EDUR 7130
Presentation 8a
Experimental Research: Control and Designs

1. Control and Confounding

A key outcome that distinguishes true experimental research from all other forms of research is the ability to provide believable evidence of causality. It is important to note, however, that one study is not enough. Causality must be established through multiple replicated studies. Quasi-experimental studies also provide strong causal evidence, but not as strong as that provided by true experimental studies due to the lack of randomly formed groups in quasi-experimental studies. Non-experimental studies that examine relationships among variables (e.g. ex post facto and correlational studies) may also provide evidence for causal relations, but because of the lack of manipulation, such evidence is necessarily weaker. When attempting to establish a causal connection between variables, it is important to establish control over confounding variables. This is true no matter what type of study conducted. Control allows one to identify those variables that contribute to variation in dependent variable scores.

Why are true experimental studies able to provide causal evidence? Two critical factors:

- **Manipulation** of the IV – the researcher determines which groups receive which treatments.
- **Control** of **confounding variables** – ability to eliminate alternative explanations for causality.

Control is the process of removing the effects of a confounding variable so one can better determine the potential effects of a manipulated factor or independent variable on a dependent variable. Control helps researchers determine which variables had an influence on the dependent variable.

Confounding variable is any variable that makes it difficult to determine if a treatment is effective, or which IV produces variation on the DV. Confounding variables confuse our ability to determine whether an experimental treatment was effective in changing or influencing the dependent variable.

Example 1 of Confounding

Conduct an experiment to learn whether cooperative learning produces better mathematics achievement than lecture.

- The IV is type of instruction (cooperative vs. lecture)
- DV is mathematics achievement
- Design:

Class	IV = Treatment	DV = Mathematics Achievement	Confounding Variable = IQ
A	Cooperative Learning	90	115
B	Lecture	70	95

DV Math Achievement Difference = 20 points

Question

What caused this 20-point difference in mathematics achievement in favor of cooperative learning?

Answer

It is difficult to separate the effects of the treatment from the effects of IQ since both groups had different treatments and different class-mean IQ scores. Thus, IQ is a confounding variable here since we don't know whether the treatment or IQ produced the 20-point difference in mathematics achievement.

Example 2

Same study as above, but the confounder is student sex classroom composition.

- The IV is type of instruction (cooperative vs. lecture)
- DV is mathematics achievement
- Design:

Class	IV = Treatment	DV = Mathematics Achievement	Confounding Variable = Sex
A	Cooperative Learning	90	All Male Students
B	Lecture	70	All Female Students

DV Math Achievement Difference = 20 points

Question

What caused this 20-point difference in mathematics achievement?

Answer

Since student sex and the treatment are perfectly confounded (no overlap of one with the other), it is impossible to know whether student sex composition or the treatment caused the 20-point difference in achievement.

Example 3

Same study as above, but student sex is **controlled** (i.e., differences by sex are eliminated by the study design).

- The IV is type of instruction (cooperative vs. lecture)
- DV is mathematics achievement
- Design:

Class	IV = Treatment	DV = Mathematics Achievement	Classroom Sex Composition
A	Cooperative Learning	90	All Female Students
B	Lecture	70	All Female Students

DV Math Achievement Difference = 20 points

Question

What caused this 20-point difference in mathematics achievement?
Why is the classroom composition of student sex controlled now?

Answer

If both groups are composed entirely of female students, then sex is not a variable in this example, it is a constant. Therefore, it logically cannot vary (covary) with variation in achievement scores. If it cannot covary with achievement scores, then it cannot be related to or impact achievement, so student sex composition is not a confounding variable.

Example 4

Same study as above, but student sex is **controlled** in a different way.

- The IV is type of instruction (cooperative vs. lecture)
- DV is mathematics achievement
- Design:

Class	IV = Treatment	DV = Mathematics Achievement	Classroom Sex Composition
A	Cooperative Learning	90	50% Female, 50% Male
B	Lecture	70	50% Female, 50% Male

DV Math Achievement Difference = 20 points

Question

Does classroom sex composition confound results in this study?

Answer

If student sex composition is equally divided among treatments, with a similar composition in both treatments, then sex would not be confounded with treatment and both groups would be equated in terms of sex. Having equated groups is an important trait of control in research.

Having **equated groups** is key to providing control in experimental research. The goal is to obtain groups that are as similar as possible so any differences observed on the DV can be attributed to treatment differences.

Question

To achieve control in Example 1, reproduced below, what must occur?

Class	IV = Treatment	DV = Mathematics Achievement	Confounding Variable = IQ
A	Cooperative Learning	90	115
B	Lecture	70	95

Answer

The key is to have a balance of IQ in both groups, then that would help control the effects of IQ upon the dependent variable. There are many methods to achieve control in a study. The basic premise is that we want our study groups to be as similar as possible, so any change on the dependent variable observed between the groups can be attributed to the manipulated IV, or the observed IV in non-experimental studies. Using the earlier example, we would want both of our classes to have similar levels of IQ, say both have an average IQ of 105, for instance. If both classes have similar IQ levels, then any difference in achievement found between the classes cannot be explained by differences in IQ.

Summary: Any variable that influences the DV could confound interpretation of results in a study, so it is necessary to control as many confounding variables as possible. Control means to remove the effects upon the DV of possible confounding variables.

2. Control Procedures

Below are four approaches to implementing control in studies. Each are explained and illustrated in turn.

- Randomly formed groups
- Subjects as their own control
- Matching
- Analysis of Covariance

2a. Randomly Formed Groups

Question

How is the use of randomly formed groups a control procedure; how does this procedure help eliminate confounding effects?

Answer

As previously noted, experimental research usually employs treatment and comparison groups, and it is important that these groups be equal or equivalent on everything except the treatments.

Random assignment of subjects to groups/treatments works because of laws of probability. Often groups randomly formed will be roughly equal on important characteristics that could lead to differences on a DV. If

treatment groups are roughly similarly on important factors that could affect the DV, then random formation has eliminated, or greatly reduced, the influence of those factors that could lead to group differences on the DV.

Example

If assignment is random, and if we have 100 people, 10 of which have IQs above 150, then it is unlikely that all 10 with 150+ IQ will be assigned to only one group.

If truly random, then we would expect about 5 of the people with 150+ IQs to be randomly assigned in one group, and 5 in the other, or some similar mix like 6/4 or 7/3. Random assignment helps make balanced groups.

Question

Does random assignment always work?

Answer

If truly random, then it is possible that all 10 people with 150+ IQs could be assigned to only one group and this would lead to confounding problems, but this outcome is very unlikely. In most cases, one will get a somewhat even mix with random assignment, especially if one has many people to work with (i.e., a large sample).

Random Assignment, Random Selection, and Random Formation

Random Assignment

- What works: Randomly assigning participants to treatment and control conditions; this results in randomly formed groups.
- What doesn't work: Randomly assigning treatments to intact, non-randomly formed groups. If the groups are not randomly formed, then randomly assigning treatments does not provide control since the benefit of randomly formed groups is not in play.

Random Selection of Participants

- What works: Randomly selecting participants for each treatment and control condition; this results in randomly formed groups just like randomly assigning participants to groups.
- What doesn't work: Randomly selecting participants who do not form treatment and control groups. For example, a group of females and a group of males were randomly selected to complete a questionnaire of job satisfaction questionnaire. Since there is no manipulation of an IV, this is an ex post facto or causal comparative study.
- Key: Understand when random selection provides control and when it does not. Randomly selecting participants to form treatment and control groups works; randomly selecting participants for a non-experimental study does not lead to the type of control discussed within the framework of an experiment.

For randomization to work, groups must be randomly formed and there must be treatment/control conditions manipulated by the researcher. Randomization works only for true experimental studies.

Example 1

Parveen, Q., & Batool, S. (2012). Effect of Cooperative Learning on Achievement of Students in General Science at Secondary Level. *International Education Studies*, *5*(2), 154-158.

Note: See presentation video for discussion of this information.

They provide a schematic for their design, but the formatting is not aligned properly so I added their design scheme in red. They explain why they chose this design – for **control** purposes.

5. Design

The design of the study was pretest posttest control group design, which is true experiment design. This design was selected because it controls many variables inflecting its external and internal validity. The design is represented schematically as ?

