STATISTICS TEACHER
SUPPORTING THE TEACHING AND LEARNING OF STATISTICS

Some Paradoxes: Puzzling or Poorly Presented?

Mark Milanick University of Missouri, Columbia; Isabella Wiebelt-Smith, Swarthmore College; and William Y Jin, Swarthmore College

For many classrooms, using counterintuitive puzzles and apparent paradoxes is a way to engage students. However, it can be challenging to find and adapt grade-appropriate examples for teaching statistical concepts. Here we present four common "paradoxes" (Will Rogers Phenomenon, Simpson Paradox, False Positive Paradox, and the Birthday Problem) that middle and high school teachers can use with their students.

The terms paradox, counterintuitive, and fallacy are related but can vary in meaning across different people. The terms intuitive and counterintuitive are particularly tricky, with meanings dependent on a person's previous experiences and training. For example, Newtonian physics was not intuitive to Newton until he was in his 40s. For the purpose of an alliterative title, we use the term paradox in the lay sense of "it doesn't seem possible." We also provide specific examples as one way to make the paradox or problem seem intuitive. That is, if a person thinks X can never happen, one specific example of X is all it takes to eliminate the truth of the "never."

For each paradox, we first give a conventional description or example that many find counterintuitive. For some classrooms, this counterintuitiveness is a great way to engage students, but it runs the risk of making some students feel incompetent. As an alternative, we provide specific numerical examples that are accessible to most middle school and high school students. Additionally, teachers can share with students that professionals in a variety of fields, including forensics, health care, law, medicine, science, and sociology, also find these puzzles counterintuitive and have been fooled by them, in part, because of the poor wording or the fractions involved.

Because the numerical examples we provide are more intuitive to most students, we offer some extensions to engage students in thinking about each problem further. For middle school students, we suggest some ways to have the students create their own examples. In doing so, the students will be able to practice critically thinking about distributions and combinations, critically evaluating data, and interpreting data. In addition, these paradoxes can be used to motivate practicing arithmetic skills such as reducing fractions, adding fractions, and finding means. For high school students, our approach may give students confidence, possibly motivating them to go into more depth by considering situations where the paradoxes may arrive in socially and politically engaging situations.

As we tried to understand why these paradoxes are counterintuitive to many, we found that in our case, the confusion was often caused because the statement or example provided incomplete information or allowed the reader to make an assumption that the provider was not making, leading the reader to an incorrect conclusion. We should note that the term, "incomplete information" is meant to include the fact that reduced fractions, decimals, and percentages can be considered incomplete in the sense that, ½ can have a different effect on the analysis than 50/100, as seen particularly with the Simpson and False Positive Paradoxes. Since ratios, rates, percentages, and means are all essentially fractions, students should learn to be careful when only the reduced fraction is presented.

Will Rogers Phenomenon

The Will Rogers phenomenon describes situations in which given two groups A and B, moving an object from one group to the other increases the mean value of both groups. This phenomenon is considered counterintuitive because it describes a situation where moving an element from one set to another can increase the mean value of both sets, even when the value of the element itself doesn't change.

As with the other examples, whether a person finds this counterintuitive depends, at least in part, upon their previous experiences. In this case, for many students, everyday life may not have given them a previous example because they have only encountered cases in which moving something from one group to another made one group better and the other worse. For some, it is counterintuitive because they are implicitly making an assumption about the distribution of values in the two groups.

One way to help students understand this is by providing specific examples. If a person believes that X is impossible, it only takes one specific example of X to disprove that belief.

One example: Suppose you have 2 groups of dogs. Group A has three dogs weighing 30, 150, and 150 pounds for a mean of 110 pounds; Group B has three dogs each weighing 10 pounds, so the mean is 10 pounds. Which dog would you move from Group A to Group B to increase the mean weight of both groups? If the 30-pound dog is moved from Group A to Group B, then Group B's mean weight increases above 10 pounds to 15 pounds and Group A's mean weight increases from 110 pounds to 150 pounds.

Another example happens when the barrier between groups is changed, as shown in Figure 1. The original cutoff between large and small fish is the dashed green line. If the cutoff between large and small fish is shifted to the solid orange line, then the mean size of both the small fish and the large fish increases.

