet 'langstudeerden' in het project betrokken, namelijk zij die uitvallen in het eerste jaar of later en zij die in drie of vier jaar het diploma behalen. Aan het uitgangspunt geen andere gegevens te gebruiken dan die reeds beschikbaar zijn in de studentenadministratie, werd vastgehouden.

De methodische vernieuwing van de voorgestelde aanpak bestaat uit een aantal onderling verbonden elementen, waarbij niet ieder element uniek is, maar het geheel een nuttig en makkelijk toepasbaar instrumentarium oplevert:

(i) Er wordt slechts gebruik gemaakt van variabelen die beschikbaar zijn in een studievoortgangsysteem. Men spreekt bij dit soort toepassingen wel van *institutional research*. Een voordeel van dit type toepassingen is dat men de gegevens van volledige cohorten kan gebruiken en niet slechts van een (soms niet-representatieve) steekproef. Een ander voordeel is dat deze werkwijze goedkoop is, omdat er geen data verzameld hoeven te worden. Een nadeel van *institutional research* is natuurlijk dat psychologische variabelen zoals leerstrategieën en persoonlijkheidseigenschappen niet meegenomen worden.

(ii) De uitkomstmaat is studiesucces of -falen. Waar in veel onderzoek als operationalisering van studiesucces kwantitatieve variabelen als 'aantal behaalde studiepunten' en/of 'het gemiddelde tentamencijfer' (zie bijvoorbeeld De Koning & Loyens, 2011; Torenbeek et al., 2011; Meeuwisse et al., 2011), dan wel dichotome variabelen als 'drop-out' of het 'behalen van de propedeuse' worden gehanteerd (bijvoorbeeld Bruinsma & Jansen, 2009), is hier op basis van het studievoortgangsysteem een ordinale variabele gecreëerd die beleidsmatig van belang lijkt. De categorieën zijn:
   - binnen één jaar met de studie stoppen;
   - later dan in het eerste studiejaar zonder diploma met de studie stoppen dan wel 'langstuderen';
   - een diploma behalen in het vierde studiejaar;
   - een diploma behalen in het derde studiejaar.

   Deze variabele wordt verder 'studiesucces' genoemd, waarbij zij aangetekend dat een deel van de categorieën gebrek aan succes beschrijven. Een andere ordinale uitkomstmaat is recent gebruikt door Annema & Ooijevaar (2011), die voor het examencohort 1998-1999 het behalen van een hogeronderwijsdiploma binnen 4 jaar, binnen 5 tot 8 jaar, of (nog) niet hanteren.

(iii) De uitkomstmaat 'studiesucces' wordt voorspeld met ordinale regressie, een generalisatie van logistische regressie (Annema & Ooijevaar gebruiken hier

het verwante multinomiale regressiemodel). Een van de resultaten van de analyse, gegeven de waarden op de verklarende variabelen (geslacht, gemiddelde examencijfers, enzovoort), bestaat uit de kansen van studenten om in een van de vier categorieën terecht te komen. Er kan hierdoor per verklarende variabele bekeken worden of deze een significante bijdrage levert aan de voorspelling van de studiesuccescategorieën. Als er sprake is van een significante bijdrage, dan wordt het interessant te bezien hoe de kansen voor studenten om in een bepaalde categorie te vallen veranderen als de waarden op de verklarende variabele veranderen. Juist de verandering van de kansen geeft een helder inzicht in het belang van individuele variabelen (gecorrigeerd voor de andere variabelen in het model). In de presentatie van de resultaten speelt de interpretatie van het model in termen van kansen een centrale rol (vergelijk Bruinsma & Jansen, 2009, die kansen presenteren in de context van survivalanalyse).

(iv) De uitkomsten van de analyses kunnen ook worden gebruikt om voor individuele studenten te bezien wat hun kansen zijn om in ieder van de vier categorieën genoemd in (i) terecht te komen. Hiervoor is een gemakkelijk te hanteren *tool* geprogrammeerd die de kansen op de vier categorieën geeft wanneer de kenmerken van de student in een Excel-spreadsheet zijn ingevoerd.

