Why is Replication so Important?

Mark Hanna 2012/12/30

One of the most important principles of the scientific method is *reproducibility*. A valid result should be able to be replicated independently, whereas an invalid result (originally achieved due to some error or perhaps just chance) will not be able to be consistently reproduced.

This is a concept that I didn't fully understand for a long time. I had reasoned that, say, doubling the sample size of an experiment should be just as good a way of confirming its results as performing the same experiment a second time with a different sample of the same size. This seemed intuitive to me, but eventually I came to understand why it is not the case.

The reason why this is the case has to do with **researcher degrees of freedom**. In an original experiment, the experimenters have the freedom to make certain choices. Some choices may have been made beforehand, whereas others are made after the study has been started. These choices may not all be made consciously, or they may be made consciously with only unconscious bias, but the fact that the choices are made at all after the experiment has begun affects the reliability of the results.

In contrast, when a well-designed experiment is being replicated, the choices have *all* been made beforehand, as the replication follows exactly the same protocols as the original study. This includes making all the same measurements and undergoing the same analysis. This reduces the researcher degrees of freedom for the replication experiment, so if the same results can be reproduced that's a good indication that the original results were accurate, whereas if they can't then it likely means they were due to some combination of bias and chance.

In some ways this can be pretty intuitive. For example, if the experimenters

were to carry out a variety of statistical analyses on their data and select the one that was most favourable to their hypothesis, then replicating the experiment with that same method analysis selected beforehand will reduce the bias from the initial decision.

Another great use for replication is in confirming the results of subgroup comparisons. For example, if I were studying the effect of a new drug on reducing blood pressure, I might perform comparisons between several subgroups and find it to be particularly effective in, say, people with type 1 diabetes. However, as more comparisons are made, the bar for statistical significance gets higher and higher. If I need to be 95% confident that my result is not due to chance for it to be statistically significant, then I can expect 1 in 20 results to appear significant by chance alone. There's a great xkcd strip that demonstrates how this can lead to unreliable results (remember to read the strip's alt text while you're there) – xkcd: Significant

The rest of this article below this point is a bit more technical. Therefore, you do not have to read it if you don't want to. However, if you want to read it, please do!

I find a scenario known as the "Monty Hall problem" (or the "Three door problem") to be an illustrative analogy of the importance of researcher degrees of freedom, especially in showing how unintuitive this importance can be. The problem goes something like this:

Imagine you're a contestant in a game show. In front of you are 3 doors, and you have to pick one of them. You have been told that behind 1 of these doors there is a car, but behind each of the other 2 doors there is a goat. You get to keep whatever is behind the door you open, so, of course, you want to win the car.

After you have made your initial choice, the host of the game show opens one of the 2 doors that you *didn't* choose and shows you that there is a goat behind it. Now, after seeing this you are given the opportunity to change your choice.

Intuitively it feels as though changing your choice would not affect your

chances of winning. After all, you know that the door you picked has a 1 in 3 chance of having the car behind it, and if you'd picked the remaining door first you'd have the same chance.

However, changing your choice at this point will *double* your odds of picking the car.

I find the easiest way to understand this is to run through each possibility in order to show the outcomes.

When you first pick a door, there are 2 possibilities – either you've picked the door with the car, or you've picked one of the doors with a goat. In 1 of every 3 attempts you will pick the correct door first and, if you don't change your decision, you'll get the car. In the remaining 2 attempts you will pick an incorrect door first and get a goat. So, the chance of picking the car if you *don't* change your decision is 1/3.

What if you do change your choice, though? 1/3 times you will have picked the car originally, so when you change your decision you will lose. However, what if the first door you picked had a goat behind it? In this case, the host will open the other door that has a goat behind it, so the only remaining door is the one with the car. This means that if you change your choice you are *twice* as likely to win, because your chance of winning would be 2/3.

This same power applies to researcher degrees of freedom. Decisions made once some or all of the data are known can have an effect on the reliability of the result.

In order to show the power of replication, let's re-imagine the Monty Hall problem. This time, you know there are 3 doors, and that behind each of them is either a car or a goat, and the same objects aren't always behind the same doors. However, you don't know if there are 2 goats and 1 car or if there are 2 cars and 1 goat.

Now, let's imagine you want to test the hypothesis that there are 2 cars and 1 goat behind the doors. In order to test this, you pick a door that you think has a car behind it. Once you've made that choice, however, you somehow find out that one of the other doors has a goat behind it, and knowing this makes you (consciously or unconsciously, it doesn't matter) change your decision to the remaining door.

In order to determine the probability that you'd pick the car, you'd need to repeat this many times. Using this approach, you'd pick the car about 2 out of every 3 attempts. This could lead you to conclude that there must be 2 cars behind the doors, instead of just one. However, we know that this is not the case! We would be able to show this by replicating the experiment.

In this case, an experiment to replicate these results might pick the same doors as the first experiment had eventually picked. Because this experiment has reduced the researcher degrees of freedom by making those choices beforehand, the result of the original experiment will be found not to be reproducible.

Of course, this analogy is exaggerated so as to make my point very obvious. In reality, the biases involved are much more subtle, but they are still there. The important thing to realise is that it is possible, even with the absolute best of intentions, to come to an unreliable result due to unconscious bias. In fact, every single decision can be made honestly and seem justified at the time, but the accumulated effect of many such choices will have an effect on the results. It's because of this that replication is so important in science.

Common choices that can affect the reliability of results by being made after the experiment has started include when to stop the experiment, how to analyse the data, and which subgroup comparisons to carry out.

The explanation that really helped me realise this was by Steve Novella in episode 373 of the SGU podcast (the relevant segment starts at 12:04),

discussing replications of the Psi Research done by Daryl Bem. He's written a post about the original research on his own blog, Neurologica Blog: Bem's Psi Research, and a post on the replications on Science-Based Medicine: The Power of Replication – Bem's Psi Research.

In a nutshell, Bem's experiments were well-designed (essentially they carried out some classic psychological experiments in reverse order) and the results were statistically significant and seemed to imply that the subjects exhibited precognition: the ability to predict supposedly random future events. However, when some of his experiments were replicated with all of the decisions made beforehand the results showed no ability better than what would be expected by chance.

Sometimes, good science gives strange and unexpected results. In some cases, such as in Bem's psi research, the results could even be called extraordinary. However, false positives can and do occur for a multitude of reasons, so in cases like this it's important to remember that "extraordinary claims require extraordinary evidence". In the face of such claims the correct course of action is neither to jump on the bandwagon nor to discard the results as false but to be on the lookout for quality replication. We live in an honest universe, and with time truth will out.