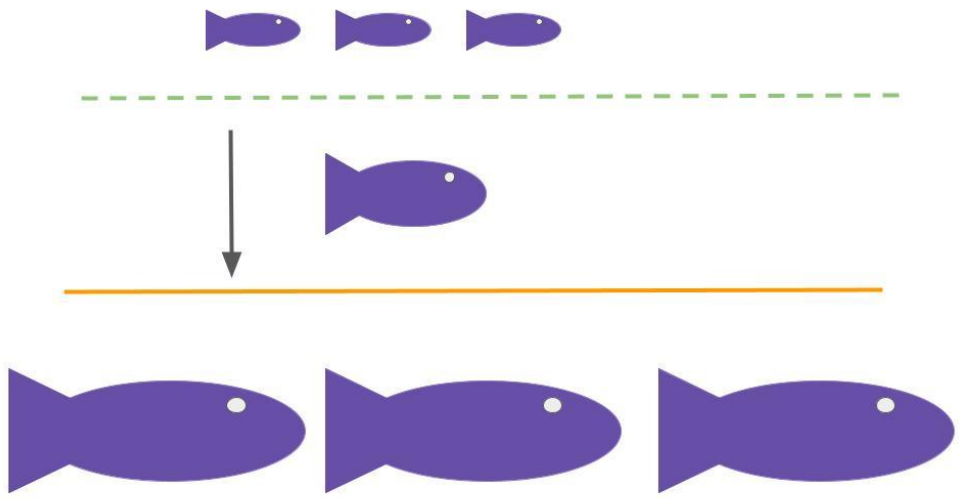

Figure 1. Example of changing the cut off yielding a Will Rogers Phenomenon

Students can be encouraged to create their own fictional examples of the Will Rogers Phenomenon. As they do so, they will probably discover that for this to occur, one must move a value that is between the means of the two original groups. High school students might even be able to prove that. Note that these conditions are necessary, but not sufficient (and having the students find examples where it is not sufficient might be an engaging activity).

In addition, the students could see if they can create examples from their class. One could divide the class into two groups (e.g., left side vs. right side, front vs. back, first half vs. second half of the alphabet for the first letter of first or last name, or birthdate in the first half or second half of the month). Then they could consider some "property" (e.g., height, number of vegetable servings per day for each person, number of pets at home). For which cases is it possible to move someone from one group to the other (e.g., from left to right) and create a Will Rogers Phenomenon where both group means change in the same direction?

High school students might be more engaged by thinking about changing cutoffs in different situations. In Figure 2, we have people with and without a difficult-to-quantify trait, represented as circles and squares, respectively. We have a surrogate approximate measure for the desired trait on a scale of 0 to 10. There is one circle and no squares at 10. There is one square and no circles at 3, and there are 5 circles and 5 squares at 6. With an initial cutoff of 9, 100% of the people above 9 have the trait and approximately 50% of the people below 9 have the trait. If the cutoff is lowered from 9 to 5, then about 50% of the people above the cutoff have the trait and 0% below the cutoff have the trait, lowering the percentages for both groups. One can do this for surrogate measures for health (e.g., blood pressure, cholesterol level, servings of vegetables) or surrogate measurements for "good" students such as GPA or standardized exam scores or other parameters.

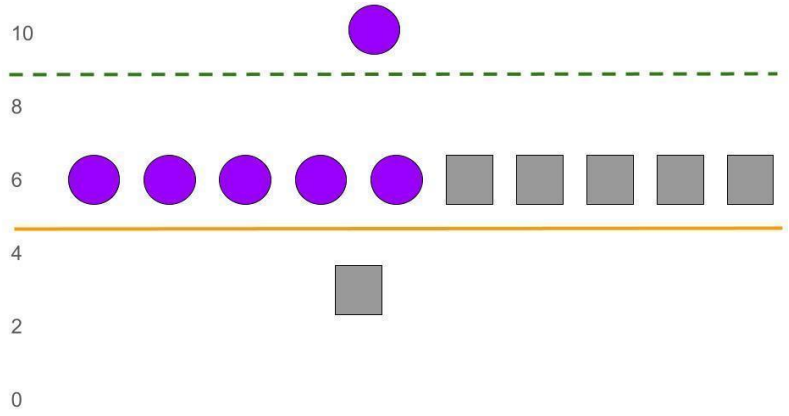

Figure 2. Another example of changing a cutoff yielding a Will Rogers Phenomenon.

