Model-based academic dismissal policies

Albers, Casper; Vermue, Carlien; de Wolff, Taco; Beldhuis, Hans

DOI:
10.31234/osf.io/6a9cz

Model-based academic dismissal policies; a case-study from the Netherlands

Casper Albers[1,*], Carlien Vermue[2], Taco de Wolff[2], and Hans Beldhuis[2]

[1]Heymans Institute for Psychological Research, University of Groningen

[2]Donald Smits Center for Information Technology, University of Groningen

[*]Corresponding author. Tel: +31 503638239, c.j.albers@rug.nl

Abstract

Many higher education institutions use a policy for academic dismissal. In the Netherlands, the academic dismissal policy is such that students with fewer credits than a certain threshold after their first year, are expelled. This article employs the beta-binomial model to assess whether this method succeeds in filtering those who have potential from those who do not and what the optimal level of the threshold is. The model considers 13,234 students in three consecutive cohorts from around fifty different bachelor's degree programmes at the University of Groningen. We found that demanding 45 out of 60 credits constitutes a fair threshold for this institution. Although a strict dismissal policy has only a minor effect on cohorts, it can have a major effect on specific groups of students. The software employed here is made available.

Keywords. academic dismissal; binding study advice; beta-binomial model; study success; exam result analysis; data-driven policy

1 Introduction

There are many undesirable ramifications when students in higher education take too long to obtain their degree. For example, low-achieving students are more expensive and they demotivate higher-achieving students. Improving the average time students take to complete a degree is important to students, because of the limited period for which they get financial support from the government and the negative effects of having a job on the side has on study performance (Beerkers et al., 2010). It is also important for educational institutions, because study progress influences society because of the demand for skilled workers with a college degree.

Such improvements can be made by understanding student's motivations. Knowledge on e.g. what causes drop-out and programme switching (Lassibille and Gómez, 2009; Breier, 2010, Belloc et al., 2010; O'Neill et al., 2011), what improves students' attitudes towards their study (Ramos and Carvalho, 2010; Lechuga, 2011), what the influence of internationalisation is (Mamiseishvili, 2012; Rientjes et al., 2012), etcetera, gives policy-makers important tools for identifying problem areas and introducing new policies to overcome these difficulties.

Better understanding of student's motivations gives policy-makers important tools for identifying problem areas and introducing new policies to overcome these difficulties. However, not all problems can be tackled by policy and it is unavoidable that there will be students with unacceptable low progress. Many higher education institutions have set up some kind of academic dismissal policy that results in expelling all students that earn fewer credits than a certain threshold. The height of this threshold is usually decided upon on the basis of ad hoc arguments and expert opinion, rather than truly evidence based.

According to Nakabo-Ssewanyana (1999), statistical data is an underestimated tool for higher education data management. In this paper we will employ a statistical model, the beta-binomial model, for modelling the performance of a cohort of students. We use this model to assist policy makers in deciding upon such a threshold. We will outline the model in the context of the Dutch 'binding study advice' system. We show that the beta-binomial model is an elegant tool for institutional data analysis. Some papers have modelled a relation between indicators of student performance and study progress (cf. Goho and Blackman, 2006; Cohen-Schotanus et al., 2006), but not in the context of whether a certain threshold score is achieved or not.

1.1 The binding study advice

In 1998, lawmakers in the Netherlands enabled higher education institutions to, under certain conditions, expel undergraduate students who underperform during their first year of study. Consequently, a majority of institutions have incorporated binding study advice (BSA) as a structured academic dismissal policy. The BSA has been under scrutiny of various peer-reviewed studies (cf. Gijbels et al., 2004; De Koning et al., 2013, Arnold, 2015, Eijsvogels et al., 2015). See Arnold (2015) for an detailed overview and history of the BSA. Our paper distinguishes itself from these works by having a different focus. The main research question is studying the effect of setting the threshold at a certain value, and advicing on what threshold constitutes a fair threshold.

All Dutch, and most European, higher education institutions employ the European Credit Transfer and Accumulation System (ECTS). Each degree programme consists of various courses with a combined study load of 60 credits per year. Most Bachelor degree programmes are three year programmes, thus consisting of 180 credits (ECTS). The BSA works in the following way (see Arnold, 2015, for a more detailed explanation and history of the BSA). Based on the number of credits a student achieves in the first year of a bachelor's degree programme, X, and the BSA threshold B, there are two possibilities:

1. $X \geq B$: *positive BSA*. The student can continue to the second year without limitations;

2. $X < B$: *negative BSA*. The student is expelled.

The choice of B is up to the university and can change from year to year. Whether a student achieves the threshold or not depends largely on his or her skills and intelligence. However, there are other factors – for example, how well a student can adjust to the transition from secondary school to university, whether a student has an off day on the day of an exam, whether a student makes some lucky guesses on a multiple choice exam. Consequently, the number of credits a student achieves, depends both on skill and chance. The skill level is a latent variable: one can only observe the combination of skill and chance. It is reasonable to assume that, for large samples, the average effect of chance is neutral.

In this study, we model the probability of achieving the BSA threshold, given a student's skill level, in which skill is measured as a number in the interval $[0, 1]$, with 1 reflecting perfect skill, and

0 reflecting no skill at all. We have chosen this parametrisation as this reflects the interpretation of skill as the probability of passing an exam; but other parametrisations (e.g. a N(100,15)-distribution, as with IQ-scores) would have been possible as well. With our scale, a threshold of $B = 30$ credits (out of 60), for example, corresponds with an average skill level of 50%. Due to this chance effect, there will be a certain amount of false positives – students scoring above B but having a skill level below $B/60$ and thus incorrectly passing – and a certain amount of false negatives – students scoring below $B/60$ but having a skill level above $B/60$, and thus mistakenly being expelled.

For a given bachelor's degree, we estimate the distribution of skill level among students, enabling the model to make inferences about both the individual and cohort levels.

Some institutions have increased the BSA threshold over the years, assuming that a higher threshold would motivate students to work harder and therefore perform better. To our knowledge, there have been no previous studies of the effect of the BSA on study performance. Stegers-Jager et al. (2011) studied a precursor of BSA and reported that study behaviour did not improve after the introduction of the dismissal policy.

We study this model based on three full cohorts of data from the University of Groningen. We study the effect the BSA threshold has both on typical students (notably poor, median and good students) individually and on the cohort as a whole. We do so because a minor change in the BSA threshold can have a large impact on individual students' probability of achieving a positive advice even though the cohort as a whole seems largely unaffected. Although written in the context of the BSA, our model applies to modelling all forms of academic dismissal, and even to modelling cohort performance in general.

