Received: 1 February 2021 | Revised: 4 June 2021 | Accepted: 7 July 2021

DOI: 10.1111/test.12287

ORIGINAL ARTICLE

WILEY

Discovering experimental design: An interactive teaching exercise using Fisher's tea-tasting experiment

Thomas R. Fanshawe

Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

Correspondence
Thomas R. Fanshawe, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK.
Email: thomas.fanshawe@phc.ox.ac.uk

Funding information
University of Oxford

Abstract

An appreciation of experimental design is an important aspect of introductory statistics teaching in a wide range of applied disciplines, including medical statistics. Understanding the impact of design decisions on the choice of analysis method and subsequent interpretation of results can help to embed the importance of statistical thinking in the experimental process. I discuss an interactive exercise, based on R.A. Fisher's celebrated "Lady Tasting Tea" experiment, that is intended to raise awareness of design issues as part of an undergraduate statistics module. The exercise used a discovery approach, with students encouraged to identify design issues and agree on solutions themselves via small group discussion, with only low-level prompting from the instructor. The value of this teaching style and possible extensions of the tea-tasting experiment to other related topics suitable for more widespread use are also discussed.

KEYWORDS

teaching, experimental design, group teaching, study design, teaching statistics

1 | INTRODUCTION

It has been said that "choosing an appropriate study design is one of the most challenging aspects of problems in applied statistics" [15]. Consequently, an appreciation of experimental design is seen as a valuable component of introductory statistics education [12], especially when in applied disciplines such as medicine, where poor design may result in studies that are scientifically inefficient or unethical [2]. Elsewhere, it has been argued that learning how to design experiments is a more useful objective than learning how to analyze data from experiments carried out by others [10].

Nevertheless, examples that successfully incorporate the experimental design topics at introductory level are difficult to find. There remains a temptation for "nuts and bolts" analysis to be prioritized ahead of design issues [19], even when the main objective is conceptual understanding rather than analytical proficiency [8]. As Bland, Altman, and Royston point out, "the function of medical statistics is not merely technical" [6]. As such, it is beneficial to show the role of statistics as an integral part of the research process, rather than as a series of numerical calculations or mathematical formulae [29].

Previous work highlights using accessible examples [17], such as those that relate to controversies in science [4]. The choice must be carefully tailored to the requirements of the course and the abilities of the students taking it. Other strategies for teaching experimental design focus on virtual experimentation or simulation, which may be better suited for applications such as engineering and the physical sciences [1,11]. Another alternative, intended to promote active learning, requires students to first design and then perform an experiment

in class before analyzing the resulting data. This approach, which may use simple or humorous scenarios to increase engagement, aims to increase students' conceptual understanding of the experimental method, although it requires more time [20,24].

This paper describes a short exercise to aid understanding of experimental design. It was delivered as part of a medical statistics module for undergraduate medical students but requires little prior knowledge and would be suitable for other audiences at a similar or lower academic level. The exercise is based around R.A. Fisher's celebrated tea-tasting experiment. After a short explanation of the historical origin of this experiment, I explain how this was developed into a teaching exercise and present the results of delivering this in class. The Discussion describes possible modifications and outlines the benefits of this type of exercise in introductory teaching.

2 | HISTORICAL BACKGROUND: FISHER'S TEA-TASTING EXPERIMENT

Since the original account [13], Fisher's tea-tasting experiment has been revisited on many occasions [3,5,7,9,16,21,22,30], and appears in the title of a history of 20th-century statistical science [25]. An accessible summary is provided by Senn [27]. A brief summary of the principal points follows.

Fisher's experiment aimed to test whether "a lady" (later named as Muriel Bristol), when drinking a cup of tea, "can discriminate whether the milk or the tea infusion was first added to the cup" [13,p. 13]. Fisher's solution is that, of eight cups of tea, four should be prepared "milk first" (MF) and four "tea first" (TF), with the order prepared at random. Importantly, he also stipulates that the subject is told the number of each type in advance and assumes that her guesses will be consistent with this, that is, that she will assign exactly four MF and the other four TF. This has been named the "double-tetrad" [16] or "octad" design [5].

Fisher specifies a null hypothesis that "the subject possesses no sensory discrimination" [13,p. 19] and calculates that, if true, correctly classifying all eight cups would occur with probability $1/C_4^8 = 1/70$ (see Data S1 for explanation). Correctly classifying three of the four MF, and hence also three of the four TF, would occur with probability $\left(C_3^4 \times C_1^4\right)/C_4^8 = 16/70$, which he states "could not be judged as statistically significant evidence of a real sensory discrimination" [13,p. 17]. Fisher dislikes allocating all eight cups at random with equal probability, as "it would occasionally occur that all the cups would be treated alike, and this, besides bewildering the subject by an unexpected occurrence, would deny her the real advantage of judging by comparison" [13,p. 27]. The performance of "the lady" in this experiment is unclear in Fisher's account, although others have stated that she could differentiate between the methods of preparation [7,27].

