What Can We Learn from Exam Grade Distributions?

Daniel R. Lewin

Chemical Engineering, Technion I.I.T. Haifa, Israel

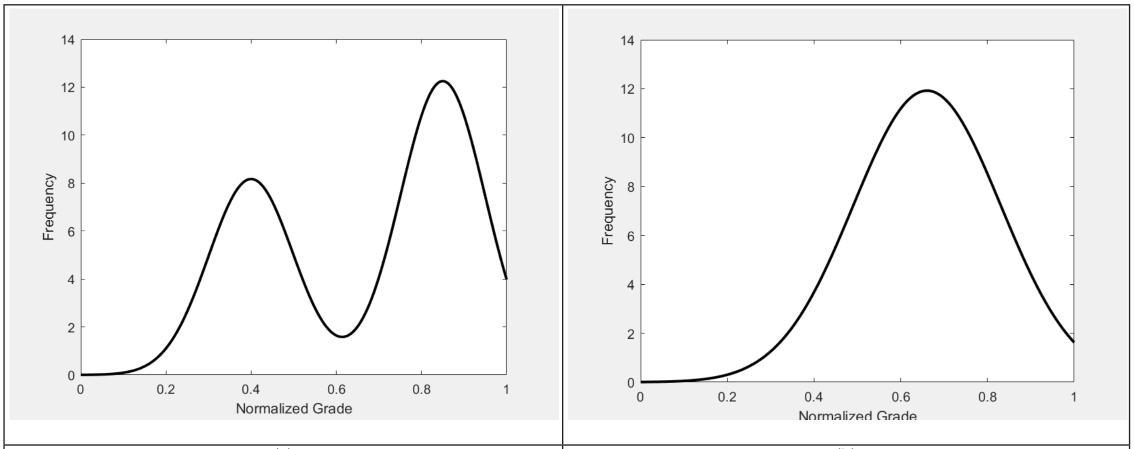
Received: 11 December 2020; Accepted: 11 June 2021

A course with good learning outcomes is one in which most of the enrolled students achieve the mastery specified in the predefined learning objectives. Since the enrolment is invariably a mix of students with heterogeneous capabilities, the class average grade is a poor indicator of how the class is divided into at least two groups, of high- and low-performers. Clearly, achieving the desired outcomes implies increasing the proportion of high-performing students and their mean grade by providing appropriately designed teaching protocols. In this paper, the actual class grade distribution is approximated by a bimodal probability distribution function, whose parameters enable the proportions and average performance of these two groups to be quantified. This paper describes the methodology to achieve this and demonstrates its usage to diagnose example exam grade distributions, as well as to provide quantification of the impact of pedagogic changes on the degree of achieving teaching objectives.

MOTIVATION

Paraphrasing Antoine de Saint-Exupéry, here is a motivation for this study: "I pondered deeply on how a heterogeneous student population, comprising of high-performing and low-performing students, can affect the overall class performance, and how it might be possible to diagnose problems by parametrizing class heterogeneity in a useful way. After some work with coloured pencils, I succeeded in making my first drawings, which looked like those in Figure 1."

"I showed my masterpieces to my peers and asked them whether my drawings of grade distributions worried them. But they answered: 'Worry? Why should anyone be worried about the distribution? The main thing is that the average grade is reasonable – who cares about the shape of the distribution?'" However, in contrast with the worrying distribution in Figure 1(a), Figure



(a) (b)

Figure 1. My first drawing of a worrying grade distribution with a normalized average of 0.65, shown in (a), and a drawing of an equivalent normal distribution with the same average, shown in (b).

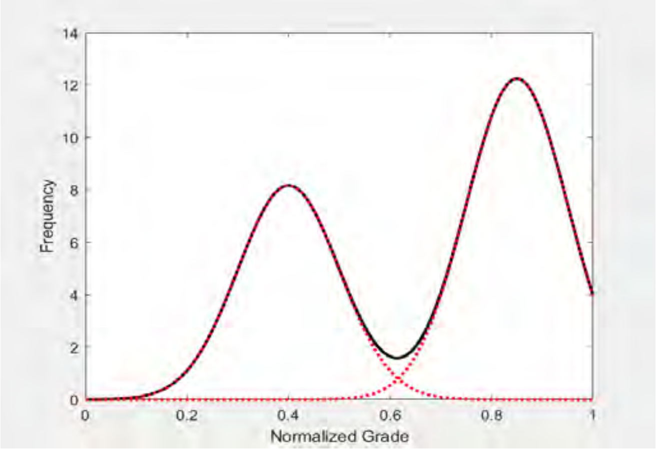


Figure 2. My second drawing of the grade distribution in Figure 1(a).

1(b) shows a normal distribution with the same mean value as that depicted in Figure 1(a). Clearly, my drawing in Figure 1(a) is not a picture of a normal distribution, but rather, of a bimodal distribution, whose average is misleading, as the actual student population is divided into two distinct groups, one in which grades are high-performers' and the other of low-performers'. In such circumstances, an unacceptably high proportion of the class score well under the computed average. Would both of the grade distributions presented in Figure 1 be equally desirable in a course outcome, even if they both have the same average grade? Accordingly, I analyzed the grade distribution in Figure 1(a) by fitting the parameters of a bimodal distribution model, producing the picture as shown in Figure 2.

Note that Figure 2 discloses that the distribution is composed of two distinct normal distributions, for the low- and high-performing students respectively, in which 40% of the class is in the low-performing group, with an average grade of 40%, with the passing grade being 55%. Regrettably, the common practice of grading on the curve simply adjusts the distribution to achieve acceptable averages. This can lead to several problematic outcomes of grade distributions such as the one illustrated in Figure 2:

a. Had the bimodal distribution's average grade been 45% and not 65%, the often acceptable "fix" of adjusting the grades to make the average, say, 70%, would simply jack up all of the grades by a factor, allowing many students who did not master the materials examined to pass the exam. This is clearly an undesirable outcome, especially so if mastery of the course in question is a prerequisite in subsequent courses.
b. Students are smart enough to recognize that if their passing grade, obtained by adjusting grades by a factor of significant magnitude may mean that in many cases, the factor itself may contribute as much to the reported grades as their own efforts. If the usage of factoring is repeated semester after semester, this can only erode the students' desire and motivation to achieve real learning (Kulick and Wright, 2008). Again, this is a highly undesirable outcome for students.
c. Ignoring the bimodal nature of the distribution overlooks the fact that a sizable portion of the class did not achieve the desired course outcomes. Would a conscientious teacher accept the situation where only 60% of the class achieved course mastery? Therefore, the practice of grading on the curve also leads to undesirable outcomes for teachers.

If we accept the premise that a teachers' duty is to encourage as many as possible of his/her students to achieve mastery in the courses being taught, then it should be of interest to monitor the *degree of engagement* of the students with the course during the entire semester, so that the proportion of the students who do not achieve mastery is reduced as much as possible. Clearly, not everything is in the teacher's control: As the saying goes – you can lead a horse to a water trough, but you cannot force it to drink. However, it is more likely that the horse will drink if you treat it well, and if you make the water taste and smell good. In the same spirit, the teacher does have significant impact on students' motivation and curiosity, and on the availability of clearly presented resources and information. The teacher's positive and patient attitude when interacting with the students in the contact time they share has a huge impact on students' engagement, as does aiming to maximize the degree to which students are participating actively with the teacher and with each other, rather than passively listening to lectures (Freeman et al, 2014; Velegol et al, 2015). Evidently, then, a key desired outcome would be to minimize the percentage of low-performing students in the final outcomes evaluations, and to measure this, one needs to be able to analyse grade bimodality, as illustrated in Figure 2. Having this analysis widely available will gauge the degree of each teacher's success and drive the lecturer to improve the teaching protocol in future course offerings to reduce the proportion of low-performing students.

This paper introduces a diagnosis method that provides a parameterized measure of the degree of bimodality present in an exam grade distribution, which can be utilized by teachers who wish to quantify the degree to which their classes contain high- and low-performers. Hopefully, the opening remarks just presented have provided a convincing motivation that this is desirable. Next, brief introductory remarks are provided, intended to review the literature regarding the occurrence and treatment of bimodal grade distributions, especially in STEM courses. Then, a simple mathematical model is introduced that can be used to quantify grade distribution bimodality, which is optimally fitted to a given grade distribution, thus generating estimates for the diagnostic parameters. This is followed by a demonstration of how multimodal distributions with more than two modes can also be usefully diagnosed with the proposed method. The approach is tested first on the diagnosis of four consecutive years of actual exam grade distributions for two mathematics courses taught in large classes to first-year engineering students at the Technion. Then, the methodology is applied to data from smaller classes to demonstrate how the diagnosis differentiates between successful and unsuccessful outcomes. Finally, as an ultimate demonstration of the practical application of the diagnosis procedure, its usage is illustrated on the analysis of the impact of pedagogic changes implemented over time in the teaching of the Technion's capstone design course in the Faculty of Chemical Engineering. The paper ends with some discussion and concluding remarks.