2 Data

The University of Groningen is one of 14 state-funded research universities in the Netherlands and is in the Top 100 of several international university rankings. It hosts close to 30,000 students in a wide range of bachelor's and master's degree programmes. During the period of data collection, the university consisted of nine faculties: medicine, law, arts, science and engineering, economics and business, social sciences, spatial sciences, philosophy and religious studies. The data set consists of data for three consecutive academic years: 2010/2011, 2011/2012 and 2012/2013.

Table 1: Breakdown of performance (number of credits obtained) of the three cohorts of students started in the first year, in both frequencies as proportions.

ECTS	Cohort					
	2010/2011		2011/2012		2012/2013	
	freq.	prop.	freq.	prop.	freq.	prop.
0	156	0.037	178	0.039	137	0.031
1 to 14	205	0.049	218	0.047	172	0.039
15 to 29	263	0.062	287	0.062	239	0.054
30 to 39	177	0.042	241	0.052	137	0.031
40 to 44	337	0.080	370	0.080	92	0.021
45 to 49	370	0.088	391	0.085	464	0.105
50 to 54	424	0.101	459	0.100	467	0.106
55 to 59	479	0.114	512	0.111	531	0.120
60	1805	0.428	1943	0.422	2180	0.493
Total	4,216		4,599		4,419	

Our data set consists of information on student performance and of information on course performance, but not on both simultaneously. For each degree programme and cohort, we have the frequency table of credits earned by students. For each course per year, we have the pass rate. Due to privacy issues in the data, we don't know which students passed which courses.

Degree programmes sometimes are discontinued and new ones are started. Therefore, the number of bachelor's degree programmes is not fixed: it was 59 in the first two cohorts and 48 in the last one. For the year 2010/2011 there is information on 4,599 first year students, for the year 2011/2012, there is information on 4,216 students, and for the year 2012/2013, there is information on 4,419 students. This information is summarised in Table 1. For each year, the average credits obtained are 47.47 (sd = 17.16), 47.19 (sd = 17.22) and 49.67 (sd = 16.19), respectively. On average, across all cohorts, students achieved 80.2% of their study load in the first year. The BSA threshold was $B = 40$ during the academic years 2010/2011 and 2011/2012 and $B = 45$ during 2012/2013. Comparing the first two years with the third year thus yields information on the hypothesis that a higher threshold motivates students to work harder. The average scores hint at this possibility, Section 5 elaborates further.

Most degrees comprise about 10 to 14 courses (2 to 4 courses per quarter) constituting the first year programme of 60 credits. The number of courses offered per cohort are 775, 784, and 657, thus the average number of courses per degree programme is 13.14, 13.29, and 13.68, respectively.

A small number of students, 5% in 2010/2011 and 2011/2012 and 6% in 2012/2013, were exempt from BSA requirements because of hardship (e.g. bereavement); we excluded these students from the analysis. These are the only students that have been excluded from the analysis. Assuming that reasons as bereavement are unrelated to skill level, this exclusion will not bias our results.

3 Methods

3.1 The basic model

We intentionally started with an oversimplified model and increased the complexity until sufficient fit was achieved. The starting model relies on two key assumptions:

1. *No between-student variation*: among all students, the probability of passing a course is the same.

2. *No between-course variation*: among all courses, the probability of passing is the same.

Together, these assumptions imply that for every course and every student, the passing probability p is the same.

The courses for most degree programmes have similar load. The optimal estimator (both OLS and ML) for this probability would simply be the average pass rate p. It is straightforward to see that this average pass rate $p = m/60$, where m is the average amount of credits obtained in the cohort under study. The following binomial distribution shows the probability of a single student passing exactly k out of K courses

$$\mathrm{P}(\text{Pass } k \text{ courses}) = \binom{K}{k} p^k (1-p)^{(K-k)},$$

thus the probability of receiving a positive advice is

$$\mathrm{P}(\text{credits} \geq B) = \sum_{k=\lceil BK/60 \rceil}^{12} \mathrm{P}(\text{Passed} = k).$$

However, not all degrees have courses of equal load. In general, a degree has K courses, with credit loads of $c_1, \ldots c_K$, with $\sum c_k = 60$. Consequently, the binomial distribution still gives the probability of passing k courses but the translation from courses to credits is less straightforward and computationally more demanding. For each course, there are two possible outcomes: passing or failing. Therefore, during the whole year, there are 2^K possible outcomes. For each of the 2^K distinct outcomes ($j = 1, \ldots, 2^K$), we created a vector $m_j = (m_j(1), \ldots, m_j(K))$ with elements $m_j(i)$ being 1 if course i was passed, and 0 otherwise. We obtained the total credits for a certain vector by computing $s_j = \sum_k c_k m_j(k)$ and the probability of this outcome occurring is $p_j = p^{\sum_k m_j(k)}(1-p)^{K-\sum_k m_j(k)}$.

Next, we created the 61 probabilities $P_0, P_1, \ldots P_{60}$ of obtaining the 61 possible cumulative scores via $P_i = \mathrm{P}(\text{credits} = i) = \sum_{s_j=i} p_j$. It immediately follows that the probability of receiving a positive advice is $\mathrm{P}(\text{credits} \geq B) = \sum_{i=B}^{60} P_i$. Like before, in this unbalanced design, the optimal estimator is $p = m/60$. To find out whether this model is useful, the assumptions need to be checked.

3.2 Assessing the assumptions

As we lack information on individual results, we can only relax either the assumption of no within-student variation, or that of no within-course variation, but not both. In this section, we study which of these two assumptions should be relaxed.

There are 743 distinct courses in our database. Most courses are offered in more than one year and various courses are part of more than one degree programme. There are 2,216 distinct combinations course×degree×year. The average number of participants per combination is 82.81 (first quartile 17, third quartile 112) and the vast majority of combinations (69.9%) concern courses with a load of four, five or six credits. For the 2,216 combinations, the average pass rate is 0.895 (sd = 0.116). When weighting student numbers, this average is 0.886 (sd = 0.098); when weighting the credit-load it is 0.890 (sd = 0.120). Very exceptionally, courses had a pass rate of 0% and whenever this occurred the number of students was either 1 or 2. For roughly 30% of the courses, the pass rate is 100%, either because of a small number of enrolled students, or because the requirement for passing the course was easily attainable (e.g. attending a series of lectures rather than an exam). Half of the remaining 1,548 courses have a pass rate between 80% and 93% (one quarter below 80% and one quarter above 93%). These statistics do indicate that differences in pass rate between courses are fairly small, at least for the majority of courses. It is reasonable, then, to assume there is equality.