(v) De data worden door de tijd heen geanalyseerd, waarbij de variabele 'het aantal behaalde cursussen' als verklarende variabele voor uiteindelijk studiesucces steeds wordt aangepast. Hierdoor kunnen ook vragen beantwoord worden als wat de kans op studiesucces is nadat de studie al een aanvang heeft genomen en er dus tentamenresultaten zijn. Andere artikelen waarin er aandacht is voor de voorspelling van studiesucces met door de studie heen veranderende gegevens zijn Willcoxson (2010) en De Koning & Loyens (2011).

In wat volgt wordt eerst inzicht gegeven in de door ons gebruikte gegevens. Dan volgt een methodesectie waarin het statistische model uit de doeken wordt gedaan. Vervolgens worden enkele resultaten gepresenteerd. Er wordt geëindigd met een discussie en aanbevelingen.

Gegevens

Om te kunnen constateren of studenten langstudeerders worden, is een gegevensbestand nodig waarin studenten tot in hun vijfde jaar ingeschreven kunnen zijn. Hiertoe zijn de gegevens van de volledige cohorten 2006 (eerste studiejaar 2006-2007) en 2007 van de Faculteit Sociale Wetenschappen van de Universiteit Utrecht uit ons studievoortgangsysteem OSIRIS gebruikt; in totaal 2134 studenten. Van deze studenten zijn er 96 buiten beschouwing gelaten omdat ze vrijstellingen hebben (89 studenten) of een moeilijk in te delen studieverloop, zodat 2037 studenten in de analyses zijn betrokken. Deze studenten zijn ingedeeld in de vier categorieën besproken in de inleiding.

Verklarende variabelen zijn geslacht (1641 vrouwen en 396 mannen), leeftijd, gemiddeld eindexamencijfer, eindexamencijfer wiskunde, vwo-examenprofiel (zoals 'economie en maatschappij'), de studierichting (algemene sociale wetenschappen (ASW, n = 274), culturele antropologie (CA, n = 162), onderwijskunde (OWK, n = 54), pedagogiek (PED, n = 516), psychologie (PSY, n = 946) en sociologie (SOC, n = 85)), inschrijvingsdatum, en het aantal behaalde cursussen.
Voorspellingen van het studiesucces zijn berekend voor momenten waarop nieuwe informatie over studenten beschikbaar komt en die beleidsmatig interessant lijken. De Universiteit Utrecht kent twee semesters die elk bestaan uit twee blokken, waarbij binnen ieder blok twee cursussen van 7,5 ECTS behaald kunnen worden. Mede gezien de datum van het voorlopige studieadvies (februari) is gekozen voor vier voorspellingen:

- voorafgaand aan de studie;
- na blok 1 (aantal behaalde cursussen is 0, 1 of 2);
- na blok 2 (aantal behaalde cursussen is 0, 1, 2, 3 of 4);
- begin van tweede studiejaar (aantal behaalde cursussen loopt theoretisch van 0 tot 8).

Overwogen is ook om de behaalde cijfers in de analyses te betrekken, maar hier is van afgezien omdat de studierichtingen onvergelijkbare cursussen hebben, en omdat studenten die niet deelnemen aan tentamens te veel *missing data* zouden krijgen die niet op eenvoudige wijze geïmputeerd zouden kunnen worden.
Voor ongeveer 300 studenten ontbrak het gemiddelde eindexamencijfer en het eindexamencijfer wiskunde, en van ongeveer 50 studenten was de inschrijvingsdatum niet bekend. Bij de groep van 300 studenten gaat het voornamelijk om studenten afkomstig uit het hbo. Deze studenten zijn als het ware *missing by design*, en het gebruiken van bijvoorbeeld multipele imputatie (Van Buuren, 2012) heeft als nadeel dat ontbrekende gegevens van de hbo-studenten worden gevuld alsof zij direct afkomstig waren van het voortgezet onderwijs. Dit zou tot niet-valide imputaties leiden. Dit probleem is als volgt opgelost: stel er is een variabele met ontbrekende waarde X; er is een dummy-variabele D aangemaakt die aangeeft of iemand een score heeft op X ($D = 1$) ofwel dat die score ontbreekt ($D = 0$). In de vergelijking (zie de methodesectie) zijn dan de variabelen D en X opgenomen. Dit leidt dan tot $b_0 + b_1D + b_2DX$, waarbij b_0 het intercept is, en b_1 en b_2 regressiegewichten. Het resultaat is dat voor de groep met ontbrekende gegevens het gemiddelde b_0 wordt geschat en voor de groep die waarden heeft op X, $(b_0 + b_1) + b_2X$, dus een afwijkend gemiddelde en een effect van variabele X. In de *missing data*-literatuur spreekt men hier van de *indicatormethode* (Van Buuren, 2012).