Simpson's Paradox

Here is a traditional incomplete wording of the Simpson's Paradox: A and B play separate games of solitaire in the morning and in the afternoon. During the morning session, A won a higher percentage of their games than B did. During the afternoon session, A again won a higher percentage of their games than B did. Yet, when combining the morning and afternoon sessions,

B won a higher percentage of their games than A. For most of us, this initially seems counterintuitive.

Simpson's paradox describes a phenomenon in which a trend appears in several groups but disappears when the groups are combined. In working out an intuitive explanation for this, we realized that part of the problem is that the wording is incomplete. Indeed, if all the information were to be presented, either verbally or visually, we think most would find it intuitive. For others, they might be adding fractions incorrectly.

For example, it can help to provide concrete information: In the morning session, A won 1 game out of 4 games that A played and B did not win the 1 game that B played. In the afternoon session, A won the 1 game that A played and B won 3 of the 4 games that B played. However, when combining both sessions, A won 2 games out of 5 played and B won 3 games out of 5 played.

Using fractions, students can identify that in the morning session, A's success rate of 1/4 is greater than B's success rate of 0. In the afternoon, A's success rate of 1 was greater than B's success rate of 3/4. However, for the combined sessions, A only had a success rate of 2/5 and B's success rate of 3/5 is clearly higher. Some students may initially assume that A's success rate is 1/4 + 1 and B's is 0 + 3/4, suggesting there is no paradox as A's success rate is still higher than B's. This illustrates a common problem when adding fractions.

When either providing students with other examples or asking them to create their own example, if using small numbers, we have found it easiest to demonstrate Simpson's Paradox to students if B's morning games have a win rate of 0%, and A's afternoon games have a win rate of 100%.

An extreme example using large numbers is:

- In Session 1, A succeeds 1 time with 999 attempts and B has no success in 1 attempt.
- In Session 2, A succeeds 1 time with 1 attempt and B succeeds 998 times with 999 attempts.

This allows us to show that in each of the sessions, A has a higher win rate than B. However, when combining the sessions, A succeeds only 2 times out of 1,000 attempts and B succeeds 998 times out of 1,000 attempts.

Figure 3 is a visual representation of the phenomenon. It shows the success rates of two people playing solitaire during morning and afternoon recess. In this figure, each game is shown as a circle. Person A is represented by the top circles with a heavy dashed outline and Person B is represented by the bottom circles with a solid outline. Winning games are represented by filled orange circles, and losing games are represented by unfilled gray circles. The morning results are shown on the left and the afternoon results are shown on the right. It is visually clear to many students that A had a higher win percentage in both the morning (0.1 vs. 0) and afternoon (1 vs. 0.99), yet for the day, B had a higher win percentage (2 out of 11 vs. 199 out of 201).

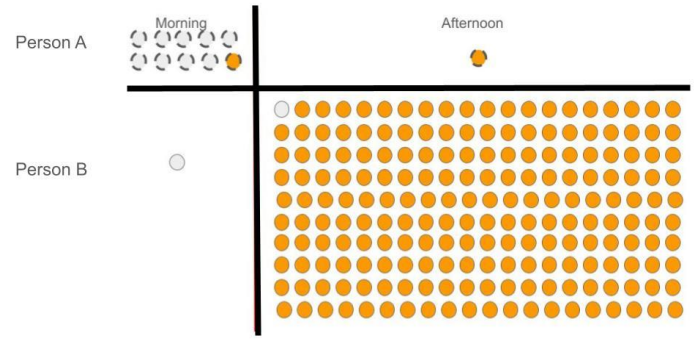

Figure 3. Simpson Paradox example illustrated

The same figure could also represent two trials of two drugs, A and B. In this figure, each patient treated is shown as a circle. A patient treated with Drug A is represented by a circle with a dashed outline, and a patient treated with Drug B is represented by a circle with a solid outline. If the patient has a positive response to the drug, its circle is filled with orange, and if it is negative, it is kept gray and empty. The first trial is shown on the left, and the second trial is shown on the right. It is visually clear to many students that the percentage of patients that responded positively to Drug A was higher than Drug B in both separate tests, but when the tests are combined, a higher percentage of patients responded positively to Drug B.

A real-life example is the incidence of mental distress among gamers. A recent study (Finseråas et al., 2022) found that recreational gamers reported less mental stress than non-gamers. The study also reported the responses for self-identified males and females. When looking at males and females separately, recreational gamers actually reported more mental stress than non-gamers. This is in part because many more of the recreational gamers self-reported as male, and female respondents were more likely to report higher levels of stress, causing the Simpson's Paradox. For the total population, recreational gamers had slightly less mental distress (0.308) than non-gamers (0.315). However, considering each gender separately, recreational gamers had more mental distress than non-gamers (for females, 0.48 vs. 0.34, for males, 0.18 vs. 0.16).

