#55 Unreal Results in Education Research

“False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant”
https://journals.sagepub.com/doi/full/10.1177/0956797611417632

“Weaning Educational Research Off Of Steroids,” by Hunter Gehlbach and Carly Robinson.
http://www.shankerinstitute.org/blog/weaning-educational-research-steroids

“Mitigating Illusory Results through Preregistration in Education,” by Hunter Gehlbach and Carly Robinson. https://www.tandfonline.com/doi/abs/10.1080/19345747.2017.1387950

“Science Isn’t Broken,” by Christie Aschwanden.
https://fivethirtyeight.com/features/science-isnt-broken/#part1 Great multi-part series on the replication crisis.

“Stereothreat” Back in 1995, Claude Steele published a study that showed that negative stereotypes could have a detrimental effect on students' academic performance. RadioLab examines whether the results stand up today. https://www.wnycstudios.org/story/stereothreat

“Corrigendum: Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results” by R. Silberzahn et al.
https://journals.sagepub.com/doi/pdf/10.1177/2515245917747646

“When the Revolution Came for Amy Cuddy,” by Susan Dominus,
https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-amy-cuddy.html

“The Extent and Consequences of P-Hacking in Science,” by Meagan Head et al.
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

Winners Take All: The Elite Charade of Changing the World, by Anand Giridharadas. Great explanation of how the rise of big money “thought leadership” fuels the problems discussed in this episode.

Jennifer Berkshire: Welcome to Have You Heard. I’m Jennifer Berkshire.

Jack Schneider: And I’m Jack Schneider.

Berkshire: And our topic today is p-hacking, otherwise known as data dredging.

Schneider: Is that really the voice you're going to use? Could you make it sound any more boring than that, Jennifer?

Berkshire: Well, Jack, you just touched on what for me was really the central challenge of this episode. I get why this issue, the replication crisis that's sweeping across the sciences, is so important, but all of the specific terms that are involved leave me feeling well, a little sleepy.

Schneider: What better way than to address the issue with verve and panache than by doing an episode of the show about it?

Berkshire: Well, I thought it would be really helpful if before we get our guest on, if you could just sort of walk me and our listening audience through some of the things that they are going to be hearing. First of all, just tell us a little bit what the replication crisis is.

Schneider: Yeah, so the replication crisis is a phenomenon that has resulted from the fact that a lot of the things that have been quote unquote found by research may actually not be real effects. They may be illusory effects, and when you then try to replicate a study, an experiment say, and the effect that you thought was real the first time does not reproduce, that's a signal that maybe the original effect was illusory. And so the reason why the replication crisis is being described as a crisis is that a lot of the findings of social science end up getting called into question if they can't be replicated.

So if you do an experiment and you find, let's say, that standing in the wonder woman pose gives you more confidence and makes you more impressive as a speaker, and then somebody tries to replicate the study and cannot replicate it, then it calls into question whether that was a real finding in the first place or if it was just due to chance. Because of course anything will happen that can. And there's a very real chance that when we do an experiment that what we've found is just due to chance and not due to any particular intervention that was done in the study.

Berkshire: You did an excellent job there, Jack, and I want to test you a little bit further. I want to see if people are going to be hearing a lot of terms in this episode.

Schneider: I feel like I should have studied for this. You did not tell me that there was an exam portion of this episode.

Berkshire: I intentionally did not tell you because I wanted to see if you actually understand what all this stuff means. So we're going to be hearing a lot about preregistration. What's that?

Schneider: So the reason why you would preregister—let's start with that—is that if you have a huge data set and you go combing through it and you want to find, you know, are there any sort of surprising correlations? You've got all this data on folks. You've got data on whether they eat breakfast every morning, you've got data on their height, you have data on their favorite movie. You can then comb through all that data to find what's interesting or surprising in here and you may find that there is a surprising correlation between people who are tall and people whose favorite movie is Beverly Hills Cop II, and you may then conclude that there is either something about that movie that makes you taller or something about being tall that makes you like that movie.