Statistisch model

De afhankelijke variabele, studiesucces, wordt genoteerd als Y met waarden $k = 1$, 2, 3 en 4, corresponderend met de vier categorieën gedefinieerd in de inleiding. Er is gebruik gemaakt van ordinale regressie (het *proportional odds*-model, zie bijvoorbeeld Agresti, 2007). Het model is

$$\text{Log } (P(Y \leq k) / P(Y > k) = b_{0k} - b_1 X \text{ (voor k = 1, 2, 3)}$$

waarbij X staat voor een vector van verklarende variabelen en b_1 voor corresponderende regressiegewichten. Er worden dus drie logistische regressies simultaan geschat, waarbij de effecten van verklarende variabelen identiek zijn in de drie logistische regressies (namelijk b_1), en slechts de intercepten b_{0k}, die de hoogte van de kansen van categorieën van studiesucces (de afhankelijke variabele) bepalen, verschillen voor de drie logistische regressies.
Een andere manier om het logistische regressiemodel te schrijven is als

$$P(Y \leq k) = \exp (b_{0k} - b_1 X) / [1 + exp (b_{0k} - b_1 X)]$$

en zo zijn voor de waarden van de verklarende variabelen in X de cumulatieve kansen $P(Y = 1)$, $P(Y \leq 2)$ en $P(Y \leq 3)$ te bepalen, waarbij $P(4) = 1 - P(Y \leq 3)$. Vervolgens zijn hieruit de kansen $P(Y = 2)$ en $P(Y = 3)$ af te leiden als $P(Y = 2) = P(Y \leq 2) - P(Y = 1)$, en $P(Y = 3) = P(Y \leq 3) - P(Y \leq 2)$. Hierbij staat bijvoorbeeld $P(Y = 1)$ voor de kans dat een student uitvalt in jaar 1, en $P(Y \leq 2)$ voor de som van kansen dat een student in het eerste jaar uitvalt ($P(Y = 1)$), of daarna uitvalt dan wel 'langstudeerder' wordt ($P(Y = 2)$).
De schattingen zijn uitgevoerd met 'IBM SPSS statistics 20, procedure PLUM', waarbij de zogenaamde 'logit-link' en verder de default-opties zijn gebruikt. Er worden kansen geschat met behulp van de formule voor $P(Y \leq k)$. Hierbij wordt onderzocht of en hoe deze kansen worden beïnvloed door verklarende variabelen uit het studievoortgangsysteem OSIRIS.
Voor de geschatte kansen zijn predictie-intervallen te bepalen met de niet-parametrische *bootstrap*methode (Efron & Tibshirani, 1993). In IBM SPSS zijn 50 bootstrapsteekproeven gegenereerd. De analyse van iedere bootstrapsteekproef levert *b*-waarden. Deze 50 vectoren van *b*-waarden zijn te gebruiken om 50 vectoren van vier kansen te genereren. Op basis van deze 50 vectoren van vier kansen zijn vier bootstrapstandaardfouten voor de kansen te berekenen. Deze zijn te gebruiken om 95%-betrouwbaarheidsintervallen te berekenen voor de vier kansen geschat op basis van de steekproef (vergelijk figuur 1, hierna).
De significantie van de parameters in tabel 3 (hierna) laat zien dat de steekproefomvang van ongeveer 2000 voor het schatten van dit model ruim voldoende is.

