Experimental Design and Analysis
HW01 Solution

Problem 1

(a)

Consider the following model

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}$$

, *where*
- μ *is the* ***overall mean.***
- α_i *is the* ***treatment effect***; $i = 1$ *if keyboard A is used,* $i = 2$ *if keyboard B is used.*
- β_j *is the* ***blocking effect*** *of the* j^{th} *manuscript, for* $j = 1, \dots, 6$.
- γ_{ij} *is the* ***learning effect***. ***(Because the order in which the keyboard was used is important.)***
- ε_{ij} *is the* ***error term***, *whose mean is 0.*

Notes about the learning effects:

Define $\gamma_{1j} := 0$ *if the used keyboards order is* (A, B), *and define* $\gamma_{2j} := 1$ *if the used keyboards order is* (B, A).
Assume $|\gamma_{ij} - \gamma_{2j}| = c \ \forall \ j = 1, \dots, 6$, *i.e., the proficiency gained from completing any manuscript is the same.*

The purpose of this typing experiment is to study $\alpha_1 - \alpha_2$ *(the differences caused by using the 2 different keyboards), which can be estimated by*

$$\overline{y}_{1\cdot} - \overline{y}_{2\cdot}, \text{ where } \overline{y}_{i\cdot} = \frac{1}{6}\sum_{j=1}^{6} y_{ij}.$$

Note that the expectation of this estimator is

$$\mathbb{E}(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) = \alpha_1 - \alpha_2 + \frac{1}{6}\sum_{j=1}^{6}(\gamma_{ij} - \gamma_{2j}),$$

which may have different values when the keyboards are used in different orders.

We consider 3 different designs below and study how ***different designs*** *influence estimation performance.*

Design 1 ***Without Randomization***						
Manuscript	1	2	3	4	5	6
Order	(A, B)	(A, B)	(A, B)	(A, B)	(A, B)	(A, B)

The statistical model under this design is $(j = 1, \dots, 6)$

$$\begin{cases} y_{1j} = \mu + \alpha_1 + \beta_j + \varepsilon_{1j} \\ y_{2j} = \mu + \alpha_2 + \beta_j + c + \varepsilon_{2j} \end{cases}$$

Consider using $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ *to estimate* $\alpha_1 - \alpha_2$, *then*

$$\begin{aligned} \mathbb{E}(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) &= (\mu + \alpha_1 + \beta_j) - (\mu + \alpha_2 + \beta_j + c) \\ &= (\alpha_1 - \alpha_2) - c \end{aligned}$$

Hence, $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ *is biased, and the absolute value of bias is* c.

Design 2 ***With Unbalanced Randomization***						
Manuscript	1	2	3	4	5	6
Order	(A, B)	(B, A)	(A, B)	(B, A)	(A, B)	(A, B)

The statistical model under this design is $(j = 1, \dots, 6)$

$$\begin{cases} y_{1j} = \mu + \alpha_1 + \beta_j + c\boldsymbol{I}_{\{j \in \{2,4\}\}} + \varepsilon_{1j} \\ y_{2j} = \mu + \alpha_2 + \beta_j + c\boldsymbol{I}_{\{j \in \{1,3,5,6\}\}} + \varepsilon_{2j} \end{cases}$$

Consider using $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ *to estimate* $\alpha_1 - \alpha_2$, *then*

$$\begin{aligned} \mathbb{E}(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) &= (\mu + \alpha_1 + \beta_j + \frac{2}{6}c) - (\mu + \alpha_2 + \beta_j + \frac{4}{6}c) \\ &= (\alpha_1 - \alpha_2) - \frac{1}{3}c \end{aligned}$$

Hence, $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ *is biased, and the absolute value of bias is* $\frac{c}{3}$, *which is decreased.*

Design 3 ***With Balanced Randomization***						
Manuscript	1	2	3	4	5	6
Order	(A, B)	(B, A)	(B, A)	(A, B)	(B, A)	(A, B)