Now it may just be that in this random sample of 100 people, we happened to find an unusually tall batch of Beverly Hills Cop II fans, and so to prevent that kind of fishing expedition, you would want to preregister your hypothesis with some sort of organization or agency like the Open Science Framework to say, we actually believe that watching Beverly Hills Cop II makes you taller. We are then going to go out, find 100 random people. Ask them what age they saw Beverly Hills Cop II, how much they liked it, and how tall they are now. And by preregistering, it keeps you from engaging in these kinds of fishing expeditions where you find these interesting but potentially not real results.

Berkshire: Very impressive, Jack. Now I just have one more question for you before we get our guest into the studio and dive into the topic of data dredging in earnest. Now, people can't see you obviously, but I just want to comment on the fact that you seem to be getting younger. Have you changed your personal regimen in any way?

Schneider: Everything has been the same. Other than that, I've been listening to a lot of Beatles songs lately and I don't know if that would have any impact… I drink the same espresso in the morning because we professors only drink espresso. We don't do drip coffee or any of the things that the hoi polloi drink. I drink cognac in the evening like professors do. And then it's kale salads all day long. So the only thing that's changed has just been my musical regimen.

[Music]

Berkshire: Well, we are joined now by Hunter Gehlbach. He's an associate professor of education at UC Santa Barbara and an expert on all things research related. Welcome Hunter. I want you to start by explaining to us what the song we just heard, and which Jack has been playing nonstop, has to do with the problem of illusory results in social science research.

Hunter Gehlbach: The story starts with really, I think, one of the most influential journal articles that has hit psychology in a long, long time and everything about the article is great except for the title,“False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.”

Schneider: I saw that on the rack at the supermarket but opted to read about Cher and her alien baby instead.

Gehlbach: Yeah, it's a little bit of a yawner. But the study itself is pretty awesome because what these guys do is they set up a study with the explicit goal of finding an illusory result— like something that just doesn't, couldn't possibly exist. And so this is how they end up with a finding, and that the causal language here, how listening to “When I'm 64 by the Beatles” makes you younger.

Schneider: I read this study and I can vouch for it and say these are all pretty standard steps that the researchers went through, but maybe you can talk through in greater detail exactly what they did because of course, you know, you can work backwards to come up with any results. But they didn't just approach this in a kind of slap dash, fly-night-way. They took very standard steps in terms of the way they were approaching this question and came up with this really sort of outlandish result. So help us understand that a bit.

Gehlbach: The steps that they go through are pretty straight forward, like really regular research practices that lots and lots of people do. So they go through and they do things, like they have a couple of different dependent variables. They bring a handful of people into the lab and sort of run them through these little tasks that they have where they are asking people to listen to different kinds of music. So they actually have this other thing that they do—they have a couple of different interventions. It may easily have been the case that it was listening to Calibo or there was some other little kid kind of song, I can't remember the name of it...

But any of those things, you know, had the results played out a slightly different way, could have made you younger. They asked for other information so they have lots of covariates. Anyway, the point being that they set up this study so that they have enough different variables in play so that as researchers, when you really thoroughly analyze your data, they had pretty much guaranteed they would be able to find a crazy finding like this.

And so lo and behold, by putting the right covariates, running the right amount of subjects, and then as soon as they found something significant, they stopped. They were actually able to come up, using the same techniques as most other researchers use, to come up with this crazy finding that listening to “When I'm 64 by the Beatles” actually makes you younger. So I don't know what your musical preferences are or are likely to be in the coming days, but I wouldn't put too much hope in this particular result.

Berkshire: The replication crisis has been getting a lot of attention in other disciplines, not so much in education, but one of the arguments you make is that there are things about how education research is done that actually make it uniquely susceptible to illusory results. Explain.

Gehlbach: So there are a few phases, well, there are many phases at which point you're making decisions as a researcher and all of them, you know, it’s pretty reasonable to make a decision just like this. And so phase number one is asking, what do I want to collect in terms of my variables? What am I interested in looking at? And one of the things that's unique about educational or maybe as compared to psychology, which is sort of my home discipline, one of the things that's unusual about education is it's really, really hard to get into schools and get permission and then get ninth graders to actually take permission slips home and get them signed and bring them back.