The discussion provided by Neyman [22] is more detailed than Fisher's original, although he primarily considers a design in which the cups are presented as four pairs, with each pair known to contain one MF and one TF. Neyman's treatment centers on specifying admissible hypotheses and determining the power function of the probability of any single classification being correct. This has been extended to multiple tasters (a "hypothetical infinite population") [30].

3 | TEACHING EXERCISE

3.1 | Context

The exercise was delivered during an introductory module in medical statistics for first- and second-year undergraduates studying for the preclinical Bachelor of Medicine degree at the University of Oxford. This module comprises 16 contact hours over 2 years, covering topics such as confidence intervals, estimation, hypothesis testing, correlation and the normal and binomial distributions (first year) and the *t*-test, Poisson distribution, linear regression, survival data and study design preparatory to critical appraisal (second year). At the time described, students were taught face-to-face in groups of 25 to 30 (around one-sixth of the annual cohort).

The objective of the module is for students to gain an understanding of statistics applied to biomedical science. The emphasis is on interpretation and application of methods rather than mathematical details. The exercise described here was carried out toward the start of the students' second year and did not form part of their summative assessment. Most other teaching sessions focus on analysis issues. Results are reported for two cohorts of students (2018 and 2019).

Ethical approval (CUREC ref. R58986/RE001) was obtained to record students' anonymous responses and use them for publication. Students were given a free choice in class to opt in or out. All students participated in the exercise, but only results from those who opted in are reported.

3.2 | Description of the exercise

The instructor asked the students to form groups of size 4 to 6 and to nominate a spokesperson to take the role of the "principal investigator." A verbal description of the objective was given:

> Muriel Bristol claims to be able to tell the difference between cups of tea made with the milk added to the cup first, and cups of tea made with the tea added to the cup first. Ronald Fisher is sceptical. The challenge is to design an experiment to test the claim.

Further specific details were kept to a minimum, although the sample size was fixed:

> You may design the experiment in any way that you think is appropriate. The only constraint is that for practical reasons a maximum of eight tastings is permitted.

Students were told that they could use any equipment they needed, such as standard tea-making equipment, cups and mugs, milk and sugar (although they were also made aware that they would not actually be performing the experiment). Groups were given 20 minutes to discuss the task and reach a consensus about the design to be used and how the results could be interpreted. The spokesperson arbitrated in case of disagreements and had responsibility for final decisions.

Groups were asked to discuss, agree and write down responses to these questions (see also Data S1):

1. How would you design the experiment? (give as much detail as possible)
2. What information would you give to the taster?
3. What results would you need to see in order to accept the taster's claim?

During these discussions, two teaching staff circulated to answer queries and prompt groups about reasons for their decisions. As this session was delivered on several occasions to different student classes, several teachers were involved in its delivery at different times. Previously, the teaching team had taken part in a practice run of the exercise (initially, while still being unaware of the possible solutions available), which clarified the questions to ask students and helped to agree how much help should be provided to them. As a result, teachers tried to avoid steering students toward any particular choice of responses during the session. Students did not have internet access during the task.

The instructor then asked each group to explain their decisions to the class, before outlining the design choices that Fisher himself preferred and his reasoning. The exercise took 30 to 35 minutes in total. As time for additional discussion during the session was limited, a written description of the various design choices selected by students was made available after the session. Anonymous questionnaire feedback was obtained at the end of the module.

4 | RESULTS

4.1 | Students' response to the task

Most groups quickly engaged with the task and required little help to complete it, although some needed a reminder to reach a consensus and record their decisions within the time allowed. Many discussions were lively, especially when there was disagreement within the group. Altogether, 61 groups gave consent for their results to be included (30 in 2018 and 31 in 2019).

4.1.1 | Responses to question (1)

Many responses showed awareness of the need to keep experimental factors constant between different tastings. These included:

- Brewing time (often specified numerically)
- Volumes and type of tea and milk (often specified numerically)
- Amount of sugar (generally, none)
- Temperature (eg, controlled using a waterbath)
- The mugs or cups in which the tea was served
- Consistency of stirring (eg, "10 stirs clockwise")
- Number, volume, and standardization of permitted sips (eg, "she must sip each cup once and spit it out")
- Time between tastings

Some groups strove to reduce the effect of extraneous factors, occasionally stretching credibility. Suggestions included providing water to "wash out" or "palate cleanse" between tastings; presenting the tea on different days or "each morning following a standardized breakfast" (cf. [22, p. 272]); or imposing sensory deprivation via a blindfold, nose clip, or noise-cancelling headphones. Groups wanting to use more than one taster (cf. [30]) were reminded that the claim related to just one individual.