Resultaten

Voorspelling van studiesucces
Tabel 1 geeft voor de verschillende verklarende variabelen de percentages in de vier studiesuccescategorieën weer. Voor wat betreft de uitval in het eerste jaar doen mannen het met 38,6% aanzienlijk slechter dan vrouwen (23,7%). Studenten ASW (39,8% uitval), Pedagogiek (35,9%) en Sociologie (34,1%) vallen vaker in het eerste jaar uit dan gemiddeld en studenten Psychologie met 17,3% uitval in het eerste jaar juist minder vaak. Uit tabel 2 blijkt dat het gemiddelde eindexamencijfer, het eindexamencijfer wiskunde en de inschrijvingsdatum samenhan-

Tabel 1 *Percentage studenten in de verschillende studiesuccesgroepen voor geslacht en studierichting (percentages).*

	Uitval in jaar 1	Uitval na jr 1/langst	Diploma in jr 4	Diploma in jr 3	Totaal	n
Man	38.6	15.9	16.9	28.5	100	396
Vrouw	23.7	11.3	21.1	43.9	100	1641
ASW	39.8	14.6	21.5	24.1	100	274
CA	24.7	16.7	37.7	21.0	100	162
OWK	27.8	16.7	14.8	40.7	100	54
PED	35.9	8.3	11.4	44.4	100	516
SOC	34.1	4.7	7.1	54.1	100	85
PSY	17.3	13.3	23.3	46.1	100	946
totaal	26.6	12.2	20.3	40.9	100	2037

Tabel 2 *Leeftijd, gemiddeld eindexamencijfer, eindexamencijfer wiskunde, inschrijvingsdatum en aantal gehaalde cursussen (gemiddelden) in de vier categorieën van studiesucces.*

	Uitval in jaar 1	Uitval na jr 1/langst	Diploma in jr 4	Diploma in jr 3
Leeftijd	19.6	19.5	18.9	19.0
Gemiddeld	6.6	6.7	6.8	6.9
Wiskunde	6.2	6.3	6.5	6.6
Inschrijv.datum	145	145	149	150
Aantal behaalde cursussen				
in blok 1	.8	1.9	2.0	2.0
in blok 2	1.3	3.7	4.0	4.0
aan einde jr 1		7.1	8.0	8.0

Inschrijvingsdatum 145 is 7 augustus en 150 is 2 augustus

gen met studiesucces. Hoe hoger de cijfers en hoe eerder de inschrijfdatum, hoe minder kans op uitval.

In de ordinale regressiemodellen bleken leeftijd, het vwo-examenprofiel van de studenten en het cohort waartoe ze behoorden geen statistisch significante bijdrage te leveren aan de voorspelling van studiesucces. In de analyse met louter gegevens voorafgaand aan de studie (zie tabel 3) waren de voorspellende variabelen 'gemiddeld vwo-eindexamencijfer', 'vwo-eindexamencijfer wiskunde', 'inschrijvingsdatum' (hoe eerder, des te beter het studiesucces), 'opleiding' en 'geslacht' significant. Na blok 1 neemt de variabele 'aantal behaalde cursussen' de rol van

Tabel 3 ***Ordinale regressiemodellen. De regressiegewichten van de eindmodellen zijn getoond. Deze regressiegewichten zijn gebruikt om tot de schattingen van kansen te komen. Positieve gewichten laten zien dat een hogere waarde op de variabele leidt tot een grotere kans om in een hogere categorie van de afhankelijke variabele terecht te komen.***

Variabelen	Vooraf	Na blok 1	Na blok 2	Na jaar 1
D gem. eindexamencijfer	-4.24 ***			
Gem. eindexamencijfer	.65 ***			
D wiskunde	-.64	-1.15 **	-1.18 **	-1.49 ***
wiskundecijfer	.15 *	.20 ***	.18 **	.21 ***
D Inschrijvingsdatum	-.96 ***	-1.96 **	-1.77 **	
Inschrijvingsdatum	.01 ***	.01 ***	.01 *	
Opl ASW	-1.03 ***	-.89 ***	-.57 ***	-.39 *
Opl CA	-.78 ***	-1.02 ***	-.90 ***	-.89 ***
Opl OWK	-.71 ***	-.05	-.14	.08
Opl Ped	-.67 ***	.05	.18	.62 ***
Opl Soc	-.03	.40	1.25 ***	1.61 ***
Opl Psy	ref	ref	ref	ref
Man	-.65 ***	-.57 ***	-.40 **	-.39 **
Vrouw	ref	ref	ref	ref
Aantal vakken	n.v.t.	3.93 ***	2.30 ***	1.48 ***

Wald toetsen *p < .05; **p < .01; ***p < .001; 'ref' geeft aan dat de categorie gebruikt is als referentiecategorie.

gemiddeld eindexamencijfer over: deze laatste variabele heeft dan geen statistisch significante bijdrage meer aan de voorspelling van het studiesucces. De voorspellingskracht van het aantal behaalde cursussen neemt door de tijd heen toe.