So the stakes are high for getting as much data out of that one collection as possible as compared to if you're in a psychology lab, where it might be relatively easy to just, you know, get a few more undergrads to come on in one afternoon and run a slightly different study. So step number one, you're collecting lots of information and because you have so many variables, you of course want to look at them. And one of the points I'd like to maybe stress throughout is, I think most of this, whether you want to call it p hacking or illusory results or data diving, it comes, I think mostly it comes from a place of, you know, really good intent. It's people trying to be thorough and taking a thorough look at their data.

Schneider: The piece that jumps out to me as being most obvious is when you just start fishing for any connection, right? So you said that the barriers that keep people out of schools actually kind of encourage folks to collect as much data as possible, sort of in one big gulp and the swallow it down and then it ends up in a spreadsheet and researchers can then begin to run their statistical software that to find connections that, you know, may or may not exist in the real world. But maybe you can talk through for folks why it is that a connection would appear between two variables on a spreadsheet, but that it might not actually reflect something that actually exists in the real world.

Gehlbach: Sure. So with the stage set in phase one that we have lots of different variables in the spreadsheet, phase two, you're certainly going to want to take a look at them all. You're curious. And one of the things about sort of basic statistical analysis is we have historically, and there are a lot of people pushing back against this part as well, but historically we've determined that something seems more like a quote real results if it meets this threshold of “p” being less than .05, so this is what a by convention most researchers in social sciences anyway deemed to be statistically significant. This sort of approximately means that the results that you're getting are occurring either by chance or not by chance. And so if there's a less than 5 percent probability that you're getting your results or some more extreme results, we feel like there's probably something unusual about your hypothesis. And typically that would be something along the lines of, well “the intervention worked.” That's why group one is different than group two.

Berkshire: I'm imagining our listeners nodding along at home because they're recalling Jackson very informative introductory seminar at the start of the episode. Okay, so now we've reached the part of our research experiment where we basically just analyze the hell out of our data. Why is that such a problem?

Gehlbach: If this is all about probabilities and trying to find something that is a result that is significant because the incentive systems for researchers are very much to find significant results. That's what gets you published in top journals. Then as soon as you try two different analyses, well your p value isn't actually below five percent, right?, because you, you kind of had two coin flips to try out. Then if you run 10 analyses, it's even more. And if you run 20 different analyses, again, with different combinations, you might compete your variables in slightly different ways, you're starting to take advantage of a running so many analyses that just by chance alone, we would expect one of them to come up as significant. And the sort of the easiest ways to way to think about it is if you ran 20 analyses, uh, you know, if our p value is set to .05, just by chance alone we'd hit that mark at least once if we ran 20 analyses.

Schneider: I've never had an issue with any of this. You know, I've done a few quantitative one with you, in fact two with you. And we preregistered our hypotheses for that. And you know, it feels like our results were non-illusory. I just wondering, do you have any experience with things going awry? Were any of the results that you found from previous studies, you know, replicated or not replicated?

Gehlbach: Really, in many ways, the way that I got into this whole issue of illusory results and how preregistration could really be helpful in terms of mitigating those is through my own work. And so I had this, what I thought and with much chagrin have to say is probably not the awesome study that I thought it was. But I set up an experiment where we went in to classrooms and gave teachers and their students in ninth grade a little get to know you survey. And so they reported out all these things from the get to know you survey and we matched them up. So each teacher, in the treatment group got matched with about half their students and half their students also found out what they had in common with their teachers. And lo and behold, at the end of the day, teachers and the students felt more similar.

The teacher student relationships were better. And by the end of the quarter, students were getting better grades, better grades if they were in the treatment group. So kind of a neat finding. But you know, these results weren't perfect. And so we did some additional analysis just to see if there was maybe another story going on in our data. So we looked more closely and found that actually where the real story was, was that the intervention didn't really make much of a difference for the white and Asian students at this school. And in conversations with the principal, it sounded like a, you know, those students typically we're pretty well served, but for the mostly Black and Latino students, this intervention looked like it really, really helped.