The frequency table of credits earned by students tells a much more dispersed story. As Table 1 shows, 44.8% of all students earned all 60 credits, 14.0% earned less than half the credits and 41.2% earned between 30 and 59 credits. If the binomial model were correct, it would be much less likely to have a score 'far away' from the mean. For example, for a degree with 12 courses of five credits each, and a passing probability of $p = .802$ (the university's average pass rate), the probability that a student would receive 60 credits is 7.1%, the probability of receiving fewer than 30 credits is 0.4% and the probability of receiving between 30 and 55 credits is 92.6%.

This dispersion is visible not only in the aggregated table for the whole university, but also in (nearly) all individual degrees. Therefore, we shall discard the assumption of no between-student variability and transform the basic model into the beta-binomial model accordingly.

3.3 The beta-binomial model

Model definition

The beta-binomial model (BBM) is an extension of the standard binomial model. Under the beta-binomial model, we assigned each student i an individual passing probability π_i, after we used the binomial distribution to compute the probabilities of passing k courses and/or achieving s credits. The BBM is well known in classical and Bayesian statistics, and has been applied occasionally in the educational and behavioural sciences (Wilcox, 1981) but, to the best of our knowledge, not in a setting such as this one.

The two steps in the beta-binomial distribution are as follows. First, for each student j, we drew the student's skill parameter π_j from the beta-distribution

$$\pi_j \sim \mathrm{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \quad 0 \leq p \leq 1,$$

where $\mathrm{B}(\alpha, \beta)$ is the beta-function with parameters α and β. The beta distribution allows for great flexibility in the shape of the distribution of student skills, as Figure 1 shows.

Next, we considered this student's skill parameter π_j as the individual probability of passing an exam (where each course has the same passing probability π_j). We applied the binomial distribution

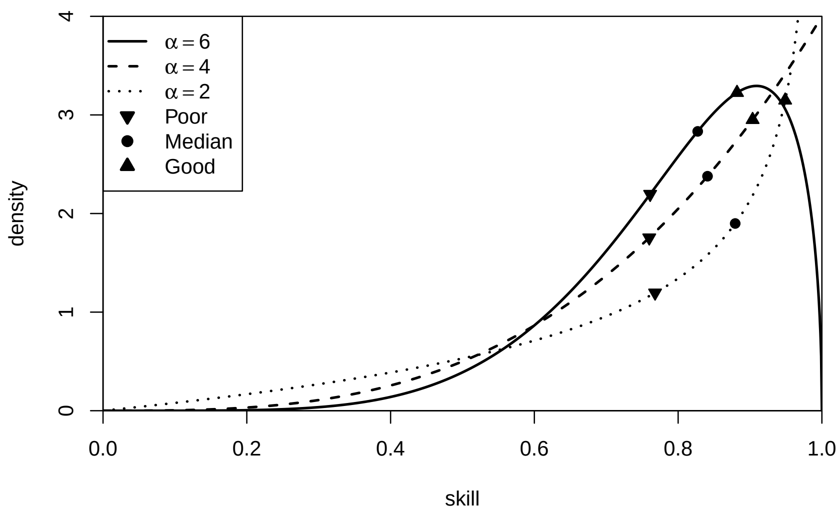

Figure 1: Three parametrisations of the beta density, all with average passing probability $\alpha/(\alpha+\beta) = 0.8$. For all three, the locations of the 'poor' (33rd percentile), median and 'good' (67th percentile) student are indicated.

in the same way as in the previous section:

$$\text{P(Student } j \text{ passed } k \text{ courses)} = \binom{K}{k} \pi_j^k (1-\pi_j)^{(K-k)}.$$

Combining both steps into one formula, yielded the probability density function of the beta-binomial distribution

$$P(\text{Student passes } k \text{ courses} \mid \alpha, \beta) = \binom{K}{k} \frac{\mathrm{B}(\alpha+k, \beta+K-k)}{\mathrm{B}(\alpha, \beta)}.$$

A beta-binomial distribution always generates values between 0 and 1. By adjusting the two parameters, both the mean value, given by $\alpha/(\alpha+\beta)$, as the dispersion of values can be varied. Thus, we can set the parameters such that the dispersion is (much) larger than for the binomial model, which is a requirement for good model fit as the previous section shows. From the data we obtained the average pass rate of all courses in a cohort. The mean skill parameter of the cohort is set to this value by defining $\beta = \frac{1-p}{p}\alpha$, because this equation implies $\alpha/(\alpha+\beta) = p$. We reparametrised the model into two parameters: the average pass rate of the cohort p and the dispersion parameter α. This does not affect the model results, but it does improve interpretability since p can be interpreted directly, whereas β can not.

The interpretation of p is clear: the average pass rate. The interpretation of α is somewhat more complicated. Small values of α indicate highly dispersed data – i.e. a degree program with very bright but also very poor students – and large values of α indicate small dispersion – i.e. the students are much more comparable. The extreme case $\alpha = \infty$ coincides with no dispersion among student skills, and is thus equivalent to the basic binomial model. The BBM is therefore a generalisation of the binomial model of Section 3.1.

Model estimation

We estimated the parameters using the method of moments, which is an efficient method of estimation in the BBM (Tripathi et al., 1994; Best et al., 2010). For a degree with all courses of equal load and notation K as the number of courses, n as the number of students in a given cohort and s_j as the credits student j earns, first we computed the first and second moment of the beta-binomial distribution:

$$m_1 = \frac{1}{n}\sum_{j=1}^{n} \frac{s_j}{60/K} \quad \text{and} \quad m_2 = \frac{1}{n}\sum_{j=1}^{n} \left(\frac{s_j}{60/K}\right)^2.$$

The average passing probability is $\hat{p} = m_1/K$, the estimator for α via

$$\hat{\alpha} = \frac{Km_1 - m_2}{K(\frac{m_2}{m_1} - m_1 + 1) + m_1}.$$

The estimator for β can also be written as a function of m_1 and m_2, but it is easier to directly compute it through $\hat{p}$ and $\hat{\alpha}$ as outlined above, thus $\hat{\beta} = \frac{1-\hat{p}}{\hat{p}}\hat{\alpha}$. Next, the model can predict the probability that the student passes exactly k out of K courses via

$$\mathrm{P}(\text{Pass } k \text{ courses} \mid K, \hat{\alpha}, \hat{p}) = \frac{\Gamma(K+1)\Gamma(k+\hat{\alpha})\Gamma(K-k+\hat{\beta})\Gamma(\hat{\alpha}+\hat{\beta})}{\Gamma(k+1)\Gamma(K-k+1)\Gamma(K+\hat{\alpha}+\hat{\beta})\Gamma(\hat{\alpha})\Gamma(\hat{\beta})},$$

with $\Gamma(\cdot)$ the gamma function. We can compute this value for single students or we can aggregate it over all students in for a complete cohort to get an estimate for the cohort as a whole. Converting this probability to, for example, the probability $\mathrm{P}(s_j \geq B)$ is straightforward.

