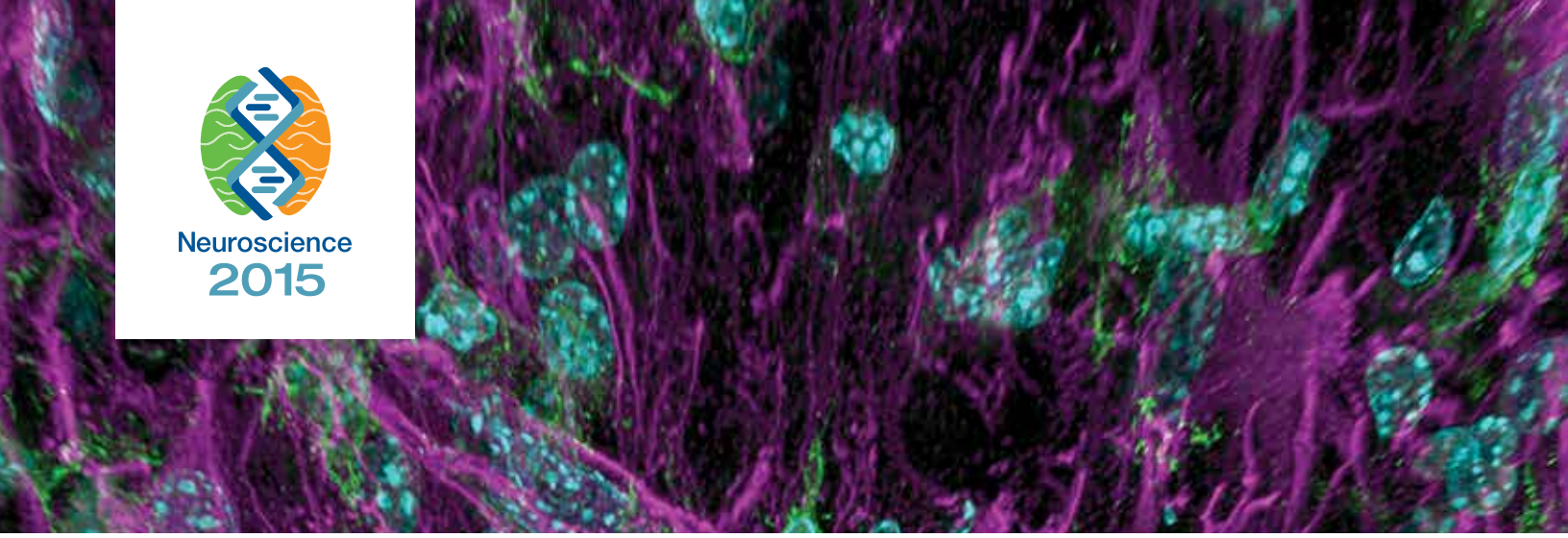


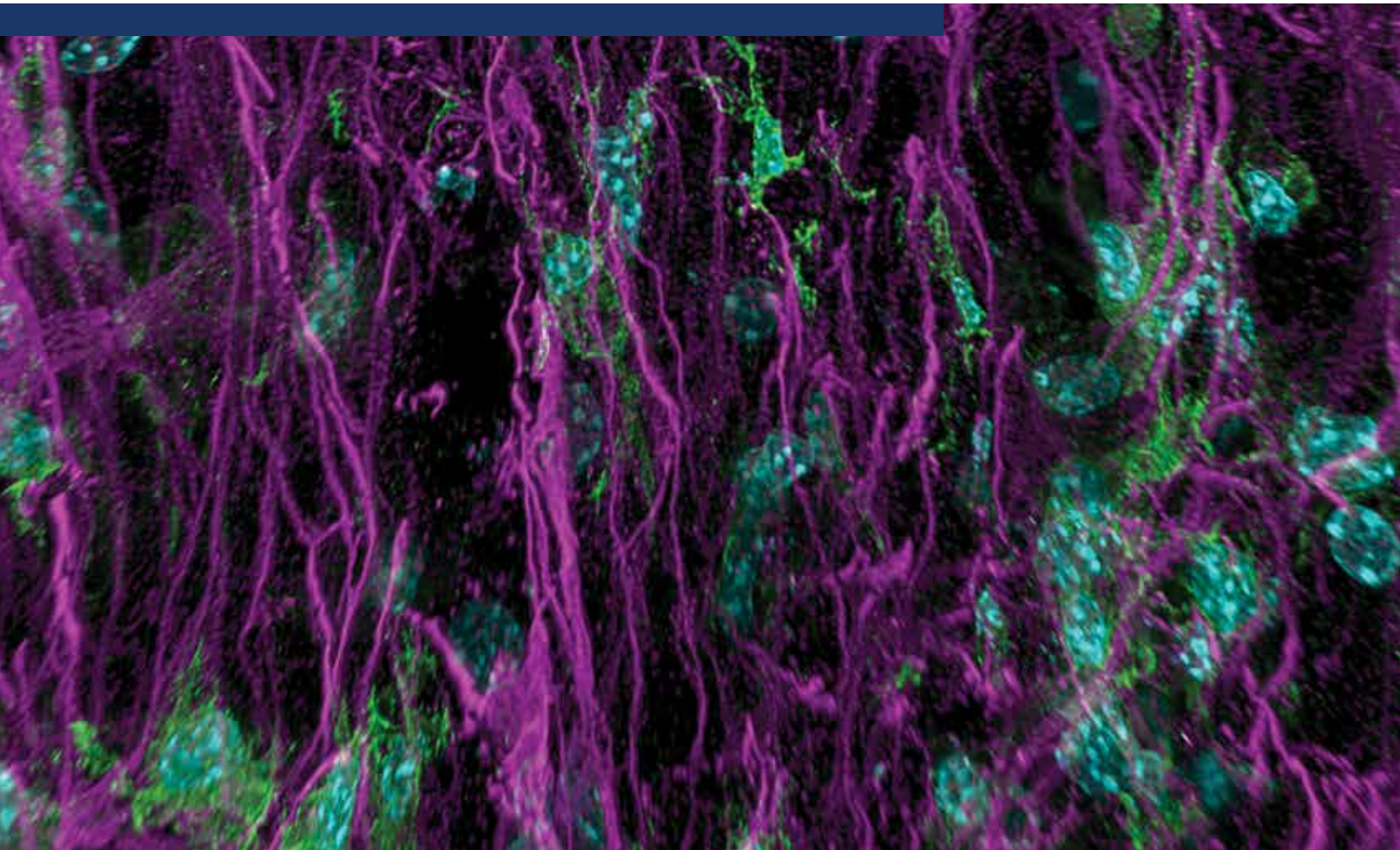
Neuroscience 2015

SHORT COURSE III

Optimizing Experimental Design for High-Quality Science

Organized by Mara Dierssen, MD, PhD; Magda Giordano, PhD; Chris J. McBain, PhD; Charles Mobbs, PhD; John Ngai, PhD; and Rae Nishi, PhD

SfN Society for Neuroscience



Short Course III

Optimizing Experimental Design for High-Quality Science

Organized by: Mara Dierssen, MD, PhD; Magda Giordano, PhD;
Chris J. McBain, PhD; Charles Mobbs, PhD; John Ngai, PhD; Rae Nishi, PhD

SOCIETY *for* NEUROSCIENCE

Please cite articles using the model:
[AUTHOR'S LAST NAME, AUTHOR'S FIRST & MIDDLE INITIALS] (2015)
[CHAPTER TITLE] In: Optimizing Experimental Design for High-Quality Science.
(Dierssen, M; Giordano, M; McBain, C; Mobbs, C; Ngai, J; and Nishi, R) pp. [xx-xx].
Chicago, IL: Society for Neuroscience.

Neuroscience 2015

Short Course 3

Optimizing Experimental Design for High-Quality Science

Organized by: Mara Dierssen, MD, PhD; Magda Giordano, PhD; Chris McBain, PhD; Charles Mobbs, PhD; John Ngai, PhD; Rae Nishi, PhD

Friday, October 16 / 1 p.m.–5:30 p.m.
Location: McCormick Place Convention Center, Room N227, Chicago, IL

