STA2020 ANOVA Notes

Ané Cloete

2024-12-17

Table of contents

Preface

This book is not an exhaustive guide for designing an experiment or conducting ANOVA's, it has been tailored specifically for the learning outcomes and methods covered in STA2020. If you are interested in a more general text on experimental design, please see:

Instructions:

- Examples
- R code
- Tips
- Footnotes

Part I

Experimental Design

1 Experiments and experimental design

There are two fundamental ways to obtain information in research: by *observation* or by *experimentation*. In an observational study the observer watches and records information about the subject of interest. In an experiment, the experimenter actively manipulates variables hypothesized to affect the response (insert small example). Although both are important ways of understanding the world around us, only through experiments can we **infer causality**.

That is, by designing and conducting an experiment properly, if we observe a result such as a change in variable A leads to a change in our response (say variable B), we can conclude that A **caused** this change in B. If we were to merely study variable B and observe that as variable A changes, B also changes without conducting an experiment, then we can only say that variable A and B are associated. We could not conclude that any change in B is due to A. It could be some other factor that is correlated with A or it could be that B caused the change in A! The key is that a well-designed experiment controls and holds constant (as best we can) all other factors that might affect the response, so we can be sure the result is caused by the variable we manipulated.

Imagine a company wants to determine whether their voluntary employee training program (the explanatory variable) increases productivity (the response). They decide to track the productivity of employees who chose to complete the training and those who did not. They note that, on average, trained employees are more productive. Can we confidently conclude that the training program caused increased productivity?

This is an observational study since no variable was actively manipulated, they merely observed and recorded the productivity

of two groups of employees. So, we cannot conclude that completing the training program increases productivity - we cannot infer causality. It could be due to many other factors, either observed or unobserved, such as maybe employees who choose to do the training program are inherently more motivated and thus productive. Can you think of any other factors?

If they actively manipulate the explanatory variable, training program, by randomly assigning employees to complete the training program or not and control other factors by ensuring the employees are as similar as possible accross the groups (i.e. conducted an experiment). Any differences in productivity between the two groups could then be ascribed to the training program. If they happen to find that the employees who were assigned the training program are more productive, they can confidently say that the program caused increased productivity (and perhaps make it compulsory for all employees!).

Experimental studies are extremely important in research and in practice. They are almost the only way in which one can control all factors to such an extent as to eliminate any other possible explanation for a change in a response other than the variable actively manipulated. In this course, we only consider experimental studies and those which aim to compare the effects of a number of **treatments**.

Here are some other reasons for conducting experiments:

1. They are easy to analyse. A well designed experiment results in independent estimates of treatment effects which allow us to easily interpret the effects. EXPAND - independent treatment effects and/or independent treatment variables?

2. Experiments are frequently used to find optimal levels of variables which will maximise (or minimise) the response. Such experiments can save enormous amounts of time and money. Imagine trying to find the optimal settings for producing electricity from coal without proper experimentation. Such a trial and error process would be extremely costly, wasteful and time consuming. In a similar vein, what if the fictional company in our previous example decided to invest a bunch of money in fine-tuning their

training program based solely on the results of an observational study. In reality though, it turns out that adjusting their hiring process to identify more keen candidates would have been much more efficient and inexpensive.

3. In an experiment we can choose exactly those settings or **treatment levels** we are interested in, e.g. we can investigate the effect of different shift lengths (6, 8 or 9 hours) on employee productivity or test specific price points (R100, R150, R200) to determine which price maximizes sales or revenue. We can actively manipulate the variable(s) to the levels we are interested in.

Experimental studies and their design are fundamental to science, allowing us to further knowledge and test theories. So lets define them more rigorously. We'll start by introducing some terminology.

INSERT WHAT THEY NEED TO KNOW / MAIN TAKEAWAYS Perhaps?

2 Terminology

Treatment factors, treatment levels and treatments:

The **treatment factor** is the factor or variable that the experimenter actively manipulates to measure its effect on the response. All factors/variables that are investigated, controlled, manipulated, thought to influence the response, are called the treatment factors. They become the **explanatory variables** (mostly categorical) in the model. For each treatment factor, we actively choose a set of **levels**. For example, the treatment factor "temperature" can have levels 10, 20, and 50°C. If temperature is the only treatment factor in the experiment, the **treatments**[1] will also be 10, 20, and 50°C.