Berkshire: So close listeners might be able to detect the sound of horses hooves in the background. That's actually my co-host typing away feverishly. He is clearly up to something. Hunter, while Jack gets up to whatever it is he's getting up to, please continue. You've got your data, you start to explore it and, well, here, you tell it.

Gehlbach: In that exploration, the data just fell into place perfectly. So it looked like we had a really, really compelling story and so naturally we wanted to try and replicate this. So we went off to a bunch of different schools and tried running different versions of this study. We made some improvements in the protocols, so we did all this work and tried it again and tried it again and again and again about seven replications all told. And although we could get people to feel more similar, we really couldn't do much beyond that. And so with the benefits of hindsight, what happened, what I realized was for those couple of hypotheses, we preregistered, which were basically around similarity, we got a couple of those but we really hit and missed on some of the other hypotheses. And so we got some because I think the intervention does make people feel more similar.

But after that it was just chance results. Just by chance we split up the data in just the right way. We could have split up the data in lots of other ways and in fact look at that, and there weren't interesting findings there. So we happened to capitalize on, just by chance, on this one way of looking at the data and that's what's happening all over educational research. People are looking at the data, finding the one way in which it happens to work out to tell a pretty compelling story and then extrapolating on that and assuming that it's true rather than trying to replicate that finding and testing out for sure whether or not it’s a real finding.

Schneider: I want to get into the question of replication and the challenges that are posed by current reporting incentives. But before that, Hunter, I want to read to you something that made me insanely jealous a couple of years ago and that came from NPR’ Hidden Brain and it was Shankar Vedantam saying “is there a way to create a connection?” I'm not doing my Shankar Vedantam impression. “Is there a way to create a connection where there isn't one and how might that change things for teachers and students? These are the sorts of questions that fascinated Hunter Gehlbach and his colleagues for the experiment he had in mind…” And he goes on and on, and then he says, “When Hunter examined the test scores of students who had been induced to see that they had things in common with their teachers, he found something astonishing. Students, especially minorities, suddenly started to perform better in class.”

And you know, I'm reading this in part to, to make myself feel better, but also because…

Gehlbach: Well Jack, I think the whole point of this is to make you feel better.

Schneider: This is, this is my weekly therapy, my biweekly therapy session. But I read things like this all of the time, right? And so this I think is a great transition actually into a conversation about reporting because this is the kind of brass ring that we scholars are striving for. Maybe not all of us, but it feels really good to find something that seems like it makes a splash that people care about it. It's a clear finding, it’s snapped up by a prestige journal and it's reported on by media outlets. And of course, that's a part of the problem as you point out in this article of yours and as others who have been looking at this phenomenon more broadly in science, have pointed out, that journals have some accounting to do for themselves with regard to findings that are illusory and that don't replicate, and that the media play a role as well. And I'm wondering if you can talk through that for us.

Gehlbach: Yeah, I mean, I think you've hit it on the head. And this was part of our problem, to be blunt. I got a bit ahead of myself and probably should have tried the replications first before getting press. So a lot of these types of studies are out there and they are incredibly appealing.

And I think the one other thing to emphasize is they make sense. Humans are inveterate storytellers. We can tell you a reason why all sorts of things occur and we can tell stories about why the opposite of those things occur. And so part of the problem is our results looked a lot like these similar studies that I had been very jealous of and always wondered why I couldn't have some of my research work out that way. I kind of believed my own schtick. I thought this story was so compelling that of course it would replicate and it looked so similar to some of these other findings that I was just really excited about it and started talking to people perhaps a little bit before the finding was warranted.

Schneider: Seems to me that part of the problem here is related to presigee, that actually there's a lot more cache in publishing something new rather than an replicating someone else's findings. This is true not just in social sciences but in humanities as well. I won't name him, but there's a historian who went around sort of replicating other people's studies and sometimes coming up with the same finding and sometimes coming up with different findings and people kind of dismissed that as hack work. And I'm thinking also of the fact that there's more prestige in having a finding then in not having a finding, right? There's more prestige in finding, you know, ‘hey, we found this song that if you listen to will make you younger’ than there is in publishing a finding, you know, ‘surprise, surprise. We found that if you listen to music, you end up, you know, approximately four minutes older at the end of the song than you were at the beginning.’ So what does prestige have to do with any of this and is anything changing around that?

