Why Is Grade Distribution Often Bimodal? Why Individualized Teaching Adds Two Sigmas to the Average Grade? And How Are These Facts Related?

Christian Servin[1], Olga Kosheleva[2], and Vladik Kreinovich[2]
[1]El Paso Community College, El Paso, Texas, USA
[2]University of Texas at El Paso, El Paso, Texas, USA

Abstract

To make education more effective, to better use emerging technologies in education, we need to better understand the education process, to gain insights on this process. How can we check whether a new idea is indeed a useful insight? A natural criterion is that the new idea should explain some previously-difficult-to-explain empirical phenomenon. Since one of the main advantages of emerging educational technologies – such as AI – is the possibility of individualized education, a natural phenomenon to explain is the fact – discovered by Benjamin Bloom – that individualization adds two sigmas to the average grade. In this paper, we provide a possible theoretical explanation for this two-sigma phenomenon. In our explanation, we use another previously-difficult-to-explain empirical fact: that the grade distribution is often bimodal – and we explain this auxiliary fact too. In view of the above, we hope that our explanations will eventually lead to a more effectively use of emerging technologies in education.

1 Formulation of the Problem

To improve education, we need to better understand the corresponding processes. How can we make teaching more effective? How can we best use emerging technologies to improve the effectiveness of education?

Many researchers and practitioners have seemingly promising ideas. As we all know, some seemingly promising ideas work, but many don't work as expected – moreover, many of the un-tested seemingly reasonable ideas actually decrease the education's effectiveness. So, trial-and-error approach is not the best way here – we do not want the students to suffer while we are experimenting. Since with the current understandings about education we still have this significant potentially negative effect, a natural idea is to come up with new understandings – that will help us better filter the ideas and thus, decrease the potentially negative effect of testing.

How do we know that a new understanding works? The same way we know that a new physical theory works – if the new understanding explains some well-observed phenomenon that was difficult to explain before, this means that this understanding indeed contributes to the body of knowledge.

Which difficult-to-explain phenomenon should we target? With computer-based techniques and tools – especially newly developed AI tools – becoming ubiquitous, one of the main challenges is to analyze how to better use these tools. One of the main advantages of these tools – in comparison with the traditional education – is that they provide a realistic path to individualized education. Of course, individualized education, where the learning approach is optimized individually for each student, is a much more effective way for a student to learn. However, in the traditional approach to education, it is not possible to attain such individualization: this would require to have almost as many teachers as students, which is not realistic. Intelligent computer-based systems promise exactly such an individualization.

And here is a related theoretical challenge. Over decades, numerous experiments confirmed that individualized education is better. This qualitative fact is easy to explain. Interestingly, there is also a quantitative aspect to this phenomenon, and this quantitative aspect is difficult to explain – that individualized education increases the average grade by two standard deviations ("two sigmas"); see [1], see also [7]. In precise terms, if we denote the mean grade for the traditional education by m and the corresponding standard deviation by σ, then the average grade of the individualized education is very close to $m + 2\sigma$. This is the phenomenon that we explain in this paper.

What we do in this paper. In Section 2, to explain the two-sigma phenomenon, we use another well-known phenomenon – that in many cases, the class grades follow a bimodal distribution, where the majority of the students perform at an average level (forming one cluster), and much fewer students perform very well (forming a second cluster); see, e.g., [5] and references therein; see also [4]. Our explanation naturally raises the next question: how can we explain the bimodal distribution? This we do in Section 3.

In line with our arguments in the beginning of this section, we hope that our explanations will eventually lead to more effective education – and, in particular, to a more effective use of emerging technologies in education.

2 Why Individualized Teaching Adds Two Sigma to the Average Grade: An Explanation

What does bimodal distribution mean? As we have mentioned, bimodal distribution means that most students are at the average-grade level, while much fewer students are at the high-grade level. Let us denote the typical average-grade level by a and the typical high-grade level by h.