RE O_1 T O_3

RC O_2 __ O_4 (Gay,2000)

R E O1 T O3
R C O2 -- O4

Published by Canadian Center of Science and Education 155

The schematic letters represent the following.

R = random assignment of units to groups
E = experimental group
C = control group
T = treatment
O = observation (i.e., test, measurement, performance, etc.), and number indicate unique observations
--- = no treatment

These symbols and schematics are presented in more detail below in the Experimental and Quasi-Experimental Designs section.

They had two experimental conditions, the treatment and control. Both groups were randomly formed as indicated by the R in the schematic above. Both were pretested (the O_1 and O_2) and following the treatment both were tested again (the O_3 and O_4). They used a pretest-posttest control group design, a true experimental study which means groups were randomly formed.

Example 2

Duckworth, A. L., White, R. E., Matteucci, A. J., Shearer, A., & Gross, J. J. (2016). A stitch in time: Strategic self-control in high school and college students. *Journal of educational psychology*, *108*(3), 329.

Note: See presentation video for discussion of this information.

They conducted three studies. The second was an experiment in which students were assigned randomly to three treatments. The first sentence below contains the two key elements of true experiment – randomly formed groups and manipulation of the IV. By randomly assigning students to the treatments, the researchers were able to address both random formation of groups and manipulation of the IV simultaneously. Note the experiment design was not depicted in symbols as done in Example 1; this is the more common approach, rarely do authors use design schematics for well known designs.

Procedure—Students were randomly assigned to one of three conditions: situation modification ($n = 44$), response modulation ($n = 35$), or no-treatment control ($n = 47$). All interventions were introduced as trying to help "students stick to study goals that they set for themselves."

2b. Subjects as Their Own Control (STOC)

With STOC, everyone in the study is exposed to all treatments (usually), not just one treatment. If everyone is exposed to all treatments, how does this provide control?

STOC works to control possible confounding variables that are static – meaning that they don't change much over time.

Example

Intelligence is thought to be relatively stable over time, so it is a good candidate for control using STOC.

Suppose we measure John and Mia's mathematics self-efficacy across a pretest and two treatments, A and B, over the span of three months (e.g., Fall semester).

	Pre-measure	Treatment	Post-measure 1	Treatment	Post-measure 2
John's Scale Score	20	A	35	B	50
Mia's Scale Score	25	B	40	A	55

Note: John's IQ = 105, Mia's = 110

Question

Did John's IQ cause the change in mathematics self-efficacy over time across pretest and treatments? What about Mia's IQ – did it cause change in her math self-efficacy?

Answer

Their IQ remains constant or generally constant, so it cannot be the reason we observe changes in math self-efficacy for John or Mia, at least of short periods of time. Using STOC, we know then that IQ is not a confounding variable since IQ will not change much for John or Mia, so IQ can be eliminated as a reason for changes in math self-efficacy.

The problem with STOC is that it cannot control for things that change about Mia or John, such as opinion or attitudes, which may be in constant flux.

Note that with subjects as their own control, multiple treatments potentially will be administered to each subject. To avoid the problem of multiple treatment interference (that is, effects of treatments are confounded with each other), one must take measurements/observations after each treatment is administered, and ideally allow enough time to elapse to diminish, if possible, the effect of the prior treatment.

For example, a treatment for Bob and Sue might look like this:

Pretest, treatment A, assess Bob, treatment B, assess Bob again,

then reverse for Sue:

Pretest, treatment B, assess Sue, treatment A, assess Sue again.

STOC Designs

(a) Single-subject Designs

Sometimes study participants may comprise very small populations (e.g., only 3 in a school), and this precludes large, randomized groups research designs. In these situations, single-subject designs can be helpful. These designs typically use a baseline observation period, a period of treatment introduction, another baseline, maybe another treatment, and so on.

For those interested in more detail, Cox (2016) introduces such designs in a short article (Cox, D. J. 2016. A brief overview of within-subject experimental design logic for individuals with ASD. *Austin Journal of Autism & Related Disabilities*, *2*, 1025.) that can be accessed on the web.

Example

Theodore, L. A., Bray, M. A., Kehle, T. J., & Jenson, W. R. (2001). Randomization of group contingencies and reinforcers to reduce classroom disruptive behavior. *Journal of School Psychology*, *39*(3), 267-277.

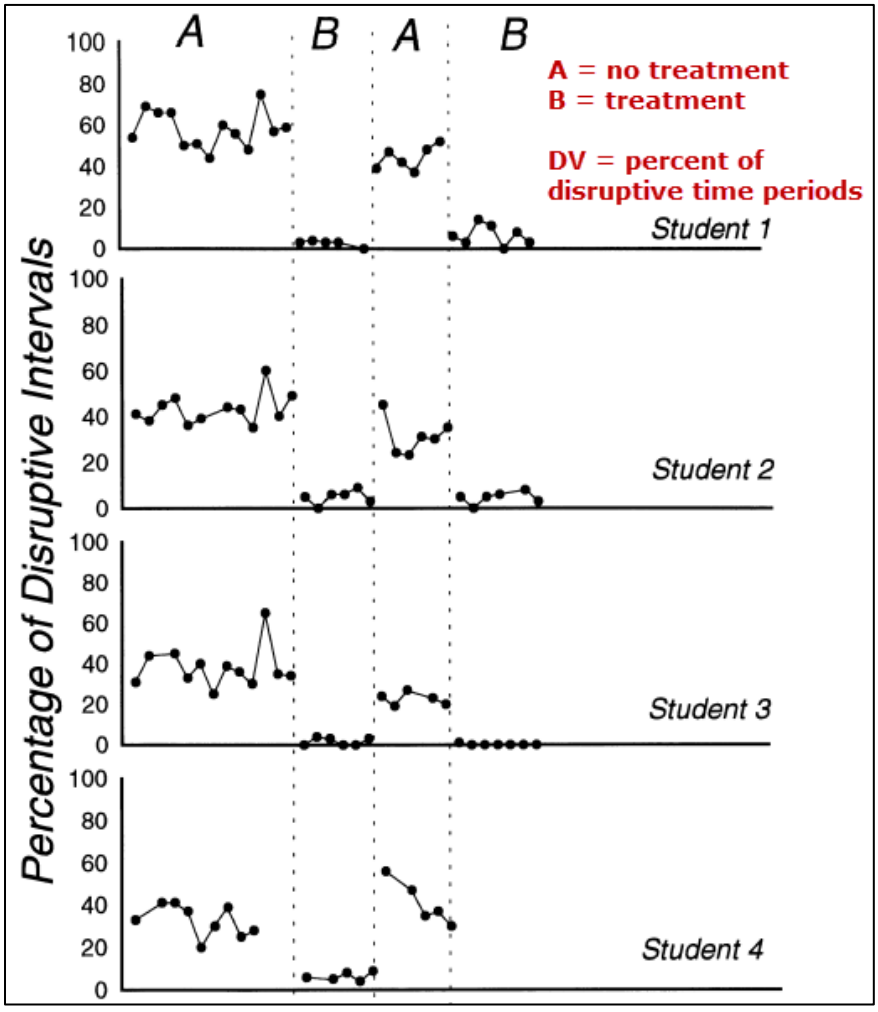
Note: See presentation video for discussion of this information.

Purpose: To explore one strategy for reducing classroom disruptions. An ABAB reversal design was used.

Results for four students are presented below.

Symbols

A = first baseline period, classroom management treatment not employed for three weeks
B = treatment period of two weeks
A = another baseline with no treatment for two weeks
B = another two-week period with the treatment



(b) Repeated Measures, or Within-Subjects Designs

Unlike single-subject designs, these use groups of individuals with repeated measurements of dependent variables over time after exposing participants to a treatment or IV. The process is like single-subject designs but focuses on groups rather than one individual at a time. Often these are longitudinal designs or involve multiple treatments. If there are multiple treatments, a plethora of designs exist to handle these types of studies. For those interested, these experimental studies are typically labeled as counterbalanced, crossover, or switchback designs. Campbell and Stanley

(1963; Experimental and quasi-experimental designs for research) provide an excellent introduction to these and other types of experimental designs. Use Google Scholar to find this publication.

Example

Bedard, C., Bremer, E., Campbell, W., & Cairney, J. (2018). Evaluation of a direct-instruction intervention to improve movement and preliteracy skills among young children: A within-subject repeated-measures design. *Frontiers in Pediatrics*, *5*, 298.