INTRODUCTION

The average is the commonly used metric to assess acceptability of exam grade distributions. This tacitly assumes that the distribution is unimodal and normally distributed, even though the students sitting the exams may exhibit heterogeneous capabilities. Especially for courses involving abstract concepts, prevalent in many STEM courses, the students in the class consist of at least two distinct groups:

1. A high-performing subset of students, who are characterized by those who ask questions in class and invest considerable time and energy each week on the course, reviewing materials and practicing example exercises. Some of the members of this subset could also be those students who are retaking the course in question, and who therefore begin the course with a better grasp of the subject-matter.
2. A low-performing subset of students, who spend relatively less time on the course, often leaving the majority of their learning to a short "cramming" period just before the final exam.

While there may be more than two classifications of students in a class, one must distinguish between at least two groups to be able to diagnose the impact of class heterogeneity on outcome-attainment. Clearly, one would be surprised if both of these groups of students were to perform equally well in the exam, and indeed, there is published evidence that in the more abstract courses, the grade distributions are often bimodal (Arthurs et al, 2019). In addition to the reasoning for the bimodality presented above, Patitsas et al (2016) list the following possible causes for two distinct student populations in a course in computer science:

a. Students with prior relevant experience (e.g., programming) and those without.
b. Students better able to grasp key concepts, and those who have difficulty with them.
c. Students that are naturally talented in the subject, and those that are not (the "Geek Gene" hypothesis).
d. Poor assessment protocols (e.g., unclear exam questions, or unfair grading that does not quantify the degree of understanding of a student).

The objective of this paper is to present a straight-forward diagnosis method to classify the degree of bimodality in a grade distribution, which can gauge the relative sizes of the high- and low-performing groups in the class, as well as estimating their mean grades and standard deviations. Note that carrying out bimodality analysis on anything other than exam grades is not advisable, as all other components of a course grade may involve cooperative or group effort, rather than individual effort (Turton et al, 2013).

METHODOLOGY

Exam grade distribution histograms can be fitted to a bimodal probability density function:

$$f(x)=\frac{p}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}+\frac{1-p}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}.$$

Eq. (1)

In Eq.(1) it is assumed that the grade distribution can be approximated by the weighted sum of two normal distributions with averages, μ_1 and μ_2, and standard deviations, σ_1 and σ_2, where p is the mixing parameter ($0 < p < 1$). As shown by the example distributions presented in Figure 3, Eq. (1) can model distributions featuring both kurtosis and left or right skewness, by appropriate selection of the model parameters. The specific parameters used to prepare the plots in Figure 3 are presented in Table 1.

The five parameters of Eq. (1) need to be selected such that the bimodal probability density function matches histograms of actual grade distributions with minimum fitting error. The parametric fit can be conveniently performed using Maximum Likelihood Estimation methods (McLachlan and Peel, 2000; Meng and Rubin, 1993), as implemented, for example in the R-package (*normalmixEM*, EM Algorithm for Mixtures of univariate normals, Benaglia et al, 2009). In this study, the optimal model parameters were determined using a genetic algorithm, thus avoiding the need for a unique initial guess of the model parameters, which can often bias the results obtained.

Table 1. Example parameter value in Eq. (1) that lead to kurtosis, left skewness and right skewness.

Feature	μ_1	μ_2	σ_1	σ_2	p
Kurtosis	0.600	0.600	0.100	0.300	0.300
Left-skewed	0.850	0.600	0.100	0.200	0.600
Right-skewed	0.850	0.600	0.150	0.100	0.300

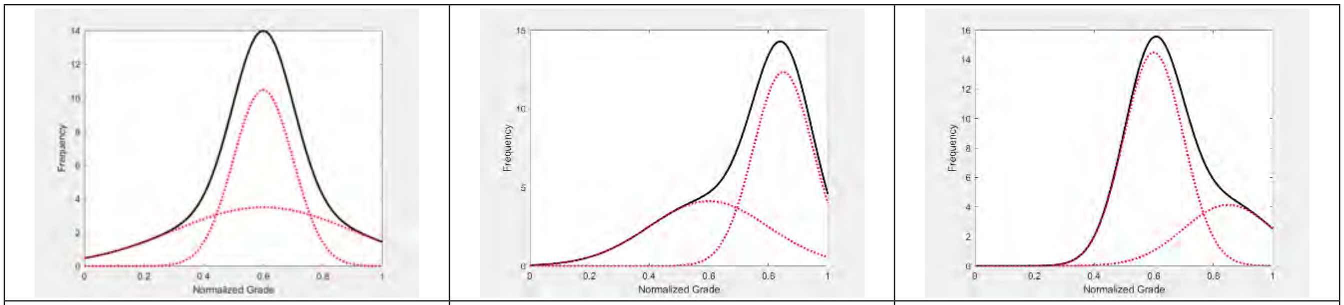
In Eq. (1), the indices 1 and 2 indicate the high- and low-performing subpopulations, respectively, with the latter consisting of the fraction 1 – p of the total population. Thus, for a unimodal normal distribution with average and standard deviation of μ_1 and σ_1 respectively, p = 1. The degree of bimodality in grade distributions can be quantified by fitting the five distribution parameters to the actual grade distribution, and then computing Ashman's D (Ashman, 1994):

$$D=\sqrt{2}\frac{\left|\mu_1-\mu_2\right|}{\sqrt{\sigma_1^2+\sigma_2^2}}.$$

Eq. (2)

Since $D > 2$ for a clean separation of the distribution into two distinct peaks, the magnitude of D is a measure of the degree of bimodality, and the value of p will indicate the proportion of the high-performing students in the class. Ideally, one would like to see values of p close to unity accompanied by relatively large values of μ_1, indicating a high proportion of high-performing students. Values of D higher that two indicate heterogeneous student populations that can be clearly distinguished by separate peaks in the grade distribution, for example as shown in Figure 2, in which $D = 4.3$. Note, however that even when Eq. (2) estimates $D < 2$, there could still be two distinct subpopulations, which will express themselves either as left-or right-skewed grade distributions, when $\mu_1 > \mu_2$, or by observable kurtosis, in cases where $\mu_1 \approx \mu_2$ (See Figure 3).

The more realistic representation of the grade distribution as a bimodal one has the distinct disadvantage that now, instead of a single average estimate, it is now necessary to interpret the implications of a five-parameter model. One possible useful work-around is to present the diagnosis result as a bubble plot, where disks of radius proportional to p are plotted in $\mu_1 - \mu_2$ space (for examples, see Figures 9 and 13).



(a) Kurtosis ($\mu_1 \approx \mu_2$) | (b) Left skewed ($p > 0.5$) | (c) Right skewed ($p < 0.5$)

Figure 3. Example distributions modelled by Eq. (1) that exhibit kurtosis and skewness, with parameters as presented in Table 1. The solid lines present the actual normalized grade distributions, and the red dotted lines plot the distributions of each term in Eq. (1).

WHAT IF THE TRUE GRADE DISTRIBUTION IS NOT BIMODAL?

It would be fair to question the reliability and purpose of fitting a bimodal distribution to examination grades. Even if it seems reasonable that there are only two levels of achievement in most classes, what would be the consequences of forcing a bimodal distribution onto a class that isn't bimodal? For example, classes could be trimodally distributed, with high-, medium- and low-scoring groups in the population. What about a class in which there is simply a rather evenly distributed set of performers, or classes in which the grade distribution is normal? What are the implications of forcing a bimodal distribution of the exam scores onto these classes?