What does "much fewer" mean? To analyze this situation, let us describe the "much fewer" in precise terms. From the commonsense viewpoint, it is reasonable to interpret "much fewer" as "fewer than fewer": n being much fewer than N means that there is some intermediate value i such that n is fewer than i and i is fewer than N. So, to describe "much fewer" in precise terms, it is sufficient to describe "fewer" in precise terms.

Fewer means that the ratio i/n is smaller than 1 – i.e., that this ratio belongs to the open interval $(0, 1)$. Since we have no information about the relative probability of different values r from this interval, we therefore have no reason to believe that some values are more probable than others. Thus, it makes perfect sense to assume that all these values are equally probable. This argument dates back to Laplace and is this known as *Laplace Indeterminacy Principle*; see, e.g., [3].

The probability distribution in which each value is equally probable is known as the *uniform distribution*. We want to select a single value as a representative of this distribution. A natural idea is to select the mean – which in this is the same as the median, and is equal to $r = 0.5$.

Thus, we can conclude that "n fewer than i" can be naturally interpreted as $n = i/2$. Similarly, "i fewer than N" can be naturally interpreted as $i = N/2$. Substituting this value i into the formula for n, we conclude that $n = N/4$. So, the proportion p_h of high-grade students is 4 times smaller than the proportion p_a of the average-grade students: $p_h = p_a/4$.

Since these two proportions should add up to 1, i.e., $p_a + p_h = 1$, we thus conclude that $p_a + p_a/4 = 1$, i.e., $5/4 \cdot p_1 = 1$ and thus. $p_a = 4/5 = 0.8$. Correspondingly, $p_h = 1 - p_1 = 1 - 0.8 = 0.2$.

Let us estimate the average grade and standard deviation for traditional education and the average grade under individualized education. Under the traditional education, 0.8 of students get the grade a while 0.2 of students get the grade h. Thus, the average grade for traditional education is $m_t = 0.8 \cdot a + 0.2 \cdot h$.

If we have a fully individualized education, education that perfectly reflects the individual features of each student, that fully unlocks the potential of each student, then all the students will be able to reach the high level h. So, under the individualized education, the average grade is $m_i = h$. Thus, individualized education will increase the average grade by the difference

$$\Delta m \overset{\text{def}}{=} m_i - m_t = h - (0.8 \cdot a + 0.2 \cdot h) =$$

$$(1 - 0.2) \cdot h - 0.8 \cdot a = 0.8 \cdot h - 0.8 \cdot a = 0.8 \cdot (h - a). \tag{1}$$

For the case of traditional education, the variance $V = \sigma^2$ – i.e., the mean value of the square of the difference between the actual grade and the average grade – is equal to

$$V = 0.8 \cdot (a - m_t)^2 + 0.2 \cdot (h - m_t)^2. \tag{2}$$

Here,

$$a - m_t = a - (0.8 \cdot a + 0.2 \cdot h) = (1 - 0.8) \cdot a - 0.2 \cdot h =$$

$$0.2 \cdot a - 0.2 \cdot h = -0.2 \cdot (h - a),$$

so

$$(a - m_t)^2 = 0.2^2 \cdot (h - a)^2. \tag{3}$$

Similarly,

$$h - m_t = h - (0.8 \cdot a + 0.2 \cdot h) = (1 - 0.2) \cdot h - 0.8 \cdot a =$$

$$0.8 \cdot h - 0.8 \cdot a = 0.8 \cdot (h - a),$$

so

$$(h - m_t)^2 = 0.8^2 \cdot (h - a)^2. \tag{4}$$

Substituting the expressions (3) and (4) into the formula (2), we conclude that

$$V = \sigma^2 = 0.8 \cdot 0.2^2 \cdot (h - a)^2 + 0.2 \cdot 0.8^2 \cdot (h - a)^2 =$$

$$0.2 \cdot 0.8 \cdot (0.2 + 0.8) \cdot (h - a)^2 = 0.16 \cdot (h - 1)^2.$$

Thus, the standard deviation is equal to

$$\sigma = \sqrt{V} = \sqrt{0.16 \cdot (h - a)^2} = 0.4 \cdot (h - a). \tag{5}$$

This explains Bloom's two-sigma increase. By comparing the increase Δm as described by the formula (1) and the standard deviation σ as described by the formula (5), we conclude that indeed $\Delta m = 2\sigma$. Thus, we have indeed explained Bloom's two-sigma phenomenon.

