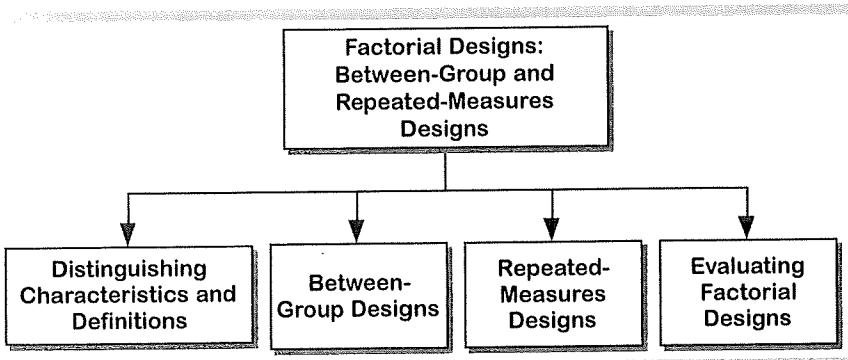
CHAPTER NINE

Factorial Designs

Between-Group and Repeated-Measures Designs



Chapter Eight introduced experimental research and demonstrated its key elements by focusing on the basic design, which includes one treatment and one control group and a pre- and posttest or posttest only, and which uses a *t*-test to establish statistical significance and to conduct hypothesis testing. This chapter introduces more complex experiments that use multistep statistical procedures.

CHAPTER OBJECTIVES

- ✓ Understand between-group and repeated-measures factorial designs.
- ✓ Understand how levels of a treatment are used in these designs.
- ✓ Understand the analysis of main effects, interactions, and covariates.
- ✓ Understand how to evaluate the quality of factorial designs.

DISTINGUISHING CHARACTERISTICS AND DEFINITIONS MAP 9.1

Factorial designs investigate the influence of two or more treatments (IVs) and/or levels of treatment in a single study. In addition, they can examine the influence of other variables.

Factors, Levels, and Effects

- *Factor* is the term applied to the main treatment or treatments (IVs), and is also used to refer to modifying or intervening variables that may influence the DV(s) in a factorial study.
- *Levels* are the varying categories within a factor.

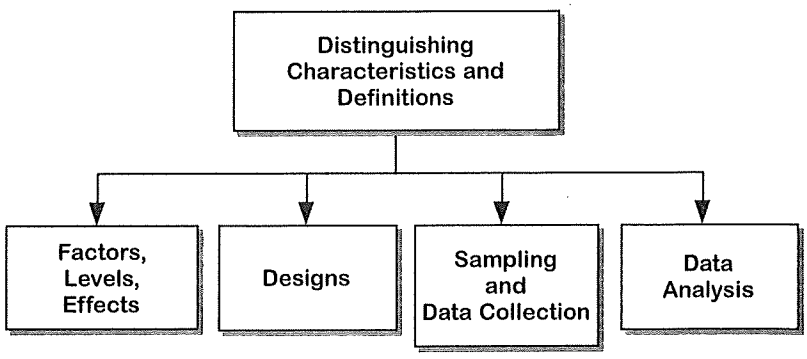
Factorial designs allow researchers to investigate both main effects and interaction effects.

- A *main effect* is the effect of each factor on the DV(s).
- An *interaction effect* is the uneven impact of the levels on the DV(s).

Designs

There are several factorial designs that fall within the following two broad categories:

- *Between-group designs* investigate differences between groups that receive varying treatments and/or levels of treatment
- *Repeated-measures designs* investigate differences within individuals who receive varying treatments and/or levels of treatment.



MAP 9.1

Sampling and Data Collection

In factorial studies, sampling and data collection are very similar to the procedures used in other experimental designs. (See Chapter Eight.)

Data Analysis

Because factorial designs consider the effect of one or more factors (IVs) and levels (and often involve two or more groups), they require inferential tests for significance and hypothesis testing that are more powerful than the *t*-test, which can only compare two groups with the same IV. The ANOVA/MANOVA and ANCOVA/MANCOVA, described in Chapter Eight, are used extensively in factorial designs to determine the statistical significance of the main and interaction effects of factors.

In factorial designs, the ANOVA and MANOVA calculate the main effect of (1) one factor with more than one level or (2) two or more factors, each with one or more levels. The ANOVA is used when there is one DV; the MANOVA is used when there are two or more DVs.