The capability of reproducing kurtosis and skewness using Eq. (1) has been demonstrated. In the case of a unimodal (normal) distribution, fitting a bimodal approximation would result either in *p* being estimated as 1, or in two distributions being derived, with approximately equal means and standard deviations. The uniformly-distributed case is not suitable for the bimodal approximation, as this exhibits no clear peaks and cannot be meaningfully fitted. To demonstrate the implications of the trimodally distributed case, consider the trimodal grade distribution function:

$$f(x)=\frac{r_0}{s_0\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m_0}{s_0}\right)^2}+\frac{r_1}{s_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m_1}{s_1}\right)^2}+\frac{1-r_0-r_1}{s_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m_2}{s_2}\right)^2},$$

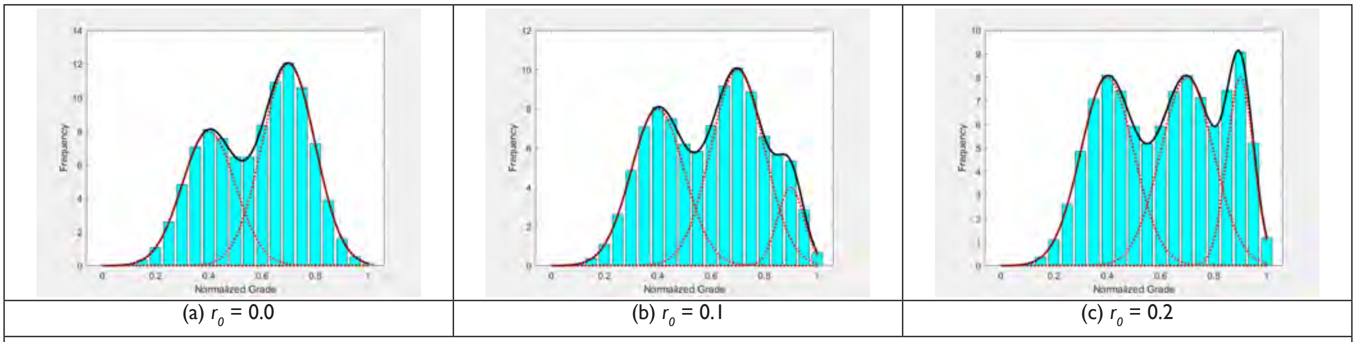
Eq. (3)

where m_i and s_i are the means and standard deviations of each of the three subpopulations, with *i* = 0, 1 and 2 indicating high-, medium- and low-scorers, respectively, r_0 is the fraction of high-scorers, and r_1 is the fraction of medium-scorers. Let us consider a specific but representative case, where means and standard deviations of the distributions are at values that enable them to be seen distinctly, for example by taking the following parametric values: $[m_0, s_0] = [0.90, 0.05]$, $[m_1, s_1] = [0.70, 0.10]$, and $[m_2, s_2] = [0.40, 0.10]$. Furthermore, taking 0.4 as the fraction of low-scorers, then $r_0 + r_1 = 0.6$, and assuming that the high-performers are not in such a proportion that they dominate the population, we investigate values of r_0 up to a value of 0.2, and test the implications of the proportion of high-scorers on the optimally fitted bimodal model parameter values in Eq.(1).

As seen in Figure 4, with the distribution frequency for a total of 100 examinees, the mean and standard deviation values selected show the effect of the value of r0 on the degree to which the distribution is trimodally distributed. The effect of increasing r0, the proportion of high-scorers, on the average grade, μ, as well as the best-fit bimodal approximation parameters (μ_1, μ_2, σ_1, σ_2 and *p*), Ashman's *D*, and the fitting error (SSE), are presented in Table 2 with two representative fitting plots shown in Figure 5. These indicate that as r_0 is increased, the five parameters of the bimodal distribution function (Eq. 1) are adjusted to match the trimodal distribution function (Eq. 3) as closely as possible. As one would expect, since the proportion of the high-performers (i.e., $r_0 + r_1$) in Eq. (3) is kept constant, the two highest peaks in Eq. (4) are represented by the second peak of Eq. (1) and therefore both μ_1 and σ_1 are increased. To minimize fitting error, μ_2 and σ_2 are slightly reduced, while *p* is increased. The essence of the trimodal distribution is captured by combining the high- and medium-scorers' contributions into the high-performer's peak in the bimodal distribution, leading to positive adjustments in the values of μ1 and σ_1. As expected, increasing the contribution of the high-scorers leads to increasing fitting errors. At some point, the fitting error will be so large that the reliability of the bimodal approximation is compromised. This will occur, for example, when trying to fit the bimodal model to a uniform distribution.

Similarly to the trimodal demonstration, distributions with *any number of modes* can be approximated using two modes, with parameters adjusted to enable the bimodal model to match the true distributions as closely as possible. The important bottom line is this: Given the most-likely eventuality that a uniform exam grade distribution will not be obtained, a bimodal distribution enables a reasonable fit to the true grade distribution that usefully models the heterogeneity of the students' grade distribution, providing diagnostics that identify the proportion of low-performers and estimates on the effective means and standard deviations of the low- and high-performers.

In the next four sections, the method is applied to the analysis of the grade distributions of several Technion final exams, noting that the Technion pass grade is 55%. In the next two sections, the diagnosis is carried out on two core mathematics courses, Algebra and Calculus given to 1st year undergraduate engineering students at the Technion, both of which involve large numbers of students. The next section analyzes grade distributions for smaller classes of students, demonstrating examples of how successful and unsuccessful teaching protocols, as measured by the exam outcomes, are flagged using the proposed diagnosis approach. To round off the paper, the final section demonstrates how the diagnosis method can be used to assess the effectiveness of changes in teaching pedagogy over time.

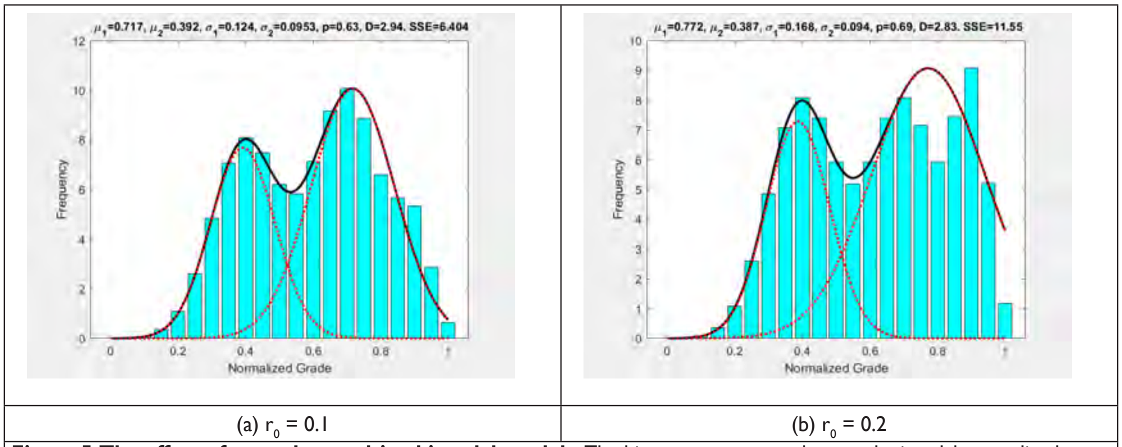


(a) r_0 = 0.0 | (b) r_0 = 0.1 | (c) r_0 = 0.2

Figure 4. The effect of r_0 on the degree of trimodality in Eq. (4) with $[m_0, s_0] = [0.90, 0.05]$, $[m_1, s_1] = [0.70, 0.10]$, and $[m_2, s_2] = [0.40, 0.10]$, and $r_0 + r_1 = 0.6$. The solid lines and histograms present the actual trimodal normalized grade distribution, and the red dotted lines plot the distributions of each term in Eq. (3).

Table 2. Best fit bimodal distribution parameters for trimodal distribution with $[m_0, s_0] = [0.90, 0.05]$, $[m_1, s_1] = [0.70, 0.10]$, and $[m_2, s_2] = [0.40, 0.10]$ and $r_0 + r_1 = 0.6$, as a function of degree of trimodality, r_0.