The other design consideration related to how to assign tastings to either MF or TF. Most groups agreed that this should be done randomly, and many groups recommended double-blinding (the experimenter and the taster), but exact choices varied. The most common choice was eight 50:50 random assignments of MF and TF, but around one-third of groups selected Fisher's design. Almost all groups considered multiple options before making their final choice. Various different designs were less frequently suggested (Table 1; I have appended names to each). The four-pair design is the

TABLE 1 Summary of designs suggested

Name of design	Description	Frequency
Completely random	Eight independent 50:50 random assignments of MF or TF	28
Double-tetrad	Four MF, four TF, random order	20
Pentad-triad	Five MF, three TF (or vice versa), random order	3
Subsampling	Random selection of eight cups from a larger preprepared group (20 cups with 10 each MF and TF, or 50 cups with 25 each MF and TF, or 100 cups with 50 each MF and TF)	3
Four pair	Four pairs, purposively designated as MF/MF, TF/TF, MF/TF, and TF/MF *or* "We would do a paired experiment...with one in each pair made in the two ways."	3
Deception	"Tell subject that she will be tasting eight cups of tea, each made one of those two ways. All cups have tea put in first." *or* "Give the patient eight identical cups of tea: all brewed either milk first or tea first."	2
Restricted random	Randomly assigned with a restriction of at least two MF and two TF	1
Random + control	One control sample of tea without milk, otherwise random MF or TF	1

same as that discussed by Neyman [22]. The subsampling design is almost equivalent to the completely random design. If the taster is aware of the design chosen, double-tetrad and pentad-triad are special cases within the "M + N" family [5].

4.1.2 | Responses to question (2)

Some groups did not initially see the reason why any information might need to be provided to the taster. When prompted that, at the minimum, enough information would be needed to ensure she could give informed consent to participate, most saw how this related to their design decision. Virtually all groups decided that the taster should not be told how many were MF and how many TF, preferring to say that either they had been selected at random (sometimes, even when this design had not been chosen) or that each could be either MF or TF, without specifying the selection method. An exception was one group who chose the paired design, stating that this design should be fully explained before the experiment began.

4.1.3 | Responses to question (3)

Many groups made the connection between the design they chose and a simple method of analysis using either the Binomial distribution or combinatorial reasoning. More details about the calculations corresponding to each design are included in Data S1.

For example, some groups argued that under the completely random design, the number X correctly identified by chance would follow Binomial ($n = 8$, $P = 0.5$). As $P(X \geq 7) = 0.035$, at least seven correct guesses might provide sufficient evidence. Some suggested that they would be convinced by six correct guesses, with or without calculating the associated probability. Groups that chose the double-tetrad design did not generally follow through Fisher's calculation, perhaps, because some groups that chose this design did not think it necessary to disclose this choice.

4.2 | Feedback and evaluation

In feedback received the first year the exercise was used, students rated the exercise as moderately enjoyable but did not appreciate its usefulness (average scores of 3.3 and 2.3 respectively, on a scale of 1-5). In free text comments, some students indicated that the exercise was too long and did not clearly relate to the rest of the course.

Before the exercise was used the next year, the teaching staff reviewed the feedback and made some changes to the way it was conducted. The time allowed for the groups to reach decisions was limited to 20 minutes and the total time for the whole exercise, including discussion, to 35 minutes. Teachers were encouraged to be more proactive in ensuring groups reached clear decisions in the time allowed, with the time limit made clearer, as during the first run some groups took 30 to 40 minutes for this step, meaning that the whole exercise filled almost a full-hour teaching session. In the closing discussion, teachers made a clearer link to other parts of

the course, and a printed summary of the conclusions was given out after the session (Data S1). The second time the exercise was used, average scores for enjoyment and usefulness rose to 4.3 and 3.5 respectively, and free text comments were almost uniformly positive, especially in relation to improving interest in the topic.

5 | DISCUSSION

The aim of this exercise was to improve engagement with, and understanding of, study design in an undergraduate medical statistics module. It was designed to demonstrate the importance of design decisions and their impact on analysis methods. Results suggest an improvement in engagement and understanding after the introduction of the exercise, which is supported by the perceptions of the teaching staff delivering the session.