Degrees with courses of varying size

In case that not all courses constitute equal credits, it is still possible to apply the model and the estimate of p is still given by $\hat{p} = m/60$, but the estimation of α is less straightforward.

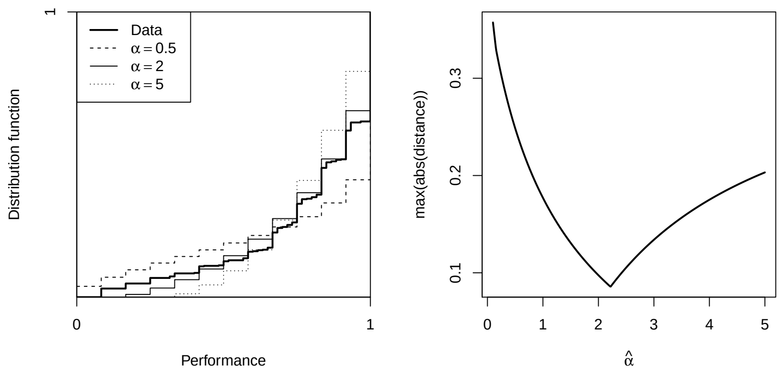

Figure 2: Visualisation of the iterative model-fitting procedure for the degree programme Business Administration in 2011/2012. Left: empirical distribution function and best fitting cumulative probability distribution functions of the BBM for various choices of $\hat{\alpha}$. Right: Relationship between the chosen value for $\hat{\alpha}$ and the maximum absolute vertical distance between the empirical and cumulative probability distribution function.

For each situation of passing k out of K courses, all $\binom{k}{K}$ combinations of k different courses are computed, with their corresponding ECTS-total and probability. Because of this extra step, we need an iterative approach to estimate $\hat{\alpha}$. Using the bisection method we explored the interval $(\underline{\alpha}, \overline{\alpha})$ for possible values for α, for a given, fixed, value of $\hat{p}$. For each value in this interval, the model fit is assessed. The α-value for which the maximal vertical distance between the probability distribution function and the cumulative empirical density function is smallest, is taken as estimator $\hat{\alpha}$. As boundaries of the interval, we chose $(\underline{\alpha}, \overline{\alpha} = (0.01, 10))$. All resulting degree programmes have their estimated α's well inside these boundaries. Figure 2 demonstrates this method and clearly shows that the choice of optimal $\hat{\alpha}$ is fairly robust. The fit, measured in maximum absolute vertical distance, does not increase severely when deviating slightly from the optimal value for α. Consequently, we computed $\hat{\beta}$ and all desired probabilities based on the estimate $\hat{\alpha}$.

Section 4 shows that this model fits the data well and that there is no need for further model modification.

Prediction through the model

As the model allows for differentiation among students, it is informative to study different profiles. Because one degree programme might be easier or harder than another, or attract a different type of students, different degree programmes might have different (average) pass rates. Therefore, these profiles are computed within a cohort. To maintain comparability among degrees, it was useful to work with a set of standardised student profiles. Section 4 uses the following three individual profiles:

1. A *median* student, with $\pi_i = \text{invBeta}(1/2 \mid \hat{\alpha}, (1-\hat{p})\hat{\alpha}/\hat{p})$, where $\text{invBeta}(\cdot)$ is the quantile function (i.e. inverse of the cumulative distribution function) of the beta-binomial distribution. This student performed better than half of his or her fellow students, and worse than the other half.

2. A '*poor*' student with $\pi_i = \text{invBeta}(1/3 \mid \hat{\alpha}, (1-\hat{p})\hat{\alpha}/\hat{p})$. This student performed worse than two-thirds of his or her fellow students, yet better than the other one third.

3. A '*good*' student who is characterised similarly through the 66.7th percentile point of the appropriate beta distribution.

Because beta distributions can be asymmetrical, the median might be quite distant from the mean. In some exceptional cases, a 'poor' student might even perform better than the 'average' student. Therefore, we used the median as the measure of the centre.

Furthermore, we explored different 'cohort scenario's'. Not only did we study the cohort with parameters that best fit the data, but we also examined the cohorts where p lay a small distance away either above or below from the 'optimal' p. In this way, we could examine the effects when a new cohort performed slightly better or worse than the previous one.

Model aggregation

Aggregating the results was achieved by taking the weighted sum of the appropriate degrees, weighting was done with respect to the number of students. Section 4 demonstrates this.

4 Results

Results for individual degrees

To demonstrate our model, we selected three degree programmes: Accountancy, History and Mathematics.

Figure 3(a) displays the data for the 2010/2011 Accountancy cohort. As expected, the BBM clearly fits the data more clearly than the binomial model. A better fit is obvious as the binomial

Table 2: Overview of parameter estimates and model fit for the beta-binomial model and the binomial model, for the degree programmes at the Faculty of Economics and Business in the 2010-2011 cohort. Model fit is measured as the sum of absolute deviations (SAD) in predicted probabilities for the intervals $[0,5),[5,10),\ldots,[55,60)$, 60 ECTS, accompanied by the χ^2 fit statistic (with 12 df) and the p-value. As the χ^2-values for the binomial model are very large, we report a lower bound.

Degree programme	n	Beta-binomial model			Binomial model			Estimates		
		SAD	χ^2	p	SAD	χ^2	p	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$
Accountancy and controlling	56	.361	11.81	.461	1.309	$> 10^5$	.000	.56	.44	.35
Business Administration	503	.208	46.97	.000	1.135	$> 10^6$	.000	.83	1.51	.32
Business Economics	139	.273	17.12	.145	0.930	$> 10^9$	.000	.68	.69	.33
Econometrics & OR	49	.438	21.12	.049	1.032	$> 10^6$	.000	.81	1.15	.28
Economics & business	98	.179	10.17	.601	1.078	$> 10^9$	.000	.72	.57	.22
Fiscal economics	21	.394	7.13	.849	1.568	$> 10^6$	.000	.61	.28	.18
Technology management	41	.504	22.07	.037	1.596	$> 10^3$	.000	.85	.94	.17

model is a restricted version of the BBM and thus has less flexibility to fit well. However, the size of the increase in model fit is a clear indication that the model really fits better, and is seen throughout all degree programmes. Table 2 shows the fit for all degree programmes of the Faculty of Economics and Business for the 2010-2011 cohort. The fit of the BBM clearly outperforms that of the binomial model. The χ^2-test for differences between the empirical and modelled frequency distributions provide huge values for the binomial model and much smaller values for the BBM. In 5 of the 7 cohorts, the estimated frequencies do not differ significantly ($\alpha = 0.05$) from the observed frequencies. Similar patterns are observed for other cohorts and at other faculties (for brevity, these results are not shown). Also in Table 2 are the estimated parameters of the binomial model ($\hat{p}$) and the BBM ($\hat{p}$, $\hat{\alpha}$, $\hat{\beta}$). The BBM not only fits better than the binomial model, but it also fits well.

