It’s Clear Not Every Exam Measures Knowledge. But Can Math Hint When Something is Wrong?

Exams have a strange way of revealing more than just what students know.

A few days ago, I completed my 32nd semester in what is now a ten-year journey at my university. Over this time, I’ve taken hundreds of exams across several faculties, and I’ve even examined a few hundred students myself. After every exam, I couldn’t help but wonder: *What should the distribution of grades really look like?* Should most people pass, or should failure be just as common? Should there be many top grades, or should they cluster somewhere in the middle? I never had a clear answer. But after one of my most recent exams, I’ve started to form some thoughts.

At Polish universities, the grading system is quite straightforward. There are six possible grades: 2, 3, 3.5, 4, 4.5, and 5. A “2” is the lowest grade and means failing the exam, while all the others are passing grades. Typically, scoring below 50% earns you a 2. To get a 3, you need between 51% and 60%; for a 3.5, between 61% and 70%; a 4 requires 71% to 80%; a 4.5, 81% to 90%; and a perfect 5 means you scored between 91% and 100%. In rough international terms, this scale spans from a failing F (2) to something like a solid A (5), with 3s and 4s corresponding to the middle range of passing grades.

In most cases, students can see how many people were eligible to take the exam and how the grades were distributed. A bar chart with percentages is usually provided to make this information clear.

Recently, I passed the exam in *Biological Bases of Behaviour*, but I’m not entirely sure whether it was thanks to actual learning or just a stroke of luck, especially when you see how strangely the grades were distributed. It was unlike anything I’d seen before.

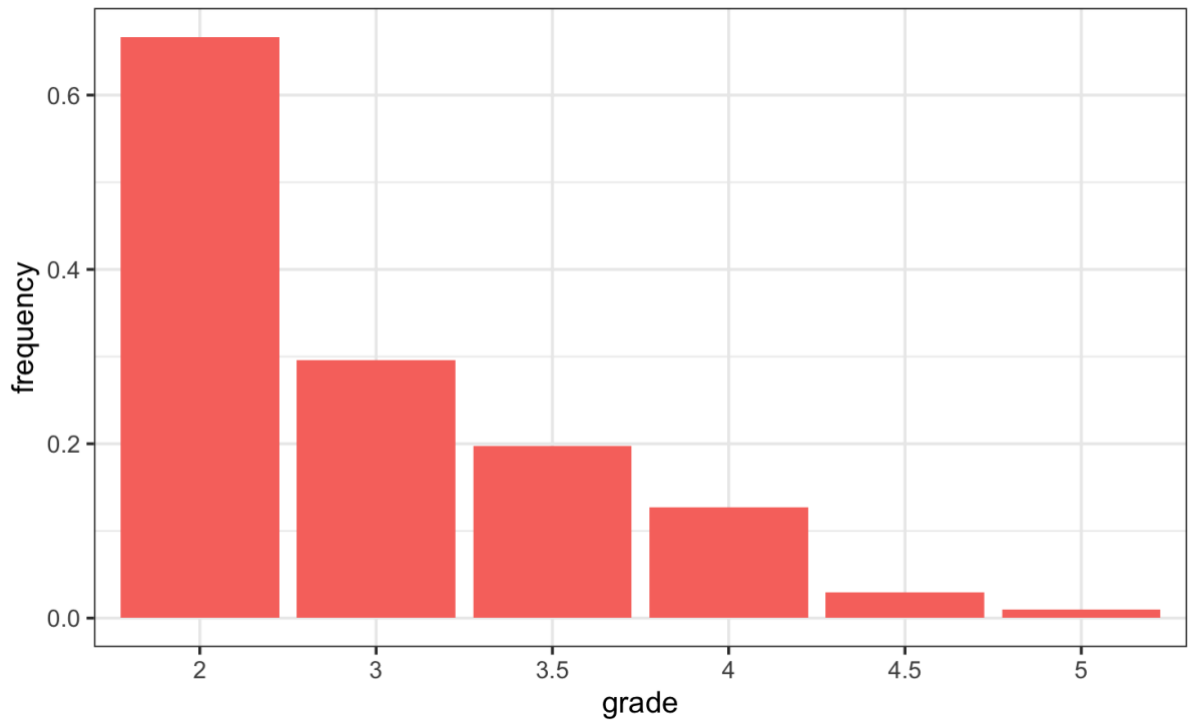

Figure 1. Distribution of grades for the *Biological Bases of Behaviour* exam. Nearly two-thirds of students failed, an unusually high rate compared to typical exam results.