TIME	TITLE	SPEAKER
12:30–1 p.m.	Check-In	
1–1:10 p.m.	Opening Remarks	Katja Brose, PhD *Cell Press*
1:10–1:40 p.m.	Rigor and Transparency	Shai Silberberg, PhD *National Institute of Neurological Disorders and Stroke*
1:40–2:20 p.m.	Rigor and Transparency Discussion	Shai Silberberg, PhD *National Institute of Neurological Disorders and Stroke*
2:20–2:35 p.m.	Break	
2:35–3:05 p.m.	Improving Experimental Design to Boost Reproducibility	Mary Harrington, PhD *Smith College*
3:05–3:45 p.m.	Improving Experimental Design to Boost Reproducibility Discussion	Mary Harrington, PhD *Smith College*
3:45–4 p.m.	Break	
4–4:30 p.m.	Data Analysis and Reporting	Ronald Landis, PhD *Illinois Institute of Technology*
4:30–5:10 p.m.	Data Analysis and Reporting Discussion	Ronald Landis, PhD *Illinois Institute of Technology*
5:10–5:30 p.m.	Question and Answer	

Table of Contents

Rigor and Transparency

Shai Silberberg, PhD

Program Director, National Institute of
Neurological Disorders and Stroke

NOTES

Publications in recent years have highlighted problems with the reproducibility of published scientific findings. Unfortunately, the popular press (and, sadly, some scientific papers) unfairly taint the scientific community by implying that poor reproducibility is, to a large extent, the result of scientific misconduct. These alienating claims, while flashy, are unfounded. Inability to reproduce published results can arise in part from the complexity of life-science research, in which it is difficult to identify and/or control all the experimental variables that impact results. There are many contributing factors to poor reproducibility. Examples include:

- **Cutting-edge science:** Complex experimental techniques developed in one lab may require extensive training before being successfully employed by others.

- **Confounding variables:** At times, variables that affect the results are unknown to the investigator and therefore not reported. This can prevent others from reproducing the results.

- **Resources:** Much has been written in recent years about misidentification of cell lines and the quality of antibodies and other experimental resources. Inability to reproduce original results can arise if labs are using different resources.

The negative impact of some, but not all, of these issues can be mitigated by careful experimental design and better transparency. This presentation will primarily focus on related contributing factors to poor reproducibility, which are major contributors to the issues surrounding reproducibility of scientific findings. These include:

Human nature: We are all prone to unintentional and unconscious bias, necessitating that scientists take every measure to minimize bias in their studies.

Deficient experimental procedures: Careful experimental design is the means to curtail bias and minimize the effects of potential confounding factors. Examples of poor experimental design will be provided.

Lack of transparency in reporting: Not all sources of potential bias can be mitigated, and therefore it is essential to transparently report how experiments are designed, conducted, and analyzed. This informs reviewers, editors, and the scientific community on potential pitfalls in the reported outcomes and conclusions.

Publication bias: The pressures to report outcomes that support the working hypothesis (i.e., "positive results"), as well as the policies of most journals to disfavor the publication of null results, have led to a vast body of unpublished research. This muted data could potentially inform future studies, improving the reproducibility of science.

The talk is primarily intended to increase awareness. It outlines the issues, highlights common design pitfalls, and provides guidance for avoiding them, but it by no means covers all topics or discusses the issues in detail. It is intended to encourage the attendees to expand their knowledge and to do their part to increase the reproducibility of science.

Rigor and Transparency

Case Study 1

An article in the *Journal of Neurochemistry* was retracted in 2013. After reading the reasons for the retraction, discuss the questions below.

"The retraction has been agreed to following the discovery of an unexpected effect of the disposable filter units on neuronal morphology. Concerns about the published data came to light following variable results in follow-up experiments investigating the mechanisms responsible for the effects reported in the article.

Further investigation revealed that most of the effects attributed to neuroserpin appear to be due to a factor or factors leaching in a volume-dependent manner from disposable filter units used to sterilize the neuroserpin. Neurobasal medium that had been filtered through a 0.2 lm syringe filter resulted in increases in neurite length and reductions in neurite diameter at 2 DIV similar to those reported with neuroserpin-containing medium. The effects were seen when 2 mL of Neurobasal medium was filtered and added to the cells, but were reduced when a larger volume of medium was filtered. Medium filtration was performed to ensure the medium containing recombinant neuroserpin was sterile. As the control medium lacking neuroserpin was already sterile and it was not anticipated that medium filtration would alter neuronal growth, medium filtration was not controlled in the study. In some experiments,

NOTES

medium for 'control' wells was filtered in a larger volume, while in others the medium was not filtered at all. Therefore, an apparent effect of 'neuroserpin' could have been caused by presence of filter leachate in the neuroserpin conditions, which was absent in the control conditions."