In a wide range of values for the BSA threshold B, the population model seems fairly smooth: an increase in five credits in B yielded a 4% to 5% decrease in the percentage of students achieving a positive advice. At the same time, Figure 3, panel b tells a different story. The figure displays, for the same cohort, the profiles of a 'poor', 'median' and 'good' student. A small increase in B could lead to a steep decrease in the probability of students achieving the threshold. For 'good' or better students – the top 33% – there was virtually no effect from the BSA threshold when $B \leq 40$; these students were practically guaranteed of a positive advise. For 'poor' and worse students – the bottom 33% – the effect of the threshold was also flat for $B \geq 30$: no matter the threshold, these

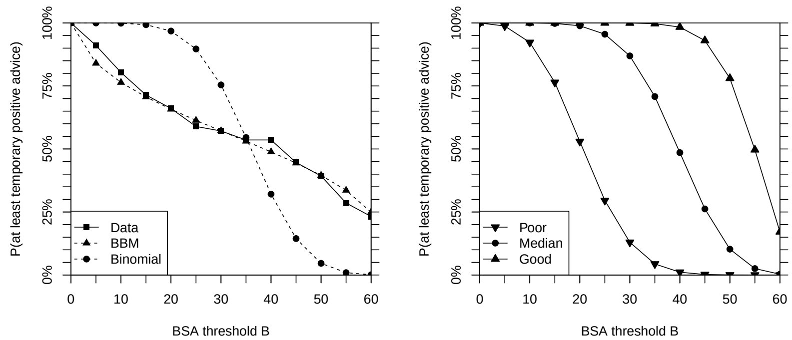

Figure 3: Accountancy, 2010/2011. Left: probabilities of obtaining at least B credits, set out for a range of B. Right: probabilities for the three standardised individual profiles according to the BBM.

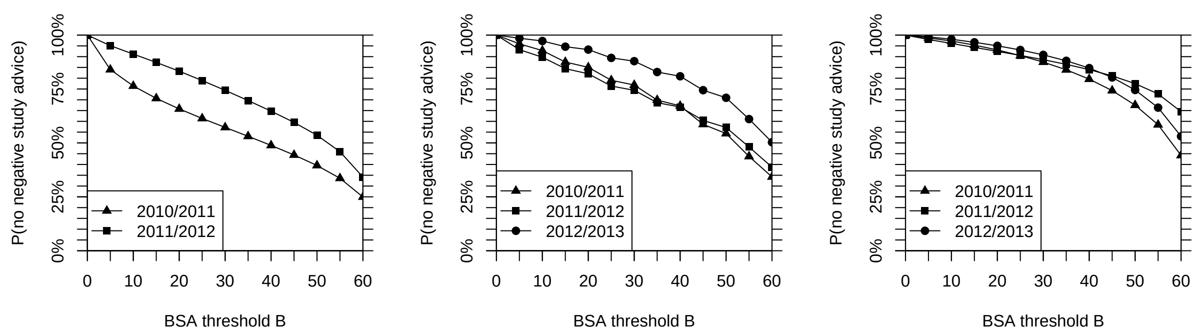

Figure 4: Population curves for the three cohorts for Accountancy (left), History (middle) and Mathematics (right). (Note that the degree in Accountancy ceased to exist in 2012/2013.)

student would not make it. Thus, for thresholds in the range that is of interest to policy makers in Dutch universities, 30 credits (50%) to 50 credits (83%), only part of the students – roughly the middle third – are affected; the others either score below the lowest acceptable threshold or score above the highest acceptable threshold. Even so, overall, a slight increase of five credits for B had only a minor effect; when focused on the group of students for whom this effect was most strongly visible, this effect naturally became much stronger.

In 2010/2011 Accountacy was a degree programme with poor results; the average pass rate p was only 55% (compared with 79% for the whole university that year). Figure 4 shows, from left to right, the cohort curves of Accountancy, History and Mathematics for all three years. The shapes are similar to that of the 2010/2011 Accountancy degree, but shift to the right, showing that students were more likely to reach the threshold.

As Figure 4 shows, there is year-to-year variation among the programmes. The variation between 2012/2013 and the other cohorts, is possibly (partially) a result of the raised threshold. The variation between the first two cohorts is purely a result of other, unknown, factors. The difference among the pass rates for individual courses within these two cohorts often exceeded 10%, even when there were no major changes to the course programme. Table 3 displays the year-to-year variation, clearly showing that the differences between the first two years, both with threshold $B = 40$ were much smaller than between the last year (with $B = 45$) and the first two. The average pass rate per exam in 2012/2013 rose about 8 percentage points compared to the two years before, corresponding to about 4.7 to 4.9 additional credits. This might suggest that the increase of five credits in BSA threshold was nearly fully compensated for by better student performance, but it is not that simple. Like all universities, the University of Groningen actively works on improving its degrees. After 2011/2012, some of the poorer performing degree programmes were cancelled, merged into similar degree programmes or restructured. Of all the 743 different courses taught within all the degrees over the three years, only 254 courses (34%) were taught in all three years under the same course name and credit load. All other courses were either new or have undergone major restructuring. For the 254 'unchanged courses', the average pass rate in 2012/2013 was 1.1 percentage point higher than in 2010/2011 and 1.2 percentage point higher than in 2011/2012. These improved results correspond to an average increase in 0.64 to 0.74 credits, respectively, covering only a minor part, 12.8% to 14.8%, of the increase of five credits in threshold. Thus, we can not infer that the change in the BSA threshold did not affect students because the students compensated the change by working harder. This compensation seems to be responsible for about one-seventh of the increase. The remainder of the increase in performance, seems to have resulted from degree restructuring as well as some random fluctuation or unexplained reasons.

Theoretically, one could attempt to build confidence intervals around inferences for population curves such as those in Figure 4(left). In practice, however, this is very difficult as this would require e.g. detailed knowledge on whether in any of the degree programmes efforts have been undertaken to increase the studyability of the programme, as this might influence the pass rate, as well as unrealistically strong assumptions, such as a fixed variability in pass rate for all degree programmes. Because of this, we do not provide confidence intervals for the population curves for

Table 3: Variation statistics of the 48–59 Bachelor degrees. Top half: mean and standard deviation of the estimated p of all degrees. Bottom half: out of the 47 degrees that were taught all three years, the number of times the cohort had the highest to lowest pass rate.

