TEACHING SCIENTIFIC INTEGRITY THROUGH STATISTICS

Marijtje A.J. van Duijn[1] and Wendy J. Post[2*]
[1]Department of Sociology and [2]Department of Special Needs Education and Youth Care
University of Groningen, the Netherlands
m.a.j.van.duijn@rug.nl

In the past years, Dutch academia was confronted with several cases of fraud. The Stapel investigation revealed that the prevailing research culture allowed questionable research practices (QRP). As a consequence, there is an ongoing debate on how to prevent academic misconduct. Teaching scientific integrity is an evident solution, although its implementation may be less obvious. In our workshops and classes we have used the principles of statistical reasoning and methodology, especially validity, to help students understand the importance of scientific integrity and the dangers and consequences of QRP. We feel that this approach is more effective than merely discussing principles of scientific integrity, such as verifiability and independence. The explanation may be that students are sufficiently aware of the ethical norms, but fail to see how they apply to or might challenge their own behavior. We will present an outline of our lectures.

INTRODUCTION

In the past years, the Dutch academic world was confronted with cases of fraud in several disciplines: social psychology (Stapel), consumer psychology (Smeesters), and medicine (Poldermans). The Stapel Investigation, conducted by three universities examining the extent and nature of the fraudulent research practices by Stapel, also revealed that the prevailing research culture allowed questionable research practice, or "sloppy science". As a consequence, there is an ongoing debate in Dutch academia as to prevent academic misconduct. For example, the Association of Universities in the Netherlands (VSNU) recently updated the 2004 guidelines on good scientific teaching and research, based on five principles for scientific integrity: scrupulousness, reliability, verifiability, impartiality and independence (VSNU, 2012). The Royal Netherlands Academy of Arts and Sciences (KNAW) published three advisory reports on scientific integrity in Dutch, one of which has been translated to English (KNAW, 2013).

Educating students about the concepts of scientific integrity is an evident first step in increasing knowledge and awareness of the consequences of academic misconduct. How to implement the teaching of scientific integrity in courses is less obvious. In this paper we propose to use basic principles and concepts taught in statistics and methodology courses in scientific integrity education. We will discuss our experiences as teachers of statistics and scientific integrity in the Faculty of Behavioral and Social Sciences of the University of Groningen, one of the Dutch universities involved in the Stapel Investigation. Our interest in the topic originates from being involved in the statistical research of this fraud case, assisting the Levelt Committee.

At the graduate level, our students are educated to apply and understand advanced statistical modeling and methodological issues like reliability and validity. Knowledge of statistical modeling, statistical reasoning and methodology should guide researchers(-to-be) in their research practice. From the findings of the Stapel Investigation, however, it became clear that the do's and don'ts in research practice were not so well-known or respected. Apparently, education in statistics and methodology or philosophy of science by itself does not guarantee good research practice.

Therefore, we are convinced that an extra step is needed: the concepts taught in statistics and methodology should be integrated in lectures on scientific integrity. Our hypothesis is that the use of the principles of statistical reasoning and validity, especially validity, in teaching scientific integrity will be more effective in preventing academic misconduct than teaching scientific integrity solely in terms of ethical norms such as scrupulousness, reliability, verifiability, impartiality and independence. An inspiring and good conceptual starting point is Maxwell and Delaney (2004, Ch. 1).

Since the Stapel Investigation we have taught workshops and short courses on scientific integrity for students at undergraduate and graduate levels as well as for researchers, in different disciplines (social psychology, sociology, special needs education, and health sciences). We used the principles of statistical reasoning as starting point and the Stapel Investigation as illustration. In

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

this paper we present and evaluate the setup and content of these courses/workshops for the different disciplines and student populations, based on our experiences and student feedback.

SETUP AND CONTENT OF THE COURSE

The Stapel Investigation

We start with the case of Diederik Stapel, a social psychologist, who "was found to have committed a serious infringement of scientific integrity by using fictitious data in his publications, while presenting the data as the output of empirical research." (Levelt Committee, 2012, p. 7). In Box 1 we give a short summary of the main findings of this investigation with respect to fraud found in publications, in order to introduce the issue of scientific integrity.