- In a between-group design with a single factor, the *one-way ANOVA* is used to analyze the data.
- In a between-group design with two or more factors, the *factorial ANOVA* is used to test all of the group differences with a single *F*-test.
- In a repeated-measures design, the *repeated-measures ANOVA* is used.

Both the ANOVA/MANOVA and ANCOVA/MANCOVA yield an *F*-value, which leads to a *p*-value that determines the overall main effects of the factors. The ANCOVA and MANCOVA calculate main effects when there is a *covariate*. The covariate may be (1) a difference in groups or individuals before the experiment begins and (2) the influence of another IV that the researchers seek to remove. They function in the same way as ANOVA/MANOVA, in that they calculate main effects and interaction effects by incorporating the covariate adjustments.

Because it may be the case that the main effect is significant and not all levels of treatment are significant, researchers want to know which treatment level or levels are the precise source(s) of the main effect. In order to do this, they perform a series of *t*-tests to isolate the source(s) of the effect. Statistical reasoning for the ANOVA/MANOVA and ANCOVA/MANCOVA procedures involves a three-step process: (1) determination of the overall main effects with a calculated *F*-score and *p*-value, (2) additional statistical tests to see whether there are interaction effects, and (3) follow-up *t*-tests and *p*-values for each of the possible combinations of treatment and control.

BETWEEN-GROUP DESIGNS

MAP 9.2

Between-group designs investigate the differences in two or more groups that (1) receive a treatment, (2) have multiple levels of one treatment, or (3) have more than one treatment with one or more levels. Between-group designs are classified as either *single-factor designs* or *factorial (multifactor) designs.*

Single-Factor Designs

Single-factor designs have one factor (IV) and two or more levels of the factor.

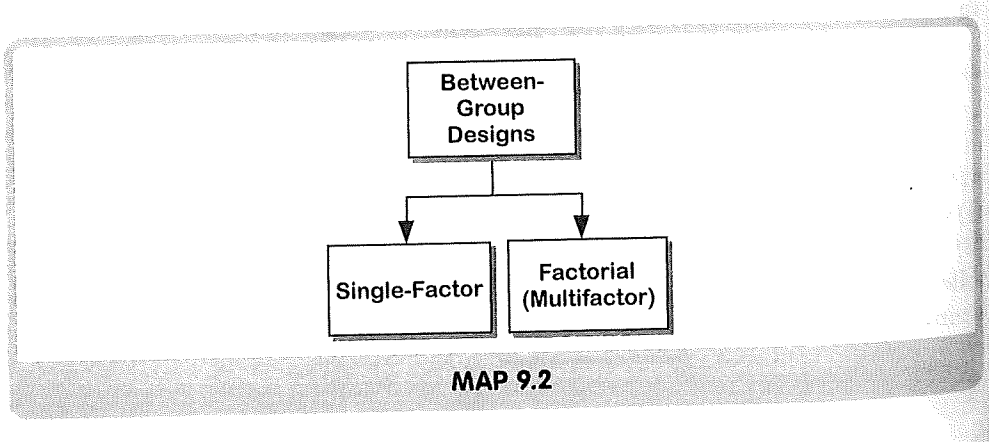
For example, a researcher could examine the effect of feedback to students on achievement and decide to have three levels of factors:

Single factor (IV) = feedback
- Level 1 = feedback as comments
- Level 2 = feedback as comments plus grades
- Level 3 = feedback as grades only

The researcher would assign the subjects in the sample to one of three groups:

Group 1 would receive comments as feedback.
Group 2 would receive comments plus grades as feedback.
Group 3 would receive grades only.

The researcher would analyze the findings by using a one-way ANOVA to calculate *F*- and *p*-values, followed by individual *t*-tests to determine which of the three levels had statistically significant results.



MAP 9.2

Factorial (Multifactor) Designs

Factorial designs (multifactor designs) have two or more factors and one or more levels for each factor.

For example, a researcher may decide to expand on the single-factor design described above by introducing the additional factor of gender. In this case, the design is called a 3×2 factorial design, because it has one factor with three levels and another factor with two levels. Factorial designs are always identified by the number of factors and the number of levels in each factor.