Een tool voor de kans op studiesucces

Met behulp van de logistische regressie is een *tool* ontwikkeld waarmee voor individuele studenten de kans op studiesucces kan worden berekend, dat wil zeggen: de kansen om in de verschillende categorieën van de variabele 'studiesucces' terecht te komen. In figuur 1 staat een afbeelding van een voorbeeld van de invoer in de *tool* (bovenin: de studentkenmerken) en de uitkomst (onderin: de kansen op studiesucces): voor een vrouwelijke psychologiestudente die als gemiddeld vwo-eindexamencijfer een 7 heeft, voor het eindexamen wiskunde een 7 heeft en ingeschreven is op 1 juli, vinden we een kans van 8% om uit te vallen in jaar 1, een kans van 6% om daarna uit te vallen of 'langstudeerder' te worden, een kans van 15% om een diploma te halen in het vierde studiejaar, en een kans van 71% om een diploma te halen in het derde studiejaar. Deze kansen zijn desgewenst optelbaar: indien de laatste twee categorieën geïnterpreteerd worden als

Figuur 1 ***Een tool waarmee kansen op studiesucces van individuele studenten kunnen worden voorspeld.***

ONDERSTEUNING MATCHINGSGESPREKKEN

VARIABELEN	**Geslacht**	Vrouw		
	Opleiding	PSY		
	Gemiddeld cijfer *(als ja, vul cijfer in)*	ja 7		
	Wiskundecijfer *(als ja, vul cijfer in)*	ja 7		
	Inschrijfdatum *(als ja, vul dag in als DD-MM-JJJJ)*	ja 1-7-2007		
			95 % betrouwbaarheid	
KANSEN	**Uitval in jaar 1**	**0,08**	*0,06*	*0,10*
	Uitval na jr 1/langstudeerder	**0,06**	*0,05*	*0,08*
	Diploma in jaar 4	**0,15**	*0,13*	*0,18*
	Diploma in jaar 3	**0,71**	*0,65*	*0,75*
		1,00		

positief studiesucces en de eerste twee als negatief, dan is de kans voor deze studente op positief studiesucces 86%. In het gele gebied zijn andere getallen in te vullen en kunnen – door de cel aan te klikken – met een keuzelijst de waarden op geslacht en opleiding veranderd worden. Uit het 95%-betrouwbaarheidsinterval van de kansen blijkt dat voor deze vrouwelijke psychologiestudente de marge voor kansen rond 10% ongeveer plus-min 2% is, en voor kansen in de buurt van 50% plus-min 5%. Voor kleinere opleidingen dan Psychologie (n = 946) worden deze betrouwbaarheidsintervallen groter: bijvoorbeeld voor Onderwijskunde (n = 54, resultaten hier niet getoond) zijn de marges ongeveer twee tot drie keer zo groot.

Met behulp van de *tool* zijn voor vrouwelijke psychologiestudenten met verschillende kenmerken de kansen op studiesucces berekend. De resultaten, weergegeven in tabel 4, illustreren hoezeer die kansen verschillen als functie van de gezamenlijke studentkenmerken. In rij 1 ziet men in de laatste vier kolommen kansen indien er geen waarden bekend zijn bij 'gemiddeld vwo-eindexamencijfer', 'vwo-eindexamencijfer wiskunde' en 'inschrijvingsdatum'. In de meeste gevallen gaat het hier om hbo-studenten. De kans op negatief studiesucces is hier 45 + 16 = 61%. We zien na blok 1, 2 en aan het begin van jaar 2 zeer grote verschillen, vooral samenhangend met het aantal behaalde cursussen. Zo is in blok 1, wanneer één van de twee cursussen is behaald, de kans op negatief studiesucces 93% terwijl die kans slechts 12% is wanneer beide cursussen zijn gehaald.