Gehlbach: Yes and no. So there’s a lot of change that's happening right now in psychology. They are moving forward with a lot of different initiatives, big, big replication trials that they're doing and finding fairly different results. In many cases there are journals in psychology that have started to offer what are called registered reports where you send in your introduction, your methods and your hypothesis and that's it. So there are no results. And so then the reviewers say like, ‘yes, this is an interesting study, let's run it and we'll publish it no matter what the results are. Or they say like, ‘eh, not a very interesting question. Don't waste your time running that study.’ And that's a very different model for publication. So the, yes, part of my answer to your question is that those things are starting to happen. The no part is that they're starting to happen in other fields, psychology is the one I'm most familiar with, but they are not starting to happen much in education.

I would maybe highlight one exception which is that AERA Open, which is a journal, has a sort of trial run on this idea of registered reports. So I've been helping out with that initiative a little bit. And one of the things that's really funny and very interesting about it is that I find reviewers are much, much more humble when they don't know what the results are. They're much less likely to say, ‘you should do this, you should do that. You should analyze the data this way or that way.’ When they don't know what the results are, they sort of say, ‘okay, your analysis plan seems fine. Go forward. Let's find out what you got.’ But the idea is very much that you’ll be able to publish results regardless of whether there are findings or not. The replication stuff really hasn't happened that I'm aware of in education. And so that's an area that needs more work and attention.

Berkshire: Hunter, you've alluded a couple of times to preregistration is a possible solution to what you call research on steroids. Can you just break down for us what that looks like?

Gehlbach: So, if we sort of conceptualize the main problem as the production of illusory results, researchers looking at their data so thoroughly that they find significant findings, but it's actually by chance alone, and how can you stop this and how can you stop it in a responsible way that doesn't cut down on researchers creativity and the possibility that we do find real findings by chance sometimes, so that's sort of the key problem. And a fairly simple solution, I think, first took place in medicine and lots of other disciplines have followed suit. And the idea is very simply to post publicly, and obviously the web makes for a very easy way to do this, exactly what your hypotheses are, exactly what your methods are, exactly the analysis you're proposing to run and then it's there and you've got to stick to it. And there can't be any ambiguity in it.

So you can't pose a general research question. You have to sort of dish out the exact analytic formula you plan to use. And if that's posted and you do that for, you know, just one or maybe a small number of hypotheses, then if you go and run the data or analyze the data and exactly the way that you said you would, and you get a finding that hits various sort of thresholds, you know, meets the p value or hits the right confidence interval, then we should really believe your study. And if you don't, then it's fairly interesting too. So that's the basic idea of preregistration.

Schneider: OK, so I’m going to shelve the Beatles albums for a little while. But I've got to ask you, I’ve read other studies that have told me that if I drink a glass of red wine every day, if I eat a little bit of dark chocolate every day, if I eat, I don't even know how to pronounce the name of those berries. Can I, can I grow to a ripe old age by feeding on these strange looking berries that Jennifer has here? And if not, how do I know when I'm being hoodwinked here?

Gehlbach: Some of the flags for when you might be finding illusory results, and I don't know the particular red wine study that you're referencing, but if indeed red wine does not make you younger, I would probably look at things like what's the sample size? So it's much easier to find chance results with a small sample. Are there a bunch of covariates that are involved? So it turns out that listening to When I'm 64 by the Beatles makes you younger only after you control for your father's age. You had to have a control variable in there. It’s a very, very important detail.

So the, the paper that Carly Robinson and I wrote a detail and list out a bunch of these things, which again, are often good practices, but when you, and, and smart things to do, but if they're not preregistered and if you start to see a few of them accumulating up, then I think it's reasonable for readers to start getting worried that there might be some illusory results in there.