3 But Why Is Grade Distribution Bimodal?

We still need to explain why the grade distribution is often bimodal. Our explanation of the Bloom's two-sigma empirical fact is based on yet another difficult-to-explain empirical fact: that the grade distribution is often bimodal. So, to make out explanation of the Bloom's two-sigma increase more convincing, it is desirable to explain the bimodality as well.

Analysis of the problem. What causes some students to perform at a lower-than-high level? The very fact that individualized education can bring all the students to the high-grade level shows that this is not the question of students ability. Since the reason is not the ability, the reason must be attitude, i.e., student's interest in the corresponding class.

For students who eventually get the high-level grade h, the interest is the largest, while for students who eventually get the average-level grade a, the interest is the smallest. It is therefore reasonable to gauge the student's interest by the student's grade g.

It makes sense to measure the level of the interest on the scale from 0 to 1, with 1 being the largest level of interest and 0 being the lowest level of interest.

It is therefore reasonable to perform a linear transformation from the a-to-h scale to 0-to-1 scale:

$$g \to \ell = \frac{g - a}{h - a}.$$

Instead of level of interest ℓ, we can talk about positive attitude $a_+ = \ell$ and about the remaining negative attitude $a_- = 1 - \ell = 1 - a_+$.

Students team together and influence each other, and this leads to polarization. Students taking the same class stick together, study together, hang out together. Usually, students who are most similar stick together. In particular, this means that students with similar attitudes stick together.

When students work together, study together, collaborate, they influence each other: students with positive attitude infect other students with their positive attitude and, vice versa, students with negative attitude infect other students with their negative attitude. When two students with mostly positive attitude affect each other, their positive attitude increases – and continues increasing until it reaches the maximum. Similarly, when two students with mostly negative attitude affect each other, their negative attitude increases – and continues increasing until it reaches the maximum – i.e., when it reaches the minimum of positive attitude.

This explains why, as a result, we have a bimodal distribution (and it also explains political polarization that have been observed in the US).

A simple mathematical model. Let us illustrate this phenomenon on a simple mathematical model. For a person A to be affected by person B's attitude – be it positive or negative attitude – we need to have this attitude in person B, and we also need person A to be amenable – i.e., have some of this attitude already. In general, the resulting effect e on A depends on the attitudes of A and B: $e = e(a_{\rm A}, a_{\rm B})$. The effect is 0 if either B has no such attitude, i.e., if $a_{\rm B} = 0$ or if A has no such attitude, i.e., if $a_{\rm A} = 0$.

To get a first-approximation description of this effect, it is reasonable to expand the dependence $e(a_{\rm A}, a_{\rm B})$ in Taylor series and keep the first non-zero terms in this expansion; this is a usual – and very successful – idea in physics applications; see, e.g., [2, 6]. In general, we have

$$e(a_{\rm A}, a_{\rm B}) = c_0 + c_A \cdot a_{\rm A} + c_B \cdot a_{\rm B} + c_{AA} \cdot a_{\rm A}^2 + c_{BB} \cdot a_{\rm B}^2 + c_{AB} \cdot a_{\rm A} \cdot a_{\rm B} + \ldots$$

Since the effect should be 0 when $a_{\rm B} = 0$, we get $c_0 = c_A = c_{AA} = 0$. Similarly, since the effect should be 0 when $a_{\rm A} = 0$, we get $c_0 = c_B = c_{BB} = 0$. Thus, the first possibly non-linear term is $e \approx c_{AB} \cdot a_{\rm A} \cdot a_{\rm B}$.