Box 1. Summary of main findings of the Stapel Investigation

*Three different types of fraud were found in the entire scientific output of Stapel, fabrication, falsification or unjustified replenishment of **data,** whole or partial fabrication of **analysis results,** and misleading presentation of crucial points as far as the **organization or nature of the experiment** are concerned. As defined in the report, the term 'data' refers to the coded raw scores as they occur in the data matrix in which the scores, fictitious or otherwise, are recorded for each research subject for all variables.*
Fifty-five publications were determined as "fraudulent" based on investigation of the data themselves, the published paper and other research material, made available by co-authors. For ten publications there was evidence of fraud. In these latter publications, no data or other research material were available, only the published paper.

Questionable Research Practices and Validity

We then discuss the secondary although possibly more important finding of the Stapel investigation: the prevailing research culture around Diederik Stapel allowed questionable research practices. We state explicitly that fraud presumes purpose, and we acknowledge that mistakes are easily made. A distinction between fraud and 'unintentional mistakes' is often hard to make and may depend on the scientific context.

In our view, mistakes might be caused by ignorance about principles of statistical modeling, reasoning and methodology and/or lack of knowledge on how to apply these properly, which may be due to working in an environment characterized by reduced scientific standards. Either way, teaching statistical and methodological concepts may help people to use the principles of statistical reasoning as framework for their research practice. In this part of the course, we link the consequences of sloppy science to validity issues. We explain the four types of validity formulated by Cook and Campbell (1979), statistical conclusion validity, internal validity, construct validity and external validity, further presented in Box 2.

Box 2. The four types of validity (Cook and Campbell, 1979)

1. **Statistical conclusion validity** deals with the question 'Is the statistical inference correct?' Issues of interest are the interpretation of test results (test statistics, *p*-values, power), and parameter estimation (confidence intervals, effect sizes) in relation to the research question or model.
2. **Internal validity** addresses the question 'Could the results suffer from bias?' This type of validity is also known as the 'third' variable problem, and might be caused by missing data, selection of respondents, regression to the mean and design issues.
3. **Construct validity** deals with the issue of measurement of (latent) constructs using a set of variables (or questionnaire items) and has a direct link to measurement theory and psychometrics.
4. **External validity** considers to which extent it is possible to generalize findings of a certain study across populations, settings and time. This question requires critical assessment of the study design and (implied) sample properties such as homogeneous versus heterogeneous, random versus convenience samples etc.

Remarkable Findings in the Stapel Investigation and Their Consequences for Validity

Next, we make a connection between the so-called remarkable findings reported in the Stapel Investigation listed in Box 3 and the threats they pose to specific types of validity. These findings are discussed in the classroom by the instructor or in smaller groups. For undergraduate students the purpose is to get acquainted with the different types of validity, and to become aware of the imperative of sound statistical analysis and methodology for scientific integrity. For graduate students and researchers, the discussion serves as an introduction to the next part of the course: asking them about experiences in making and preventing mistakes in research practice.

Box 3. Remarkable findings in the Stapel Investigation

Remarkable findings in data
- Respondents intractable (no unique code) or more/less than reported
- Data inconsistent with questionnaire and/or reported data collection
- No missing data
- Inconsistent data organization (variable names, variable labels)

Remarkable findings in scale construction
- Inconsistent operations on variables (e.g. recoding)
- Reliability of constructed scale too low (and lower than reported in the paper)
- Scale construction deviates from experimental design (discarding items without justification or unreported adjustment of existing or validated scale)

Remarkable findings in statistical analysis
- Analysis results cannot be replicated (no syntax available, inadequate analysis)
- Analysis does not match study design (discarding experimental conditions or respondents and outliers)

Remarkable findings in reported results
- Incomplete description of experiment
- Incorrect or incomplete descriptive statistics
- Statistical analysis: incorrect degrees of freedom, p-values (rounding or inconsistent with reported values of test statistics)

Discussion in Small Groups

This part of the workshop is designed for junior (PhD-level) up to senior researchers. We introduce three topics with (sub)questions in research practice related to data collection and management, linking research hypothesis, study design, and statistical analysis, and reporting the research (see Box 4).