Note: See presentation video for discussion of this information.

Purpose: Provide instruction on development of preliteracy and motor (print) skills for 3 to 4-year-old children.

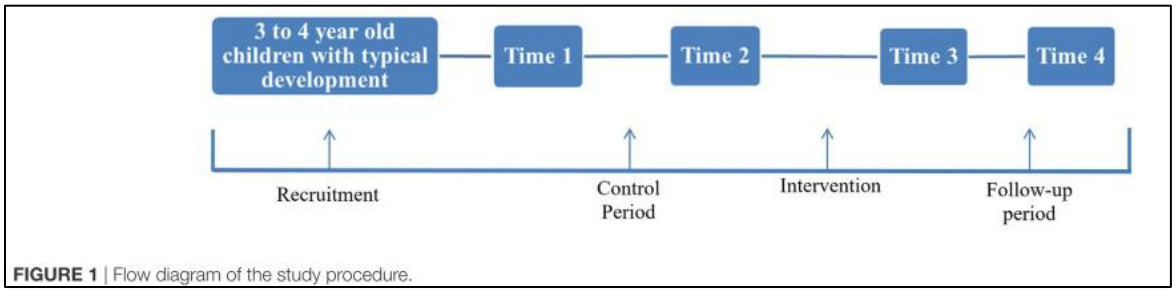


FIGURE 1 | Flow diagram of the study procedure.

The means for A and B at times 3 and 4 were higher, statistically (significantly) than at time 1 and time 2 – see below.

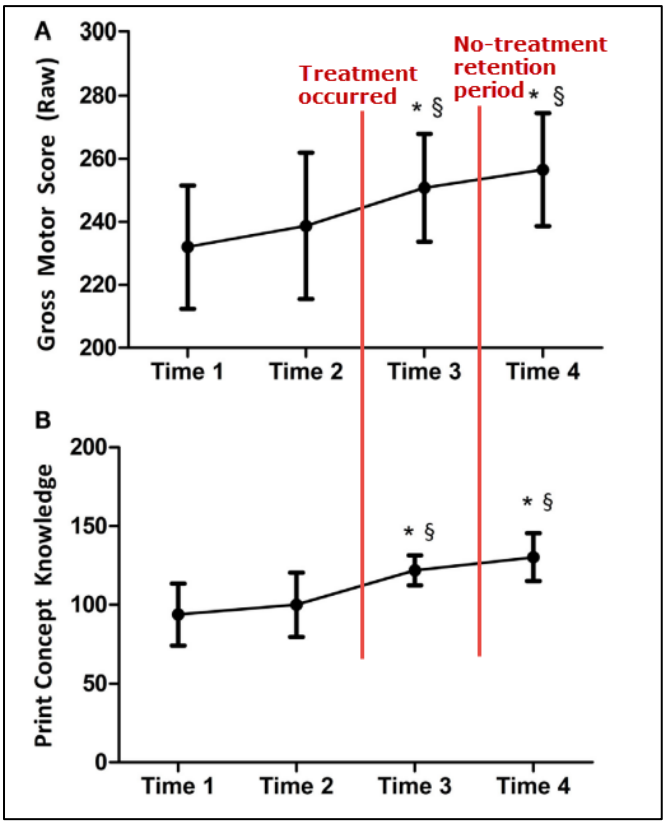
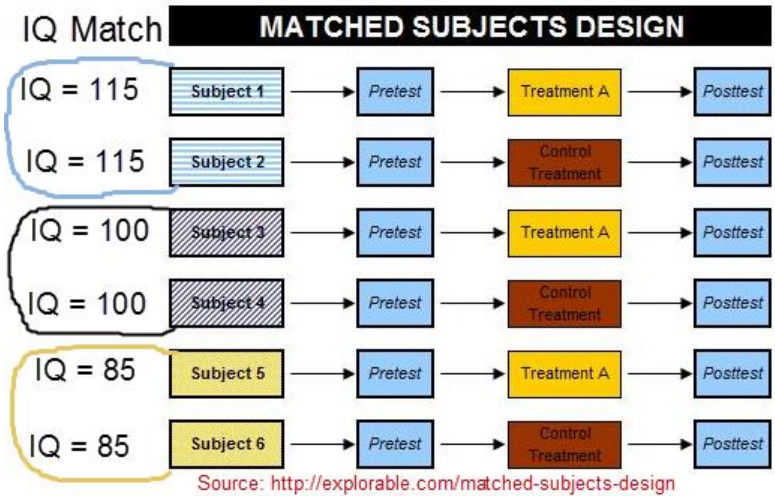


FIGURE 2 | Change over time in the primary outcome measures. **(A)** Gross motor skills; **(B)** print-concept knowledge; and **(C)** alphabet knowledge; *statistically significantly different from time 1; §statistically significantly different from time 2.

The design employed by Bedard et al. (2018) is poor because there is no control or comparison group to allow one to judge natural growth vs. growth due to the treatment. Notice the increase from time 1 to 2 – this is natural growth since no treatment occurred between these two time periods. The plotted change from time 2 to 3 is not visually overwhelming like with the ABAB design shown earlier. Having and plotting a control performance across these four times would greatly illuminate differences due to the treatment, if such a difference exists.

2c. Matching, and Group and Subgroup Matching

Both are essentially the same. With matching one links individuals (or groups of individuals) between two or more groups (e.g., students with IQs between 110 and 115 are matched in each group).



An important drawback with matching is that the more variables one wishes to match, the more likely one will have difficulty finding matches. Therefore, this procedure is limited to only a few key variables for control if done by hand. There is, however, a statistical approach called **propensity score matching** that enables one to match on several variables. Propensity score matching is complex and not covered in this course, but for those interested, a short primer can be found at Wikipedia, linked below.

https://en.wikipedia.org/wiki/Propensity_score_matching

How does group and sub-group matching work?

Suppose we have two groups and wish to match on sex and IQ. In Group 1 (Cooperative Learning) we separate males, and then within males we create four sub-groups based upon IQ. We do the same with the comparison group (Computer Instruction). See table below for an illustration. We create subgroups rather than strict one-to-one matches. These subgroups may contain 3, 5, 10, or any acceptable number and it is not critical that subgroup sizes be the same, although rough similarity is often sought.

	Treatment Group 1: Cooperative Learning		**Treatment Group 2: Computer Instruction**	
Sex	Males	Females	Males	Females
IQ Categories	75 & below	75 & below	75 & below	75 & below
	76 to 85	76 to 85	76 to 85	76 to 85
	86 to 95	86 to 95	86 to 95	86 to 95
	96 to 105	96 to 105	96 to 105	96 to 105

Example 1

Ahmad, S. I., Leventhal, B. L., Nielsen, B. N., & Hinshaw, S. P. (2020). Reducing mental-illness stigma via high school clubs: A matched-pair, cluster-randomized trial. *Stigma and Health*, *5*(2), 230.

Note: See presentation video for discussion of this information.

Purpose: To learn whether high school club participation would reduce mental-illness stigma for students.

Ahmad et al. identified 42 high schools in northern California and used a "matched-pairs" design to match schools on demographic variables: school size, public vs other, student body diversity, and percentage of students receiving reduced-price lunches. This example illustrates **group matching**, where groups are schools of students.

> Thus, from 2015–2017 in Northern California, we utilized a matched-pairs design—a subset of a randomized block design—with each "pair" incorporating two similar schools, one of which started its club at the beginning of the school year and the other midway through. We matched schools based on relevant school demographic variables and then randomly assigned one school in a pair to the immediate and the other to the delayed club condition. Measures (nearly identical to those used in the prior quasi-experimental study) were collected at the beginning, middle, and end of the school year. Our multilevel model design and analyses account for both school- and student-level variation (Flay et al., 2005).

> Instantiating a matched-pair, cluster-randomized design, study staff matched schools based on overall school size (*N* of student body), public versus "other" school status, student body diversity (as indexed by the proportion of Hispanic youth), and the percentage of the student body receiving reduced-price lunches. For example, two large urban high schools with high diversity and relatively high socioeconomic disadvantage could form a pair; two smaller parochial schools could form another pair. Paired schools were then randomly assigned to the immediate ($n = 23$) or delayed ($n = 23$) conditions. However, four schools declined to participate in the study after being assigned to the delayed condition, so the final count of immediate and delayed schools was 23 and 19, respectively.

