Implementing assessment innovations in higher education

Boevé, Anna Jannetje

Chapter 6

Natural Variation in Grades and its Implications for Assessing the Effectiveness of Educational Innovations in Higher Education

Note: Chapter 6 is submitted as

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

6.1 Introduction

Due to increasing performance-based accountability systems in higher education (Alexander, 2000; Liu, 2011), universities have to keep track of student performance as one of many indicators of quality and effectiveness. To achieve this, lecturers need to demonstrate that the results of student evaluations are taken seriously, and to show how changes when necessary, improve the teaching and learning environment. As a result courses are evaluated every year and lecturers keep track of how different cohorts of students perform in subsequent years. At the same time, lecturers also need to evaluate the success of implemented changes or educational innovations, where an important criterion is often the extent to which student performance has improved. This is difficult to measure in practice, however, since variation in test scores across different years may be due to different factors, including differences in exam difficulty, all sorts of cohort differences, and the effect of educational innovations. Using a Randomized Controlled Trial (RCT) to study the causal effects of an educational innovation is usually practically unfeasibly, and alternative designs are needed (Carey & Stiles, 2015; West, et al., 2008). Thus, comparing course results across years is possible, but it is not an easy task.

To disentangle different sources of variation in this context, the aim of this study was to gain insight into the amount of variation in course grades and pass-rates between years across different courses. These variations constitute "naturally expected variability", variability that is bound to exist and is not due to specific interventions. An important advantage of understanding the extent of "naturally expected variability" of exam scores is that lecturers, management, and researchers can anticipate effect sizes necessary to evaluate the success of educational changes. This is especially important in field studies in educational practice, which are often dependent on quasi-experimental designs at best. In this study we will both conduct an analysis on variation in course grades and pass rates and we will provide an example of how this information can be used in a research setting.

6.1.1 Prior Research

There is a long history of research into grading throughout all levels of education (Brookhart et al., 2016). In the early twentieth century, a lot of research focused on the variability and reliability of grades in primary and secondary education, while research on grades in higher education has focused a lot on course evaluations (Brookhart et al., 2016). There is some research on the variation of grades in higher education, mainly focused on student Grade Point Average (GPA). Kostal, Kuncel, and Sackett (2016) found evidence of GPA inflation between the mid 1990's and 2000's, and argued that instructor leniency must be an important source of the observed grade inflation. Other research on GPA in higher education focused on reliability, with Beatty (2015) finding that student GPA in the first year of college, and over the entire college period is highly reliable and did not vary much between institutions. While the focus on student GPA in research has been necessary and fruitful, research on the variability of college grades from a course perspective is lacking.

Important research has also been conducted at the primary and secondary level of education. Hollingshead and Childs (2011) showed that there was more variation over time for small schools relative to large schools in large-scale research on the percentage of students above a cut-score in Canadian primary education. School mean grades are another common

aggregate measure that is often used to consider school performance in primary education. Wei and Haertel (2011) showed that ignoring the clustering of students in classes within schools led to biased reliability and standard errors of school mean grades. In the context of secondary education, Luyten (1994) showed that there was both systematic variation in mean grades across years for specific subjects, as well as systematic variation in mean grades between courses.

The above research has important implications for the context of understanding the variability of grades in higher education. Given the more limited time, resources, and expertise of lecturers to ensure equal exam quality every year, pass rates and mean grades may vary more in higher education compared to primary and secondary education standardized testing. On the other hand, the massification of higher education may contribute to smaller standard errors given larger classes compared to primary and secondary education. The clustering of grades is an important factor to take into account as demonstrated by Wei and Haertel (2011). While research in higher education has often considered student GPA, the clustering of grades within years within courses has not been investigated. Similar to secondary education as investigated by Luyten (1994), students in higher education also take different courses taught by different teaching staff. Thus, grades in higher education are also expected to vary between courses, as well as within courses across different years.

While there is little large-scale research on course grades in higher education, course grades are often used in small-scale field studies to investigate various changes or innovations in the learning environment, with sometimes firm conclusions. Therefore, in the present study we examined the variation in course grades and pass-rates in higher education and illustrate how this information can be used to better compare course mean grades across different years.

6.2 Method

6.2.1 Data

Fully anonymized administrative records containing assessment results from the academic years 2010/2011 through 2015/2016 from the University of Groningen, the Netherlands were analyzed for the present study. The university administration provided assessment records for all first-year courses at all nine faculties of the university at that time. This research classifies as document-research for which no ethical approval was necessary according to the guidelines of the ethical committee at the University of Groningen.