Literally, two-thirds of the students failed. In my entire academic journey, I've never seen anything like it. Usually, fewer than 20% of students fail an exam, and even a 30% failure rate is already considered very high under normal circumstances. There are exceptions, of course. In the first year of mathematics, there were a few subjects where around 50% of students failed. But that was different. There were several such subjects in the same year, and a large group of students simply never showed up at the faculty, so they failed automatically by absence. If I remember correctly, that group made up about 30–40% of all students entitled to take the exam.

That's why the chart above feels like a true outlier, especially since results from other exams taken by the same group of students weren't nearly as bad. Could it really be that so many students simply didn't study enough? I find that hard to believe. Especially considering there were 213 students eligible to take the exam. So I decided to look into it more closely and run a few calculations, turning to a bit of probability and statistics for answers.

There is a very powerful theorem in mathematics that can offer some interesting insights into this situation: the Central Limit Theorem. In simple terms, it explains how random events, when repeated many times, tend to produce a predictable, bell-shaped pattern.

The *Biological Bases of Behaviour* exam was a test with 37 questions, each offering about four possible answers worth 1 point. There was also one open-ended question

worth 3 points. To simplify the analysis, let's set aside the open question for now, as it rather wouldn't change the overall reasoning or the general pattern of results in any major way.

We can assume that a student marks the correct answer with some probability. It's hard to say exactly how the learning process translates into that probability, but it's much easier to describe what happens when a student hasn't learned at all. If they don't know the right answer, they simply guess, so the chance of marking it correctly is 0.25.

This leads to a simple model: each exam question can be treated as a random variable that takes the value of 1 if the student answers correctly and 0 otherwise. In this case, the probability of 1 is one-fourth, and the probability of 0 is three-fourths. Under this assumption, and thanks to the Central Limit Theorem, the overall exam result can be seen as a normally distributed random variable, representing the mean of all those question-level random variables, with values ranging from 0 to 1. A normal distribution simply means that most results cluster around the average, forming the familiar bell-shaped curve.

So what would have happened if the students had guessed all the answers? Nothing surprising. Almost all of them would fail, and with a group of 213 students, maybe one would barely manage to pass with the lowest positive grade, a 3. So it's probably not as simple as that.

However, here's the interesting part: according to a well-known statistical tool, the chi-squared test (which checks whether observed results differ significantly from what a model predicts), there is no significant difference between this theoretical model, where almost everyone fails, and the actual results from the exam.

The most unrealistic part of the model above is that, according to its assumptions, every student behaves like a machine, marking answers completely at random without even reading the questions. This is obviously not true. In reality, students read the questions, think them through, make decisions, and have at least some knowledge.

In the case of *Biological Bases of Behaviour*, students might have picked up bits of relevant knowledge from high school or other sources, like basic facts about the nervous system or hormones. Perhaps not enough to confidently pass the exam, but still something. All of this suggests that the probability of choosing the correct answer should be higher than 0.25.

Moreover, the group of students shouldn't be seen as entirely homogeneous. Some students almost certainly didn't study at all, while others worked really hard and likely learned a lot. And then there are those who, even without perfect preparation, can reason their way to the correct answer by using context or clues hidden in how the

question is phrased. This natural mix of different types of students inevitably shapes the overall distribution of exam results.

So, what if we imagined three types of students in this group? Let's say about 50% have a 0.3 probability of choosing the correct answer, 35% have a probability of 0.5, and the remaining 15% have a probability of 0.7.

Why these numbers? Simply because. Why not? I don't have the detailed data needed to fit a proper machine learning model, so I am speculating a bit, using a simple mixture model just to see what would happen. A mixture model is basically a way to combine several different groups with their own probabilities into one overall description of the results.

Surprisingly, even this rough theoretical model produces a grade distribution that closely mirrors the actual exam results. The small differences could easily come from random variation or from the fact that I excluded the open-ended question from the model, or both.