Discussion points

- What led to the erroneous conclusion that neuroserpin affects neurite outgrowth?
- Was the experiment poorly designed?
- Is this an example of scientific bias?
- How would you design the experiment to avoid such a mistakes?
- What general lessons can be learned from this real-life experience?

Rigor and Transparency

Case Study 2

At a lab meeting Dr. Proof discussed the reviews of a manuscript submitted to a high-profile journal. The cover letter from the Journal Editor suggested that the manuscript will be accepted for publication if the lab can demonstrate that the fraction of morphologically modified cells in histological sections is significantly reduced after treatment. Dr. Proof requested that the additional analysis be conducted independently by the two students listed as co-first authors on the manuscript, and that someone from an adjacent lab be asked to mask the samples.

Over lunch, the two distraught students grumbled that Dr. Proof doesn't trust them and that by requesting that the samples be masked by the adjacent lab she is publically degrading them. Both students felt that it was a mistake to join her lab!

Discussion points

- Do you agree with the students that Dr. Proof treated them unfairly?
- What reasons could have led Dr. Proof to request that the experiment will be conducted blind by two independent students?
- Should this be common practice?

Improving Experimental Design to Boost Reproducibility

Mary Harrington, PhD

Tippit Professor in Life Sciences, Smith College

NOTES

Our goal as scientists is to make important discoveries. We should celebrate most, not when we publish a paper in a top-tier journal, but when our fiercest competitor replicates our findings. If our work is not reproducible, it cannot be used to lead to further discoveries; in fact, it can slow the pace of progress while other scientists spend their time attempting to replicate our experiments.

Are studies within the field of neuroscience leading to findings that are reproducible? Some indications suggest not. Neuroscience studies are routinely under-powered (average statistical power is estimated to be 8 to 31 percent; Button et al., 2013). Many of the positive effects reported in our literature may not be valid. Other studies indicate that results from drug development studies are replicated 25 percent of the time, at most (Begley and Ellis, 2012; Prinz et al., 2011). Analysis of published studies reporting findings that were not reproduced indicates that several elements of good experimental design were not reported and may not have been followed. Many journals are now adopting standards that examine submissions for adherence to these principles of good design. We will focus on three design elements:

- consideration of sex as a variable,
- use of randomization, and
- application of blinding where possible.

In FY2016, NIH grant applications for preclinical research will be required to "explain how relevant biological variables, such as sex, are factored into research design and analysis." Scientists will be required to provide "strong justification from the scientific literature, preliminary data, or other relevant considerations…for applications proposing to study only one sex." These new guidelines were based in part on analysis of published papers that showed the predominance of preclinical studies only using one sex and also based on instances of many drugs being withdrawn after approval because of unforeseen effects on women. We will discuss these findings, as well as the new guidelines.

When assigning animals or cells to treatments, the use of randomization allows distribution of the effects of uncontrolled variables in an unbiased manner. Random does not mean haphazard; we will discuss techniques to utilize this approach for experiment design. Randomization is an important factor that can distinguish studies that lead to replicable and translatable findings.

While we are familiar with clinical studies being conducted single or double-blind, often preclinical studies are not using blinding. We will consider the impact of using blinding in collecting observations with qualitative elements. Conducting analysis blind can be another approach to guard against bias.

Most scientists know about these three design elements. Here, we will consider reports summarizing many published studies that do not report experimental designs employing these. This talk and our subsequent discussion will illustrate the importance of keeping our experimental designs as strong as possible to better enhance the reproducibility of neuroscience investigations.

References cited:

Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012 Mar 28;483(7391):531-3. doi: 10.1038/483531a. PubMed PMID: 22460880.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013 May;14(5):365-76. doi: 10.1038/nrn3475. Epub 2013 Apr 10. Erratum in: Nat Rev Neurosci. 2013 Jun;14(6):451. PubMed PMID: 23571845.

Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011 Aug 31;10(9):712. doi: 10.1038/nrd3439-c1. PubMed PMID: 21892149.

Improving Experimental Design to Boost Reproducibility

Case Study 1 – A discussion about sex

Sarah was pleased to see that Dr. Black greeted her with a wide, warm smile.

"Your performance last week was superb, Sarah! That was quite an impressive qualifying exam!"

She ducked her head and smiled shyly. "Thank you, Dr. Black."