They compared the matched samples on demographics and stigma measures for participants and found highly similar results before implementing the experimental treatments. The table below shows the level of similarity between pairs. These comparisons are helpful for showing the matched-pairs are similar at the study outset. Results of the experimental treatment, comparing the two groups, follow with other analyses, but are not shown here.

Table 2
Demographic and Baseline Stigma Information by Intervention Group

	Immediate schools		Delayed schools		
Variable	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*p* value
Demographics					
Female (%)	258 (75)		259 (76)		*ns*
Class standing[a]	344	2.7 (1.00)	345	2.8 (1.01)	*ns*
GPA	269	3.6 (.56)	286	3.58 (.55)	*ns*
Parent education[b]	339	3.6 (1.49)	334	3.6 (1.47)	*ns*
Stigma measures					
Knowledge[c]	285	3.7 (.30)	266	3.8 (.31)	*ns*
Attitudes[c]	285	4.1 (.35)	267	4.2 (.33)	*ns*
Social Distance[c]	283	4.2 (.58)	264	4.3 (.60)	.07
Positive Actions[d]	283	.54 (.24)	262	.57 (.23)	.07

Note. Sample sizes vary due to missing data (cases excluded analysis by analysis). *ns* = not statistically significant.
[a] For class standing: 1 = freshman, 2 = sophomore, 3 = junior, and 4 = senior. [b] For primary parent education: 1 = *did not complete high school*; 6 = *medical degree/doctorate*. [c] Mean scores are calculated on a 5-point scale; higher scores indicate more accurate knowledge, positive attitudes, and greater willingness to interact with a person with mental illness (less social distance). [d] Mean scores are calculated as a proportion of 16 possible positive actions (range 0–1). Higher scores indicate more actions taken.

Example 2

Suldo, S. M., Savage, J. A., & Mercer, S. H. (2014). Increasing middle school students' life satisfaction: Efficacy of a positive psychology group intervention. *Journal of happiness studies*, *15*(1), 19-42.

Note: See presentation video for discussion of this information.

Purpose: Determine whether a 10-week group wellness-promotion intervention could increase middle school students' sense of life satisfaction for those who were dissatisfied prior to the experiment.

The authors used propensity score matching to match 40 sixth-grade students, 20 in the intervention group and 20 in control group, which had a delayed intervention. The authors were able to create two groups that matched closely on several variables important to the study (i.e., possible confounders that were controlled via matching). See the table below for variables and their mean scores between groups.

2.5 Overview of Data Analysis Plan

Prior to analyses, 2 students (1 from the control condition, 1 from the intervention condition) identified as univariate outliers (across-time changes in an indicator of SWB were >3 *SD* from the sample mean on change scores) were removed. Because *t*-tests of between-group differences indicated that at baseline the 27 students randomly assigned to the intervention group had lower life satisfaction scores and reported significantly more externalizing symptoms of psychopathology than the 26 students assigned to the control group, propensity score matching (see Fan and Nowell 2011, for an overview) was used to identify subsets of students in the intervention and control groups with approximately equivalent baseline scores. To calculate propensity scores representing the likelihood of being assigned to the intervention condition based on baseline characteristics, baseline life satisfaction, positive affect, negative affect, internalizing, and externalizing scores were entered as predictors of assignment to the intervention condition in a logistic regression analysis, with predicted probabilities of being assigned to the treatment group saved as propensity scores. Absolute differences between the propensity scores of students in the intervention and control conditions were calculated using the *dist* macro (Kosanke and Bergstralh 2004a), and the *vmatch* macro (Kosanke and Bergstralh 2004b) was used to match students based on these absolute differences. The *vmatch* macro was specified with a maximum absolute difference of .20, .15, and .10 between

Results of their experiment did not produce the results they sought. The mean scores and a graph of those scores, shown below, reveal that the control group scored similar to the intervention group.

Table 3 Means on mental health and achievement outcomes by matched intervention groups

Variable	Intervention ($n = 20$)		Control ($n = 20$)	
	M	*SD*	*M*	*SD*
T1: Life satisfaction	4.04	.90	4.14	1.03
T1: Positive affect	3.69	.70	3.63	.85
T1: Negative affect	2.20	.63	2.22	1.02
T1: Internalizing psychopathology	19.05	9.71	18.30	11.14
T1: Externalizing psychopathology	10.55	5.59	10.30	7.99
T2: Life satisfaction	4.39	.87	4.01	1.23
T2: Positive affect	3.92	.72	3.78	.84
T2: Negative affect	2.24	.72	2.15	1.07
T2: Internalizing psychopathology	14.25	10.13	15.40	11.51
T2: Externalizing psychopathology	9.25	6.33	9.65	9.30
T3: Life satisfaction	4.53	1.13	4.59	.91
T3: Positive affect	4.00	.65	4.10	.96
T3: Negative affect	1.88	.65	1.83	.79
T3: Internalizing psychopathology	11.65	9.42	12.55	10.21
T3: Externalizing psychopathology	8.30	5.56	6.70	7.18
Propensity	.48	.16	.51	.17

Time 1 - before intervention

Means similar

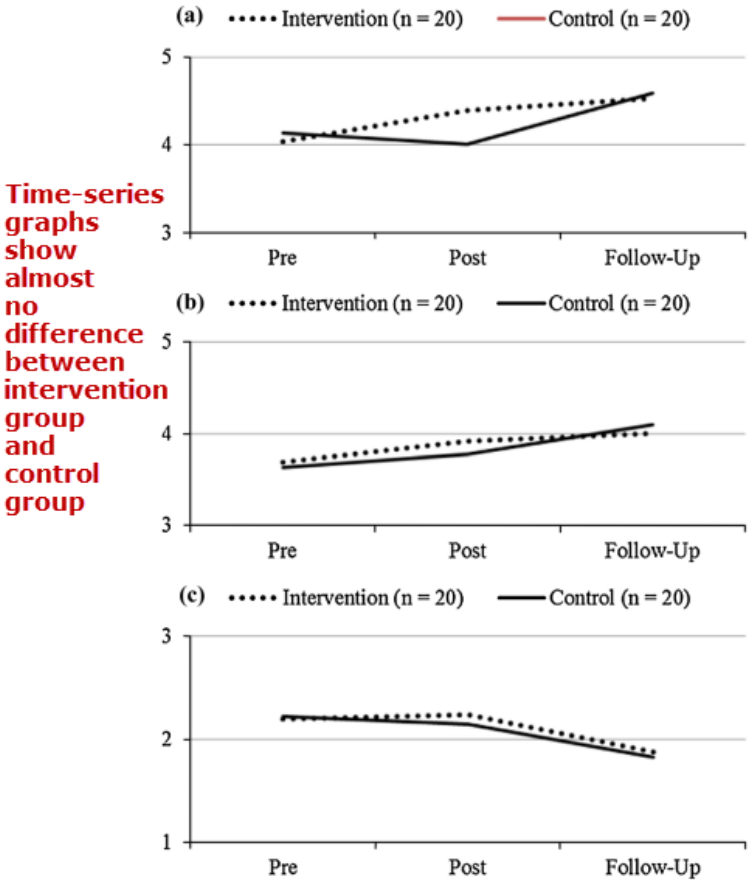
Time 2 - one week after intervention

Means not similar

Time 3 - six months after intervention

Means similar or control group better

T1 Time 1 (Baseline); *T2* Time 2 (Post-intervention); *T3* Time 3 (6-month follow-up). Life satisfaction was assessed by the Students' Life Satisfaction Scale; Positive affect and negative affect were assessed by the Positive and Negative Affect Scale for Children; Internalizing psychopathology and externalizing psychopathology were assessed by the Youth Self Report form of the Child Behavioral Checklist



Time-series graphs show almost no difference between intervention group and control group

Fig. 1 Participants' mean levels of subjective well-being at baseline, post-intervention, and follow-up. **a** Life satisfaction. **b** Positive affect. **c** Negative affect

2d. Analysis of Covariance (ANCOVA)

ANCOVA was introduced in the inferential statistics presentation. Recall that with ANCOVA, the DV mean scores are adjusted to account for differences between groups on the covariate, which is typically a quantitative IV. The covariate serves to statistically equate groups if there are differences and thereby offers a control mechanism. The previous control mechanism – random groups, STOC, and matching – provide control by design. ANCOVA provides control by statistical, mathematical, adjustment.