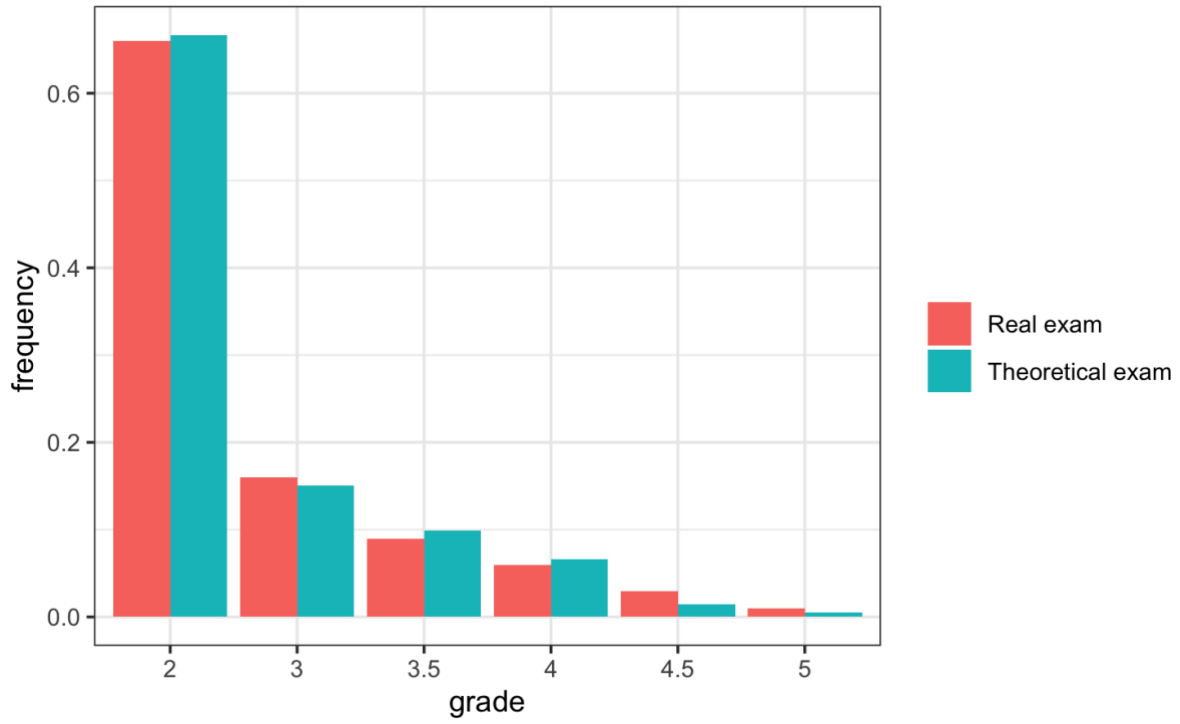

Figure 2. Theoretical grade distribution based on three groups of students with different probabilities of answering correctly. Despite being a simple, speculative model, it closely resembles the actual exam results.

What does this really mean? Does it suggest that the questions on the *Biological Bases of Behaviour* exam could just as well have been about quantum physics or the political situation in Uganda, and the grade distribution would still look almost the same? Maybe not quite that extreme, but we cannot fully rule out such a scenario. After all, topics like quantum physics or politics are not so obscure that, in a group of 200 people, no one

would be able to recognize an obviously incorrect answer with some probability, especially after a semester of lectures or at least skimming through a few presentations or articles.

So what really happened? This model suggests that between 50% and 85% of the students were unprepared for the exam, and those who passed did so mostly thanks to luck. Around 15 percent were prepared well enough to earn grades of 3.5 or 4, which in Poland are described as “passable” and “good”, respectively; and their exact results likely varied due to chance.

Think about it: over 200 adults at a reputable Polish university, all with prior experience passing exams (at least from high school) would all simply neglect a subject like biology, which is known to be challenging? I’m not convinced. Clearly, something was wrong with this exam.

Maybe the examiner prepared misleading questions. Maybe the lecture content didn’t match the exam requirements. Or maybe the examiner simply wanted to amuse himself. I don’t know. I’ve examined a few hundred students myself, and I know that sometimes there’s a temptation to fail many people just because you have the power to do it. But you have to resist that urge, because failing students is not the purpose of a university. (For the record, I’ve never failed more than 35% of my students). :)

I believe that studying is a matter of cooperation between the lecturer, who provides knowledge, and the student, who receives it. A grade distribution like the one described above is definitely a failure. But whose failure is it? In my opinion, the students are not to blame this time, especially since grade distributions in other subjects look quite regular. I swear, after ten years across three majors and three faculties, covering both social sciences and exact sciences, I have never seen a distribution like this one from *Biological Bases of Behaviour*.

It makes me sad that some lecturers so easily shrug off responsibility and blame the students by saying, “They didn’t learn enough”. I’m pretty sure something like that happened here. A true educator should have the humility to admit when they have failed; whether by setting inappropriate requirements or delivering poor lectures.

I hope my fellow students will eventually pass this exam. At the same time, as an academic, I feel ashamed of those lecturers colleagues who take this approach. I hope there are very few of them, because a university should be a place for sharing knowledge, not a place to prove who is smart and who is not.

I may never know what the ideal grade distribution should look like. But after this experience and with a little help from the Central Limit Theorem. I’m certain about one thing: an exam where most students fail says more about the exam than about the students.

Universities should be places where knowledge is shared, not hoarded. Where exams guide and inspire, not punish. And I hope there are fewer and fewer moments when we have to ask whether it was really the students who failed, or the educators.