The statistical model under this design is $(j = 1, \dots, 6)$

$$\begin{cases} y_{1j} = \mu + \alpha_1 + \beta_j + c\boldsymbol{I}_{\{j \in \{2,3,5\}\}} + \varepsilon_{1j} \\ y_{2j} = \mu + \alpha_2 + \beta_j + c\boldsymbol{I}_{\{j \in \{1,4,6\}\}} + \varepsilon_{2j} \end{cases}$$

Consider using $\overline{y}_{1.} - \overline{y}_{2.}$ *to estimate* $\alpha_1 - \alpha_2$, *then*

$$\begin{aligned}\mathbb{E}(\overline{y}_{1.} - \overline{y}_{2.}) &= (\mu + \alpha_1 + \beta_j + \frac{3}{6}c) - (\mu + \alpha_2 + \beta_j + \frac{3}{6}c) \\ &= (\alpha_1 - \alpha_2)\end{aligned}$$

Hence, $\overline{y}_{1.} - \overline{y}_{2.}$ *is unbiased. This desirable property can be attributed to the good design.*

(b)

Design 4						
Manuscript	1	2	3	4	5	6
Order	(A, B)	(A, B)	(A, B)	(B, A)	(B, A)	(B, A)

I will ***NOT*** *use this for the study. I may do randomization again.*

If we assume $|\gamma_{ij} - \gamma_{2j}| = c \ \forall \ j = 1, \dots, 6$, *as in* ***part (a)***, *then this design is acceptable.*

However, if the learning effects may diminish over time, i.e., $|\gamma_{1j} - \gamma_{2j}| := c_j$ *is no longer constant, then* $c_1 > \cdots > c_6$ *when the manuscripts are typed sequentially as shown above.*

Then the statistical model under this design and the assumption that learning effects will diminish over time is $(j = 1, \dots, 6)$

$$\begin{cases} y_{1j} = \mu + \alpha_1 + \beta_j + c_j \boldsymbol{I}_{\{j \in \{4,5,6\}\}} + \varepsilon_{1j} \\ y_{2j} = \mu + \alpha_2 + \beta_j + c_j \boldsymbol{I}_{\{j \in \{1,2,3\}\}} + \varepsilon_{2j} \end{cases}$$

Consider using $\overline{y}_{1.} - \overline{y}_{2.}$ *to estimate* $\alpha_1 - \alpha_2$, *then*

$$\begin{aligned}\mathbb{E}\left(\overline{y}_{1.} - \overline{y}_{2.}\right) &= \left(\mu + \alpha_1 + \beta_j + \frac{c_4 + c_5 + c_6}{6}\right) - \left(\mu + \alpha_2 + \beta_j + \frac{c_1 + c_2 + c_3}{6}\right) \\ &= (\alpha_1 - \alpha_2) + \left(\frac{c_4 + c_5 + c_6}{6} - \frac{c_1 + c_2 + c_3}{6}\right) \\ &< (\alpha_1 - \alpha_2)\end{aligned}$$

Hence, $\overline{y}_{1.} - \overline{y}_{2.}$ *is biased, and the absolute value of bias is* $\dfrac{c_1 + c_2 + c_3}{6} - \dfrac{c_4 + c_5 + c_6}{6}$, *although the design is balanced.*

Problem 2

(a)

Note that there are 2 factors and each of them has 2 levels , so there are $2^2 = 4$ treatments in total.
There are many differences between the observations from the two groups, as explained in detail below.

Group 1:
Each treatment is only applied to one unit and the measurement of each unit will be read by both students.
So, they made 8 observations on 4 units, with repetitions.

Group 2:
Each treatment is applied to two units, each of which is measured and read once by one student.
So, they made 8 observations on 8 units, with replications.