In the 3×2 factorial design of the effect of feedback, the researcher would assign subjects in the sample to one of six groups, each representing a different combination of factors and levels:

Group 1: Males/comments only
Group 2: Females/comments only
Group 3: Males/comments and grades
Group 4: Females/comments and grades
Group 5: Males/grades only
Group 6: Females/grades only

The researcher would use a factorial ANOVA to calculate *F*- and *p*-values, followed by individual *t*-tests to determine which of the three levels had statistically significant results. The advantage of a 3×2 factorial design is that it allows a researcher to test more than one hypothesis, to build additional variables into the design as factors, and to test interaction effects. The disadvantage is that it requires a large sample.

Below are four examples of between-group designs, with increasing levels of complexity.

1. Single-Factor Design with One-Way ANOVA

Gencosman, T., & Dogru, M. (2012). Effect of student teams–achievement divisions technique used in science and technology on self-efficacy, anxiety and academic achievement. *Journal of Baltic Science Education, 11*(1), 43–54.

This study examined one factor with one treatment group and two control groups. The experiment was designed to compare one treatment approach (cooperative learning strategies) against two comparable learning strategies deemed to be control conditions. Students were randomly assigned to all three conditions, and each group completed a curriculum unit in physics on "Force and Motion." Both control conditions contained lectures and whole-group discussions/questioning. In addition, the Control 2 condition used a constructivist approach, which encouraged

individual development of conceptual knowledge. (The researchers considered their design to have two control groups, but the Control 2 condition may be considered a treatment group.)

Factor = Teaching strategies

Treatment group = Cooperative learning (student teams)

Control 1 = Lectures and whole-group discussions

Control 2 = Constructivist approach, lectures, and whole-group discussions

There were three DVs:

DV 1 = achievement

DV 2 = self-efficacy

DV 3 = test anxiety

Each DV was measured on a pretest and a posttest. The study's design is illustrated in Figure 9.1.

The treatment, cooperative learning, was an approach carefully designed to reflect DVs. The results for self-efficacy showed a statistically significant difference among the three groups ($F = 10.37$, $p < .0001$), as did the results for test anxiety ($F = 10.79$, $p < .0001$) and the results for academic achievement ($F = 3.53$, $p < .03$).

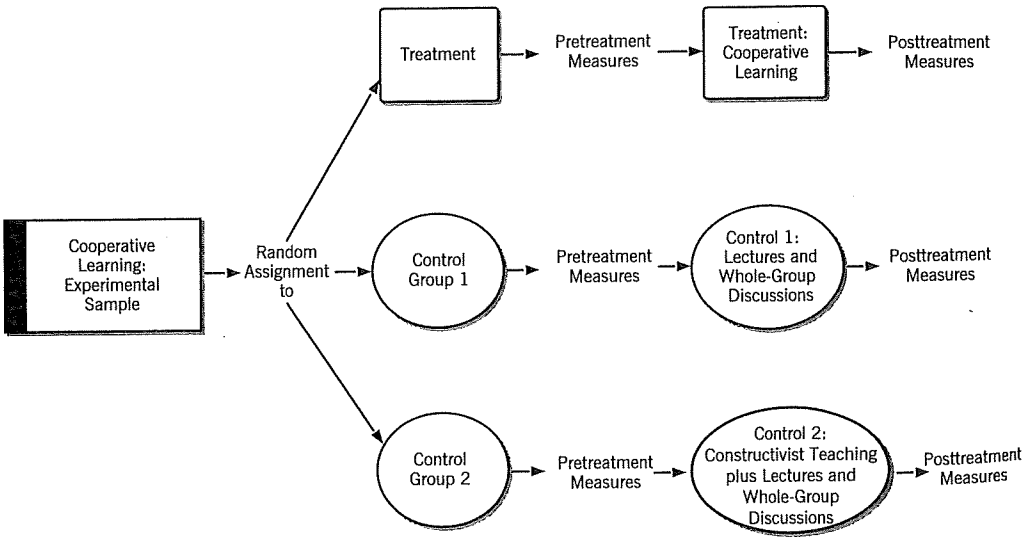


FIGURE 9.1. Design for the Gencosman and Dogru (2012) experiment.

2. Single-Factor Design with One-Way ANCOVA

Nesbit, J., & Adesope, O. (2011). Learning from animated concept maps with concurrent audio narration. *Journal of Experimental Education, 79,* 209–230.