This control mechanism is not as robust as the other techniques and is more suspectable to misleading results if used inappropriately. ANCOVA adjustments can be misleading because groups that differ on the covariate may also differ in other important ways that are not measured in the study. Thus, any differences on the DV observed between groups, even after the covariate adjustment, must be viewed cautiously and one must not assume that those DV differences are the direct result of the IV studied. Many authors warn against using ANCOVA when groups are not randomly formed. This warning is important, but I believe ANCOVA can be used productively if one understands and acknowledges that ANCOVA adjusted DV differences must be viewed with skepticism rather than with certainty.

Example 1 - Adjustments

Group	IQ (Covariate)	Treatment	Achievement DV
1	130	A	80
2	100	B	70

Question

Which way will the DV mean scores be adjusted for each group (assuming a positive correlation between IQ and achievement)?

Answer

Group 1's score of 80 will be adjusted down since they started with the higher IQ of 130, and group 2's scores of 70 will be bumped up since they started with the lower IQ score of 100.

Example 2 - Adjustments

Group	IQ (Covariate)	Treatment	Achievement DV
1	93	A	65
2	93	B	85

Question

Which way will the DV mean scores be adjusted for each group (assuming a positive correlation between IQ and achievement)?

Answer

No adjustment to the DV since both groups have the same covariate mean score of IQ = 93.

ANCOVA serves as a control procedure because it allows one to control for covariates by **statistically equating groups**. The other control procedures offer better control because they create **equated groups by design**, not by statistical adjustment.

As noted, one must be cautious with ANCOVA, however, since it is not always clear if the adjustment is legitimate or appropriate for some situations or contexts. A classic example of possible misleading or inappropriate use of ANCOVA is with intact, already-formed groups. With such groups there may be many pre-existing group differences that are not measured and therefore cannot be control statistically. If a possible confounding variable is not measured, it cannot be used in ANCOVA as a covariate. Only variables with scores – those that are measured – can be included in ANCOVA.

Example 1

Moh'd Al-Migdady, A., & Qatatsheh, F. (2017). The effect of using Crocodile mathematics software on Van Hiele level of geometric thinking and motivation among ninth-grade students in Jordan. INSTRUCTIONAL TECHNOLOGY, 27.

Note: See presentation video for discussion of this information.

Purpose: To learn whether two types of mathematics instruction differ in student performance.
Instructional IV:

CMG = Crocodile Mathematics Group

NCMG = Non-Crocodile Mathematics Group

DV 1 = Geometric test performance
DV 2 = Motivation to learn mathematics
Covariates = Pretest scores of the DV (test scores and motivation)
Sample = 9th grade students in Jordan

Table 4

Two way analysis of covariance summary table of students' motivation to learn Mathematics

Significant

Source of variation	Sum of Squares	Degrees of Freedom	Mean of Squares	F	P Value	η2
Covariate	11142.437	1	11142.437	56.90	0.00	0.3424
Group	1635.999	1	1635.999	8.354	0.005	0.0502
Gender	64.540	1	64.540	0.330	0.568	0.00198
Group× gender	401.353	1	401.353	2.050	0.156	0.01233
Errors	14686.816	75	195.824			0.4513
Overall	13244.329	79				

*P< 0.05

Not significant

The table above shows the ANCOVA for motivation to learn mathematics. Both the covariate and the instructional groups, Group, are significant which means the covariate, the pretest motivation scores, are related to posttest motivation scores, as one might expect. The Group variable represents the instructional strategies, and since that is significant, it means there are group differences in adjusted posttest motivation scores.

To learn how these differences manifest, the authors reported descriptive statistics in the Table 3, further below. The relevant means are posted in the table immediately below. It is highly likely that pretest and posttest motivation scores correlated positively, so if one group started higher then their posttest score will be adjusted down and the group that started lower will have their posttest score adjusted up. That occurred for this experiment, see below.

Group	Pretest Motivation	Posttest Motivation	Adjusted Posttest Means
Crocodile Math Software	124.22 (Higher)	136.90	131.57 (Adj Down)
Non-crocodile Math Software	117.90 (Lower)	122.15	126.97 (Adj Up)

The **adjusted posttest means**, also called **estimated marginal means** by some, provide a what-if scenario. What if both groups started with the same level of pretest motivation, say an average group motivation score of 121.06 (which is the mean of the two pretest means)? What would be their predicted posttest mean scores?

That is what the adjusted means, or the estimated marginal means, provide, an estimate of posttest mean scores if both groups started with a pretest mean score that was equal. The original mean difference was 136.90 –122.15 = 14.75, but the adjusted mean difference is 131.57 – 126.97 = 4.60, a more modest difference. This tells us, according to the ANCOVA adjustments, that if both groups started with the same level of motivation, their posttest mean difference would be only 4.60 points instead of the observed mean difference of 14.75. This suggests that while the Crocodile software helps boost motivation, the gain is not as large as the raw means suggest.

Table 3
Descriptive statistics for the pre-and the post-tests of students' motivation to learn mathematics

			The Pretest		The Posttest		The Adjusted Mean of the Posttest
Group	Gender	Number	Mean	Standard Deviation	Mean	Standard Deviation	
CMG	Male	19	115	14.04	128	22.11	133.075
	Female	21	132.57	11.79	143.9	17.86	131.127
	Total	40	124.22	15.33	136.9	21.30	131.577
NCMG	Male	19	116.11	19.50	124.21	19.61	127.47
	Female	21	119.52	18.11	120.29	13.67	126.519
	Total	40	117.9	18.62	122.15	16.65	126.97

Higher
Adj Dow
Lower
Adj. Up

Example 2

Rogers, M. A., Wiener, J., Marton, I., & Tannock, R. (2009). Parental involvement in children's learning: Comparing parents of children with and without Attention-Deficit/Hyperactivity Disorder (ADHD). *Journal of school psychology*, *47*(3), 167-185.

Note: See presentation video for discussion of this information.

Purpose: Examine parental involvement with their children's learning, ages 8 to 12, and compare that involvement between children with ADHD and children without ADHD.

IV = Parents with children with ADHD vs. parents with children without ADHD
DV = many were examined, see table below.
Covariate = Parents' education level

Rogers et al.'s results are reported in Table 2 below. Note that they do not provide a complete ANCOVA summary table, but instead report on the F ratio, effect sizes (which we have not covered), and the **adjust means** which they label as **estimated marginal means**. In general, the results show that parents, and specifically fathers, tend to perform worse in their roles if their children have ADHD.

Nowhere in Table 2 is ANCOVA mentioned. How might a reader know that ANCOVA was employed? The **adjusted means** (**marginal means**) and the F ratios are the clues that informs readers ANCOVA was used. If the adjusted means were not presented, but the F ratios were included, then one would likely conclude an ANOVA was used.

The authors did not report the original, unadjusted means for each DV or the parental education level means. This is a mistake, I think. To provide more information is better since it allows readers to judge whether the adjustments seem reasonable.

Table 2
Differences between ADHD and non-ADHD participants on parental involvement variables.

	ADHD [a]				Non-ADHD [b]				F (1,101)	Effect size (η^2)
	Minimum	Maximum	Mean [c]	*SD*	Minimum	Maximum	Mean [c]	SD		
Role beliefs	3.80	5.90	4.93	.48	3.30	6.00	5.05	.52	.86	.01
Parental efficacy	1.60	5.20	3.83	.89	2.20	6.00	4.43	.87	7.10**	.07
General school invitations	2.00	6.00	4.52	.91	3.33	6.00	5.17	.66	14.13**	.13
Specific teacher invitations	1.00	4.80	2.39	.98	1.00	4.20	1.96	.80	7.40**	.07
Specific child invitations	1.00	6.00	2.89	.94	1.00	4.00	2.76	.67	2.10	.02
Knowledge and skills	2.60	6.00	4.71	.80	3.40	6.00	5.10	.60	1.83	.02
Time and energy	1.40	6.00	4.34	.98	2.40	6.00	4.76	.96	4.04*	.04
Mothers' active participation	3.23	4.85	4.09	.43	2.15	4.85	3.94	.59	2.02	.02
Mothers' academic pressure	1.30	4.40	2.75	.65	1.20	3.90	2.53	.63	1.45	.02
Mothers' encouragement	2.90	5.00	4.14	.44	3.60	4.90	4.27	.35	.46	.01
									F (1,69)	
Fathers' active participation	2.09	4.82	3.56	.62	3.09	5.00	3.96	46	6.53**	.09
Fathers' academic pressure	1.67	3.67	2.91	.55	1.67	4.44	2.58	.61	6.60**	.10

*$p \leq .05$. **$p \leq .01$.
[a] N=53.
[b] N=48.
[c] Estimated marginal means.