If we manipulate more than one factor (e.g., temperature and pressure), we have two treatment factors. When several treatment factors are manipulated, the experiment is called factorial and the **treatments** are all possible combinations of the factor levels. If we have pressure levels "low" and "high," there are 6 treatments in total:

In the figure above, there are two treatment factors: Temperature (on the y-axis) and Pressure (on the x-axis). The axis ticks represent the levels of each treatment factor, and the blocks within the grid represent the treatments, which are specific combinations of the levels of Temperature and Pressure. Each treatment is labeled with the corresponding combination of levels (e.g., '50, Low' or '10, High').

[1] The terminology of treatments can be traced back to 1920's when it was first applied by Ronald Fisher in the agricultural sciences. He is often refered to as the Founder of Statistics! Have a look at the very first application of ANOVA here and also a very nice article describng the history of statistics and his contribution to the field.

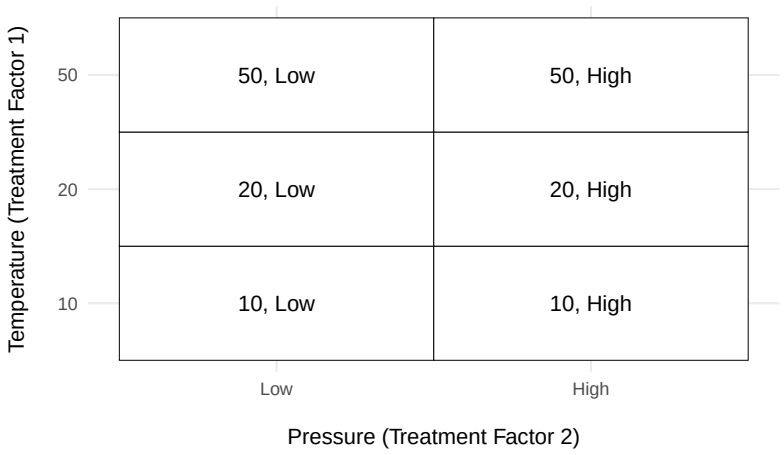

Figure 2.1: Visualization of how treatments are formed as combinations of treatment levels.

Example 1

Three groups of students, 5 in each group, were receiving therapy for severe test anxiety. Group 1 recieved 5 hours, group 2 received 10 hours and group 3 received 15 hours. At the end of therapy each subject completed an evaulation of test anxiety. Did the amount of therapy have an effect on the level of test anxiety?
The three groups of studnets received the scores on the Test Anxiety index (TAI) at the end of treatment shown in the table below.

Group 1	Group 2	Group 3
48	55	51
50	52	52
53	53	50
52	55	53
50	53	50

When faced with a text like this, it is useful to identify the treatment factors, their levels and the treatments, as well the response. Clearly, from the question, we are interested in the effect of therapy on test anxiety. A statement like this can generally be read as the effect of the treatment factor on the response. Nowhere is another treatment factor mentioned, so we only have one in this example. What are the levels of therapy we set? The levels are 5, 10 and 15 hours of therapy and since we only have one factor these are also the treatments. Let's summarise this as follows:

- **Response:** Test Anxiety
- **Treatment Factor:** Therapy
- **Treatment Levels:** 5, 10, and 15 hours of therapy
- **Treatments:** 5, 10, and 15

Experimental and observational unit

The **experimental unit** is the entity (e.g. material, object, or individual) to which a treatment is assigned or that receives the treatment. By contrast, the **observational unit** is the entity from which the response is recorded. This distinction is very important because it is the experimental units which determine how often the treatment has been replicated and therefore the precision with which we can measure the treatment effect. In the methods that we cover in this course, we require that in the end there is only one 'observation' (response value) per experimental unit. If several measurements have been taken on an experimental unit, we will combine these into one observation, typically by taking the mean. Very often, the experimental unit is also the observational unit.

example-box

For See Example (**example-box?**) for more details, what are the experiemental units? To determine this, revisit the text of Example 1 and ask yourself: what entity received the treatments or to what were treatments applied? Most of you, will probably answer the students and this is correct. Each student

received the respective treatment (number of hours in therapy) assigned to their group and so there are $5 \times 3 = 15$ experimental units.