"And now," he continued, rubbing his hands together, "you can get to work full-time on your experiments. I hope we can hammer out all the final details today, and then we can order the animals to arrive before the holiday."

NOTES

Sarah settled into a chair, and immediately brought up her concern.

"Dr. Black, I have been wondering why you suggested we do this experiment with only male mice. You seemed to think it was a good idea to use both males and females in my grant proposal for the exam. Why is this different?"

"Well, it is mainly the money, Sarah. If we use both males and females then we would have to use twice the number of mice."

"But don't we need to know if this drug will help women as well as men?"

"Oh yes, certainly. But we are a long way from that point Sarah. We need to first see if there is any effect at all. And your proposal to test a full range of five doses expands our work a little, but this seems like a very good idea, in light of the work coming out of the Williams lab."

"Why choose male mice in that case? Could we instead do this first study in female mice? You brought up the estrus cycle in my exam, and Gomez just published a study showing there was no difference in variability of response as measured by infarct volume in females vs. males."

Dr. Black rubbed his chin as he considered her suggestion. "Now when you say that, it makes me think that the effect might be different in females; they could show a different size effect, or a different dose-response relationship. I suppose that is something we really are going to need to know." He sighed. "But I still don't know how we would pay for that. The grant was cut 18 percent, you know."

"Well, we calculated that each group needs to have 24 mice. Could we make that 12 females and 12 males? It would not be powered to let us analyze by sex, but if the results were very different we could see that."

"I don't think that is a good idea," he advised. "It would be better to have the sample size adequate to allow us to analyze by sex if we are going to include both sexes."

Silence fell and they both sighed.

Sarah volunteered "How about if we hold off for another week? I would like to see if I can work out a way for us to include gender in this design."

"Well, that would be OK with me" agreed Dr. Black. "And I expect you probably would like to be at home for the holiday. We can order the mice when you get back."

She nodded.

"How is your mother doing?" he asked kindly.

"She is doing well," Sarah replied, tears stinging in her eyes. "Thank you Dr. Black. I am glad her boss knew that it was a stroke and called 911 right away."

Discussion points

- What options would you suggest for Sarah to consider to allow her experiment to include both sexes?
- Does Sarah have a good justification for including both sexes?
- Is it ethical to use twice as many animals in order to see if the effect might vary with sex? Is it ethical to constrain the experimental design because of budgetary concerns?

Improving Experimental Design to Boost Reproducibility

Case Study 2 – Another experiment for Meagan

"Breathe in through your nose."

Meagan's mind quieted. She felt her abdomen expand with the air.

"Breathe out through your mouth. Make some noise. I want to hear you!"

The release of air carried with it some of her tension. She felt her shoulders relax a bit more.

"One more time and then we will move into *savasana*, corpse pose."

With a last steady breath, Meagan let her body lay still, supported by the floor, palms up, legs limp. This yoga class was just what she had needed. Her muscles felt much more relaxed. Her mind, on the other hand, was still not responding to the orders to be quiet. She kept thinking back to the conversation with her PI Gordon.

The weekly check-in with Gordon had been going pretty well she thought. They had discussed the

NOTES

problems she was having with the new antibody and agreed on a plan to troubleshoot the odd staining pattern in her recent experiments. Gordon had years of experience in immunocytochemistry. Meagan loved how involved he would get with that aspect of her experiments, even doing some of the bench work when she was learning the process in her first few weeks.

Soon, some of her data was going to be included in the manuscript Gordon was submitting to the British Journal of Pharmacology. Even though she was third author, she was still pretty psyched – this would be her first publication! It was going to be so great to show that paper to her parents. But the pit in her stomach returned when she remembered the rest of today's conversation. The journal submission checklist was on his screen, and he was waiting to complete it as she turned to the recent section of her lab book. His next question startled her: "Did you randomly assign the mice to groups in this experiment?"

"Random? No, the groups were knock-out and wildtype. I used littermate controls."

Gordon frowned. "Yes, I know that. I am indicating that you used matching to control for litters. But did you randomly select these mice from our colony?"

"I guess you could say that. Yes. They were the only male mice in the colony that were in the 2-3 month age window we wanted for this experiment. I used them all because Jo would be away for the month for her wedding; that's where our "n" of 24 came from." Megan swallowed nervously, knowing that Gordon was not pleased about the other grad student taking so much time off. She hoped that her own extra work on the weekends was scoring some points with him, even though her mother would have given her more points for going out on a date once in a while.