The exercise followed a discovery approach: students were given little information initially, so that they could make their own decisions from first principles. In this sense, it resembles a problem-based learning activity on a small scale [26]. From a learning perspective, the process of discussion and reaching decisions is at least as important as the decisions themselves.

It aligns with recommendations about the importance of interactive exercises to develop conceptual and statistical reasoning for teaching statistics to students in applied disciplines [8,14]. With limited time available, using an example that is easy to understand without existing knowledge was important, even though in many ways, the tea-tasting experiment might be seen as an atypical one. That the context of this particular example was not directly related to medical science did not appear to be a barrier to participation.

The diverging opinions that arose over design choices were a microcosm of discussions that arise when designing real-life experimental studies. Many student groups were able, unprompted, to appreciate the importance of several points that have provoked debate since Fisher's original account. For example, some interpreted the original wording of the taster's claim to mean that she could always distinguish between MF and TF, rather than simply that her probability of a correct classification was greater than 0.5. Consequently, two groups saw virtue in preparing all eight cups the same way (the "Deception" design), arguing that if the taster could discriminate perfectly, an unexpected permutation should not faze her.

Two practical points should be noted before implementing this exercise. Firstly, if it is delivered by different teachers of varying levels of experience, a practice run for the teaching group is useful, as this informal training can help to highlight issues that the students might themselves face when attempting the exercise. Secondly, it is helpful to agree beforehand how much the teachers should intervene, as opposed to letting students making all decisions without prompting. A useful strategy during discussions is for the teacher to ask the designated spokesperson what decisions have been reached and then to ask the rest of the group if they all agree with these decisions. If, as often occurs, there is not a complete consensus, this provides an opportunity for the group to clarify the reasons for their choices.

Some limitations of the evaluation of this exercise should be noted. As it was difficult to add a direct assessment of student understanding to the existing assessment schedule, many of the measures presented are by student self-report rather than objective measures of statistical literacy. It was not possible to conduct a randomized comparison between different student groups to assess its effectiveness. The exercise has only been used in a face-to-face teaching format. Although it may be possible to adapt it for remote teaching, for example, using online breakout discussion rooms, this has not been tested.

This exercise can address only a limited number of design issues that teachers might wish to discuss. It would probably not be suitable for topics such as optimal and factorial designs. It was also not intended to address the teaching of hypothesis testing and *P*-values or misconceptions relating to these topics [28]. However, it provided a useful way for students to gain familiarity with randomization and blinding ahead of a later class dedicated to medical study design (randomized controlled trials, cohort, and case–control studies). It also provided an opportunity to learn a key general point relating to the connection between design and analysis—that the latter is dependent on the former and therefore that both should be considered at the outset. This helps to reinforce the importance of statistical thinking throughout the research process, rather than solely at the analysis stage.

Extensions are possible to suit different audiences. For example, the exercise could be used to demonstrate the concepts of the power function [22] and sample size, as well as the hypergeometric distributions. This cohort already had awareness of the binomial distribution, so the exercise provided an opportunity to reinforce this knowledge. For students less familiar with probability or analytical issues, the third question might be removed but discussion of design issues would remain relevant.

One article describes using the experiment as motivation to apply Bayesian reasoning, without explaining how this would be best implemented [18]. For more advanced students, this might entail specifying a prior distribution on the interval [0,1] or [0.5,1] for the probability that the taster makes a correct classification and using this in conjunction with the likelihood to derive the posterior. This

could be used to illustrate the effect of the choice of the prior, if students have different levels of scepticism about the original claim and are able to express this via appropriate prior distributions.

More generally, an application-focused approach is more likely to be appreciated by students and sustain engagement with the subject [23]. For this student group, real-world applications are usually more effective than examples that are abstract or use computer simulation. My experience teaching group exercises is that examples that might hold intellectual interest to a statistician are not always well-received as teaching examples at undergraduate level. A certain amount of trial-and-error is therefore inevitable, and small adjustments were required both after a practice run and after the first time this class was delivered.

ACKNOWLEDGEMENTS

I would like to thank the teaching team and other researchers from the Nuffield Department of Primary Care Health Science, University of Oxford, who helped to test and deliver the teaching exercise, and the students who participated and provided feedback.