There is an argument to be made that it is not clear whether the students received therapy on their own or that the groups of students received therapy together. In that case, treatments were applied to groups of students and so there would be three experimental units. This will usually be clear from the text, but we'll use this scenario to illustrate some concepts as we go.

We also need to know what the observational units are. The text states that at the end of therapy, each student completed an evaluation to determine their level of test anxiety. So the response, test anxiety, was measured on the student level which means students are the observational units. In the first scenario, the students are both the experimental units and observational units. But this would not be the case if groups are the experimental unit.

We also require that there is only one observation per experimental unit, the first scenario meets this requirement. For the second scenario, we have 5 observations per group and so we would have to take the mean of these values to end up wth one response value per group.

Let's add to the summary assuming students are the experimental units:

- **Experimental unit (no):** Student (15)
- **Observational unit (no):** Student (15)

Homogeneity of experimental units

When the set of experimental units are as similar as possible such that there are no distinguishable differences between them, they are said to be **homogeneous** (a fancy word for saying they are of the same kind). The more homogeneous the units are, the smaller the experimental error variance (natural variation between between observations of the same treatments) will be.

It is super important to have fairly homogeneous units because it allows us to detect differences between treatments more easily.

Blocking

If the experimental units are not fairly similar but are heterogeneous (the opposite of homogeneous), we can group them into sets of similar units. This process is called **blocking** and the groups are considred "blocks". We compare the treatments within each block as if each block is its own mini-experiment. This way we account for the differences between blocks and can better isolate the effect of the treatments.

Example 2.2 EDIT THIS STILL

Imagine you're testing the effectiveness of two marketing strategies (A and B) to increase sales at a chain of coffee shops. The coffee shops are located in different neighborhoods, where factors like income levels might influence sales. To prevent these differences from skewing the results, you group the coffee shops into "blocks" based on neighborhood income level (e.g., low, medium, high).
Within each block, you randomly assign coffee shops to either Strategy A or Strategy B. This approach allows you to compare the strategies while controlling for variability caused by differences in neighborhood income levels.
Without blocking, would you be able to confidently attribute differences in sales to the strategies alone? Likely not, as any observed differences could be due to neighborhood-specific factors rather than the strategies themselves.

Replication and pseudoreplication

If a treatment is applied independently to more than one experimental unit it is said to be **replicated**. Treatments must be

replicated! Making more than one observation on the same experimental unit is not replication, but *pseudoreplication*. Pseudoreplication is a common fallacy (REF?). The problem is that without true replication, we don't have an estimate of uncertainty, of how repeatable, or how variable the result is if the same treatment were to be applied repeatedly.

In Example 1, if experimental units were the groups and we didn't take the average of the observations per group, we would have pseudoreplication as each student would not be an independent replicate of a treatment - effectively, we have only applied each treatment once. You might notice that we then only have one true replicate per treatment group and this is problematic. To get an estimate of uncertainty, we would have to repeat this experiment a few more times to get more than one proper replicate.

The first scenario, however, did not have this problem and each treatment was replicated five times. After going through all this, we have the following summary:

- **Response:** Test Anxiety
- **Treatment Factor:** Therapy
- **Treatment Levels:** 5, 10, and 15 hours of therapy
- **Treatments:** 5, 10, and 15
- **Experimental unit (no):** Student (15)
- **Observational unit (no):** Student (15)
- **Replicates:** 5

> Tip
>
> Creating a summary like this, is a handy exercise for any experiment you come across, and we'll keep doing it for every experiment in this book. As we go along, we'll also add information about the type of experiment that was conducted.

3 The three R's of experimental design

Experimental Design is a detailed procedure for grouping, if blocking is necessary, experimental units and for how treatments are assigned to the experimental units. There are three fundamental principles, known as the 'three R's of experimental design' which are at the core of a good experiment. The following section might feel a bit repetitive, but these concepts cannot be emphasised enough.

3.1 Replication

Let's define it again: replication is when each treatment is applied to several experimental units. This ensures that the variation between two or more units receiving the same treatment can be estimated and valid comparisons can be made between treatments. In other words, replication allows us to separate variation due to differences between treatments from variation within treatments. For true replication, each treatment should be **independently** applied to several experimental units. If this is not the case, treatment effects become confounded with other factors.