Berkshire: That was Hunter Gehlbach. He's an associate professor of education at UC Santa Barbara and the author of a recent paper called, and this is a little bit of a mouthful, Mitigating Illusory Results Through Preregistration in Education. And he wrote that with Carly Robinson. And Jack and I will be right back to wrap things up.

[Music]

Berkshire: So Jack, one of my favorite podcasts other than ours, of course, is called Reply All. And they do this regular feature where the sort of boss of the show, Alex Blumberg, who is very well known, comes in to the studio and there's something from the Internet that he doesn't understand. The implication is that because he's a gentleman of some years, that things are happening, particularly on Twitter, that he just doesn't get it all and then the young sort of tech savvy experts break it down for him and then at the end of the show he has to repeat back to them what he has learned.

Schneider: Oh my God, I see what you've set me up to do here.

Berkshire: Well no, we heard from you at the beginning of the episode. I'm actually giving myself this challenge. So I thought what I would do is I would summarize for our listeners what this episode was about and what you know, why these issues are important. And you can tell me if I got it right.

Schneider: I am so sad that I left my buzzer at home. That would bring me such satisfaction.

Berkshire: Okay, so there is a replication crisis in science, social science including education research. And what that means is that when people go back and they try to replicate the findings of big studies, big or small studies, they can't do it. And why can't they do it? Well, one reason is that findings that were originally found to be statistically significant, to have a particular p value, turn out to be illusory. And we talked about all the reasons why this could be the case in education, right? That people have to grab all this data in one gulp. They're under pressure to find something. And so one possible solution is for people to preregister their hypotheses, lay out what they expect to find. And then that should help with the problem of people proclaiming results that turned out to be illusory. Did I use enough of the words?

Schneider: I think you used all the words. This was a great round of men magnetic poetry. I would make one change to that.

Berkshire: Oh, unfortunately, we're out of time.

Schneider: Thanks folks. I'm Jack Schneider. And that would be that, some stuff, if you replicated it, would replicate right. That the crisis is not that nothing would replicate, and that everything has just been a kind of illusory result, but that lots of the biggest splashiest findings don't hold up and I would say preregistration is not a silver bullet, that actually there are lots of other ways to check that. And that you know, one of them is by actually having an ideologically diverse research team because of course, you know it, if what you're finding aligns with what you hoped to find, you're not going to dig as hard as if it just doesn't make sense to you at all. Why you're finding, what you're finding. So I think there are a lot of other ways to kind of check against illusory results, but certainly preregistering your hypothesis and holding yourself to that so that you're not just casting a huge net and hauling in whatever kinds of real or fictitious results come in with that. That I think is a good first step.

Berkshire: Well, as our regular listeners know, we end each episode urging them to follow us into the weeds. It's how we support the podcast. And my question this week was, 'how could we possibly get any weedier than we did in this episode?' And Jack said, 'oh, I, I have a way.'

Schneider: It can get so much weedier! We are going to talk about the implications for qualitative research for all of this and the relationship between qualitative and quantitative research and the ways that human judgment shapes all of it. Jennifer's eyes have totally glazed over. This is so the opposite of your usual pitch. Like, Hey, we're going to talk about this super exciting topic and then you sort of dangle the keys to the car and then throw them over the paywall. Um, I think it was a pinata last time, this week it's a car. But I think it's gonna be super fun.

Berkshire: Well, this is the holiday time and just a reminder that if you're feeling particularly generous, we would love your support. It helps us keep the podcast going, helps us pay our outstanding producer. You can go to Patreon.com and search for Have You Heard or go to www.haveyouheardblog.com and you will find all the different ways that you can support us.

Schneider: And if you are trying to engage in a less consumerist holiday season, you can support the podcast by telling your friends and neighbors about us, by engaging with our twitter handle @haveyouheardpod or by taking the lessons you've learned from this episode and going out into the world and acting on them.

Berkshire: In the meantime, thanks for joining us for another statistically significant episode. I'm Jennifer Berkshire.

Schneider: And I'm Jack Schneider. Do not edit that out. I really liked that.