This study investigated a series of hypotheses related to the impact of concept maps on undergraduate students' recall and understanding of the course topic of attribution theory. Three different forms of expert maps were created by the research team: (1) plain, black-and-white concept maps; (2) color-coded concept maps; and (3) animated concept maps, using visual icons. "Each participant was randomly assigned to one of four groups in a single factor, between-subject experimental design" (p. 218). The research hypotheses were as follows: (1) Participants receiving either the plain concept maps or the color-coded concept maps would outperform students who received the plain text only; (2) participants receiving the color-coded maps would outperform the students receiving plain maps; and (3) participants receiving the animated concept maps would outperform the students who received plain text, and students receiving color-coded maps would outperform students who received plain concept maps on measures of the central ideas and supporting details.

Factor = Concept mapping

Group 1 = Black-and-white concept maps (plain maps)

Group 2 = Color-coded concept maps (color maps)

Group 3 = Text-animated concept maps (prior text)

Group 4 = Control (plain text)

There were two DVs: free-recall short-term memory and achievement on a knowledge examination. Each DV was measured with a separate ANCOVA. The researchers used an ANCOVA in order to adjust for differences in scores on a midterm examination.

Covariate = Midterm exam scores

DV 1 = Free-recall short-term memory

DV 2 = Knowledge test (multiple-choice test)

Figure 9.2 represents this design visually.

The overall results of the ANCOVA were found to be statistically significant for the main effect of concept mapping on the knowledge measure ($F = 11.06$, $p < .001$). In addition, the researchers conducted *t*-tests to determine the specific sources of the main effect. There were statistically significant differences for color-coded concept maps versus text, and for plain concept maps versus text. However, they found no statistically significant differences between students who used color-coded concept maps and students who used plain, black-and-white concept maps.

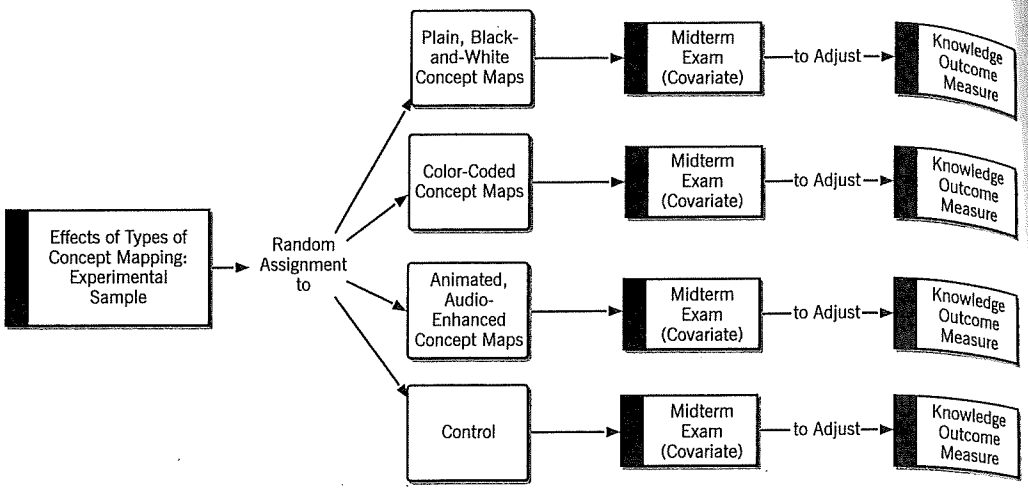


FIGURE 9.2. Design for the Nesbit and Adesope (2011) experiment.

3. *Factorial Design with 3 × 2 ANOVA*

Lim, K. Y., Lee, H. W., & Grabowski, B. (2009). Does concept-mapping strategy work for everyone?: The levels of generativity and learners' self-regulated learning skills. *British Journal of Educational Technology, 40*(4), 606–618.

The researchers conducted a 3 (treatment) × 2 (self-regulated learning) factorial analysis of concept-mapping strategies. They were interested in the effects of three levels of the concept-mapping treatment on two levels of students' self-reported self-regulated learning, as well as their effects on student knowledge of a topic. The participants were randomly assigned to three groups, each receiving a different concept-generating strategy. The three levels of treatment were these: Level 1 = expert-generated, Level 2 = partially student-generated, and Level 3 = fully student-generated.