Table 6.1 shows the faculties by both the full faculty name and an abridged short description that will be used in the remaining text. Table 6.2 shows the mean (*sd*) grade and pass rate per faculty. All courses from the first year of all bachelor degree programs were included. We only used first-year courses since these are obligatory and prerequisite introductory courses for further specializations later in the bachelor degree programs. Using these courses, a good picture could be obtained from the results of complete cohorts. In addition to the full cohorts of enrolled students, second- and third-year students from other bachelor degree programs may also take first-year courses in order to complete a minor. These students were also included in the data analyzed. The data analyzed had the following structure: an anonymous student-identifier, a course-code, a faculty code, date of examination, examination attempt, and examination result in the form of a grade or pass/fail.

Table 6.1 Number of assessment observations per faculty in each year, with mean (*SD*) grade and overall pass rates.

Full faculty name	Short name	*N* assess-ments	*N* year-courses	*N* unique courses	*N* unique students	Mean grade (*SD*)[a]	Mean pass rate[b]
Arts	Arts	65,798	1094	358	9,270	6.74 (0.74)	.80
Behavioural & Social Sciences	Social	73,563	427	112	8,155	6.45 (0.66)	.77
Economics & Business	Economy	83,952	354	115	9,879	6.25 (0.66)	.74
Law	Law	36,953	147	43	5,785	6.18 (0.74)	.72
Medical Sciences	Medicine	26,385	221	74	3,945	6.65 (0.61)	.80
Philosophy	Philosophy	6,301	110	36	1,388	6.73 (0.62)	.83
Science & Engineering	Science	68,209	622	139	6,709	6.67 (0.79)	.80
Spatial Sciences	Spatial	11,676	104	30	2,023	6.44 (0.65)	.71
Theology & Religious Studies	Theology	2,256	126	33	428	7.10 (0.70)	.92
Total		375,093	3205	940	47,582	6.61 (0.74)	.78

[a]Mean grade (*SD*) is computed as the mean (*SD*) of the mean grades per course
[b]Pass rate is computed as the mean of the pass rates per course

Table 6.2 Mean grade and overall pass rate for each cohort (disregarding faculty)

Cohort	Mean grade (*SD*)	Overall pass rate
2010	6.66 (0.78)	.79
2011	6.57 (0.74)	.78
2012	6.66 (0.76)	.79
2013	6.57 (0.76)	.78
2014	6.60 (0.70)	.78
2015	6.61 (0.70)	.80

In the data cleaning process, after removing empty rows and duplicate records, we selected main course results (excluding partial assessment records kept by some faculties), first-attempt results (excluding re-sits), and excluded exemption records, resulting in a total of 375,222 assessment records. Subsequently, courses were excluded if they consisted of only one student, as these have no within-course variation (n = 129). The final data consisted of a grand total of N = 375,093 assessment records from 940 unique courses (see Table 6.1 for

further details per faculty). In the appendix, tables A6(a-c) show the distribution of assessment records across faculties and cohorts, and by number of cohorts per course in the data.

The total number of students in the data equaled N = 40,087, whereas the total number of unique faculty-student combinations was N = 47,582. These numbers imply that some students took first-year program courses in more than one faculty, for example because they were enrolled in two programs simultaneously. The total number of unique student-year combinations was N = 58,612. This means that some students took courses from first-year bachelor degree programs within the same faculty in different years. Common reasons for students taking first-year program courses in multiple years include: delayed study program due to illness, unforeseen circumstances, double-degree enrollment, and following a minor-program from another bachelor program at the same faculty as the main degree of enrollment. It is important to stress that only a student's first course enrollment and assessment result were included in the data, thus there were only unique student-course combinations: a student-course combination cannot occur more than once in the data.

6.2.2 Measures

The variation in student performance was operationalized by variation in student grades and by whether students passed or failed an exam. As most continental European countries, in the Netherlands a number grading system is employed. For most courses (specific to each year), 96.8% (N = 3101) gave grades on a scale ranging from 1 to 10 where grades of 6 and higher represent a pass. Sometimes grades are given with decimals; for the present study all grades were rounded to a single integer. A small part of the courses (specific to each year) 3.2% (N = 104) only recorded whether the student passed or failed an exam, thus providing a dichotomous result.