Statistic	2010-2011	2011-2012	2012-2013
Number of degrees	59	59	48
Average pass rate	.780	.784	.860
SD pass rate	.193	.156	.179
Highest	8	8	32
Middle	22	19	7
Lowest	18	21	9

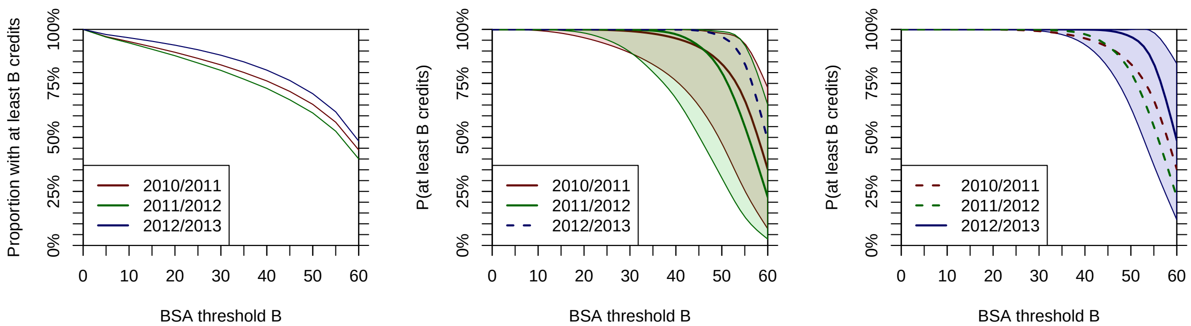

Figure 5: Details for the faculty of Economics and Business. Left: the population curves (lines) for the three cohorts. Middle: (smoothed) individual curves for the two cohorts with $B = 40$. The shaded areas range from 'poor' students (33rd percentile) to 'good' students (67th percentile). Thick curves represent median students. Right: (smoothed) individual curves for the cohort with $B = 45$.

our data.

Aggregating towards faculty and university level

The University of Groningen has approximately 60 different degree programmes. Applying the BBM to all the degrees would provide 60 population figures and 60 individual figures, such as those in Figure 3(right). This information is too detailed for someone who is interested in examining the performance of a faculty or the university as a whole. We condensed the information by aggregating over degrees through weighted sums.

Figure 5 shows detailed results for the Faculty of Economics and Business, see Table 2 for an overview of this faculty's degree programmes and their sizes. We computed the detailed results for the other faculties as well; summarized results are presented in Table 4. Most faculties had similar

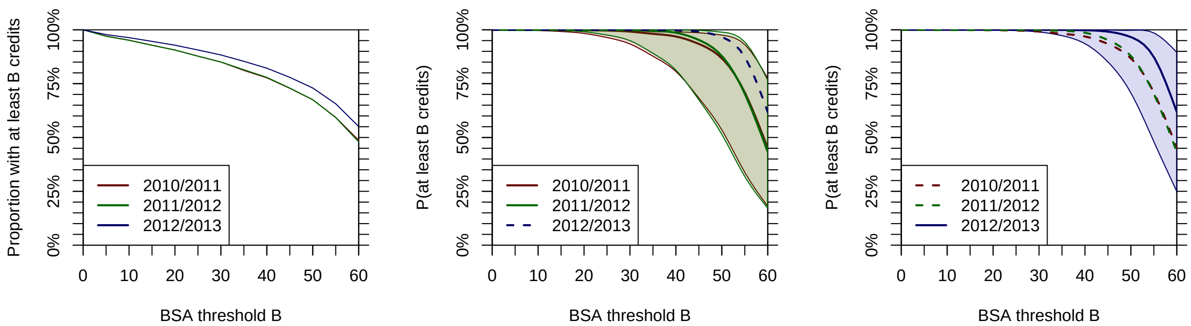

Figure 6: This figure contains the same type of information as Figure 5 but applies to the whole university. Note that, due to the fact that the results of the first two cohorts are so comparable, the population curves overlap almost entirely.

curves, with the exception of Medical Sciences – which seemed to perform better – and Philosophy – which seemed to perform worse. Medical Sciences performed better because this faculty selects its students from a large pool of applicants ('numerus fixus'), whereas the other faculties more or less must accept anyone provided they have an appropriate high school diploma. Philosophy performed worse because it is very popular as a 'second degree' that eager students can take in addition to their main degree. Because the BSA regulations apply only for the main degree, such students are allowed to score below the threshold in Philosophy and consequently drag the performance of the cohort down.

Figure 6 and Table 4 display the predictions according to the model on an aggregated university-wide level. From Figure 6(mid/right) we learn that the differences between cohorts with the same threshold B are *much* smaller than between two cohorts with a different threshold. Given that the results of each cohort are based on the data of over four thousand students, hundreds of different courses, and 48 to 59 different degree programmes, the probability that this finding is due to sampling variation is minimal.

5 Discussion

In this study, we used the beta-binomial model for analysing academic success under the Dutch academic dismissal policy, BSA. Our model is descriptive in the sense that it provides insightful summary of the performance of a cohort of students, given the detailed frequency table of credits

Table 4: Summary per faculty and aggregated to the university as a whole. For the three cohorts (the first two with a threshold of $B = 40$, the third with $B = 45$) it computes the probability for a median student to obtain a positive advice, as well as the proportion of the total cohort achieving this advice; for three potential levels of the BSA-threshold. Reported values are percentages.