The second factor was a categorical variable, student self-regulated learning. The researchers used a survey to divide students in each group into two categories: Level 1 = high self-regulated learning, and Level 2 = low self-regulated learning. In the end, therefore, six groups of students were compared: two groups (high and low self-regulated learning) using expert maps; two groups (high and low self-regulated learning) using partially learner-generated maps; and two groups (high and low self-regulated learning) using fully learner-generated maps.

Factor 1= Type of concept mapping

Group 1 = Expert-generated concept map

Group 2 = Partially learner-generated concept map

Group 3 = Learner-generated concept map

Factor 2 = Self-regulated learning

Group 1 = Low self-directed learning

Group 2 = High self-directed learning

DV = Achievement on knowledge and concepts

The design is illustrated in Figure 9.3.

The researchers analyzed the effect of each factor and its levels on knowledge acquisition, as measured by a multiple-choice test. The means and standard deviations for each factor and level are presented in Table 9.1.

The results of the 3 × 2 ANOVA indicated an overall difference in the level of concept mapping ($F = 4.43$, $p = .014$) and a statistically significant difference for the factor of self-regulated learning ($F = 7.95$, $p = .006$). Both the treatment with concept mapping and the factor of self-regulated learning therefore produced significant effects. Follow-up t-tests isolated these effects. For the concept mapping, the condition in which students fully generated their own concept maps performed better than the other conditions. In addition, the group with higher scores for self-regulated learning outperformed the group with lower scores. The test for the interaction of the IVs was not statistically significant.

4. Factorial Design with 3 × 2 MANOVA with Interaction

Carrier, S. (2009). Environmental education in the schoolyard: Learning styles and gender. *Journal of Environmental Education*, *40*(3), 2–12.

The researcher studied the impact of environmental education lessons by comparing "activities conducted in the schoolyard with traditional classroom activities involving elementary school boys and girls" (p. 2). In this study, the overall statistical test was a MANOVA, because four DVs were measured and analyzed: (1) achievement, (2) attitudes toward the lesson, (3) behaviors, and (4) comfort level with learning outdoors. The experiment examined the treatment variable (classes conducted in the schoolyard on environmental activities) versus the traditional classroom instruction, and added the mediating variable of gender (boys vs. girls).

IV = Treatment (outdoor learning included or not)

Group 1 = Use of outdoors (schoolyard) for instruction

Group 2 = Control (indoor, classroom only)