6.2.3 Analyses

Most research on student grades in higher education has focused on student GPA, as the main outcome of interest. In order to examine the variation in outcomes across years and between courses in the present study we focused on course grades. This means that a nested structure was assumed, which is depicted in Figure 6.1. The illustration of the different nesting structures of interest to the present research on course grades, compared to research on student GPA illustrate that the same data can be assigned to different levels and that both models are essentially incomplete. In the common perspective of student GPA, the lowest level observations are not independent as each student does not take a new set of courses, but rather some students take the same set of courses. Similarly, in the present study, courses in particular years do not all have a new set of students, but rather some course-years share a common set of students. This complexity in higher education assessment data is an important challenge for researchers, but beyond the scope of the present study to solve definitively. A work-around for this problem, feasible due to the very large sample size, is as follows.

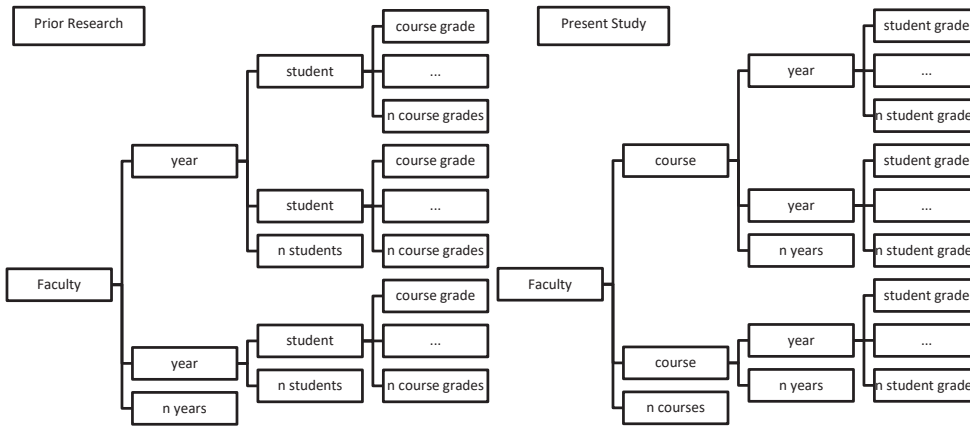


Figure 6.1. Conceptual visualization of the assumed nesting structure in prior research on student GPA (left), and the nesting structure of interest to the research question in the present study (right).

To avoid violation of independence assumptions, the analyses in the present study were repeated for 25 samples of the data where only a single assessment result was included for each student. In the first step, therefore, a single assessment result was sampled at random for each student. For students with a single assessment in a particular year and faculty, the probability of inclusion of this result would be 1. These records would therefore always be included, which may bias the findings. Therefore, a second step was added where a random 75% of the assessments selected in the first step were included.

6.2.4 Models

We constructed two models: the first model concerned the variation in mean grades and, thus, is applicable to 96.6% of the data. The second model concerned variation in the pass rate. As, obviously, a grade can always be converted into a pass/fail-statement, this model is applicable to the full data set.

Model for mean grades

The variation in course grade results was examined by estimating an intercept-only multilevel model (Snijders & Bosker, 2012; Hox, 2010) with three levels for student grades as follows:

$$Y_{ijk} = \gamma_{000} + v_{00k} + u_{0jk} + e_{ijk}, \quad (6.1)$$

where a particular grade Y_{ijk} for student i in year j in course k is modeled by the expected value γ_{000}, with a random error component for the course level (v_{00k}), a random error component for the year level (u_{0jk}) and a residual error component (e_{ijk}). All random components are assumed to be normally distributed around zero. As shown in Figure 6.1, courses are also nested in faculties. However, the number of nine faculties was too small to include as a separate level (Hox & Maas, 2005). In order to explore whether there were differences in

mean student performance per faculty, we included faculties as fixed effects, with the faculty of Arts as the reference group. In addition, we examined the proportion of variance at the year and course-level within each faculty by separately estimating the model shown in Equation 6.1 for each faculty.

The variance decomposition at different levels was investigated in the following way for student grades. First, we examined the total proportion of variance between courses and years as

$$\rho_{course.year} = \frac{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}{\sigma^2_{e_{ijk}} + \sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.2)$$

where $\sigma^2_{e_{ijk}}$ denotes the remaining variance in grades at the lowest level, $\sigma^2_{u_{0jk}}$ denotes the variance between years, and $\sigma^2_{v_{00k}}$ represents the variance between courses. The residuals of each level are assumed to have a normal distribution, around 0. Next we examined what proportion of the higher level variation is specific to the year level by:

$$\rho_{year} = \frac{\sigma^2_{u_{0jk}}}{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.3)$$

Model for pass rates

To model the pass rates, a couple of additional steps were required. To examine variation in pass-rates, we modeled the log-odds of whether an assessment result was a pass (1) or a fail (0) as

$$\pi_{ijk} = \text{logistic}(\gamma_{000} + v_{00k} + u_{0jk}) \quad (6.4)$$

where π_{ijk} indicates that an assessment *i* in year *j*, in course *k* yielded either a pass or a fail which is assumed to have a binomial distribution, with an expected value of Y_{000}, a random error component across years (u_{0jk}), and with a random error component across courses (v_{00k}). After estimating this model, a second model was estimated to explore whether the mean log-odds of passing differed in each faculty. As in the analyses of grades, dummy-variables for each faculty were specified with the faculty of Arts as the reference faculty. In order to explore whether the amount of course- and year- level variance in log-odds of passing varied across faculties, the intercept only-model in Equation 6.4 was also repeated for each faculty separately.