Faculty	Cohort	'Median student'			'Population'		
		$B = 40$	$B = 45$	$B = 50$	$B = 40$	$B = 45$	$B = 50$
University	2010	96.9	93.5	86.8	77.7	72.8	67.5
	2011	98.7	95.3	87.9	77.9	73.0	67.6
	2012	99.8	99.2	96.6	82.2	77.9	73.0
Medical Sciences	2010	100.0	99.6	98.2	87.7	84.0	81.1
	2011	100.0	99.9	99.2	89.9	86.7	84.0
	2012	100.0	99.7	98.3	91.4	88.4	85.9
Religious Studies	2010	99.0	95.2	88.2	71.6	67.1	62.9
	2011	93.0	89.4	87.4	87.3	82.8	77.8
	2012	100.0	100.0	100.0	91.0	88.7	86.3
Mathematics & Natural Sciences	2010	98.2	94.8	89.7	81.5	75.8	70.0
	2011	99.2	96.6	91.9	83.1	78.1	72.5
	2012	99.6	98.4	94.9	82.6	78.6	73.6
Social & Behavioural Sciences	2010	96.2	95.4	91.5	77.6	73.1	67.5
	2011	99.8	98.8	94.0	79.2	74.0	67.5
	2012	100.0	99.9	99.1	83.6	79.4	74.1
Arts	2010	97.3	92.4	84.0	76.2	70.9	65.5
	2011	97.3	91.3	82.2	75.0	69.4	64.4
	2012	99.5	98.6	96.1	78.7	74.2	69.6
Spatial Sciences	2010	99.8	98.8	94.7	83.0	77.9	71.2
	2011	98.1	93.5	81.6	77.1	70.2	61.6
	2012	99.9	99.4	96.2	82.2	77.0	70.3
Economics & Business	2010	95.8	91.8	84.0	76.2	71.3	65.3
	2011	97.8	92.4	80.1	72.7	67.4	61.3
	2012	99.9	99.4	96.6	81.2	76.4	70.2
Law	2010	93.7	85.5	71.2	66.5	61.5	56.5
	2011	98.7	94.9	85.9	69.6	65.2	60.7
	2012	99.8	98.7	94.2	76.8	72.0	66.9
Philosophy	2010	93.7	79.3	51.6	65.0	58.2	50.0
	2011	95.0	83.4	60.3	62.9	57.5	51.2
	2012	100.0	100.0	99.3	74.6	71.2	67.3

obtained by that cohort. By comparing the descriptive summaries of different degree programmes, different faculties or different cohorts, one obtains useful information. Furthermore, the model can be used to carry out 'what if'-analyses: e.g. 'what would be the consequences if we raise or lower the BSA-threshold by, say, 5 ECTS?'. By taking subsequent cohorts into account, the 'natural variation' that occurs between cohorts is taken into consideration, something that not always receives the attention it deserves (Boevé et al., 2018).

We have shown that our model provides a very good fit for the data (Table 2). This model is relevant because of the growing interest in systems that filter the students that lack the potential (to finish their degree in time) out of degree programmes as early as possible.

In essence, our model is a dimension reduction technique. The raw data consists of the course pass rates and the relative frequency distribution of ECTS scores and this, essentially, has 60 degrees of freedom. Our model provides a smoothed estimate of the frequency distribution based on only two variables, p and α, a reduction of 58 variables. We showed that this two-dimensional approximation still is a very accurate description of the sixty-dimensional data.

One of the main merits of our model is that it clearly distinguishes between the effect that a small change to the BSA threshold can have on a cohort versus that on individual students. It turns out that a threshold of 40 to 50 credits (out of a maximum of 60 credits) dismisses not only poor students – as is the intention of the BSA system – but also median students; in some cases even good students have a concrete risk of expulsion due to bad luck rather than lack of skill. This remark has important practical ramifications: policy makers should use caution in putting the BSA threshold too high. In addition, it is unavoidable that a small proportion of poor students will achieve a positive advice, even when the threshold is high. A parallel with statistical hypothesis testing is evident; one can alter the testing procedure (significance level, power, etc.) but there will always be a trade-off between false positives and false negatives. By applying a sensible line of thought, it is possible to minimise the false negative error rate for a given acceptable false positive error rate. This is comparable to e.g. medical decisions: when using a mammogram to screen for breast cancer, the medical specialist also takes into account the false positive and false negative rates.

The same applies to the BSA; by applying a sensible line of thought, it is possible to minimise the number of students who are incorrectly not filtered out and those who are incorrectly filtered out. The clear distinction in effect on a cohort versus specific groups of students is a valuable argument in the evaluation of academic dismissal policy. In general, the use of a model with empirical data is crucial to evidence based policy (Nakabo-Ssewanyana, 1999; Reale 2014).

Correct filtering – that is, deciding upon acceptable false positive and false negative error rates – is the primary role of the policy makers of the educational institutes. Our model provides the policy makers with a tool to make an informed decision. Whatever the height of the threshold, there will be false positives and false negatives. Our model quantifies the amount of false positives and negatives, the policy maker should then decide which distribution of false positives/false negatives is 'best'.

Our model does not directly take into account a 'learning effect'; raising the threshold might pressure the students to work harder and achieve better results. Consequently, it is possible to diminish the strong effects on individual students. Such pressure effect was also found by Arnold (2015) and Eijsvogels et al. (2015); the introduction of a BSA, whatever the level of the threshold, raises the average number of ECTS obtained (although correlational data alone are insufficient for causal claims). Without correcting for adjustments in the study programmes, we would find clear evidence of that effect in our data as well. However, after correcting for adjustments in course programmes, our data still hint at such a softening effect, albeit that this effect is only of a small size. It has to be noted that data from only one cohort following a threshold change are not enough to infer from with a high degree of certainty. These data might also be affected by the on-going efforts of a university to improve the education it provides. Successful improvements also yield better results.

To gain knowledge on the future performance, we did a follow-up study. Here, we studied for the three cohorts described in this paper the average performance of all students in the second and third years of the degree programme. Obviously this data is censored: only those who passed the BSA-threshold were allowed into the second and third year. Table 5 shows the average performance in years 2 and 3, split out by performance in year 1. The performance in years 2 and 3 is systematically lower than that in year 1, which can be attributed to regression towards the mean. This table does not provide evidence that the thresholds held by this university were too high. At the same time, the table also shows that the performance of students in subsequent years is also not too low. As a rule of thumb, the university aims for obtaining at least 45 ECTS per year on average, as

Table 5: Performance of students in the second and third year of the Bachelor programme, measured in average number of ECTS obtained. Students in the 2012-2013 cohort with fewer than 45 ECTS were expelled and thus no second year performance is recorded for them. The last two rows depict the number of students in each follow-up cohort, and the correlation of their second/third year performance with their first year performance. All reported correlations have corresponding p-values below 0.01.

	Year 2			Year 3		
Performance Year 1	Cohort 2010-2011	Cohort 2011-2012	Cohort 2012-2013	Cohort 2010-2011	Cohort 2011-2012	Cohort 2012-2013
40–44 ECTS	36	35	(NA)	40	40	(NA)
45–49 ECTS	42	41	37	46	42	43
50–55 ECTS	46	46	42	47	45	47
55–59 ECTS	50	51	48	51	46	52
60 ECTS	57	58	58	58	54	61
N	3306	3520	3493	2764	3081	3217
Correlation	.41	.45	.42	.29	.26	.33

this implies finishing a three year programme within four years, which is considered an acceptable duration. Table 5 shows that year 1 performance is clearly associated with later performance, but that the association is far from perfect. Policy-makers should be reluctant with dismissing students purely on their academic performance in the first year.