Another recent real-life example of Simpson's paradox concerns the recent COVID-19 pandemic (von Kugelgen et al., 2021). When examining the preliminary fatality rates of COVID infections, the rate was lower in Italy than China for each age group, but the total fatality rate for Italy was higher than for China. For example, the fatality rate for the whole population of China was half that for Italy. However, for ages 50-59, China had a 5 times higher fatality rate than Italy, and for ages 60-69, China had a 1.6 times higher fatality rate than Italy. The authors of the study suggest that this is due, in part, to the different distribution of people in different age groups in the two countries, but they also felt that a difference in testing patterns or social contacts for different ages could also have played a role.

To further engage, students could look for or create situations where Simpson's Paradox occurs. For example, for sports statistics, such as batting averages or points per game, Player A may have higher numbers in the first and second half of the season than Player B, but for the whole season, Player B has the higher numbers. As shown in our examples, finding a case of Simpson's

Paradox in real life can be challenging, but one approach that can help is to think about potential confounding variables.

False Positive Paradox

One hundred percent accurate tests are considered the gold standard. However, often it is too expensive, inconvenient, risky, or invasive to do a 100% accurate test. So a screening test is done, which is not 100% accurate. A false positive on a screening test is when the subject tests positive but does not have the condition being tested for. For example, in medicine, one might have a positive screening test result for a particular cancer, but further tests indicate the patient does not have cancer. The false positive paradox occurs when a test is highly accurate, yet the subject with a positive test result is still unlikely to have the condition.

A major problem of this paradox is the wording. For conditions or diseases, the term "test accuracy" is about how well the test separates those that have the condition or disease from those that don't. This is different from "test predictivity." However, "test accuracy" is often interpreted in the lay sense of how well the test predicts whether the person has the condition or disease.

A clearer and more complete statement would be that screening tests are evaluated as the ratio of those who have the condition and test positive divided by all those that have the condition (i.e., test sensitivity). In contrast, when determining the probability that one actually has the condition if they tested positive (i.e., a true positive), the fraction has the same numerator but there is a different denominator, that is, all those that have tested positive.

In the past false positives have been a problem for rare congenital diseases. Phenylketonuria (PKU) is the most common inborn error of amino acid metabolism in the U.S. Note all the qualifiers in that sentence—not only inborn, but also metabolic-involving amino acids (there are only 20 amino acids). Yet, the March of Dimes reports PKU is a rare condition: it only occurs in about 1 in 10,000 babies born in the US. For more about rare diseases, consult this mini-course by the National Organization for Rare Disorders.

Suppose the test sensitivity is 100%, which means that all who have PKU test positive. Furthermore, suppose the test specificity (or the percentage of those who do not have PKU that test negative) is very high, 99%. Table 1 shows that 100 people will test positive, even though only 1 in 10,000 people have PKU. These 100 undergo further screening and eventually all but one of their parents discover the baby does not have PKU; 99% were false positives. This type of scenario highlights the stress the parents may feel from a false positive test result contrasted with the assurance that the screening test will detect cases of PKU.

	have PKU	do not have PKU
PKU test positive	1	99
PKU test negative	0	9900

Table 1. Two by two table of PKU testing outcomes.

Once the concepts are understood, it may be appropriate to provide the standard (jargon) terms used in the literature:

- **Sensitivity** is the ratio of those that test positive and have the condition to all those that have the condition.
- **Specificity** is the ratio of all those that test negative and do not have the condition to all those that do not have the condition.
- In contrast,
- **Positive Predictive Value** is the ratio of those that test positive and have the condition to all those that test positive—the same numerator as sensitivity but a different denominator.
- **Negative Predictive Value** is the ratio of all those that test negative and do not have the condition to all those that test negative—the same numerator as specificity but a different denominator.

To engage the students further, they can look for examples of false positive rates that can cause possible concerns, or different people can weigh the risks of the adverse effect of a false positive with the adverse effect of missing a true positive. For example, shortly after 9/11, there was an attempt to screen for terrorists by the color of their shirt. The number of terrorists is very small compared to the number of airline passengers, so no matter what color one would pick, the chance that the person wearing that color was NOT a terrorist would be overwhelmingly high (see the excellent discussion by the Cato Institute). Screening for diseases has a similar problem as screening for security and safety. At what point should the anxiety and extra testing of having too many false positives outweigh the benefit of catching a few more people with the disease? Criminal profiling can also have a similar problem.