The logg-odds are not straightforward to interpret, but can be transformed back to probabilities using the relation $p = e^{\pi}/(1 + e^{\pi})$. In each multilevel model with dichotomous outcomes, the variance of the lowest level = is scaled to 3.290 (which $\pi^2/3$, Snijders & Bosker, 2012). This means that in each model for binary outcomes using the logistic link, the residual variance is the same. To examine the variance in log-odds of passing at higher levels, the proportion can be decomposed as:

$$\rho_{year} = \frac{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}{3.290 + \sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.5)$$

$$\rho_{course} = \frac{\sigma^2_{u_{0jk}}}{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}} \quad (6.6)$$

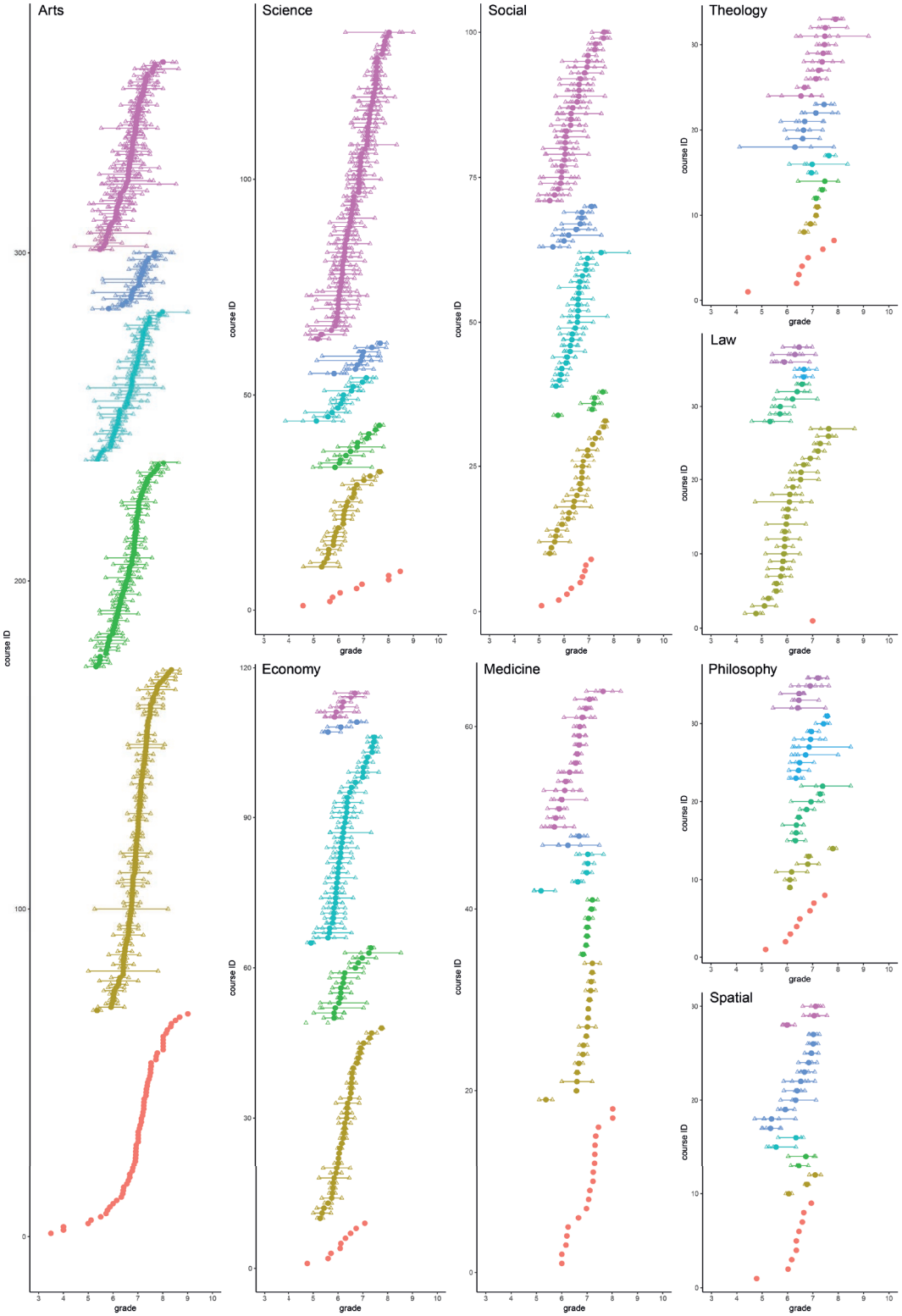
6.2.5 Software

All analyses were conducted in *R* (*R* Core team, 2017, version 3.4.1), using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015, version 1.13). Full maximum likelihood estimation was used to estimate the model deviance, in order to be able to compare the intercept-only model with the model including fixed-effect dummy variables for the different faculties.

6.3 Results

To depict the variation in mean course grades Figure 6.2 shows the overall mean course grade, and the mean course grade for each year within a course for all faculties included in the data.

Figure 6.2. The variation of mean course grades within and between each course in each faculty included in the data, with colors indicating the different number of cohorts for each course (6 to 1 years from top to bottom). Each line represents the distance between the lowest mean year grade and highest mean year grade for each course, triangles representing mean grade in each year, and closed circles the mean grade for each course.


Arts
Science
Social
Theology
Law
Economy
Medicine
Philosophy
Spatial
course ID
grade

6.3.1 Course Grades

Table 6.3 shows the model results for the intercept only model, and the model with faculty included as a dummy variable in the analyses. Overall, about 17% of the variation in grades can be attributed to systematic variation between courses and years. When adding faculties as a fixed effect by means of dummy variables to the model there is a statistically significant reduction in the model deviance (Δ deviance = 107.58, $df = 8$, $p < .001$), implying better model fit. Variation in mean grades between faculties explains about 10% of the variance between courses, which is about 1% of the total variance. The size of the variance