Gordon's jaw muscles twitched as he turned, and his stern voice startled her. "Meagan, did you at least randomize the mice when you assigned them to drug or vehicle treatment?"

"No sir," she replied slowly. "I didn't think of that."

He gave a deep sigh and pushed his chair away from the computer screen. "Well, that is a shame."

Meagan gulped and looked down at the page in her lab book showing mice 1-12 assigned to drug and 13-24 assigned to vehicle. She looked up and gazed out the dusty office window at the city street below. "How much difference does this make?" she asked.

"Well, these new reporting guidelines mean that we have to explain that in the manuscript, and that will be seen as a weakness," Gordon said. "But I guess that is just where we are. What do you think about running the experiment again? This time I can help you design it to include randomization."

"I can do that Gordon. We have more animals in the colony now. I don't mind working long hours to get this run quickly." Her voice dropped. "I am really sorry that I forgot to randomly assign the mice to treatments."

"Don't worry," Gordon assured her. "It will strengthen our paper if we can show that we can replicate the findings in a separate group of mice."

Discussion points

- What would have been the best way for Meagan to use randomization in her experiment? How important is this in her experimental design?

- In what other ways might she improve her experiment when she runs it again?

Improving Experimental Design to Boost Reproducibility

Case Study 3 – Frank works on his figure

"Only 10% of you will likely find secure positions as tenure-track professors."

Frank shifted uncomfortably in his seat. This was the most depressing lunchbag talk yet. Who needs to hear this downer talk when graduate school is hard enough?

"But there are many routes to success in science. We just need to redefine success."

"Yeah, as a bunch of failures working in some dumb industry job," Frank thought. "That's where most of you are likely to end up, and it is not my idea of success." He slouched further in his chair and considered the plan for the rest of the day. The prep this morning had gone well. He bet he would be recording decent cells until 8 or 9p.m. That should bring his sample up to n=20. His analysis last night showed that they were getting closer to statistical significance. Adding the final seven drug-exposed cells should do the trick for this experiment. That would mean he didn't have to record more this

NOTES

week, and he could finalize the analysis and prepare graphs for the paper. The controls he recorded last week were lovely, classic whole cell recordings with several traces that he had already selected for the figure. Now he needed a similarly outstanding example of the drug effect for the second part of that figure. He hoped he would get that this afternoon; the recordings earlier in the week had not looked so good, but the preps were just adequate. The prep this morning was super, so his expectations were high.

He was a really good electrophysiologist. He looked around the room. Honestly, he was the best graduate student in this program. His adviser, Hannah, said he had good hands. She had high expectations for his work, and he enjoyed being able to deliver on that. He knew that the preliminary data in her grant application was meager, and the effect of the drug at that time depended on dropping cells recorded by the rotation student, Bill. He imagined her smile when he showed her the completed figure for publication at the end of the week. A significant effect with n=20 will put the crowning touch on her submission to *Nature*. Make that *their* submission to *Nature*.

…..

Washing up at 8:30p.m., Frank happily thought of the recordings from the afternoon. He had held several cells an extra-long time, reapplying drug until (hurray!) he achieved those traces he needed for the figure. He had at least two super examples of the effect Hannah had described in her prior work. He knew that she already had a high opinion of him, but this was going to really nail it. He smiled to himself and headed out to his car. Friends were meeting at Rafters for wings and beers. For once, he was in a great mood to hang out with fellow graduate students.

The smell of stale beer and the noise of multiple screens feeding sports blather to the jovial crowd hit him as he walked into Rafters. He saw his friends in a booth by the wall and headed over. Sliding into the booth next to Meagan, he greeted Ichiro and Carlos. Meagan, Ichiro, and Carlos resumed their conversation after greeting him.

"There is no way I could do my experiments blind" said Ichiro. "How in the world could I do that? It's a bit hard to overlook that my knock-in mouse is obese. "

They all laughed as skinny Ichiro puffed out his cheeks.

"Good point," said Carlos. "There really isn't much way around that when you have to score social interactions."

"But Gordon is insisting everyone in the lab incorporate blinding in their experiments," said Meagan. "And it should work for my mice because they aren't obviously different. I was thinking that I might be able to get someone to give temporary labels to my mice each day, just for testing. I bet Josie would help with this. I just need five minutes of her time each morning when I will do testing."