Confounding means that is not possible to separate the effects of two (or more) factors on the response, i.e. it is not possible to say which of the two factors is responsible for any changes in the response. This is what happened in the Example 1 when groups are the experimental units. With only one observation per experimental unit, the effect of therapy is confounded with the experimental unit or the effect of group on test anxiety. The reason why this is a problem is that any difference between the treatments could be due to any differences between the groups

and not just the number of therapy hours. The same would be true if we only had one student per group, why? Take a moment to think about this.

Consider the first row of the data from Example 1. It looks like the student in group 2 scored the highest, followed by group 3 and then group 1. So does longer therapy sessions lead to higher test anxiety? Likely not! With only one student per treatment, we are not able to say that any differences in the response are due to the treatments. It could be due to any differences between the individuals. Maybe the student in group 3 tends to score higher on anxiety tests regardless of the treatment, or perhaps the student in group 1 was unusually calm that day. Without replication, these individual differences could mask (or mimic) the true effects of the treatments.

By replicating the treatments across multiple students, we can average out these individual differences and gain a clearer picture of whether therapy duration truly impacts test anxiety. With five students per group, we might observe that group 1 consistently scores lower than group 3. This consistency would provide stronger evidence that the treatments, and not just individual variation, are responsible for the observed differences. So by replication, we can compare within treatment variation to variation between treatments.

Treatment 1	Treatment 2	Treatment 3
48	55	51
50	52	52
53	53	50
52	55	53
50	53	50

ß

Example 3.1

Maybe the co2 uptake data?

3.2 Randomisation

Randomisation refers to the process of randomly assigning treatments to experimental units such that each experimental unit has equal chance of receiving a specific treatment. Randomisation ensures that:

1. There is no bias on the part of the experimenter, either conscious or unconscious, when assigning treatments to experimental units.
2. No experimental unit is favored to receive a particular treatment.
3. Possible differences between units are equally distributed amongst treatments. If there are clear differences between units, then blocking should be performed and randomisation occurs within blocks. We'll talk more about this in Chapter INSERT
4. We can assume independence between observations.

Randomisation is not haphazard. In statistics (and here in the context of experimental design), randomisation has a specific meaning: namely that each experimental unit has the same chance of being allocated any of the treatments. This can be done using random number generators such as with software packgaes, dice or drawing number from a hat (provided the number have been shuffled adequately and have equal chance to be picked).

Let's have a look at randomisation in R. Suppose we have 4 treatments (`A`, `B`, `C`, and `D`) and 32 experimental units. There are no differences between the units, so we don't have to block, and we can equally split the units across the treatments, which means we have 8 units per treatment, i.e., 8 replicates. In R, we first create a long vector of 8 `A`s, 8 `B`s, 8 `C`s, and 8 `D`s called `all.treat`. Then shuffle the vector to obtain a randomisation using the function `sample`.

```
# repeat the vector A, B, C, D 8 times
all.treats <- rep(c("A","B","C","D"), times = 8)
```

```r
# permutation of all.treats (sample withut replacement)
rand1 <- sample(all.treats)

# example output
rand1
```

```
 [1] "C" "C" "B" "C" "A" "D" "B" "C" "A" "D" "D" "A" "A" "B" "B" "B" "B" "D" "A"
[20] "C" "D" "A" "A" "C" "D" "A" "C" "B" "C" "D" "B" "D"
```

Experimental unit 1 recipes the first treatment that appears as the first element in the shuffled vector, experimental unit 2 receives the second and so on.

Notes on randomisation?

3.3 Reduction of Unexplained Variation (Blocking)

Unexplained variation (or experimental error variance or within treatment variance) is largely due to inherent differences between experimental units. The larger this unexplained variation, the more difficult it becomes to detect treatment differences (a treatment signal). To minimise experimental error variance we can control extraneous factors (i.e. keeping all else constant) and by choosing homogeneous experimental units. Otherwise, we can **block** experimental units to reduce the variation.

Blocking variables are nuisance factors that might affect your response or introduce systematic variation in the response and we are typically, not interested in these. Often, they are factors that cannot be randomised, e.g. biological sex of a person, time of day, location of a warehouse etc. We control the effect of such variables on the response by blocking for them so that we can investigate the possible effect of a variable that we are interested in. Usually, in a **complete block** experiment, there are as many experimental units per block as there are treatments, so that each treatment is applied once in every block. Treatments are randomized to the experimental units in the blocks. We can

then compare the effects of treatments on similar experimental units, and we can estimate the variation induced in the response due to the differences between blocks. This variation due to blocks can then be removed from the unexplained variation.

