

REGRESSION TO THE MEAN

I had one of the most satisfying eureka experiences of my career while teaching flight instructors in the Israeli Air Force about the psychology of effective training. I was telling them about an important principle of skill training: rewards for improved performance work better than punishment of mistakes. This proposition is supported by much evidence from research on pigeons, rats, humans, and other animals.

When I finished my enthusiastic speech, one of the most seasoned instructors in the group raised his hand and made a short speech of his own. He began by conceding that rewarding improved performance might be good for the birds, but he denied that it was optimal for flight cadets. This is what he said: “On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver. The next time they try the same maneuver they usually do worse. On the other hand, I have often screamed into a cadet’s earphone for bad execution, and in general he does better on his next try. So please don’t tell us that reward works and punishment does not, because the opposite is the case.”

This was a joyous moment of insight, when I saw in a new light a principle of statistics that I had been teaching for years. The instructor was right—but he was also completely wrong! His observation was astute and correct: occasions on which he praised a performance were likely to be followed by a disappointing performance, and punishments were typically followed by an improvement. But the inference he had drawn about the efficacy of reward and punishment was completely off the mark. What he had observed is known as *regression to the mean*, which in that case was due to random fluctuations in the quality of performance. Naturally,

he praised only a cadet whose performance was far better than average. But the cadet was probably just lucky on that particular attempt and therefore likely to deteriorate regardless of whether or not he was praised. Similarly, the instructor would shout into a cadet's earphones only when the cadet's performance was unusually bad and therefore likely to improve regardless of what the instructor did. The instructor had attached a causal interpretation to the inevitable fluctuations of a random process.

The challenge called for a response, but a lesson in the algebra of prediction would not be enthusiastically received. Instead, I used chalk to mark a target on the floor. I asked every officer in the room to turn his back to the target and throw two coins at it in immediate succession, without looking. We measured the distances from the target and wrote the two results of each contestant on the blackboard. Then we rewrote the results in order, from the best to the worst performance on the first try. It was apparent that most (but not all) of those who had done best the first time deteriorated on their second try, and those who had done poorly on the first attempt generally improved. I pointed out to the instructors that what they saw on the board coincided with what we had heard about the performance of aerobatic maneuvers on successive attempts: poor performance was typically followed by improvement and good performance by deterioration, without any help from either praise or punishment.

The discovery I made on that day was that the flight instructors were trapped in an unfortunate contingency: because they punished cadets when performance was poor, they were mostly rewarded by a subsequent improvement, even if punishment was actually ineffective. Furthermore, the instructors were not alone in that predicament. I had stumbled onto a significant fact of the human condition: the feedback to which life exposes us is perverse. Because we tend to be nice to other people when they please us and nasty when they do not, we are statistically punished for being nice and rewarded for being nasty.

TALENT AND LUCK

A few years ago, John Brockman, who edits the online magazine *Edge*, asked a number of scientists to report their “favorite equation.” These were my offerings:

success = talent + luck

great success = a little more talent + a lot of luck

The unsurprising idea that luck often contributes to success has surprising consequences when we apply it to the first two days of a high-level golf tournament. To keep things simple, assume that on both days the average score of the competitors was at par 72. We focus on a player who did very well on the first day, closing with a score of 66. What can we learn from that excellent score? An immediate inference is that the golfer is more talented than the average participant in the tournament. The formula for success suggests that another inference is equally justified: the golfer who did so well on day 1 probably enjoyed better-than-average luck on that day. If you accept that talent and luck both contribute to success, the conclusion that the successful golfer was lucky is as warranted as the conclusion that he is talented.

By the same token, if you focus on a player who scored 5 over par on that day, you have reason to infer both that he is rather weak *and* had a bad day. Of course, you know that neither of these inferences is certain. It is entirely possible that the player who scored 77 is actually very talented but had an exceptionally dreadful day. Uncertain though they are, the following inferences from the score on day 1 are plausible and will be correct more often than they are wrong.

above-average score on day 1 = above-average talent +
lucky on day 1

and

below-average score on day 1 = below-average talent +
unlucky on day 1

Now, suppose you know a golfer's score on day 1 and are asked to predict his score on day 2. You expect the golfer to retain the same level of talent on the second day, so your best guesses will be "above average" for the first player and "below average" for the second player. Luck, of course, is a different matter. Since you have no way of predicting the golfers' luck on the second (or any) day, your best guess must be that it will be average, neither good nor bad. This means that in the absence of any other information, your best guess about the players' score on day 2 should not be a repeat of their performance on day 1. This is the most you can say:

- The golfer who did well on day 1 is likely to be successful on day 2 as well, but less than on the first, because the unusual luck he probably enjoyed on day 1 is unlikely to hold.
- The golfer who did poorly on day 1 will probably be below average on day 2, but will improve, because his probable streak of bad luck is not likely to continue.

We also expect the difference between the two golfers to shrink on the second day, although our best guess is that the first player will still do better than the second.

My students were always surprised to hear that the best predicted performance on day 2 is more moderate, closer to the average than the evidence on which it is based (the score on day 1). This is why the pattern is called regression to the mean. The more extreme the original score, the more regression we expect, because an extremely good score suggests a very lucky day. The regressive prediction is reasonable, but its accuracy is not guaranteed. A few of the golfers who scored 66 on day 1 will do even better on the second day, if their luck improves. Most will do worse, because their luck will no longer be above average.

Now let us go against the time arrow. Arrange the players by their performance on day 2 and look at their performance on day 1. You