Birthday Problem

The birthday problem (or paradox), which may be familiar to teachers, involves the following question: How many people need to be in a room so that the probability that at least two of them have the same birthday is about 50%? The answer of 23 surprises many people. While the mathematics of the answer is interesting, we want to talk about a different aspect of the problem: why the answer was not intuitive to us. As we were modeling the problem with smaller numbers, we realized that our intuition was thinking about this problem: In a room with 22 people that don't share a birthday, a 23rd person arrives. And our intuition, correctly, led us to this line of reasoning: The probability is 23 out of 365, 6.03%, or about 1 out of 16.5 that the 23rd person shares a birthday with one of the previous 22. Another way to think about this is that the probability that the person does NOT share a birthday with anyone else in the room is 1 - (22/365).

Given this insight, we developed another way to present the birthday paradox. We realized that demonstrating with a 50% probability is problematic; in a class of 23, there is little use in surveying a class to show how common a shared birthday is as half the time there is no match, and the students are not convinced of the surprisingly common result. So, we prefer an example asking how many people are needed for a more than 90% chance of at least one match. Since the problem is general and not unique to 365, we suggest rather than asking about whether the birth month and day of month match, just ask if the day of the month matches. We demonstrate this

with a deck of 30 cards, numbered 1 to 30, making the simplifying assumption that all 12 months have 30 days. In this game, one draws a card, writes down the number, and returns the card to the deck and shuffles. This is repeated. The game stops (and the player "wins") when the same number is drawn twice as recorded on the sheet.

Suppose we have 30 cards numbered 1 to 30. We want to calculate the probability of NOT having a match.

Turn 1. Because this is the first turn and no cards have been previously picked, the probability of not having a match on Turn 1 is 1. You pick a card, say 1.

Turn 2. On this pick, there is a 29 out of 30 chance of NOT picking a 1 and therefore not having a match on Turn 1 AND Turn 2. Let's say you got a 2.

Turn 3. There is a 28 out of 30 chance of you not picking a match (a 1 or a 2) ON THIS TURN. BUT the probability of not picking a match on Turn 1 AND Turn 2 AND Turn 3, is 1 * 29/30 * 28/30 = 0.90. Say you got a 3.

Turn 4. On this pick, there is a 27 out of 30 chance of NOT picking a match (1 or 2 or 3) ON THIS TURN. BUT the probability of not picking a match on Turn 1 AND Turn 2 AND Turn 3 AND Turn 4 is 1 * 29/30 * 28/30/ * 27/30 = 0.81

And so on.

If one gets to Turn 15, our intuition is right in that the probability of not picking a match ON TURN 15 is about 15/30. However, the probability of not picking a match on all turns up to 15 is almost 0, that is, by this point you are almost guaranteed to pick at least one match and win. These probabilities for each turn are summarized in Figure 4.

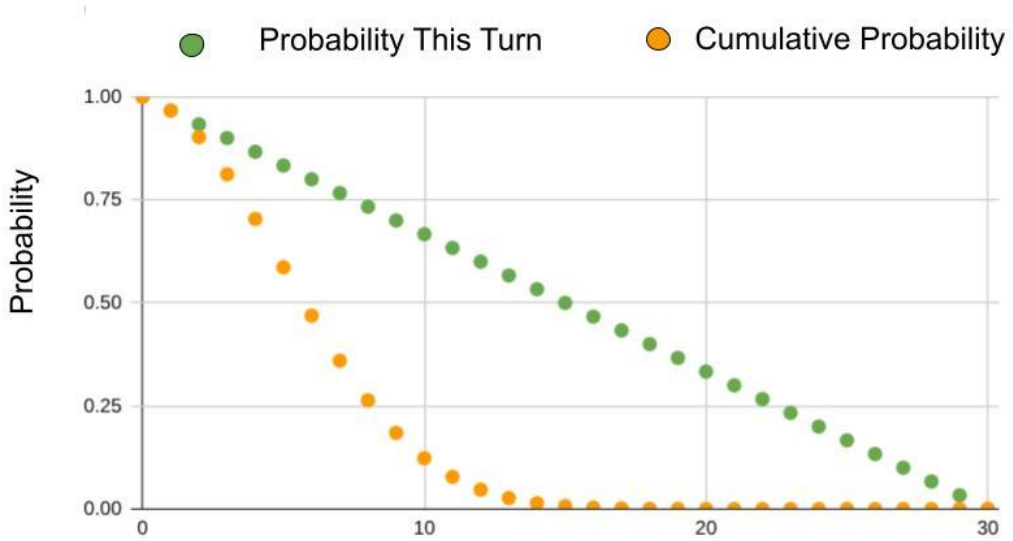

Figure 4. Probability of a match for this turn and cumulative probability.