IV = Gender

Group 1 = Girls

Group 2 = Boys

DV = measures of knowledge, understanding, and attitudes

DV 1 = Achievement

DV 2 = Attitudes toward the lesson

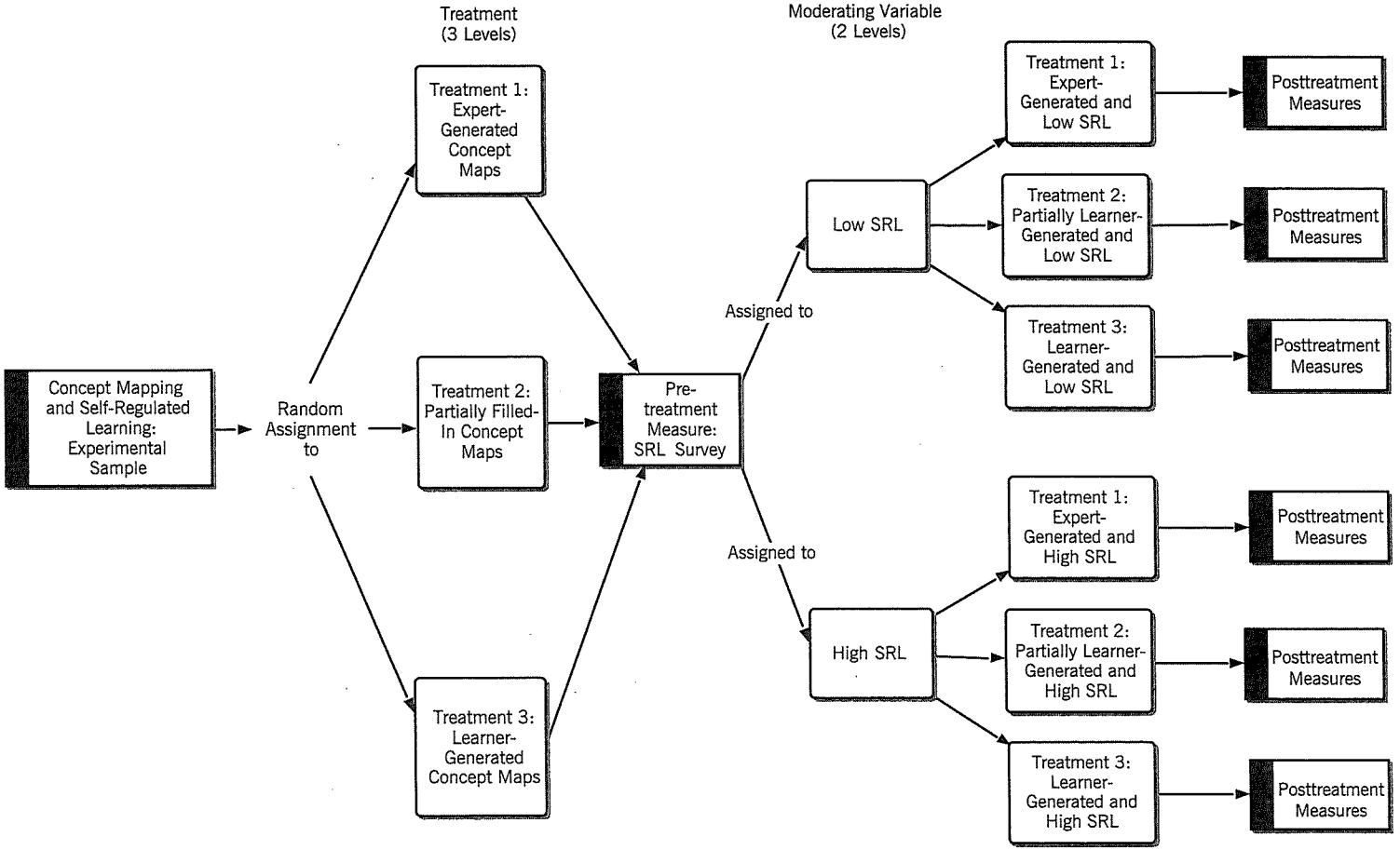


FIGURE 9.3. Design for the Lim, Lee, and Grabowski (2009) study.

TABLE 9.1. Means and Standard Deviations for Concept-Mapping Study (Lim et al., 2009)

Self-regulated learning skill level	Levels of concept map generativity		
	Expert-generated	Partially generated by learners	Fully learner-generated
Low	20.23 (8.2)	22.53 (8.6)	23.0 (9.7)
High	22.78 (9.4)	24.92 (8.4)	31.0 (5.8)

Note. Standard deviations are in parentheses.

DV 3 = Behaviors

DV 4 = Comfort level with learning outdoors

Figure 9.4 illustrates the design.

The researcher hypothesized the interaction effect that boys and girls would respond differently, and the results showed statistically significant differences for the interaction of treatment with gender ($F = 4.12$, $p < .01$). This means that boys and girls rated their experiences of schoolyard learning activities differently. Follow-up analysis was necessary, and it showed that comparisons of boys' treatment versus control scores on all four DVs were statistically significant, whereas there were no significant differences when girls in the treatment and control groups were compared.

The best way to describe the interaction is with a visual representation of the group means. In Figure 9.5, the scores are disaggregated into male and female groups, and the mean scores are placed on the graph. The graph shows the changes in the boys' scores between experimental and control groups for environmental attitudes,