"Geez, I hope Hannah doesn't get the same bee in her bonnet, right Frank?" Carlos smiled. "I can just see us trying to run back and forth to each other's rigs to select drug or vehicle and using a random numbers table to select which one to give next."

"Oh, yeah," said Frank, thoughtfully. He wondered if his day's experiments might have turned out differently if he hadn't known what he was applying.

Discussion points

- Describe the best procedure for Frank to use blinding in his experiment.
- Could Ichiro use blinding in his experiment?
- Suggest revisions to Frank's experimental protocol to incorporate blinding, control for order effects, and randomization.

Data Analysis and Reporting

Ronald Landis, PhD

Nambury S. Raju Professor of Psychology,
Illinois Institute of Technology

NOTES

Recent years have seen increased attention focused on the quality of our scientific literature. In particular, there have been numerous concerns raised about whether our publication practices encourage (either explicitly or implicitly) less than accurate reporting of results. The emphasis of this presentation will be on evaluating the scenarios in which inaccuracies work into our research and on developing sensitivity to these situations with the goal of reducing such mistakes.

We can think about inaccuracies in reporting of results as being associated with either overt or covert issues. Specifically, overt issues are those that are self-evident to a reader of the work, while covert issues are invisible to the reader. In addition, we can think about issues that occur prior to the study being conducted, issues that arise during the data analysis, and issues that arise following data analysis or when writing or otherwise presenting the results. We can put these together to create a 2-by-3 matrix (with examples) as follows:

	Overt	Covert
Pre-Study	Power Analysis	Power Analysis
Analysis	Incorrect degrees of freedom	Missing Data, Violation of Assumptions, Outliers, etc.
Presentation	p-hacking Figures, Tables don't match text	HARKing, Misrepresentation of results

The session will begin with a discussion of the typical hypothesis testing approach, with particular emphasis on concepts of Type 1 and Type 2 Error, as well as statistical power.

I will then move to a discussion of how hypothesis testing rewards/encourages some poor behaviors, using the preceding 2-by-3 matrix as the overarching structure for discussing some specific practices.

The purpose of the presentation is to ensure that all participants have a strong grasp of the underlying foundations of null hypothesis testing (including key terms and concepts) and have an appreciation for the types of errors that commonly occur when conducting and reporting statistical analyses. The ultimate goal of the session is to help participants be fully aware of the ethical issues related to conducting and reporting statistical tests.

Data Analysis and Reporting

Case Study 1

John is collaborating on a particular project with a colleague, Marcia. John approached Marcia with the project because of Marcia's statistical expertise. After data collection, John asked Marcia to run the appropriate analyses necessary to test the study hypotheses. One of these tests involved an application of structural equation modeling. Marcia ran all of the necessary tests and sent a Results section back to John for incorporation into a manuscript ultimately submitted for publication. Because John did not feel competent with the more advanced analyses (i.e., the SEM output), he simply took the material Marcia sent him and included it in the submitted paper. Unfortunately, when Marcia was running the analyses, she incorrectly specified the model to be tested. Thus, the reported results were not possible given the hypothesized model (i.e., the degrees of freedom weren't correct). The paper was ultimately accepted and the editor and reviewers did not catch the error.

Discussion points

- What steps should John and/or Marcia have taken to prevent this error from occurring?
- What are some likely consequences of this error for John and Marcia?
- What are some likely consequences of this error for the journal and editorial team who published the paper?
- What are some likely consequences of this error for the field?
- How does your thinking about this case change if Marcia deliberately reported results from a different model because they were more favorable?

Data Analysis and Reporting

Case Study 2

Paula is leading a research team. The team has recently finished a very large data collection effort involving a large number of cases and variables. Paula suggested that the team make every effort to evaluate as many relationships between variables as possible even though there were no specific a priori

NOTES

hypotheses. As a result, the team computed more than 100 correlations. Of those computed, there were four correlations that were statistically significant at the $p < .05$ level. Paula told the research team that she thought they should work on writing the front end of their paper to emphasize the relationships that were statistically significant.

Discussion points

- If you were a member of Paula's research team, what reactions would you have to her suggestion?
- What are some of the more critical ethical issues associated with Paula's suggestion for her, her team, and the field?
- Would your reactions to this case differ if 40 of the correlations had been statistically significant? Why or why not?