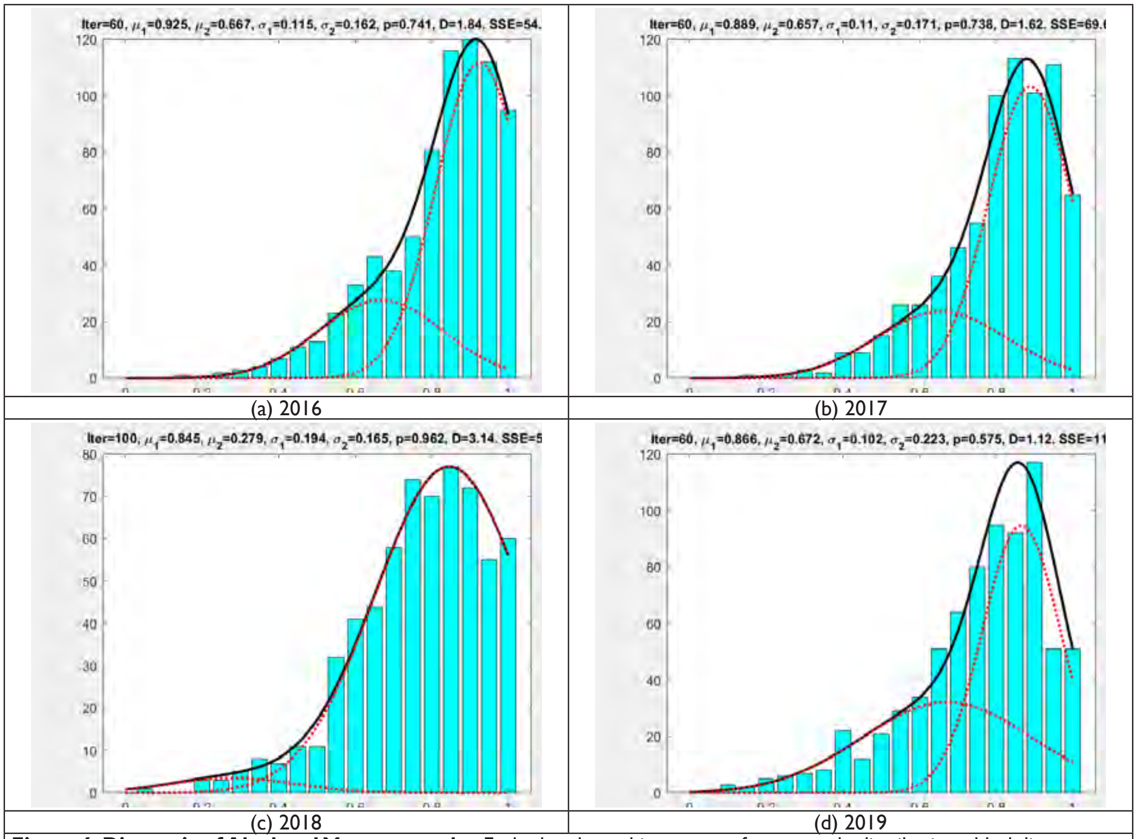
r_0	μ	$μ_1$	$μ_2$	$σ_1$	$σ_2$	*p*	*D*	*SSE*
0.00	0.58	0.700	0.400	0.100	0.100	0.600	3.00	10^{-2}
0.05	0.59	0.709	0.398	0.106	0.099	0.606	2.97	3.79
0.10	0.60	0.717	0.392	0.124	0.095	0.630	2.94	6.40
0.15	0.61	0.734	0.389	0.145	0.095	0.648	2.82	8.71
0.20	0.62	0.772	0.387	0.168	0.094	0.690	2.83	11.6



(a) r_0 = 0.1

(b) r_0 = 0.2

Figure 5. The effect of r_0 on the resulting bimodal models. The histograms present the actual trimodal normalized grade distribution, the solid lines show the best fit bimodal model, with distributions of each term in Eq. (1) shown by the red dotted lines.



(a) 2016

(b) 2017

(c) 2018

(d) 2019

Figure 6. Diagnosis of Algebra 1M exam results. Each plot shows histograms of exam grade distributions, black lines indicating f(x) as predicted by Eq. (1), and red dotted lines showing the high- and low-performing subpopulation contributions to f(x). The abscissa is the normalized exam grade while the ordinate is the number of students in each histogram bin, in steps of 5% of the total grade.

Table 3. Bimodal distribution parameters for Algebra 1M course grades

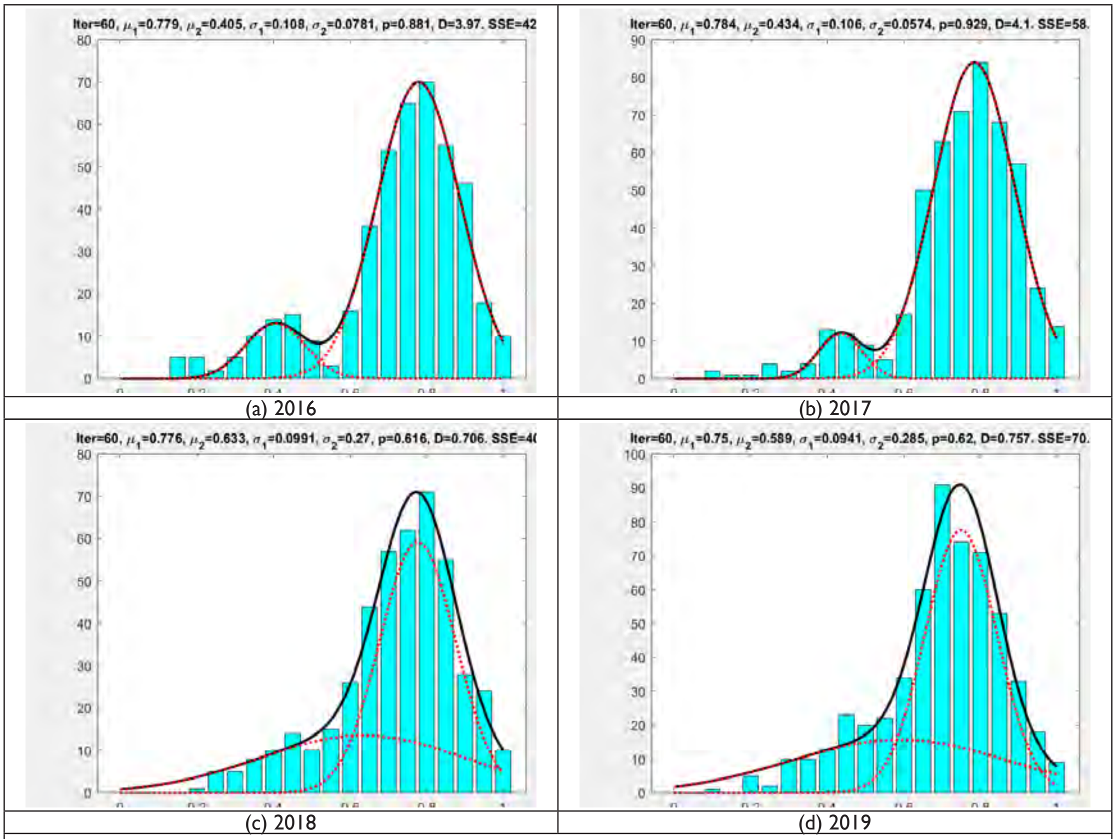
Year	N	$N_{<55}$	μ	σ	$μ_1$	$μ_2$	$σ_1$	$σ_2$	*p*	*D*
2016	752	46(6%)	0.80	0.15	0.93	0.67	0.12	0.16	0.74	1.84
2017	719	46(7%)	0.79	0.15	0.89	0.66	0.11	0.17	0.74	1.62
2018	632	54(9%)	0.75	0.17	0.85	0.28	0.19	0.17	0.96	3.14
2019	719	94(13%)	0.74	0.18	0.87	0.67	0.10	0.22	0.58	1.12

https://doi.org/10.20429/ijsotl.2021.150207

Diagnosis of Algebra 1M Grade Distributions

Exam grade distributions for Algebra 1M for four academic years from 2016 to 2019 are shown in Figure 6. Table 3 shows a summary of the number of students who took each exam, the number of students who failed, also expressed as a percentage of the class, as well as averages and standard deviations. Also shown in the table are the resulting fitted bimodal model parameters as well as Ashman's *D*. As an example, the first entry in Table 3 indicates that for the 2016 exam, in which 752 students participated, the failure rate was 6%, with an average grade of 80% and a standard deviation of 15%. This distribution is best fitted by a bimodal model, in which 74% of the total students belong to a high-performing group with a mean grade of 93%, with a small group of low-performing students, including only 26% of the total population, with a mean grade of 67%.

From Figure 6, it is apparent that the distributions are all left-skewed and comprise a majority of high-performing students. Of the four years, the last one presented, for 2019, indicates the lowest proportion of high-performing students, only 58%, but the effect of the 42% low-performers on the overall class performance is less significant because the low-performers obtained average grades closer to those of the high performers. Another thing to notice is the increased variance of the low-performing group in the last year of data, which leads to a more pronounced spreading of the skew. Nonetheless, bimodality is not observable as separate peaks in any of the histograms for this course, and contributes mainly to explaining the left-skewness of the distributions. This course is the most populated course given at the Technion, and the performance over the diagnosed period represents a consistently good outcome, though there is evidence of a slight degradation in performance over time, as indicated by the gradually lowering values of μ_1, as well as either lower values of μ_2 or lower values of *p,* both of which lead to lower values of the average grade, μ.



(a) 2016

(b) 2017

(c) 2018

(d) 2019

Figure 7. Diagnosis of Calculus 1M exam grades. Plots prepared as in Figure 6.