Estimated Marginal Means, i.e., Adjusted Means - adjusted for Parents' educational level

4. Experimental and Quasi-Experimental Designs

For experimental designs, there are only a few symbols to comprehend.

O = observation (e.g., test administered, judges evaluate, count made, observers record, etc.),
X = treatment (e.g., experimental manipulation such as different instruction, different drugs, etc.),
R = Randomly formed group (i.e., groups created randomly)
N, NR, or simply left blank = Non-randomly formed group

If trying to determine which design was used in a study,

- first note whether groups were randomly formed. If yes, that corresponds to the R in the design. If no, then that corresponds to the N or NR (non-random) in the design, or simply no R.
- A pretest, if given, is represented by O, which follows with a treatment X then a post-test, so another O. Sometimes the O's have subscript numbers to differentiate when and who was observed or tested, e.g., O_1, O_2, O_3, etc.

A few designs are presented below, but for more designs and examples, please refer to chapter 20 in the supplemental text found in the syllabus by Cohen, L., Manion, L., & Morrison, K. (2018; Research Methods in Education. Routledge.) or the Campbell and Stanley (1963) text mentioned above under Subjects as their Own Control section.

4a. True Experimental Designs

(a) Pretest-posttest Control Group

This design is symbolized as follows.

R O X_1 O	or	R O X_1 O
R O X_2 O		R O O

There are two groups since there are two rows (although more groups are possible), and each group is randomly formed since each row begins with R. The next column is O which signifies a pretest since it occurs before the treatment. The symbols X_1 and X_2 represent treatment 1 and treatment 2, or treatment 1 and the control condition. Sometimes the lack of an X signifies the control condition as shown in the second schematic above. Following the treatments are the final set of observations which represent the posttests.

Example: Pretest-posttest Control Group

Watkins, P. C., Cruz, L., Holben, H., & Kolts, R. L. (2008). Taking care of business? Grateful processing of unpleasant memories. The Journal of Positive Psychology, 3(2), 87-99.

Note: See presentation video for discussion of this information.

Purpose: To learn whether grateful processing can bring "closure to unpleasant emotional memories."

Procedure

Participants were recruited by introducing the study in classes at least a day before the study took place. In the announcement, potential participants were informed that they would be recollecting an unpleasant open memory, and they were given a brief description of an open memory. This announcement served two purposes. First, it gave subjects a clear idea of what they would be participating in and thus reduced any coerciveness. Second, we hoped that priming subjects about recalling an open memory would help them recall a significant open memory in the actual study.

All measures were administered in a group setting. In the pretest session the study was described again, and participants were given an informed consent form, which they proceeded to read and sign. We then administered our pretest measures. We first administered the positive and negative affect measure (PANA; Diener & Emmons, 1984), followed by the WBSI, the GRAT-R, and the RSS. When all

Participants were provided with four sheets of lined paper to write about their topic. A few subjects actually used up their sheets, and thus were encouraged to write on the back of the sheets provided. Following each writing session subjects completed the PANA, and following the third session all participants completed the posttest measures. Posttest and 1-week follow-up measures were identical to that at pretest, with the following exceptions. First, we did not administer the WBSI or the RSS after pretest. Second, we administered the GRAT-R at the end of the posttest and follow-up sessions. Finally, we did

Participants were randomly assigned to one of the three writing conditions and wrote for 20 minutes for 3 days on their assigned topic. In some classes this occurred on consecutive days, whereas in other classes 1–3 days separated the writing sessions. This timing difference was simply due to the different manner in which the courses were scheduled. In the control condition, participants were instructed to write about 'your plans for tomorrow. Think about what you would like to do and how you will probably spend your day.' If they ran out of things to write about, they were to write about 'a description of your shoes.' As

Often researchers will inform readers of the design they used (e.g., we employed a pretest-posttest control group design…). Sometimes this information is not presented so readers must look for **clues** to identify the type of design.

In the Procedure section, Watkins et al. explain that they administered several **pretest** measures.

Clue 1 – a pretest is present.

The authors also explain that following the pretest measures, participants were **randomly assigned** to one of three conditions. Following discussion of the **three conditions**, one being a control condition, **posttest** measures were administered, then administered again 1 week after the initial administration of the posttest.

Clue 2 – groups are randomly formed.
Clue 3 – there are three experimental conditions, so three groups.
Clue 4 – two posttest periods, repeated measurement of the posttest.

Together Watkins et al.'s design follows the schematic shown below.

R O X_1 O O	or	R O X_1 O O
R O X_2 O O		R O X_2 O O
R O X_3 O O		R O O O

(b) Posttest-only Control Group

This design is the same as the pretest-posttest control group design, except that no pretest is employed.

R X_1 O	or	R X_1 O
R X_2 O		R O

Why eliminates a pretest? If random assignment works to equate groups, then it is likely, if the groups are large enough, that the random process will control confounding variables, so no pretest is needed to check for group equivalence. This design is useful if a pretest could alter performance on the posttest such as through priming so participants know what to expect on the posttest or make them more aware during the treatment.

Example: Posttest-only Control Group

Vohs, K. D., Mead, N. L., & Goode, M. R. (2006). The psychological consequences of money. science, 314(5802), 1154-1156.

Note: See presentation video for discussion of this information.

Purpose: To test whether money changes "people's motivation (mainly for the better) and their behavior toward others (mainly for the worse)."

Below the authors explain the 5th experiment in a series of 9 experiments discussed in this article. This presentation makes the following clear.

- Pretest = none mentioned
- Treatments = three, after playing Monopoly
    - one group will have won high value money ($4,000)
    - the second group will have won low value money ($200)
    - control group will have won no money ($0)
- Posttest, DV = number of pencils participants helped collect after a pencil box was spilled

Together, this information can be symbolized in the follow schematic.

R X_1 O	or	R X_1 O
R X_2 O		R X_2 O
R X_3 O		R O

In Experiment 5, we wanted to give money-primed participants a helping opportunity that required no skill or expertise, given that the help that was needed in the two previous experiments may have been perceived as requiring knowledge or special skill to enact. The opportunity to help in the current experiment was quite easy and obvious, in that it involved helping a person who spilled a box of pencils.

Participants were randomly assigned to one of three conditions that were manipulated in two steps. Each participant first played the board game Monopoly with a confederate (who was blind to the participant's condition) posing as another participant. After 7 min, the game was cleared except for differing amounts of play money. Participants in the high-money condition were left with $4000, which is a large amount of Monopoly money. Participants in the low-money condition were left with $200. Control condition participants were left with no money. For high- and low-money participants, the play money remained in view for the second part of the manipulation. At this step, participants were asked to imagine a future with abundant finances (high money), with strained finances (low money), or their plans for tomorrow (control).

Next, a staged accident provided the opportunity to help. A new confederate (who was blind to the participant's priming condition) walked across the laboratory holding a folder of papers and a box of pencils, and spilled the pencils in front of the participant. The number of pencils picked up (out of 27 total) was the measure of helpfulness.

Vohs et al. found that that more money primed less helpfulness. Those with more money were less helpful (collected fewer pencils), than those with low money or no money.

Summary: Two examples of true experimental designs were presented; however, many variations exist such as multiple treatments, multiple observations before and/or after a treatment, counterbalanced or (crossover/switchback) designs, and no observations prior to a treatment. See supplemental textbooks for more examples.

4b. Quasi-Experimental Designs

(a) Non-equivalent Control Group Design

While this design mimics the pretest-posttest control group design, a true experiment, this is a quasi-experimental design because it lacks randomized groups. This design is symbolized as follows.

O X_1 O	or	N O X_1 O	or	NR O X_1 O
O X_2 O		N O O		NR O O

The lack of R indicates groups were not randomly formed; sometimes this fact is symbolized using N or NR (non-randomized). Due to the lack of randomly formed groups, results from this design have less certainty since it is unclear whether experimental groups were essentially equal, or equated, on possible confounding variables at the outset of the study. Otherwise, this is a strong design. One approach to address group equivalence is to examine several premeasures that may be relevant to the DV prior to execution of experimental treatments. If results of these premeasures show the groups are similar, that will add to confidence that the results obtained can be trusted, although certainly replication studies are needed to ensure trust in the findings.