components may be underestimated due to the violation of independence, as shown in the mean variance estimates over 25 replications (see Table 6.3). Running a separate intercept only model for the different faculties shows that the amount of total course and year variation ranges between 11 to 20% (see Table 6.4). Furthermore, of the higher-level amount of variance, Table 6.4 also shows that the proportion at the year-level ranges from 25% to 52%.

Table 6.3 Estimates of the fixed effects, random effects, and model deviance for course grades

	Intercept only model	**Model including faculty fixed effect**	**Intercept only model 25 samples**
Fixed effects (*SE*)			
Intercept Y_{000}	6.59 (0.02)	6.76 (0.03)	
$D_{Theology}$		0.25 (0.11)	
D_{Law}		-0.60 (0.10)	
$D_{Medicine}$		-0.01 (0.08)	
$D_{Science}$		-0.16 (0.06)	
$D_{Economy}$		-0.51 (0.06)	
D_{Social}		-0.30 (0.07)	
$D_{Philosophy}$		-0.09 (0.11)	
$D_{Spatial}$		-0.36 (0.11)	
Random effects			Mean (*SD*)
Courses $\sigma^2_{v_{00k}}$	0.32	0.28	0.36 (0.02)
Years $\sigma^2_{u_{0jk}}$	0.15	0.14	0.16 (0.01)
Grades $\sigma^2_{e_{ijk}}$	2.27	2.27	2.48 (0.01)
Deviance	1,294,263	1,294,156	
Δ Deviance		107	
$\rho_{course:year}$	.17	.16	
ρ_{year}	.31	.34	

Table 6.4 Variance partition of grades at the different levels for each faculty

Faculty	Residual variance	Year-variance	Course-variance	$\rho_{course:year}$	ρ_{year}
Theology	1.44	0.21	0.14	.20	.41
Law	2.87	0.21	0.33	.16	.38
Medicine	1.34	0.09	0.24	.19	.29
Science	2.36	0.15	0.44	.20	.25
Arts	1.92	0.16	0.26	.18	.38
Economy	2.59	0.12	0.25	.13	.33
Social	2.23	0.15	0.26	.16	.37
Philosophy	2.31	0.14	0.13	.11	.52
Spatial	1.50	0.11	0.27	.20	.30

6.3.2 Pass rates

Based on the model with the full data, Table 6.5 indicates that about 40% of the variance in the log-odds of passing is at the year and course level. Of the higher-level variance, about 23% is due to differences between years within courses. When taking 25 subsamples of the data, so that the independence assumption is not violated the variance components are smaller. Table 6.6 shows that there is considerable variability between faculties in the amount of variance in log-odds at the year and course level, with estimates ranging from 22% to 74%. Furthermore, the relative amount of variance at the year-level within a course rather than between courses, also varies considerably, from 5% to 70%. It is important to note that these percentages of variability at the log-odds level do not translate easily to percentages at the pass-or-fail level, which will be made clear in the application.