EXAMPLE

Blocking also offers the oopurutnity to test treatments over a wider range of conditions, e.g. if I only use people of one age in my experiment (say students) I cannot generalize my results to older people. However, if i use different age blocks I will be able to tell whether the treatments have similar effects in all age groups or not.

Lastly, if blocking is not feasible, randomization will ensure that at least treatments and nuisance factors are not confounded.

> "Block what you can, randomize what you cannot."
>
> — Box, Hunter & Hunter (1978)

4 Designing an Experiment

When planning an experiment we need to decide on:

- treatment factors and their levels
- the response
- experimental material / units
- blocking factors
- number of replicates

Some of these will be determined by the research question and how experimental units are assigned to treatments are determined by the design. The design that will be chosen for a particular experiment depends on the **treatment structure** (determined by the research question) and the **blocking structure** (determined by the available experimental units).

Here are two ways the treatments can be structured:

1. **Single factor**: the treatments are the levels of a single treatment factor.
2. **Factorial**: when more than one factor are of interest, then the experiment is said to be a factorial experiment. The treatments are constructed by crossing the treatment factors like we did in Figure **??** such that the treatments are all possible combinations of the treatment levels. For example, if factor A has a levels and factor B has b levels, there are $a \times b$ treatments. Such an experiment would then be called an $a \times b$ factorial experiment.

The blocking structure is determined the set of experimental units chosen or available for the experiment.are there any structures/differences that need to be blocked? Do I want to include experimental units of different types to make the results more general? How many experimental units are available in each block? For the simplest design in this course, the number of

experimental units in each block corresponds to the number of treatments. This is called a complete block experiment. There are several other blocking structures, such as incomplete blocks and blocks with missing values, all with specific analysis which we will not cover here.

In this course, we cover two basic designs: Completely Randomized Designs (CRD) and Randomized Block Designs (RBD). For both designs, the treatment structure can be single or factorial. Where they differ is in terms of the experimental units and how randomization occurs.

Completely Randomized Designs (CRD)

When all experimental units are fairly homogeneous, a CRD is used. Treatments are randomized to all experimental units.

Randomized Block Design

This design is used when all experimental units are not homogeneous or blocking is required to control a nuisance factor. The treatments are randomized to the units within blocks.

Part II

Single Factor Completely Randomised Designs

5 Introduction

Completely Randomized Designs (CRDs) are the simplest experimental designs. They are used when experimental units are uniform enough. We expect them to react simirlary to a given treatment; we have no reason to suspect that a group of experimental units might react differently to the treatments. We also don't expect any effects (besides possibly a treatment effect) to cause any systematic changes in the response. In other words, we don't have to block for nuisance factors.

Remember experimental design is the procedure for how experimental units are grouped and treatments are applied. We have already said that there are no blocks in CRDs. So randomisation occurs without restriction and to all experimental units. More generally, the a treatments are randomly assigned to r experimental units, such that each experimental unit is equally likely to receive any of the treatments. This means that there are $N = r \times a$ experimentnal units in total. We only consider designs that are *balanced* meaning that there an equal number of experimental units per treatment, i.e. a treatment is applied to r units.

5.1 Example: The effect of social media multitasking on classroom performance.

Can we really multitask? I remember as a student I thought I could multitask in lectures, while studying or driving and listening to a podcast. It felt like I was paying attention but in hindsight I can't remember those podcasts well, I know I had to revisit lectures and restart studying sessions. This extends beyond student life, where in the average workspace, tasks are interspersed with social media or email checks and notifcations

(). I think most of us are almost always a little bit tempted by our cellphones when we study or work.

So, if we live in an age of perceived multitasking and getting distracted by our phones and devices, what are the efffects of social media multitasking on our academic performance?

Example 5.1

Two researchers from Turkey, Demirbilek and Talan (2018), conducted a study to investigate this question. Specifically, they examined the impact of social media multitasking during live lectures on students' academic performance.
A total of 120 undergraduate students were randomly assigned to one of three groups:

1. **Control Group:** Students used traditional pen-and-paper note-taking.
2. **Experimental Group 1 (Exp 1):** Students engaged in SMS texting during the lecture.
3. **Experimental Group 2 (Exp 2):** Students used Facebook during the lecture.