ORCID

Thomas R. Fanshawe https://orcid.org/0000-0002-9928-8934

REFERENCES

1. C. M. Anderson-Cook and S. Dorai-Raj, *An active learning in-class demonstration of good experimental design*, J. Stat. Educ. **9** (2001), 1.
2. D. R. Appleton, *What statistics should we teach medical undergraduates and graduates?* Stat. Med. **9** (1990), 1013–1021.
3. D. Basu, *Randomization analysis of experimental data: the Fisher randomization test*, J. Am. Stat. Assoc. **75** (1980), 575–582.
4. K. A. Bennett, *Using a discussion about scientific controversy to teach central concepts in experimental design*, Teach. Stat. **37** (2015), 71–77.
5. J. Bi and C. Kuesten, *Revisiting Fisher's 'Lady Tasting Tea' from a perspective of sensory discrimination testing*, Food Qual. Prefer. **43** (2015), 47–52.
6. J. M. Bland, D. G. Altman, and J. P. Royston, *Statisticians in medical schools*, J. R. Coll. Phys. Lond. **24** (1990), 85–86.
7. J. F. Box, *R. A. Fisher and the design of experiments, 1922-1926*, Am. Stat. **34** (1990), 1–7.
8. T. E. Bradstreet, *Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning*, Am. Stat. **50** (1996), 69–78.
9. S. J. D. Chadwick and H. A. F. Dudley, *Can malt whisky be discriminated from blended whisky? The proof. A modification of Sir Ronald Fisher's hypothetical tea tasting experiment*, BMJ **287** (1983), 1912–1913.
10. G. W. Cobb, *One possible frame for thinking about experiential learning*, Int. Stat. Rev. **75** (2007), 336–347.
11. P. L. Darius, K. M. Portier, and E. Schrevens, *Virtual experiments and their use in teaching experimental design*, Int. Stat. Rev. **75** (2007), 281–294.
12. R. G. Easterling, *Teaching experimental design*, Am. Stat. **58** (2004), 244–252.
13. R. A. Fisher, *The design of experiments*, Oliver and Boyd, Edinburgh, 1935.
14. J. Garfield and D. Ben-Zvi, *Helping students develop statistical reasoning: implementing a statistical reasoning learning environment*, Teach. Stat. **33** (2009), 72–77.
15. S. M. Gore, *Teaching experimental design: Prescribed by a medical statistician*, J. R. Stat. Soc. D **33** (1984), 243–247.
16. N. T. Gridgeman, *The lady tasting tea, and allied topics*, J. Am. Stat. Assoc. **54** (1959), 776–783.
17. S. M. Hiebert, *Teaching simple experimental design to undergraduates: Do your students understand the basics?* Adv. Physiol. Educ. **31** (2007), 82–92.
18. D. V. Lindley, *The analysis of experimental data: The appreciation of tea and wine*, Teach. Stat. **15** (1993), 22–25.
19. M. MacDougall, H. S. Cameron, and S. R. J. Maxwell, *Medical graduate views on statistical learning needs for clinical practice: A comprehensive survey*, BMC Med. Educ. **20** (2020), 1.
20. S. Montangero, F. Vittone, S. Olderbak, and O. Wilhelm, *Exploration of experimental design and statistical methods using the* stick-on-the-wall spaghetti *rule*, Teach. Stat. **40** (2018), 40–45.
21. R. Morton, *On the efficiency of Fisher's tea-tasting designs*, J. R. Stat. Soc. B **37** (1975), 49–53.
22. J. Neyman, *First course in probability and statistics*, Henry Holt and Company, New York, NY, 1950.
23. D. Nolan and T. P. Speed, *Teaching statistics theory through applications*, Am. Stat. **53** (1999), 370–375.
24. L. Pyott, *Tennis anyone? Teaching experimental design by designing and executing a tennis ball experiment*, J. Stat. Data Sci. Ed. **29** (2021), 22–26.
25. D. Salsburg, *The Lady tasting tea: How statistics revolutionized science in the twentieth century*, Henry Holt and Company, New York, NY, 2001.
26. H. G. Schmidt, J. I. Rotgans, and E. H. J. Yew, *The process of problem-based learning: what works and why*, Med. Educ. **45** (2011), 792–806.
27. S. Senn, *Tea for three: Of infusions and inferences and milk in first*, Significance **9** (2012), 30–33.
28. A. Vallecillos, *Understanding of the logic of hypothesis testing amongst university students*, J. Math.-Didakt. **21** (2000), 101–123.
29. C. J. Wild, *Embracing the 'wider view' of statistics*, Am. Stat. **48** (1994), 163–171.
30. R. F. Wrightson, *The theoretical basis of the therapeutic trial*, Acta Genet. Stat. Med. **4** (1953), 312–343.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: T. R. Fanshawe, *Discovering experimental design: An interactive teaching exercise using Fisher's tea-tasting experiment*, Teach. Stat. **43** (2021), 140–145. https://doi.org/10.1111/test.12287