We think the birthday problem can be more intuitive by providing more information. For example, one can start with the intuition: In a room with 22 people that don't share a birthday, a 23rd person arrives. The probability is 22 out of 365, or about 1 out of 16.6 that the 23rd person shares a birthday with one of the previous 22. However, the probability that the first 22 people don't share a birthday is just over 50%, because as each person arrives after the first, a match may occur, and the chance of matching accumulates as each person is added.

The insights from the birthday problem can also be applied to rare diseases. In the U.S. the legal definition of a rare disease is a condition that affects fewer than 200,000 people. With a population of 333 million, this means that a singular rare disease strikes fewer than 1 in 1,600 people. So, the chance that a person-with-rare-disease-A meets a randomly selected person that has the same rare-disease-A is less than 1 in 1,600. In contrast, because there are 5,000 to 10,000 rare diseases, meeting someone with some rare disease is about 8%! Thus, considering the probability of having a singular rare disease is not the same as taking into account all rare diseases, much like careful analysis of the birthday paradox reveals the importance of each additional person in the room adding to the probability of at least one birthday match. To phrase both situations using a parallel construction, the probability that you have the same birthdate (month/date) as a random person is 1 in 365. But the probability that at least two people in a room of 23 have the same birthday is about 1 in 2. The probability that a person with rare disease A meets a random person that has the same rare disease A is about 1 in 1,600. But the probability that a person with rare disease A meets a random person with SOME rare disease is about 1 in 12.

Conclusion

We presented four different paradoxes and shared different teaching techniques, such as rephrasing questions, thinking of analogies, and considering extreme situations, to support student intuition. For example, for some students, a more complete phrasing can counteract the counterintuitive nature of these paradoxes. We note the term, "more complete phrasing" also includes the fact that reduced fractions, decimals, and percentages can be considered incomplete in the sense that, ½ can have a different effect on the analysis than 50/100, as seen particularly with the Simpson and False Positive Paradoxes. Since ratios, rates, percentages, and means are all essentially fractions, students should learn to be careful when only the reduced fraction is presented. Decimal fractions and simplified fractions hide information—without knowing the size of the denominator, results can appear counterintuitive, but when all the information is provided, often the result is not surprising. In the case of the Birthday Paradox, the reader is making assumptions that the presenter was not. It may be the case that leaving out the critical information was not intentional, but it can still lead to confusion.

The strategy of examining an extreme case, either very large or very small, can also make the possibility of a paradox more obvious to students without the need to do complicated arithmetic calculations or use a calculator. This strategy of examining an extreme situation is not only useful for mathematical and statistical puzzles but also in many areas of science; for example, in biology, studying creatures that live in extreme conditions can lead to insights into physiology.

As illustrated, these different teaching techniques can be used within the context of common paradoxes to help support middle and high school students' understanding, confidence, and abstract reasoning skills.

References

Dickinson, J.A., Pimlott, N., Grad, R., Singh, H., Szafran, O., Wilson, B.J., Groulx, S., Thériault, G., & Bell, N.R.. 2018. Screening: When things go wrong. *Canadian Family Physician* 64(7):502-508. PMID: 30002025; PMCID: PMC6042667.

Finserås, T.R., Sivertsen, B., Pallesen, S., Leino, T., Mentzoni, R.A., & Skogen, J.C. 2022. Different typologies of gamers are associated with mental health: Are students DOOMed? *International Journal of Environmental Research and Public Health* 19(22):15058-71. DOI: 10.3390/ijerph192215058

Stone, W.L., Basit, H., & Los, E. 2023. Phenylketonuria. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535378/

von Kugelgen, J., Gresele, L., & Scholkopf, B. Feb 2021. Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects. *IEEE Transactions on Artificial Intelligence* 2(1):18-27. DOI: 10.1109/TAI.2021.3073088