As we have mentioned, communicating with a person whose attitude is positive increases the person's positive attitude, so we should have $c_{AB} > 0$.

So, when two students with similar attitudes a_+ and a_- affect each other:

- each student's positive attitude increases by $c_{AB} \cdot a_+^2$, to $a_+ + c_{AB} \cdot a_+^2$, and
- each student's negative attitude increases by $c_{AB} \cdot a_-^2$, to $a_- + c_{AB} \cdot a_-^2$.

The sum of attitudes should be 1, so we should normalize the resulting values by dividing them by the sum of these two values. As a result, the new positive attitude becomes equal to

$$\frac{a_+ + c_{AB} \cdot a_+^2}{a_+ + c_{AB} \cdot a_+^2 + a_- + c_{AB} \cdot a_-^2}.$$

This process continues until the influence stop affecting the attitude, i.e., until the new value of the attitude becomes equal to the original value

$$\frac{a_+ + c_{AB} \cdot a_+^2}{a_+ + c_{AB} \cdot a_+^2 + a_- + c_{AB} \cdot a_-^2} = a_+.$$

This is clearly true when $a_+ = 0$. If $a_+ > 0$, we can divide both sides of this equality by a_+. Then, if we substitute $a_- = 1 - a_+$ into this formula and multiply both sides by the denominator, we conclude that:

$$1 + c_{AB} \cdot a_+ = 1 + c_{AB} \cdot a_+^2 + c_{AB} \cdot (1 - a_+)^2.$$

Subtracting 1 from both sides and dividing both sides by $c_{AB} \neq 0$, we conclude that $a_+ = a_+^2 + (1 - a_+)^2$. If we open parentheses and move all the terms to one side, we get $2a_+^2 - 3a_+ + 1 = 0$, i.e., $(2a_+ - 1) \cdot (a_+ - 1) = 0$. Thus, we have either $a_+ = 1$ or $a_+ = 0.5$.

One can check that the second case is unstable: collaboration with someone whose a_+ is slightly larger than 0.5 will eventually lead to $a_+ = 1$, while collaboration with someone whose a_+ is slightly smaller than 0.5 will eventually lead to $a_+ = 0$. Thus, the only two stable situations are $a_+ = 0$ and $a_+ = 1$ – which is exactly the bimodal distribution that we observe.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References

[1] Benjamin S. Bloom, “The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring”, *Educational Research*, 1984, Vol. 13, pp. 4–16.

[2] Richard Feynman, Robert Leighton, and Matthew Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.

[3] Edwin T. Jaynes and G. Larry Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[4] Christian Servin, Dan Padilla, Olga Kosheleva, and Vladik Kreinovich, "A Generative Model for Assessing Computing Education: The Fuzzy Bloom's Taxonomy Approach", *Proceedings of the NAFIPS International Conference on Fuzzy Systems, Soft Computing, and Explainable AI NAFIPS'2024*, South Padre Island, Texas, May 27–29, 2024, to appear.

[5] Karamjeet K. Singh, Tara Allohverdi, and Steffen P. Graether, "Changing Bimodal Grade Distributions – A Missed Opportunity?", *International Journal of Higher Education*, 2022, Vol. 11, No. 5, pp. 70–75.

[6] Kip S. Thorne and Roger D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2021.

[7] Laurah Turner, Matthew Kelleher, Andrew Zahn, Eric Warm, David Furniss, Anoop Sathyan, Weibing Zheng, Seth Overla, and Kelly Cohen, "Fuzzy Education: The potential for an Agentic AI System to Advance Precision Medical Education using Large Language Models, Fuzzy Logic and Shapley Values", *Proceedings of the NAFIPS International Conference on Fuzzy Systems, Soft Computing, and Explainable AI NAFIPS'2024*, South Padre Island, Texas, May 27–29, 2024, to appear.