Over a three-week period, participants attended lectures on Microsoft Excel. Pre-tests and post-tests were administered to measure learning outcomes.

The analysis of experimental data is determined by the design. The design dictates the terms that we will include in our statistical model and so it is crucial to be able to identify the design and blocking and treatment factors. It is also important to check that randomisation has been done correctly and determine the number of replicates used. In the previous chapter we started doing this by creating a summary of the design and we do the same here. From the description of the study, it is clear that:

- **Response Variable:** Academic performance, as measured by test scores.
- **Treatment Factor:** Level of social media multitasking.

- **Treatment Levels (Groups):** Control, Exp 1, and Exp 2.

Students were randomly assigned to one of the three groups, and performance was measured for each individual. Although this may seem obvious, they only took one measurement per student, so we don't have to worry about pseudoreplication. This setup indicates that the students are both the experimental units and the observational units in this study. With a total of 120 experimental units and three treatments, the experiment has 40 replicates. Since only one treatment factor was investigated, and no blocking was performed, this is classified as a **single-factor Completely Randomized Design (CRD).** Here is the study breakdown:

- **Response Variable:** Academic Performance
- **Treatment Factor:** Level of Social Media Multitasking
- **Treatment Levels:** Control, Experimental 1 (SMS), Experimental 2 (Facebook)
- **Treatments:** Control, SMS multitasking, Facebook multitasking
- **Experimental Unit:** Student (120)
- **Observational Unit:** Student (120)
- **Replicates:** 40 students per group
- **Design Type:** Single-Factor Completely Randomized Design (CRD)

Before we continue, now is the time to note that we won't be using the real data collected in this experiment. It wasn't available but I have simulated data to match their results. I've also made some other modifications such as the original study included 122 students but to ensure a balanced design I include only 120.

5.2 Exploratory data analysis

Before we start any analyses, two things need to be done. First, we start by exploring our data to get familar with the format and to get a feel for any patterns. In R, we read in the data set and then peform a few fommands to check the dataset:

```
multitask <- read.csv("multitask_performance.csv")
nrow(multitask) # check number of rows
```

```
[1] 120
```

```
head(multitask) # check first 5 rows
```

```
    Group Posttest
1    Exp1 86.39427
2    Exp1 64.19996
3    Exp2 52.75394
4 Control 67.81147
5    Exp1 52.39911
6    Exp1 56.58150
```

```
tail(multitask) # check last 5 rows
```

```
      Group Posttest
115 Control 77.94344
116 Control 63.58444
117    Exp1 55.17758
118    Exp2 67.16150
119    Exp2 32.58373
120    Exp2 49.58119
```

```
summary(multitask)
```

```
    Group              Posttest
 Length:120         Min.   :23.38
 Class :character   1st Qu.:52.67
```

```
 Mode  :character   Median :65.01
                    Mean   :63.59
                    3rd Qu.:76.32
                    Max.   :98.78
```

The dataset consits of 120 rows (each row representing a student) and two columns (`Group` and `Posttest`). The first column, `Groups`, contains the treatment the student was assigned and the `Posttest` column contains the response measure. Using the functions `head` and `tail`, we can look at the first and last 5 rows and the function `summary` provies us with a descrption of each column. We do this to check that R has read in our data correctly (you can view the whole data set by running `view(multitask)` as well). The summary tells us that the `Group` column is of the class "character". For our analysis, we want it to be read as a factor:

```
multitask$Group <- as.factor(multitask$Group)
summary(multitask)
```

```
     Group       Posttest
 Control:40   Min.   :23.38
 Exp1   :40   1st Qu.:52.67
 Exp2   :40   Median :65.01
              Mean   :63.59
              3rd Qu.:76.32
              Max.   :98.78
```

Now, we can see that there are 40 replicates per treatment group, confirming that the experiment was balanced. I have assumed that based on the resuts shown that the `Posttest` scores were stored as percentages and using the sumamry we can quickly checked whether there are any observations that are not on the appropriate scale. Looks good so far!

5.3 Model checking

Demirbilek and Talan (2018) had several research questions,

Demirbilek, Muhammet, and Tarik Talan. 2018. "The Effect of Social Media Multitasking on Classroom Performance." *Active Learning in Higher Education* 19 (2): 117–29.