Tabel 4 ***Kansen op verschillende momenten op studiesucces voor vrouwelijke psychologiestudenten met verschillende kenmerken***

	Studentkenmerken				*Kans op studiesucces*			
	Cursussen gehaald	*Gem. cijfer*	*Wisk. cijfer*	*Inschr. datum*	*Uitval in jr 1*	*Uitval na jr 1/langst.*	*Diploma in jr 4*	*Diploma in jr 3*
Vooraf	*n.v.t.*	*geen*	*geen*	*geen*	.45	.16	.19	.21
	n.v.t.	*7*	*7*	*1-7*	.08	.06	.15	.71
	n.v.t.	*6*	*6*	*15-8*	.26	.14	.23	.37
Na blok 1	*2*	*	*7*	*1-7*	.04	.08	.20	.68
	1	*	*6*	*15-8*	.80	.13	.05	.02
Na blok 2	*4*	*	*7*	*1-7*	.03	.09	.23	.65
	3	*	*6*	*15-8*	.27	.36	.24	.13
Begin jaar 2	*8*	*	*7*	*	N.v.t.	.11	.28	.61
	7	*	*6*	*	N.v.t.	.40	.38	.22

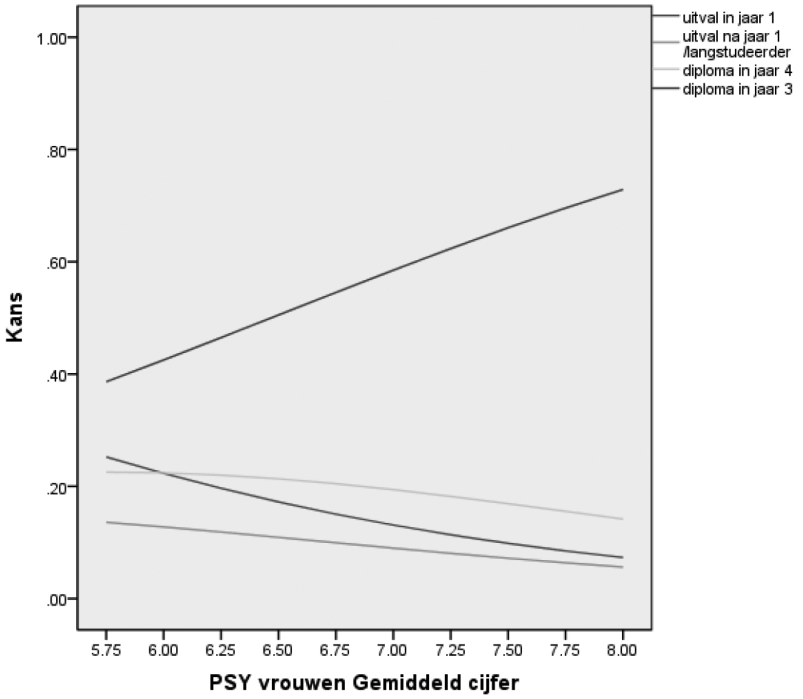
*Deze variabele is op dit moment niet meer relevant.

De invloed van individuele studentkenmerken op de kans op studiesucces kan ook grafisch worden weergegeven. Figuur 2 en figuur 3 tonen hoe de kansen samenhangen met bepaalde studentkenmerken bij het constant houden van andere kenmerken (voor die andere kenmerken hebben we de gemiddelde waarde in de populatie ingevuld). Figuur 2 laat zien dat, voorafgaand aan de studie, het gemiddelde vwo-eindexamencijfer een sterke voorspeller is van het binnen 3 jaar afstuderen: die kans stijgt van ongeveer 40% bij een gemiddeld cijfer 5,75 naar ongeveer 70% bij een gemiddeld cijfer 8. Figuur 3 laat, na afloop van het eerste studiejaar, een sterke daling zien van de kans om uit te vallen of 'langstudeerder' te worden wanneer meer cursussen zijn gehaald en tegelijkertijd een sterke stijging van de kans om in het derde jaar het diploma te behalen. Opvallend is dat ook wanneer 6 van de 8 cursussen gehaald worden (per 2013 het criterium voor het bindend studieadvies in Utrecht) de kans om naderhand uit te vallen dan wel 'langstudeerder' te worden, nog altijd ongeveer 70% is.