Table 6.5 Estimates of the fixed effects, random effects, and model deviance for the log-odds of passing

	Intercept only	Model with faculty as fixed effect	Intercept only model over 25 samples
Fixed effects (*SE*)			
Intercept Y_{000}	1.80 (0.05)	1.82 (0.07)	
$D_{Theology}$		1.73 (0.29)	
D_{Law}		-0.40 (0.23)	
$D_{Medicine}$		0.21 (0.18)	
$D_{Science}$		0.11 (0.14)	
$D_{Economy}$		-0.44 (0.14)	
D_{Social}		-0.05 (0.15)	
$D_{Philosophy}$		0.04 (0.24)	
$D_{Spatial}$		-0.59 (0.27)	
Random effects			Mean (*SD*)
Courses $\sigma^2_{v_{00k}}$	1.73	1.62	0.92 (0.33)
Years $\sigma^2_{u_{0jk}}$	0.51	0.51	0.30 (0.02)
Deviance	366,947	366,885	
Δ Deviance		65	
$\rho_{course:year}$	.41	.39	
ρ_{year}	.23	.24	

Table 6.6 Coefficients for the random intercept models on the grades and log-odds of passing for each faculty

Faculty	Year-variance	Course-variance	$\rho_{course:year}$	ρ_{year}
Theology	2.60	6.94	.74	.27
Law	0.20	3.45	.53	.05
Medicine	0.14	2.24	.42	.06
Science	1.01	2.91	.54	.26
Arts	0.47	0.84	.28	.36
Economy	0.21	1.36	.32	.13
Social	0.54	1.95	.43	.22
Philosophy	0.24	0.70	.22	.25
Spatial	1.40	0.60	.38	.70

6.3.3 Application

Consider the following scenario, with intentionally simplified numbers: A course instructor is interested in implementing a new teaching method. It is not possible to do a randomized experiment, and the instructor would like to compare the results of the previous-year, that is the results prior to the implementation of the new teaching method, with that of the current year, that is, the results after implementing the changes. In both years, n = 50 students participate, and the GPA for both years is 6.00 and 6.50, respectively. In both years, the standard deviation of grades is 1.00. A standard t-test shows that the increase in GPA is highly significant $t(98) = 2.50$, one-sided $p = .007$). Concluding that, thus, the new teaching method is beneficial is misleading, as the regular year-to-year variations are not taken into account. To infer a significant increase in GPA after an educational intervention, the increase in GPA should not just be significantly above zero, but significantly above regular values obtained from year-to-year variation.

The variance partitioning of grades and year-variation in the present study can be informative: based on the estimated proportion of variance across years, a course-instructor can estimate the 95% CI around the difference between two cohort mean grades as follows:

$$0 \pm t^* \times \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \times (\sigma^2_{year} + \sigma^2_{residual})} \quad (6.7)$$

where n_1 and n_2 are the number of students participating in the course both years, t^* is the critical t-value with $n_1 + n_2 - 2$ degrees of freedom. The course-level variance is excluded here since the result in both years is for the same course. For a random course the year-level variance component of the overall model can be used based on the intercept-only model. It is also possible to use a faculty-specific variance component if the faculty is known. Figure 6.3 shows the 95% confidence interval around the mean grade for different possible numbers of students in each cohort, based on the estimated variance components of the overall model. From this figure, it is clear than an increase of 0.5 in GPA for a course with 50 students per year is non-significant. For larger courses, for example, with 200 students per year, a 0.5 increase would be a significant effect.

Similarly, Equation 6.7 can also be used to estimate the 95% confidence interval around the log-odds of passing. In contrast to the application of mean grades this is, however, dependent on the intercept (i.e. the log-odds of the average pass grade), while the application for grades is equivalent regardless of the mean expected grade. Figure 6.4 shows the same 95% confidence interval after transforming the log-odds interval back to the probability of passing. Say a lecturer observes that the original cohort had a pass rate of .86, and observes a pass rate of .90 in the course with the new lecture method, Figure 6.4 shows that you need at least 150 students per year for this difference to be significant at the 5% level.