Table 4. Bimodal distribution parameters for Calculus 1M course grades

Year	N	$N_{<55}$	μ	σ	μ_1	μ_2	σ_1	σ_2	*p*	*D*
2016	438	65 (15%)	0.70	0.18	0.78	0.41	0.11	0.08	0.88	3.97
2017	501	49 (10%)	0.73	0.16	0.78	0.43	0.11	0.06	0.93	4.10
2018	445	55 (12%)	0.70	0.16	0.78	0.63	0.10	0.27	0.62	0.71
2019	549	96 (17%)	0.68	0.16	0.75	0.67	0.09	0.25	0.51	0.44

https://doi.org/10.20429/ijsotl.2021.150207

Table 5. Bimodal distribution parameters for three test courses.

Course	N	$N_{<55}$	μ	σ	μ_1	μ_2	σ_1	σ_2	p	D
"The Good"	200	13 (7%)	0.74	0.13	0.84	0.64	0.08	0.11	0.61	2.09
"The Bad"	288	159 (55%)	0.53	0.18	0.63	0.48	0.20	0.13	0.50	0.91
"The Ugly"	126	121 (96%)	0.37	0.12	0.40	0.39	0.20	0.10	0.32	0.06

Diagnosis of Calculus 1M Grade Distributions

Exam grade distributions for Calculus 1M for four academic years from 2016 to 2019 are shown in Figure 7. Table 4 shows a summary of the number of students who took each exam, the number of students who failed, also expressed as a percentage of the class, as well as averages and standard deviations. Also shown in the table are the resulting fitted bimodal model parameters as well as Ashman's *D*.

As shown in Figure 7, the first two years of data (2016 and 2017) are clearly bimodal (*D* values close to 4), although the low-performing students constitute a small percentage of the class (7-12%). The bimodality is less obvious for the last two years of data (2018 and 2019), and the distributions look very similar to the general pattern observed in the Algebra grade distributions, but with slightly lower average grades. It is noted that in the last two years, the distributions of low-performing grades have larger variances than in the first two years, causing the left-skewed overall distribution to feature a flattened tail. It should also be noted that the students who take Calculus 1M constitute most of the class that takes Algebra 1M each year, and yet, the grade distributions are qualitatively different. It is clear that the calculus course is found to be the more difficult of the two by many students, and like the Algebra course, there is evidence of a slight deterioration of performance over time, caused by the same changes in the estimated model parameters as observed for the grades of Algebra 1M.

Diagnosis of Three Test Courses: The Good, the Bad and the Ugly

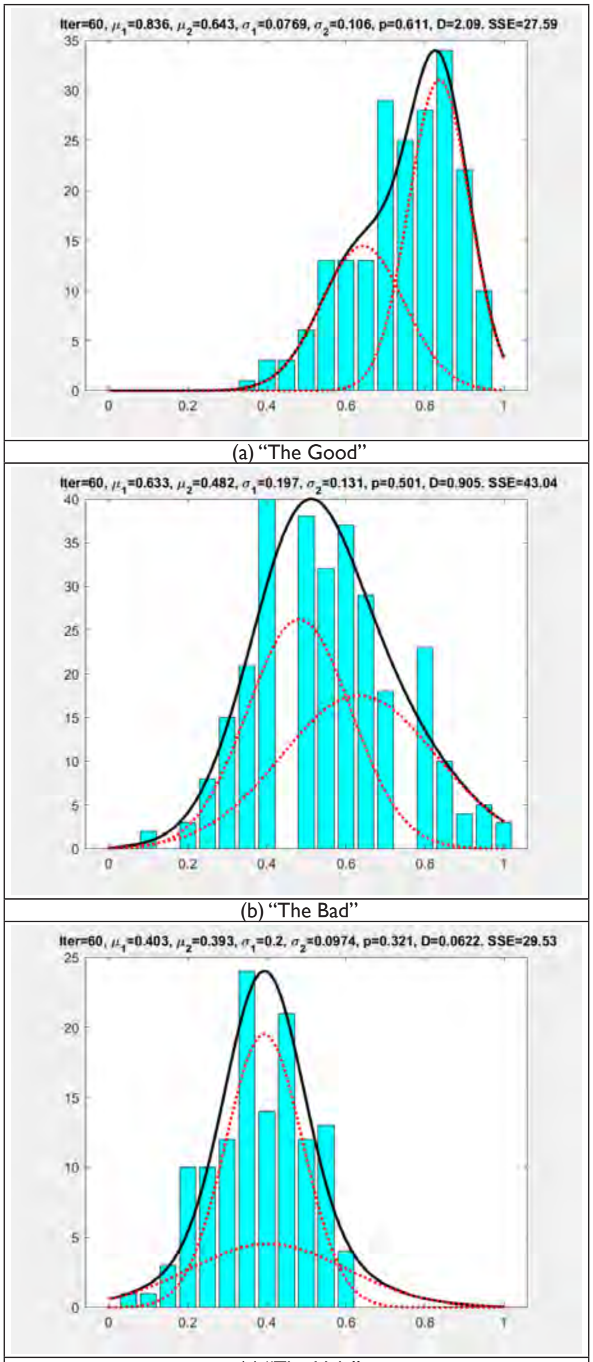
The previous two sections have demonstrated the usage of the diagnosis technique on the exam results for extremely large classes. It would be fair to question whether the diagnosis is capable of providing useful information when dealing with smaller data sets. This section demonstrates the diagnosis of three representative engineering core courses, whose results are shown in Figure 8, with a summary of parametric data in Table 5. The three courses are diagnosed in the following order: the good, the bad, and the ugly.

The Good:

The final exam for this course appears to have achieved a good outcome, with a failure rate of only 7%, in which the grade distribution is left-skewed, as shown in Figure 8(a). The diagnosis tool reveals that 61% of the class are high-performers, with an average grade of 84%. The remaining low-performing students attained an average grade of 64% with a relatively tight distribution (σ_2 = 0.11), which explains the low failure rate. Of the three exams diagnosed in this section, this is the one where the student population did the best, as clearly indicated by the estimated bimodal model parameters. So, a good outcome here, both for the lecturer and for the students. Note also that the average grade is relatively high, and comparable to the results obtained from the diagnosis of the algebra and calculus grades presented previously.

The Bad:

Even without recourse to the diagnostic tool, it is clear that the lecturer of this course has a problem, given the fact that 55% of the class has failed the exam. The results from the diagnostic tool, summarized in Table 5 and shown in Figure 8(b), indicate that the high- and low-performing groups have average grades that are slightly above and slightly below the fail grade, respectively. The fact that it is difficult to differentiate between high-



(a) "The Good"

(b) "The Bad"

(c) "The Ugly"

Figure 8. Diagnosis of three representative exam grade distributions. Plots prepared as in Figure 6.

and low-performers is an indication that most of the students have not mastered the course. Whatever is done in this case to mitigate the low percentage of passing students, whether it is by a factor to jack up the grades or by accounting for classwork, cannot disguise the fact that the class as a whole has not demonstrated mastery. The lecturer needs to seriously take stock and implement changes to the course protocols, to significantly improve the mastery achieved by students in the course.

The Ugly:

This exam result is close to the worst-possible scenario, with the distribution data indicating a 96% failure rate (gevalt/oy vey!). As confirmed in Table 5, it is further noted that the diagnostic tool returns a result indicating that the two student populations have almost the same average grades, accompanied, of course, by a value of *D* close to zero, from which it can be concluded that the distribution is close to a unimodal normal one, exhibiting some kurtosis, as clearly shown by the data and model fit in Figure 8(c). Unfortunately, the average grade is about 40%, which is consistent with the resulting high failure rate. The diagnosis is clear: all of the students in the course were low-performers. The lecturer of this course has a serious problem; the only way to mitigate the unacceptable failure rate is to account for the grades of homework assignments, as well as a large enough factor to jack up the grades sufficiently. The unquestionable outcome in these circumstances, is that most of the students would be given passing grades, even though the majority have not achieved mastery in the final exam. Given the failure of virtually the entire class, all aspects of this course need thorough revision: (a) As in "The Bad," the lecturer should consider appropriate modifications to the way the course is taught, and in particular, the support system provided to all students; (b) Given the failure of essentially the entire class, it may be prudent to more closely align the final exam with the course's learning objectives.

To put the above results in perspective, the values of μ_1, μ_2 and p estimated for the above three cases are presented in the bubble plot shown in Figure 9, which indicates bubbles of diameter proportional to the value of *p* plotted on the $\mu_1 - \mu_2$ plane. For comparison, the data estimated for the series of algebra and calculus exams have also been positioned on the same plot, noting that the bubbles are colour-coded: red for Algebra 1M, blue for Calculus 1M, and green for the three Test Courses.