Example: Non-equivalent Control Group Design

Sun, K. T., Lin, Y. C., & Yu, C. J. (2008). A study on learning effect among different learning styles in a Web-based lab of science for elementary school students. *Computers & Education*, *50*(4), 1411-1422.

Note: See presentation video for discussion of this information.

Purpose: Compare science achievement between virtual-lab instruction and classroom instruction for 5th grade students.

The authors explain that 4 classes were used, two were assigned to the virtual lab and the other two served as the control group. Their wording appears below. Their original wording was confusing and could be read that individuals were assigned to the two conditions, which might create randomly formed groups, but they also wrote that classes were randomly assigned to treatments. It is clear, however, in the next paragraph that a quasi-experimental design was used, so groups were not randomly formed.

> in Kaohsiung City using random sampling. Sixty-five students from two of the classes were assigned to the experimental group where they received information-integrated Web-based lab teaching, while the other 67 students from the other two classes were assigned to the control group where they received traditional classroom teaching.
>
> *2.4.2. Experimental design*
>
> This study adopted a quasi-experimental design method to examine how the Web-based lab influences the effectiveness of teaching the natural sciences in elementary schools (Best & Kahn, 1989; Cohen, 1997). This experiment applied the nonequivalent-control group design to evaluate teaching effectiveness. Following a

They also provide a diagram illustrating their design, and this shows that pretests and posttests were used.

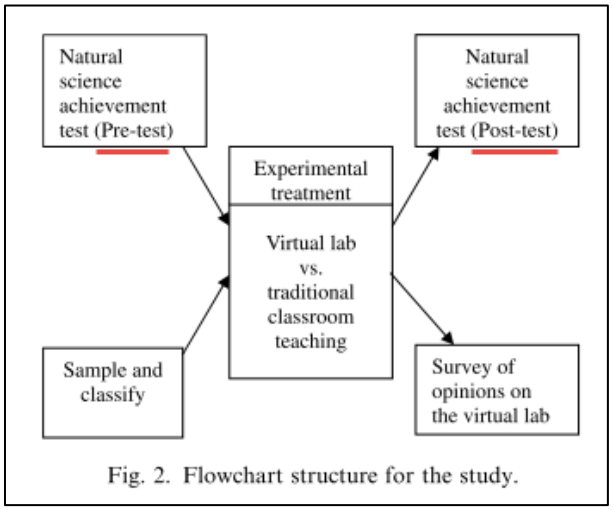


Fig. 2. Flowchart structure for the study.

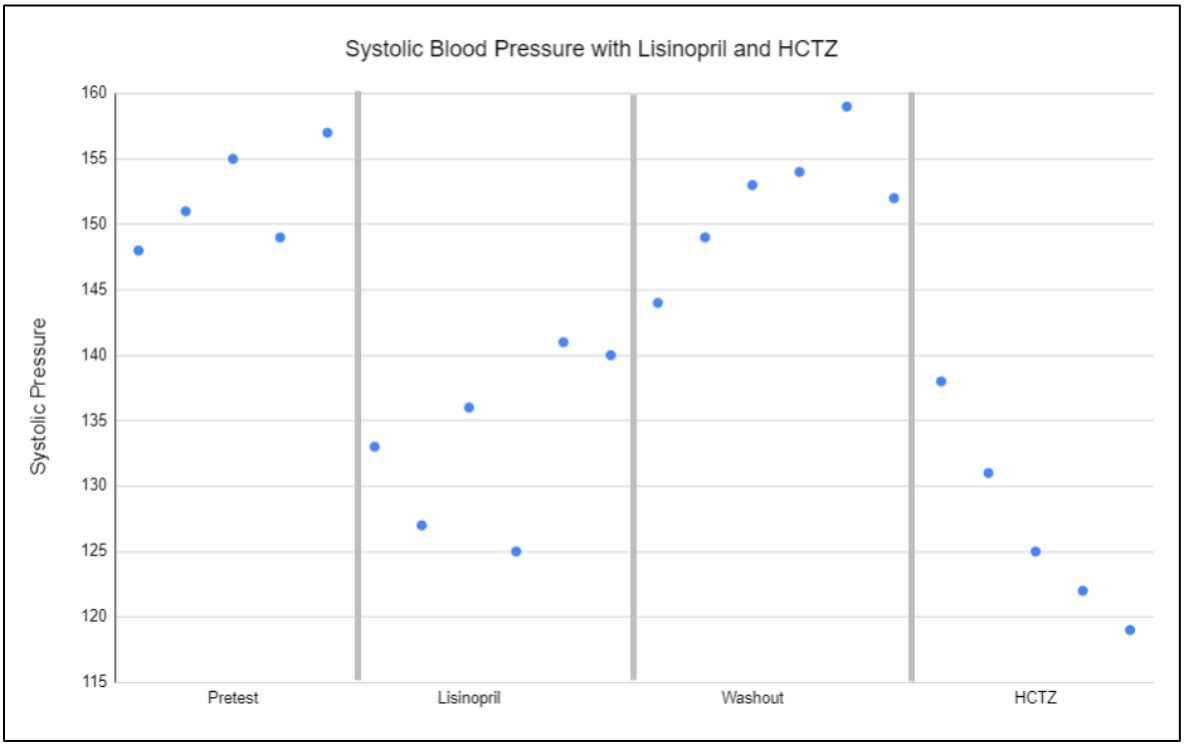
In sum, their design follows the classic nonequivalent control group design: O X_1 O and O X_2 O.

(b) Counterbalanced Designs (also called Cross-over and Switchback Designs)

These designs enable one to test multiple treatments on all participants. There is no separate experimental and control group. Instead, each group experiences both experimental and control conditions. Since groups are not randomly formed, this design is quasi-experimental. It would be possible to form a true experimental counterbalanced design by randomly forming groups, although that is rarely the case. These designs are popular because fewer groups and fewer participants are needed to test all conditions.

A serious drawback to these designs is the possible carryover effect from different treatments. If mathematics students learn addition with treatment A, then teaching addition with treatment B is now confounded because of the carryover effect from treatment A – students already have learned addition via treatment A.

These designs work well for treatments that have little carryover effect and dependent variables that tend to return to a normal state in time. Blood pressure would be one example. In drug trials, if one takes drug A to reduce blood pressure, over time drug A leaves the body, maybe one week to 10 days, so there is no carryover effect. After this washout period, drug B can now be tested on the participant to learn how well it reduces blood pressure. It is important to take regular readings of the DV between treatments to ensure the DV returns to a normal state. For blood pressure, one would take pressure readings several days before drug A, several days while using drug A, several days after drug A during the washout period, and several days while taking drug B. Graphically it would look very much like a single-subject study as illustrated below.



The basic schematic for a counterbalanced design follows. Three time periods were used, but this can be altered for 2 to whatever is a reasonable number of treatments for the phenomenon investigated.

Group	Time 1	Time 2	Time 3
A	X_1 O	X_2 O	X_3 O
B	X_2 O	X_3 O	X_1 O
C	X_3 O	X_1 O	X_2 O

Additionally, pretests/washout measures can be added if useful, as shown below.

Group	Time 1	Time 2	Time 3
A	O X_1 O	O X_2 O	O X_3 O
B	O X_2 O	O X_3 O	O X_1 O
C	O X_3 O	O X_1 O	O X_2 O

Example: Counterbalanced Design

Hill, L., Williams, J. H., Aucott, L., Milne, J., Thomson, J., Greig, J., ... & MON-WILLIAMS, M. A. R. K. (2010). Exercising attention within the classroom. *Developmental medicine & child neurology*, *52*(10), 929-934.

Note: See presentation video for discussion of this information.

Purpose: "To investigate whether increased physical exercise during the school day influenced subsequent cognitive performance in the classroom."

Procedure

The experiment had a counterbalanced design. Within each school, the two classes of similar size at each primary level were randomly designated to either group A or group B by a blinded member of staff. Given that schools did not stream these parallel classes (e.g. for sex or ability), this form of randomization made any between-group differences in age, sex, or IQ very unlikely. Group A received the CEP in week 1 and no CEP in week 2. Group B had this order reversed.