To illustrate how the confidence interval of the log-odds varies depending on the expected intercept, Figure 6.4 shows the interval for different possible numbers of students given three different intercepts, based on the quantiles of pass-rates in the present data. This figure shows that whether a certain increase from year 1 to year 2 in pass rate is

significant depends on the pass rate of year 1. For instance, in a course with 100 students, a 5 percentage point increase from 60% to 65% is not significant, whereas the same increase from 90% to 95% would be. In general, for pass rates closer to 1 (or to 0), smaller increase in pass rate can be significant than for pass rates closer to 50%.

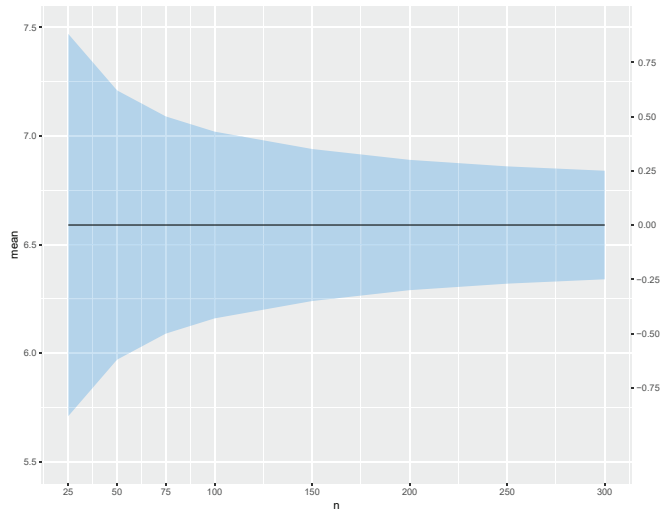


Figure 6.3. The 95% confidence interval around the predicted mean grade when $n_1 = n_2 = N$. The horizontal line is placed around the observed mean grade in the data set (left vertical axis), but the width of the confidence interval would be the same when placed around another mean value. The right vertical axis displays the deviation from this mean value.

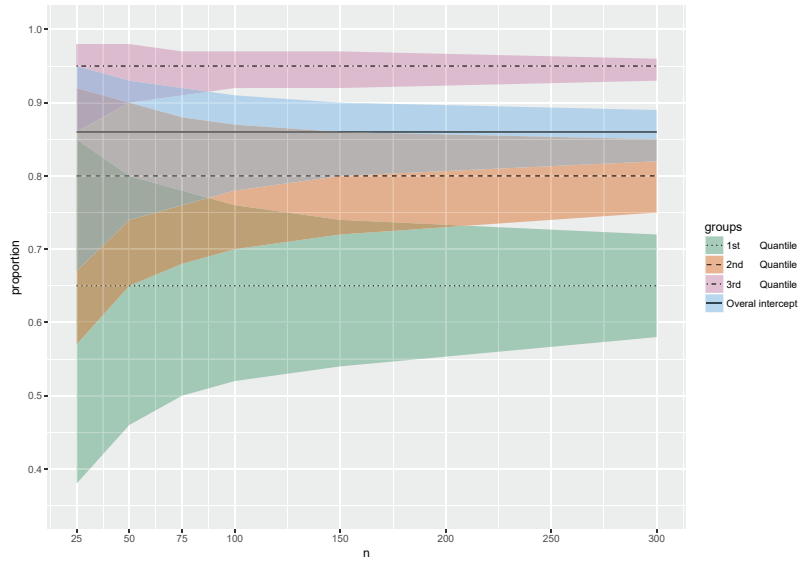


Figure 6.4. The 95% confidence interval of the probability of passing based on the overall model intercept (mean of .86), and the quantiles of mean pass-rates, when $n_1 = n_2 = N$.

6.4 Discussion

The aim of the present study was to explore the extent to which assessment results, both in terms of grades and passing, vary between years within courses and the extent to which they vary across courses within different faculties. Disregarding the different disciplines of the different courses, the present study found that about 17% of the variation in grades was at the year- and course- level. Of this variation, about 30% was due to variation within courses across different years, whilst the remaining 70% was due to systematic variation between courses. Despite the high reliability of student GPA as demonstrated by Beatty et al. (2015), the present study showed that year-over-year variations in grades may be considerable.

When examining the log-odds of whether an assessment result was a pass or a fail, we found that approximately 40% of the variance was at the year- and course- level, with 25% of this variation across different years within courses, and 75% between different courses overall. When accounting for different disciplines (faculties) in the data, the amount of variation between courses decreased slightly from 17% to 16% in terms of course grades, and from 41% to 39% in the log-odds of passing.