Figure 9 provides a quick visual diagnosis of the Test Courses' exam results, allowing for bimodality in the grade distributions. It is noted that the bubble plots for both Algebra 1M and Calculus 1M are clustered on the top right of the plot area, indicating relatively large values of both μ_1 and μ_2, with the results for the former being slightly better than the latter. As for the three Test Courses, one notes that the scores for "The Good" are comparable to those of Algebra 1M and Calculus 1M, while both "The Bad" and "The Ugly" are immediately identified as clearly problematic results.

Diagnosing the Impact of Teaching Pedagogy in the Capstone Design Course

As described by Lewin and Barzilai (2021), the capstone process design course at the Technion has been periodically evolving over the last 15 years, with some of the changes having been made to update and/or streamline the curriculum. However, two of the changes involved updating the teaching pedagogy with a view of improving the degree to which the learning outcomes are achieved by the students, namely:

a. Transforming the recitations, in 2011, from lecture-based sessions, where the teaching assistant largely demonstrated solutions of example problems to students, to active tutorials in which most of the time is allocated for students to solve problems for themselves.
b. Transforming the course to flipped format, in 2015, in which the lectures have been moved to an on-line format, including built-in quiz questions, which have become the students' home assignments. The lecturer's meetings with students have been converted to active problem-solving sessions with students' participation. The active tutorial component implemented in 2011 has been retained and rounds off each week's activity.

The two principle desired course outcomes, which have not been changed over the 15 years, are the degree of success of teams of students on a competitive design project, and of the demonstration of students' individual mastery of the taught materials as measured by their exam grades. The exam data for the 15-year span from 2005-2020 therefore makes an interesting

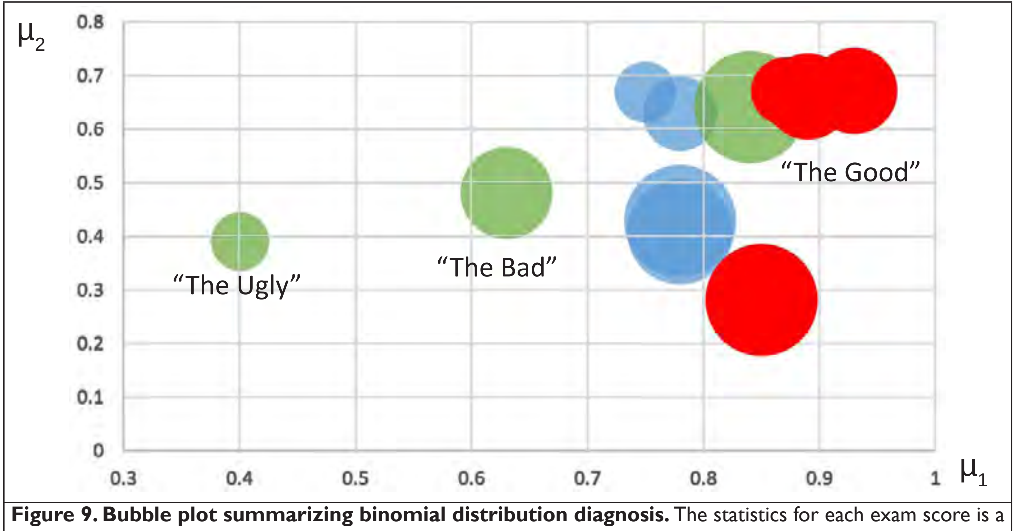
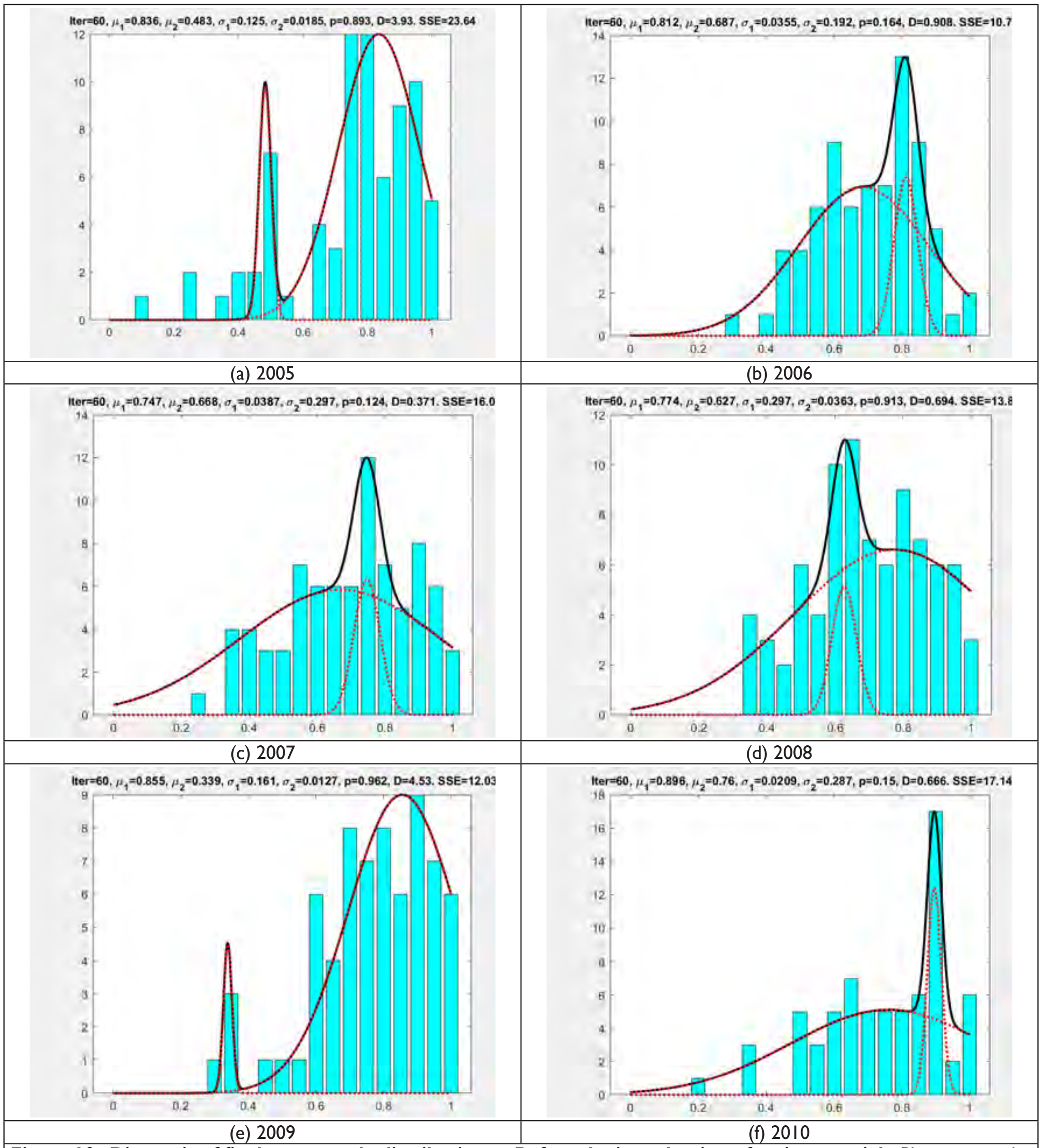


Figure 9. Bubble plot summarizing binomial distribution diagnosis. The statistics for each exam score is a bubble centred on the $\mu_1 - \mu_2$ plane, with diameter proportional to *p*. The bubbles are colour-coded: red – Algebra 1M, blue – Calculus 1M, green – Test scores.

case study for the proposed diagnosis method. Some details of these 15 years of data are in order: (a) The same teacher taught all classes (the author) during the entire period; (b) Classes met at approximately the same time (mid-day) and day of the week during the entire period; (c) The learning outcomes and topics taught and examined remained almost constant during the entire period, as documented in Lewin and Barzilai (2021); (d) The demographics of enrolled students have not changed significantly during the entire period. The two hypotheses to be tested are as follows:

H1: The transformation from lecture-based recitations to active tutorials improved the degree to which classes of students achieved the desired learning outcomes.

H2: The transformation of the course from lecture-based to the "flipped classroom" improved the degree to which classes of students achieved the desired learning outcomes.