(b)

We now consider in what situation would one group obtain more accurate results than the other.
There are 3 sources of variation:

$$\begin{cases} \textit{Unit-to-Unit Variation} & (\boldsymbol{U2UV}) \\ \textit{Student-to-Student Variation} & (\boldsymbol{S2SV}) \\ \textit{Measurement Error Variation} & (\boldsymbol{MEV}) \end{cases}$$

Group 1 *can study the differences between measurements made by the two students on each unit,so they have much information about **S2SV**. However, they lack information about **U2UV** because they do not have replications.*

Group 2 *have much more information about **U2UV** than **Group 1**, because they have more experimental units.*
*Because they do not have repetition, the information about **U2UV** will be aliased with that about **MEV**.*
*Because each unit is only measured and read by only one student, it provides less information about **S2SV** than **Group 1**.*

We may face only the following two cases:

$$\begin{cases} 1. \quad \boldsymbol{S2SV} \gg \boldsymbol{U2UV} \\ 2. \quad \boldsymbol{S2SV} \ll \boldsymbol{U2UV} \end{cases}$$

1. $\boldsymbol{S2SV} \gg \boldsymbol{U2UV}$

When $\boldsymbol{S2SV} \gg \boldsymbol{U2UV}$, the main source of variation is student.
*If each unit is measured by two students under the same treatment, impact of large **S2SV** can be mitigated by repetitions.*
*Hence, the method adopted by **Group 1** is better when $\boldsymbol{S2SV} \gg \boldsymbol{U2UV}$.*

2. $\boldsymbol{S2SV} \ll \boldsymbol{U2UV}$

When $\boldsymbol{S2SV} \ll \boldsymbol{U2UV}$, the main source of variation is unit.
*If each treatment is applied to more units, impact of large **U2UV** can be mitigated by replications.*
*Hence, the method adopted by **Group 2** is better when $\boldsymbol{S2SV} \ll \boldsymbol{U2UV}$.*

(c)

***Yes**, student should be treated as a block factor.*

Since different students may have different habits or tendencies that may affect the experiment, and we are not interested in these student effects, we should treat student as a block factor.

Problem 3

(a)

This design has some potential problems, as explained below.

1. Setting and reading is done by the same student at some runs.
If one does not want to set again, she may increase the reading value deliberately.

2. The "setting-and-reading-done-by-one-person" runs are performed by student A more times than by B.
This leads to an unbalanced design and breaks orthogonality.

3. The runs with treatment $(+,+)$ are performed solely by student A.
These outcomes may easily be biased due to the habits or tendencies of student A.

(b)

*To solve **Problem 1**, we should treat (**Setting,Reading**) as a block factor with levels (A,B) and (B,A).*

*To solve **Problem 2** and **Problem 3**, each treatment combination should be carried out under both settings (**Setting,Reading**) $= (A,B)$ and (**Setting,Reading**) $= (B,A)$ an equal number of times.*

The following design is superior to the original one:

A 2^2 full factorial experiment in a randomized complete block design			
Run (Illustrative)	***Factor 1***	***Factor 2***	***(Setting,Reading)***
1	+	+	(A,B)
2	+	−	(A,B)
3	−	+	(A,B)
4	−	−	(A,B)
5	+	+	(B,A)
6	+	−	(B,A)
7	−	+	(B,A)
8	−	−	(B,A)

*The run orders should be **randomized within each level of the block factor**.*

Hence, many run orders could be used in practice. Two examples are given below.

A 2^2 full factorial experiment in a randomized complete block design			
Run (Practical)	***Factor 1***	***Factor 2***	***(Setting, Reading)***
4	+	+	(A, B)
2	+	−	(A, B)
1	−	+	(A, B)
3	−	−	(A, B)
7	+	+	(B, A)
8	+	−	(B, A)
5	−	+	(B, A)
6	−	−	(B, A)

A 2^2 full factorial experiment in a randomized complete block design			
Run (Practical)	***Factor 1***	***Factor 2***	***(Setting, Reading)***
7	+	+	(A, B)
5	+	−	(A, B)
6	−	+	(A, B)
8	−	−	(A, B)
1	+	+	(B, A)
2	+	−	(B, A)
4	−	+	(B, A)
3	−	−	(B, A)