Discussie en aanbevelingen

De hier gepresenteerde benadering in het onderzoek naar studiesucces is praktisch relevant, omdat ze gebruik maakt van een ordinale categorische variabele die het eenvoudig mogelijk maakt kansen te schatten voor elk van de categorieën van studiesucces. Daardoor kan eenvoudig te bedienen programmatuur worden ontwikkeld, die ook nog eens eenvoudig te begrijpen resultaten geeft (zie figuur 1). De bediening is eenvoudig omdat slechts van een gering aantal studentkenmerken gebruik gemaakt wordt. De uitkomsten zijn eenvoudig te begrijpen omdat dit kansen op studiesucces zijn.
Belangrijke vraag is hoe van de resultaten van onze analyses in de praktijk gebruik gemaakt kan worden. Hoe moet bijvoorbeeld een studieadviseur omgaan met een

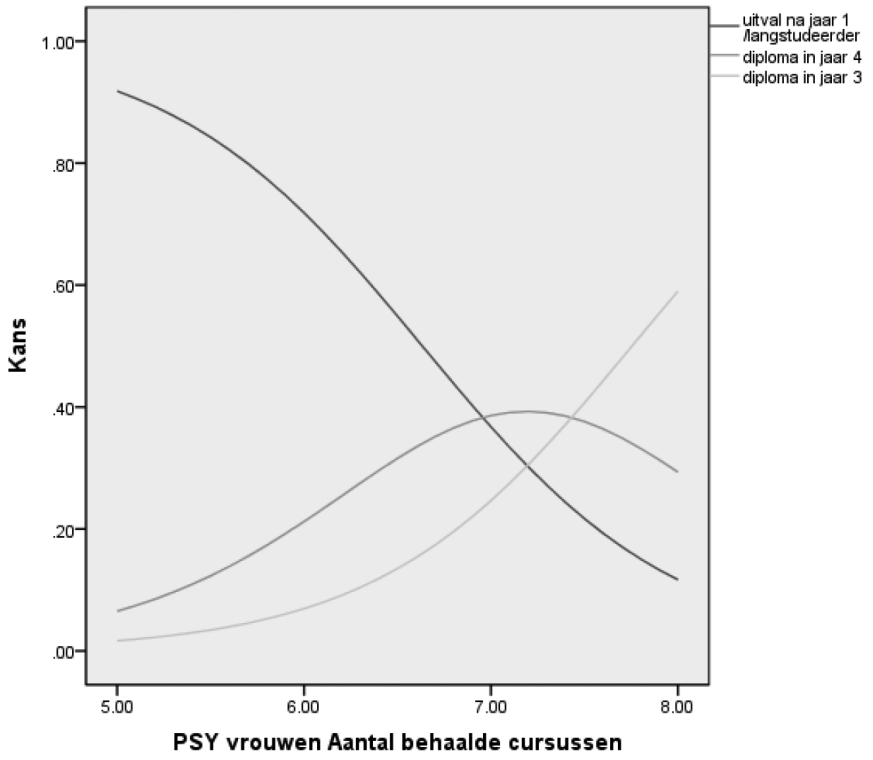
Figuur 2 *Kansen op studiesucces van vrouwelijke psychologiestudenten voorafgaande aan de studie als functie van gemiddeld vwo-eindexamencijfer*



student die een kans van 40% heeft om zijn bachelorstudie niet in vier jaar af te ronden? Hoe moet in een matchingsgesprek worden omgegaan met het gegeven dat mannen aanzienlijk meer kans hebben dan vrouwen om 'langstudeerder' te worden of uit te vallen? Dit onderzoek geeft hier uiteraard geen antwoord op. Het lijkt verstandig kwantitatieve gegevens een rol te laten spelen bij het vormen van een oordeel. Tegelijkertijd is het ook niet raadzaam om het verhaal van de individuele student bij de geschatte kansen te negeren.

De gegevens die gebruikt zijn voor de schattingen zijn relatief oud. Onderzocht moet worden of de *tool* aangepast dient te worden omdat studenten, hun gedrag en hun studie veranderd zijn, bijvoorbeeld als gevolg van de invoering van een scherper bindend studieadvies of de langstudeerboete. Dit kan jaarlijks door de *tool* aan te passen op grond van recentere gegevens over een heel cohort. Wachten op de studieresultaten van een geheel cohort leidt onvermijdelijk tot het hanteren van relatief oude gegevens. Er kan gebruik worden gemaakt van recentere gegevens, maar dan dient de concessie gedaan te worden dat voor studiesucces geen vier categorieën meer worden gehanteerd. Indien bijvoorbeeld het cohort 2011 wordt gebruikt, dan kunnen slechts twee categorieën worden gebruikt, namelijk 'eerste jaar niet gehaald' en 'eerste jaar wel gehaald'. In deze laatste categorie zijn

Figuur 3 Kansen op studiesucces van vrouwelijke psychologiestudenten na studiejaar 1 als functie van aantal behaalde cursussen.



dan drie categorieën samengenomen die in dit onderzoek nog werden onderscheiden.