We ask the participants to discuss their experiences in making and preventing mistakes in all stages of the research process, and their opinion on preventing such ‘problems’ and what they would advise to colleagues, research group, department, graduate school, and university. In groups containing a mixture of no more than six senior and junior researchers - if possible joined by a statistician or methodologist - (a selection of) the questions are discussed and reported back to the wider audience, followed by a general discussion.

CONCLUSION

A first evaluation suggests that our approach works. In view of the recent fraud cases, academic misconduct is a topic of great interest to students and colleagues, who enthusiastically participated in the course and workshop. The classroom and small group discussions are lively and never too long. The interaction between researchers, whether junior, senior, from different departments or research areas, is affable.

We noticed a difference in our audiences in level of sophistication and openness. Undergraduate students, new to the research process, are naïve in their expectations and reactions. Junior researchers are idealistic and are genuinely shocked by the remarkable findings presented.

Senior researchers react less strongly, through their experience with the research process and – adjusted – expectations with respect to validity. It is clear that they know the ‘right answers’ to all the questions posed in the workshop or have found ways to deal with them. Statistical reasoning does not seem the guiding principle. They have developed certain routines and rather use

Box 4. Small group discussion topics and (sub)questions

Topic 1: Data collection and management, measurement scale construction
How to decide on the exact data to collect?
Which sample size? Selection of respondents (criteria for inclusion)? Variables/concepts limited to the research question?
How to store all research material and original (raw) data?
Respondent identification? "Cleaning" data, saving (cleaned) data in new file? Keeping track of missing data, deciding on or defining outliers (criteria for exclusion)?
How to decide on the exact variables to measure?
Use existing (validated) or new scales? What to do when reliability is (too) low? When and how to use and report scale construction?
Topic 2: Relation between research hypothesis, study design, statistical hypothesis, analysis
How to decide on the exact data to collect?
Pilot study? Collecting more data after first results? Focused on the statistical analysis to be performed?
How to decide on statistical analysis?
Strictly keep to prespecified analysis plan? Further analyses? Danger of verification bias?
How to perform the analysis?
Simple or complex? When (and whom) to ask for help?
Topic 3: **Reporting the research**
What to report exactly?
Non-significant results (or rounding p-values)? Report "complications"? Outliers, adjusted analysis, 'new' (unexpected) results? (Using the useful concepts by Wilkinson et al., 1999.)
How to deal with limited space for (detailed) information in papers and/or PhD theses?
What is sufficient information on data and analysis for replicability of results (experiment)?
How to deal with the publication part, i.e. journals, editors, reviewers?
What is wise in handling criticism, does the possibility of supplementary materials help?

accepted statistical procedures (or tricks). Junior researchers are still struggling to find the right answers and procedures and are open to advice.

It is our impression that the openness of (especially senior) researchers with respect to ethical concerns or their experiences with sloppy science depends on the research context. The level of independence of researchers seems an important explanatory variable in how they view their own academic conduct as an individual responsibility or as part of group behavior, governed by a specific research tradition to which they have to conform.

To conclude, we feel that we have found a way to make students and researchers more aware of the usefulness of statistical reasoning and issues of validity to make sound decisions in performing and reporting research, and at the same time to instigate an open and safe discussion about the important topic of good and questionable research practices.

[*] both authors contributed equally to this work

REFERENCES

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin Company.

KNAW (2013). *Responsible research data management and the prevention of scientific misconduct.* Amsterdam: KNAW.

Levelt Committee (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Available at http://www.commissielevelt.nl

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

VSNU (2012). *The Netherlands code of conduct for scientific practice. Principles of good scientific teaching and research (2nd edition).* The Hague: VSNU.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.