In line with the findings of Luyten (1994) in the context of secondary education, the present study found that the proportion of course-level variation was larger than the variation within courses across years. However, exploring discipline specific differences in the amount of variation at the course- and year-level revealed substantial differences between faculties. The overall amount of higher-level variation varied from 11%-20% concerning grades, and ranged from 22% to 72% for the log-odds of passing. Of the higher-level variance, the proportion of variance across years ranged from 25%-52% for grades, and for the log-odds of passing, the proportion of variance across years relative to the higher-level variance ranged from 5%-70%.

The implications of the findings in this study are severe, as was shown in the application section. In educational literature, innovations are often judged effective based on a direct comparison of two cohorts, without taking this the “naturally expected variation” into account. Disregarding the general fluctuation in course grades over time leads to a severe increase of false positives as innovations may incorrectly be labeled as effective. Whereas for a course with 50 students a difference in grade mean of 0.5 *SD* before and after intervention would be considered highly significant (p = .007) when disregarding this variation, the difference actually is non-significant at the α = .05 threshold. At least 75 students are needed to get the p-value below.05.

In line with the findings of Hollingshead and Childs (2011), as the number of students increase, the uncertainty around both the mean course grade and pass-rates decrease. This study demonstrated that, even with large sample sizes, conclusions about cohort differences should be taken with caution. For instance, for a large course, with 300 students per year, an increase in pass rate from 65% to 70% is not even significant. When ignoring the natural variation, this difference would be highly significant.

As evaluations of educational innovations ignore this natural variation, it is to be expected that the number of false positive findings is very large. A practical recommendation to avoid this, is as follows. Based on the number of students in a course, one can use Figure

6.3 to find the value δ which is the maximum value of the difference in mean grades in two consecutive years, $m_2 - m_1$, which would be non-significant. For instance, with $n = 50$, $\delta = 0.62$. Rather than testing for a significant difference between both means (H_0: $\mu_2 - \mu_1 = 0$, with the standard *t*-test), one can then test whether the difference between both means is significantly larger than δ or not (H_0: $|\mu_2 - \mu_1| < \delta$). For this, one can use equivalence tests (Schuirmann, 1987; Lakens, 2017). To claim a successful educational innovation, the difference between grade means should significantly exceed δ, rather than just significantly exceeding 0. When the interest lies in the pass rate, rather than the grade mean, a similar approach can be employed using Figure 6.4.

6.4.1 Limitations

The present study was focused on assessment in higher education. As always in data analysis, not all potentially relevant variables are measured. Some faculties offer multiple bachelor degree programs, and there may be systematic variation between bachelor programs within the same faculty. As type of bachelor program was not recorded in our dataset, we could not take this level into account in our analyses. Also the effect of individual lecturers could not be taken into the model as this information was not part of the data set.

Another limitation in the present study was that it is unknown to what extent courses were taught in the same way or by the same lecturers in different years. Major education innovations did happen, but not without assigning a new code to the course (thus treating both courses separately). The variance across years however, likely does include lecturer experimentation with perhaps new technology or assessment methods. Note that the average grade did not increase significantly over the years (Table 6.2).

The main limitation of this study has to do with the generalizability of the results. The present study examined the grades and pass-rates of first-year courses in higher education at a single university, the University of Groningen. Given the large amount of information, the estimated variance components could be informative for other institutions, especially those using a number grading scale. Although it is unknown to what extent the numerical findings in this study are representative for other universities, it is obvious that also at other places a considerable part of grade variation can be labeled as 'natural variation'. Thus, the message that many 'significant' findings in assessing educational interventions are actually false positives holds, but further research is needed to assess *how* many of these findings are false.

Furthermore, higher education institutes can employ the model introduced by us on their own assessment records. If they find more natural variation at their institute than we did in our study, an even larger grade increase is required for a successful intervention. Reversely, with less natural variation, smaller increases can be labeled as successful.

Finally, the present study demonstrated that assigning observations to different levels is sometimes not straightforward (see, e.g., Hox, 2010). For research on student grades in higher education, the focus has often been on the student with the interest in explaining why individuals differ in their achievement, here the focus was on how courses differed in achievement across different years.

6.4.2 Conclusion

The goal of this study was three-fold: (i) introducing a model for assessing "natural variablity" in grades in higher education, (ii) estimating the parameters in this model based on a large (n = 375,093) data set from a single university, (iii) showcasing the consequences of ignoring this natural variation in studying whether an educational intervention yields a significant increase in grades and/or pass rate. The assessment records of higher education institutes contain valuable information when examined from a course perspective rather than from a student perspective. Understanding the variation in course results across years can help lecturers and institutions to evaluate the impact of innovations at a cohort level, while reducing the risk of false positives when grades between two subsequent cohorts are compared.