We noemden in de inleiding als nadeel van *institutional research* dat psychologische variabelen niet meegenomen worden. Aanvullende gegevens voor een beperktere groep studenten – bijvoorbeeld verzameld in een *survey* – zijn aan studievoortgangsgegevens te koppelen, waarbij voor de overige studenten de ontbrekende gegevens dan met *missing data*-technieken zijn te imputeren. Zo kan worden nagegaan of door het opnemen van dergelijke variabelen een versterking van de voorspellende waarde van de *tool* kan worden gerealiseerd.

Referenties

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New Jersey: Wiley.

Annema, A. & Ooijevaar, J. (2011). De invloed van eindexamenresultaat op succes in het hoger onderwijs. *Tijdschrift voor hoger onderwijs, 29,* 65-81.

Brogt, E., Sampson, K.A., Comer, K., Turnbull, M.H. & McIntosh, A.R. (2011). Using Institutional Research Data on Tertiary Performance to Inform Departmental Advice to

Secondary Students. *Journal of Institutional Research of the Australasian Association for Institutional Research, 16* (2). Published online.

Bruinsma, M. & Jansen, E.P.W.A. (2009). When will I succeed in my first- year diploma? Survival analysis in Dutch higher education. *Higher Education Research & Development, 28* (1), 99-114.

Buuren, S. van (2012). *Flexible imputation of missing data.* Boca Raton: CRC Press.

Campbell, J. (2007). *Utilizing Student Data within the Course Management System to Determine Undergraduate Student Academic Success: An Explorative Study.* Purdue: Unpublished PhD thesis, Purdue University.

Cassidy, S. (2011). Exploring individual differences as determining factors in student academic achievement in higher education. *Studies in Higher Education, 36,* 989-1001.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Jansen, E.P.W.A. & Bruinsma, M. (2005). Explaining achievement in higher education. *Educational research and evaluation, 11,* 235-252.

Koning, B. de & Loyens, S. (2011). Studiesucces in de bachelor, deel 1: Generation Psy, studiefactoren en studiesucces. *Beleidsgerichte studies hoger onderwijs en wetenschappelijk onderzoek, 138.* Rotterdam: Risbo.

Meeuwisse, M., Wensveen, P. van & Severiens, S. (2011). Studiesucces in de bachelor, deel 3: Tijd om te studeren: een onderzoek naar tijdbesteding en studiesucces in leeromgevingen. *Beleidsgerichte studies hoger onderwijs en wetenschappelijk onderzoek, 138.* Rotterdam: Risbo.

Onzenoort, C.H. (2009). *Als uitval opvalt: studie-uitval in het hoger beroepsonderwijs.* Academisch proefschrift. Oisterwijk: Uitgeverij BOXPress.

Torenbeek, M., Suhre, C., Jansen, E. & Bruinsma, M. (2011). Studiesucces in de bachelor, deel 2: Studentfactoren, curriculumopzet en tijdbesteding als verklaringen. *Beleidsgerichte studies hoger onderwijs en wetenschappelijk onderzoek, 138.* Rotterdam: Risbo.

Visser, K. (2011). *Studiesucces WO bachelor-opleidingen Psychologie Nederland*. Amsterdam, UvA, Faculteit Psychologie, Interne rapportage.

Whitmer, J., Fernandes, K. & Allen, W.R. (2012). Analytics in Progress: Technology Use, Student Characteristics, and Student Achievement. *Educause Review on Line* Published on Monday, August 13, 2012.

Willcoxson, L. (2010). Factors affecting intention to leave in the first, second and third year of university studies: a semester-by-semester investigation. *Higher Education Research & Development, 29,* 623-639.

Yu, C.H., DiGangi, S., Jannasch-Pennell, A. & Kaprolet C. (2010). A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. *Journal of Data Science,* 8, 307-325.