(a) 2005

(b) 2006

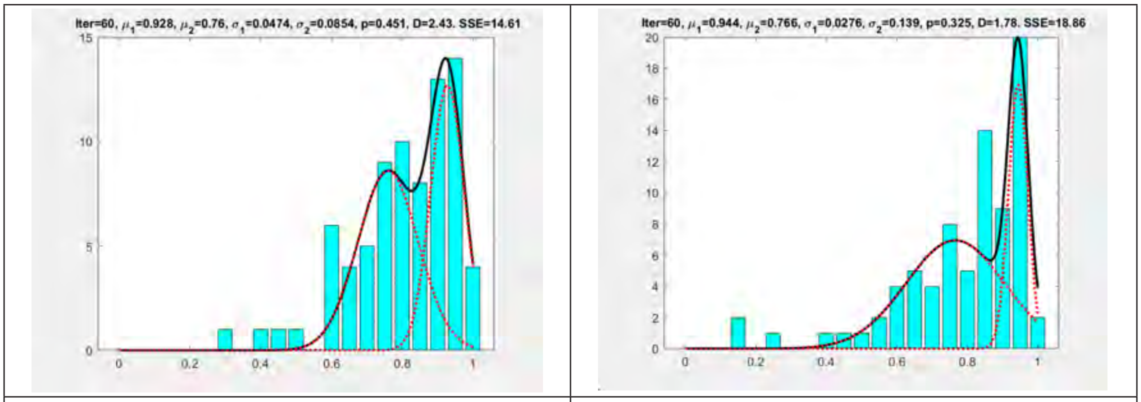
(c) 2007

(d) 2008

(e) 2009

(f) 2010

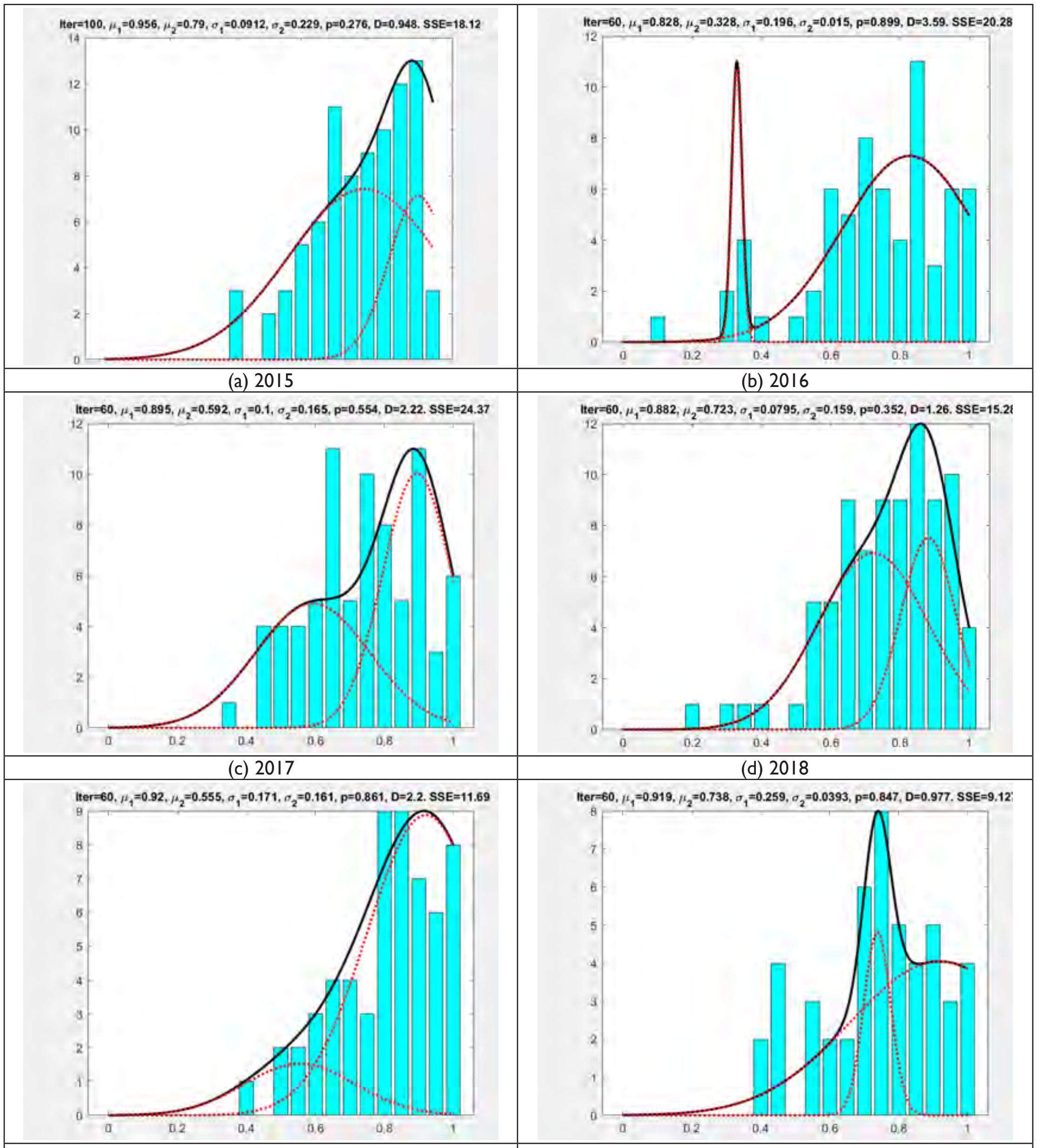
Figure 10. Diagnosis of final exam grade distributions – Before the introduction of active tutorials. Plots prepared as in Figure 6.



(a) 2011

(b) 2012

Figure 11. Diagnosis of final exam grade distributions – After the introduction of active tutorials. Plots prepared as in Figure 10.



(a) 2015

(b) 2016

(c) 2017

(d) 2018

(e) 2019

(f) 2020

Figure 12. Diagnosis of final exam grade distributions (Flipped Course). Plots prepared as in Figure 10.

Table 6. Analysis of Sample Process Design Final Exam Grades, 2005-2020.

Year	N	N<55	μ	σ	μ_1	μ_2	σ_1	σ_2	*p*	*D*
2005	77	16 (21%)	0.73	0.21	0.84	0.48	0.12	0.02	0.89	3.93
2006	77	11 (15%)	0.69	0.15	0.81	0.67	0.04	0.19	0.16	0.91
2007	81	17 (21%)	0.68	0.18	0.75	0.67	0.04	0.12	0.12	0.37
2008	85	15 (18%)	0.68	0.17	0.77	0.63	0.30	0.04	0.91	0.69
2009	68	7 (10%)	0.75	0.24	0.86	0.34	0.16	0.01	0.96	4.53
2010	70	9 (13%)	0.74	0.18	0.90	0.76	0.02	0.29	0.15	0.67
2011	77	4 (5%)	0.78	0.15	0.93	0.76	0.05	0.09	0.45	2.43
2012	79	7 (9%)	0.77	0.18	0.94	0.77	0.03	0.14	0.33	1.78
2013	68	6 (9%)	0.74	0.14	0.80	0.70	0.20	0.05	0.91	0.66
2015	85	7 (8%)	0.76	0.15	0.96	0.79	0.09	0.23	0.28	0.95
2016	66	9 (14%)	0.71	0.20	0.83	0.33	0.20	0.02	0.90	3.59
2017	87	10 (11%)	0.72	0.16	0.90	0.59	0.10	0.16	0.55	2.22
2018	84	7 (8%)	0.75	0.16	0.88	0.72	0.08	0.16	0.35	1.26
2019	58	4 (7%)	0.79	0.15	0.92	0.56	0.17	0.16	0.86	2.20
2020	48	6 (13%)	0.73	0.17	0.92	0.74	0.26	0.04	0.85	0.98

Given that the proposed diagnostic tool enables the classification of an overall grade distribution by high- and low-performers, it can be used to address these two hypotheses, and provide insights into the impact of changes in the teaching pedagogies on the course's exam outcomes. Figures 10-12 present exam grade distributions for the years 2005 – 2020, spanning a period before and after the introduction of active tutorials in 2011, and since the introduction of flipping in 2015. Table 6 presents the resulting fitted bimodal model parameters as well as Ashman's *D,* obtained after diagnosis of the exam grades. The results lead to the following observations:

1. There has been a gradual improvement in the average grades scored in the final exam of the design course, as well as a reduction in the failure rates. These changes have become most pronounced after the switch to active tutorials in 2011. This result alone supports *Hypothesis H1*, that switching to active tutorials has improved outcomes.
2. There is a degree of bimodality in all of the results shown in Figures 10-12 and Table 6, with the improvements reported gradually resulting from increased averages of both high- and low-performers. Again, the most significant improvements occurred after active tutorials were introduced, supporting *Hypothesis H1*.
3. There are large variations in possible class performance, as confirmed by the large swing in the estimated binomial distribution parameters. It is noted that the variations have been somewhat attenuated after the introduction of flipping.