Throughout this paper, we worked with the same BSA-threshold B for all degree programmes. Although this now is current practice in Groningen, it is not a requirement. Should a university decide to do so, they can set the threshold different for certain degrees. This does not change how one can apply our model (except that aggregation over degrees with different thresholds introduces additional complexity for interpretation). The reason that the University of Groningen chose a fixed B for all programmes is threefold: (i) a unified approach leads to easier acceptance and implementation, (ii) easier communication towards (prospective) students, (iii) to avoid inequality between students.

Data on student characteristics, such as nationality, gender, age, parents' educational level, etc., were not available to us (because of privacy-related issues). Should one have such information, it can, however, easily be incorporated into the model. One can for instance straightforwardly compare the performance of the male students in a cohort with that of the female students, simply by making separate plots for both genders. Alternatively, such variables can be included as covariates into the model.

A limitation of this paper is that we only report the results based on the data of one single university. The quantitative results presented here apply to the University of Groningen and cannot directly be generalised to other academic institutions. It would certainly be interesting to compare the results for the University of Groningen, presented in this paper, with those of other universities. To achieve this, the model presented in this paper, and the accompanying software, could directly be applied by other institutions.

In this paper, we focused on a descriptive approach to our model. A limitation of the method of moments estimation applied in this paper is that derivation of standard errors and confidence intervals is not straightforward. Assessing the variability of the model, or its parameters, is a complicated technical issue. However, Wilcox (1979) has shown that the method of moments estimation of α and β (and thus, also of p) is $O(\sqrt{n})$ (with n the number of students in the degree programme), which implies better estimation when the sample size increases. Furthermore, the individual curves for poor and good students, presented in Figures 4, 5 and 6, could be seen as a construction of a 33%-confidence interval for a random student. In this paper we showed that the BBM convincingly fits better than a regular binomial model. As future research, it would be interesting to study whether the estimation can be further improved by using alternative parameter estimation techniques (Wilcox, 1979), alternatives to the BBM (Wilcox, 1981), or a random effects model (cf. Snijders and Bosker, 2012). Furthermore, by including student characteristics such as age, parental education level and nationality into the model – when such information is available – one can study whether academic dismissal policies like the BSA treat all students essentially in the same way or not.

Software: the software developed for this study is available, free of charge, from www.casperalbers.nl. A short manual is provided with this software.

References

Arnold, I.J.M. 2015. The effectiveness of academic dismissal policies in Dutch university education: an empirical investigation. *Studies in Higher Education*, 40(6), 1068–1084.

Beerkers, M., E. Mägi, L. Lill. 2011. University studies as a side job: causes and consequences of massive student employment in Estonia. *Higher Education*, 61:679–692.

Belloc, F., A. Maruotti, and L. Petrella. 2010. University drop-out: an Italian experience. *Higher Education*, 60:127–138.

Best, D.J., J.C.W. Rayner and O. Thas. 2010. Four tests of fit for the beta-binomial distribution. *Journal of Applied Statistics*, **37**:9, 1547 – 1554.

Boevé, A.J., C.J. Albers, R.R. Meijer, H.J.A. Beldhuis and R.J. Bosker. 2018. On Natural Variation in Grades at Higher Education, and its Implications for Assessing Effectiveness of Educational Innovations. *Submitted.*

Breier, M. 2010. From 'financial considerations' to 'poverty': towards a reconceptualisation of the role of finances in higher education drop out. *Higher Education*, 60:657–670.

Cohen-Schotanus, J., A.M.M. Muijtjens, J.J. Reinders, J. Agsteribbe, H.J.M. Van Rossum and C.P.M. Van der Vleuten. 2006. The predictive validity of grade point average scores in a partial lottery medical school admission system. *Medical Education*, 40(10): 1012–1019.

Eijsvogels, T.M.H., Goorden, R., Van den Bosch, W., Hopman, M.T.E. 2015. The binding study advice in medical education: a 2-year experience. *Perspectives on Medical Education*, 4(1), 39–42.

Gijbels, D., J. van der Rijt and G. van de Watering. 2004. Het bindend studie-advies in het wetenschappelijk onderwijs: worden de juiste studenten geselecteerd? *Tijdschrift voor Hoger Onderwijs*, 22(2): 62–72.

Goho, J. and A. Blackman. 2006. The effectiveness of academic admission interviews: an exploratory meta-analysis. *Medical Teaching*, 2(4), 255–340.

De Koning, B.B, S.M.M. Loyens, R.M.J.P. Rikers, G. Smeets and H.T. van der Molen. 2013. Impact of binding study advice on study behavior and pre-university education qualification factors in a problem-based psychology bachelor program. *Studies in Higher Education*, iFirst article, 1–13.

Lassibille, G. and M.L.N. Gómez. 2009. Tracking students' progress through the Spanish university school sector. *Higher Education*, 58:821–839.

Lechuga, V.M. 2011. Faculty-graduate student mentoring relationships: mentors' perceived roles and responsibilities. *Higher Education*, 62: 757–771.

Mamiseishvili, K. 2012. International students persistence in U.S. postsecondary institutions. *Higher Education*, 64: 1–17.

Nakabo-Ssewanyana, S. 1999. Statistical data: the underestimated tool for higher education management. *Higher Education*, 37: 259–279.

O'Neill, L.D., B. Wallstedt, B. Eika and J. Hartvigsen. 2011. Factors associated with dropout in medical education: a literature review. *Medical Education*, 45: 440–454.

Ramos, M. and H. Carvalho. 2011. Perceptions of quantitative methods in higher education: mapping student profiles. *Higher Education*, 61:629–647.

Reale, E. 2014. Challenges in higher education research: the use of quantitative tools in comparative analyses. *Higher Education*, 67: 409–442.

Rientjes, B., S. Beausaert, T. Grohnert, S. Niemantsverdriet and P. Kommers. 2012. Understanding academic performance of international students: the role of ethnicity, academic and social integration. *Higher Education*, 63:685–700.

Snijders, T.A.B. and Bosker, R.J. 2012. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second edition. London: Sage.

Stegers-Jager, K.M., J.C. Cohen-Schotanus, T.A.W. Splinter and A.P.N. Themmen. 2011. Academic dismissal policy for medical students: effect on study progress and help-seeking behaviour. *Medical Education*, **45**: 987 – 994.

Tripathi, R.C., R.C. Gupta and J. Gurland. 1994. Estimation of parameters in the beta binomial model. *Annals of the Institute of Statistical Mathematics*, **46**(2): 317 – 331.

Wilcox, R.R. 1979. Estimating the parameters of the beta-binomial distribution. *Educational and Psychological Measurement*, **31**: 527–535.

Wilcox, R.R. 1981. A Review of the beta-binomial model and its extensions. *Journal of Educational and Behavioural Statistics*, **6**(1): 3–32.