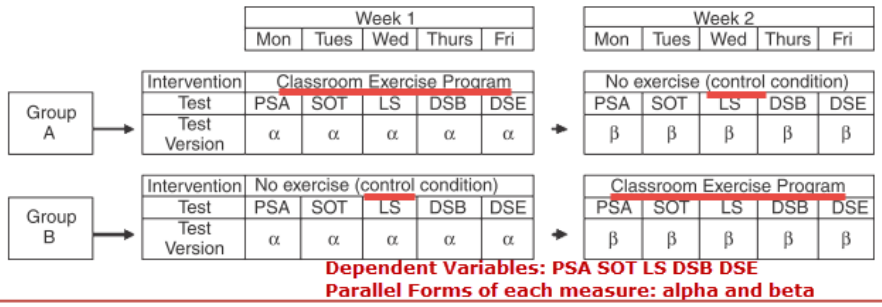


Figure 1: Experimental design: group A had physical exercise in week 1 and no physical exercise in week 2; group B had the opposite order. Two versions of each psychometric test were administered, α and β, to reduce practice effects on retesting. PSA, paced serial addition; SOT, size-ordering task; LS, listening span; DSB, digit-span backwards; DSE, digit-symbol encoding.

Their design involved two groups and two conditions, treatment and control. They provided a schematic above, and it is simplified below.

Group	Time 1	Time 2
A	X_1 O	X_2 O
B	X_2 O	X_1 O

Summary: Many quasi-experimental designs exist, and Campbell and Stanley (1963; see earlier reference above) provide an excellent introduction to design examples. Additional designs include time series, equivalent time samples, equivalent materials samples, several variations of separate-sample pretest-posttest designs, multiple time series, institutional cycle designs, and regression discontinuity.

5. Self-Test: Experimental Designs

Question

What is the name of this design?

R O X1 O
R O X2 O

Answer

The design above is known as a “pretest-posttest control group”.

Example

A teacher randomly assigns book A to half her students, and book B to the other half. After teaching a lesson using both books, a posttest is administered to learn of any achievement differences resulting from the books. What type of design is this and what is the schematic for it?

Answer

R X1 O
R X2 O

This is a “posttest only control group” design, why? There are randomly formed groups, R; no pretest; two treatments (X1 and X2); and a posttest O.

Example

A researcher wished to learn whether the perception of a person's height depends on that person's perceived status. A random sample of army inductees was selected and randomly divided into four groups. An actor gave a short address to each group separately, extolling the joys of army life. For the first group, the actor was dressed as a private; for the second, as a sergeant; for the third, as a captain; and, finally, for the fourth group, as a colonel. The inductees were asked to complete a questionnaire evaluating the speech. Among the questions was one asking for an estimate of the lecturer's height. Which design was used what is the schematic?

Answer

R X1 O
R X2 O
R X3 O
R X4 O

There is no pre-measure (pre-test), there are randomly formed groups, and there is a post-measure (post-test).

As above with the first example, this is a “post-only control group” design.

Example

A researcher is interested in how students interpret and respond to feedback in class on their academic performance (e.g., teacher’s comments, graded tests, papers). The researcher chooses to use "attribution theory" to guide the research. Essentially this theory specifies that each student has an attributional tendency that governs the way in which he will respond in feedback situations. That is, one student may attribute his B grade to his having studied carefully, another to his good fortune, and a third may conclude that the teacher likes him. The researcher wants to know whether giving all students consistently high grades will change their attributional patterns. A random sample of 50 students is drawn from a local middle school. Half are assigned, in a random fashion, to a treatment condition where the teacher has agreed to give them high grades regardless of their performance, and the other half are graded in a usual manner. The researcher sends a graduate assistant

to the various schools to administer his attribution instrument (25 statements regarding feelings toward evaluation that students are to rate on a 5-point Likert scale) before and again after the study is completed. During the course of the study the researcher volunteers to serve as a teacher's aid or para-professional in the school. The researcher observed the students and noted their comments and behaviors after receiving the teacher's comments and grades. The researcher took notes on students from both conditions (those who received high grades and those who were evaluated normally). After the study was completed, the researcher tallied the scores obtained on the attribution instrument and found that the group who received consistently high grades showed marked differences from their initial attribution scores. After examining the notes obtained while serving as an aid, the researcher also noted that the students who received consistently high grades developed different attribution patterns from those who were graded in the normal manner.

Which design was used and what is the schematic?

Answer

Consider the evidence provided:

(1 and 2) A random sample of 50 students is drawn from a local middle school. Half are assigned, in a random fashion, to a treatment condition where the teacher has agreed to give them high grades regardless of their performance, and the other half are graded in a usual manner. So we have randomly formed groups, and which treatment students received was manipulated by the experimenter.

(3) The researcher sends a graduate assistant to the various schools to administer his attribution instrument (25 statements regarding feelings toward evaluation that students are to rate on a 5-point Likert scale) before and again after the study is completed. So we have a pretest and a posttest (testing before and after the treatment).

We have for group 1:

Random Formation (R) Pretest (O) Treatment (X1) Posttest (O)

and for group 2:

Random Formation (R) Pretest (O) Different Treatment (X2) Posttest (O).

R O X1 O
R O X2 O

As these symbols indicate, the designed used was a "pretest-posttest control group" design.

Example

A teacher randomly assigns book A to her first-period class, and book B to her fourth-period class. Prior to using the books, a pretest was administered to assess current achievement levels. After teaching a lesson using both books, a posttest is administered to learn of any achievement differences resulting from the books. What type of design is this and what is the schematic for it?

Answer

No randomly formed groups – intact classes were used, so no R in the schematic. A pretest was used, two treatments (A vs. B), and a posttest administered. This would be a "non-equivalent control group design. "

O X_1 O or N O X_1 O
O X_2 O N O O

6. Internal and External Validity: Introduction

Validity, as used here, refers to fidelity of control to eliminate the effects of confounding variables (**internal validity**), and whether results of experiments or non-experimental studies (i.e., correlational and ex post facto) can be applied to other populations and settings distinct from the study participants and settings (**external validity**). These forms of validity are different from the **measurement validity** which focuses on measured scores and was covered earlier in the form of reliability and validity.

There is a movement among researchers to use validity as an all-encompassing construct that ranges from test scores to experimental study control and applicability to other settings and populations. I disagree with this approach because I believe it creates confusion where none should exist.

In short, when my presentations refer just to **validity**, I am referring specifically to **measurement validity** – reliability and validity of test scores, scale scores, etc. I sometimes also call this **test validity**.

Internal validity is different and does NOT focus on **measurement validity**, rather, **internal validity** is that aspect of a study that deals with determining which variables caused changes on the dependent variable. It is a focus on control – determining which variables affected the dependent variable.

For example, after an experiment the researcher proclaims, with confidence, that the treatment caused changes on the DV and there were no confounding variables – this is an issue of **internal validity**. The higher the level of internal validity for a study, the better one's ability to pinpoint which variables caused changes on the dependent variable. So internal validity is directly related to control, the better the control in a study, the higher the level of internal validity.

External validity represents the extent to which results from a study can be **generalized** to other people or other settings. The higher **external validity** for a study, the better the results of the study can generalize to other people and settings.

Thus, a study in Bulloch county schools may not generalize to schools in Atlanta or Montana. However, results of the study in Bulloch county may generalize well to other local counties near Bulloch.

For example, maybe one finds that reciprocal peer tutoring causes a 10% gain in achievement over lecture for high school students in Bulloch county—if this study had high **external validity**, then the results would generalize to other high school students elsewhere in GA or across the USA.

The way to increase **external validity** is to include a variety of people, settings, measures, etc. Note that geography is not really the issue, rather, cultural differences, attitudinal differences, SES differences, and things like that often limit **external validity** of studies.

Summary

One of the more common mistakes is to confuse **internal validity**, **external validity**, and **test validity** (content, predictive, concurrent, construct, etc.). Note that **test validity** is a different issue which concerns accuracy of scores from an instrument. Be sure to understand the difference among these.

Here's a quick recap:

- If the issue deals with validity of scores from a test or instrument, then that would be **test validity**.
- If the issue is focused on determining which IV affected a DV, then that is **internal validity**.
- If the issue focuses on whether results from a study can generalize to other people or settings, that would be **external validity**.