To facilitate elucidation of the results, the estimated values of μ_1, μ_2 and *p* are presented in the bubble plot shown in Figure 13, which shows bubbles of diameter proportional to the value of *p* plotted on the $\mu_1 – \mu_2$ plane. Note the bubbles are colour-coded, with the period before the introduction of active tutorials (Phase I) shown in black, those for the period between the introduction of active tutorials but before flipping (Phase II) shown in grey, and the period after flipping (Phase III) shown in white. It was hoped that the bubble chart would show a clear separation between the three phases of the course's pedagogic evolution: before and after the introduction of active tutorials, and after the introduction of flipping. Instead, it confirms that annual grade results have significant variation, even when analyzing each period of the course's evolution, leading to overlap. This implies that the inherent capabilities of each cohort have a significant impact on the results. Even so, some features are quite distinct:

1. The six exam results for the period before active tutorials were introduced, the black bubbles representing the years 2005-2010, are clustered on the lower-left, that is, with rel-

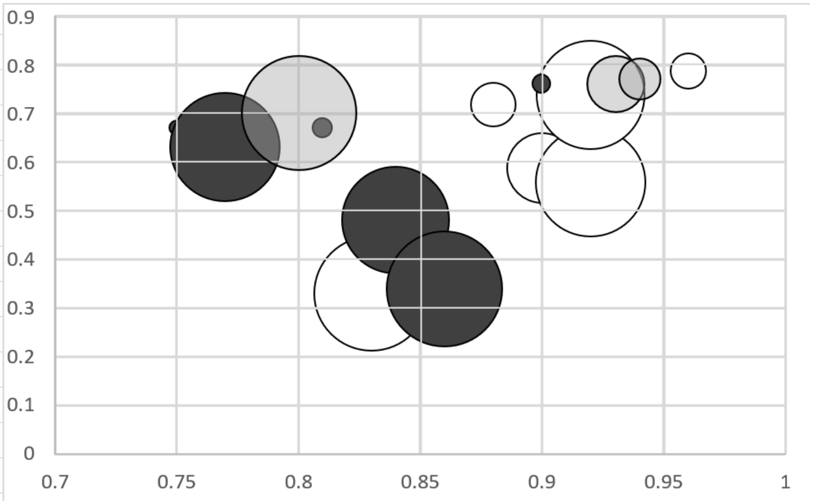


Figure 13. Bubble chart summarizing binomial distribution diagnosis. The statistics for each year are centred on the μ1 – μ2 plane, with the bubble diameter proportional to p. The bubbles are colour-coded according to period: black – Phase I: 2005-2010 (before active tutorials), grey – Phase II: 2011-2013 (before flipping), white – Phase III: 2015-2020 (after flipping). Ref: Lewin and Barzilai (2021).

atively low values of both μ_1 and μ_2. The instances in which relatively high values of μ_1 have been obtained have relatively low values of p. Table 5 indicates that the failure rates in this period are relatively high, averaging at 16%.

2. The results for the relatively short period between the introduction of active tutorials and before switching to flipping, indicated by the grey bubbles representing the years 2011-2013, have significant scatter. Even so, the average failure rate in this period was only 8%, indicating that, even though erratic and prone to bias depending on the nature of the class, there was a significant drop in the failure rate compared to the situation before introducing active tutorials. It is apparent from these findings that the move to active tutorials have led to significantly better outcomes, thus strongly supporting Hypothesis H1.
3. Most of the exam results for the period after flipping was introduced (five out of the six in all), indicated by the white bubbles representing the years 2015-2020, are clustered on the top right, that is, with relatively high values of both μ_1 and μ_2. There is still some scatter, but it is less pronounced than for the exam results from active tutorials alone. The average performance of the six years with flipping is at about the same level as that over the three years with active tutorials alone, noting that the failure rate in this period averages at 10%. However, flipping has achieved a uniformly higher performance, as indicated by the relatively tight clustering observed in the bubble plot of Figure 13, with a more consistent proportion of the class achieving higher performance, as indicated by the larger bubbles in the cluster on the right of the plot, compared with those of the other two phases. This indicates that the transformation from active tutorials to full flipping has advantages, thus supporting Hypothesis H2, that switching to the "flipped class" paradigm has improved outcomes.

CONCLUSIONS & RECOMMENDATIONS

Ideally, one would like to see a unimodal grade distribution with a reasonable average grade and a relatively low variance, indicating that most of the class has achieved course mastery. This result is unrealistic, as the class is generally split into at least two distinct groups, of high- and low-performing students. When a class is heterogeneous with respect to student capabilities, the desired mastery is mostly achieved by the high-performers, with the low-performers being left behind, leading to a bimodal grade distribution, and consequently, to higher failure rates. In this regard, the diagnostic method that has been demonstrated here provides a "post-mortem" indication of the number of low-performers in the student population, and the degree to which they will affect the overall failure rates. As demonstrated by the large number of computed examples in this paper, the most direct way in which a grade distribution exhibits acceptable mastery is when its diagnosis indicates relatively large values of p, the proportion of the total population in the high-performing group, accompanied by a reasonably high values of μ_1, the average grade of the high-performers. Grade distributions characterized by lower values of p, may still achieve acceptable performance if the average grade of the low-performers, μ_2, is relatively high. This analysis can be repeated into the future and used as a diagnostic aid to assist in tailoring the teaching protocols to reduce the extent of the problem – we would like run our courses to result in relatively large values of p and μ_1, or in relatively large values of μ_1 and μ_2.

Other researchers may make use of the proposed diagnosis methodology to flag problematic learning outcomes and to justify changes to teaching pedagogy that will increase the proportion of engaged students. Unacceptable diagnosis results are an indication that more work is needed to improve course teaching protocols. As demonstrated by the last extended study in this paper, continuous monitoring of the exam outcomes over time enables the diagnosis method to provide indications of the degree of success of the pedagogy used. This returns us to the importance of continuously monitoring the degree to which the students in a class are engaged in their learning *during the semester*, which is necessary to ensure good outcomes. This implies the need to make students learning as effective as possible by applying active learning methods, and possibly increasing the remedial assistance provided, especially for the benefit of the students who engage the least.

REFERENCES

Arthurs, N., Stenhaug, B., Karayev, S., & Piech, C. (2019). "Grades are not Normal: Improving Exam Score Models using the Logit-Normal Distribution". *Proceedings of the 12th International Conference on Educational Data Mining* (EDM 2019), Lynch, C. F. Merceron, A., Desmarais, M. and Nkambou, R. (eds.), 252 – 257.

Ashman, K. M, Bird, C. M., & Zepf, S. E. (1994). "Detecting Bimodality in Astronomical Datasets". *The Astronomical Journal.* 108, 2348–2361.

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009) "Mixtools: An R package for analyzing finite mixture models." *Journal of Statistical Software*, 32(6):1-29.

de Saint-Exupéry, A. (1945). *The Little Prince*, William Heinemann Ltd., London.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. S., Okorafor, N., Jordt, H., & Wenderoth, M. P. (2014). "Active Learning Increases Student Performance in Science, Engineering, and Mathematics," *PNAS*, 111 (23): 8410-8415.

Kulick, G., & Wright, R. (2008). "The Impact of Grading on the Curve: A Simulation Analysis," *International Journal for the Scholarship of Teaching and Learning*, 2(2):1-19. https://doi.org/10.20429/ijsotl.2008.020205

Lewin, D. R. & Barzilai, A. (2021) "Teaching Process Design to Chemical Engineering Undergraduates – an Evolution," *Chemical Engineering Education*, 55(3):157-172.

McLachlan, G. J., & Peel, D. (2000) *Finite Mixture Models*, John Wiley & Sons, Inc., New York.

Meng, X.-L., & Rubin, D. B. (1993) "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika* 80(2): 267-278.

Patitsas, E., Berlin, J., Craig, M. & Easterbrook, S. (2016). "Evidence that Computer Science Grades are not Bimodal" Proceedings of ICER'16. http://www.cs.toronto.edu/~sme/papers/2016/icer_2016_bimodal.pdf

Turton, R., Bailie, R.C., Whiting, W. B., Shaeiwitz, J. A., & Bhattacharyya, D. (2013). *Analysis, Synthesis, and Design of Chemical Processes*, Pearson Education International, Upper Saddle River, NJ.

Velegol, S. B., Zappe, S. E., & Mahoney, E. (2015). "The Evolution of a Flipped Classroom: Evidence-based Recommendations," *Advances in Engineering Education*, 4(3):1-35.