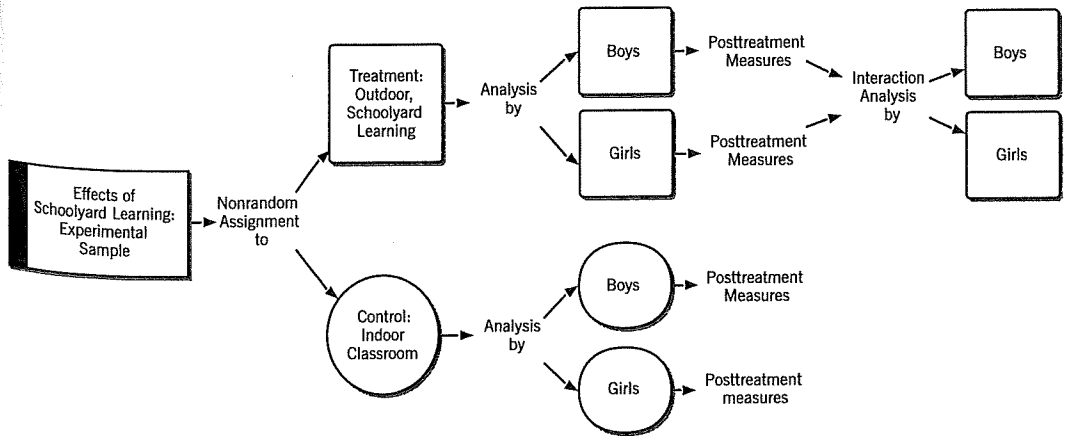


FIGURE 9.4. Design for the Carrier (2009) experiment.

FIGURE 9.3. Design for the Lim, Lee, and Grabowski (2009) study.

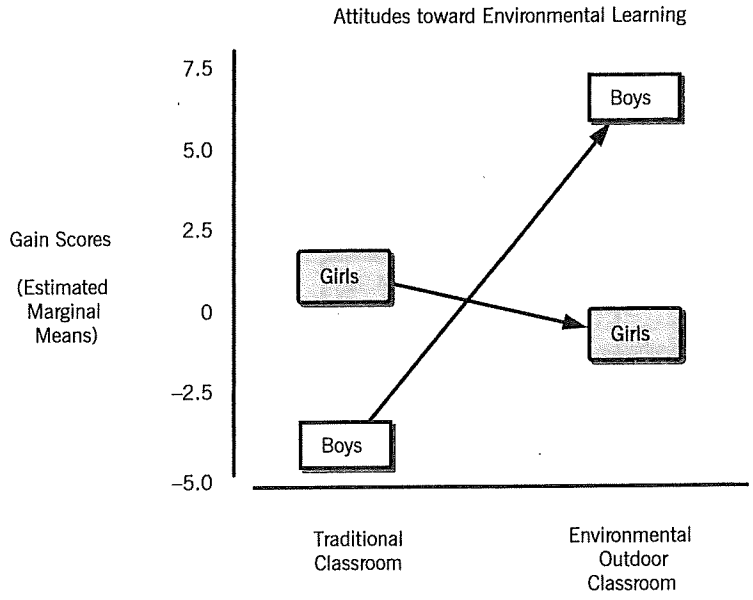


FIGURE 9.5. Interaction effects in the Carrier (2009) study.

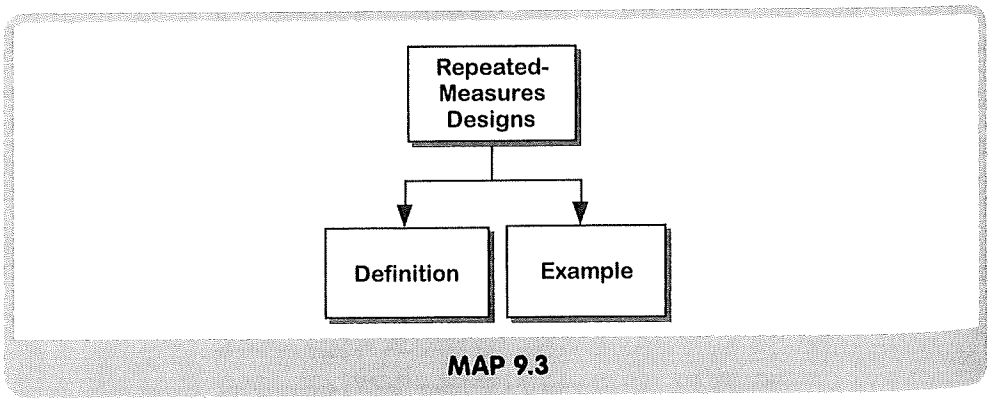
but shows little change in the girls' scores. The crossover of lines indicates the inconsistent effect of the treatment on males and females; this is the interaction effect.

REPEATED-MEASURES DESIGNS

MAP 9.3

Repeated-measures designs investigate differences in individual subjects in a group after treatments are introduced and the outcomes are repeatedly measured over time.

Other terms for repeated-measures designs are *time series designs* and *within-subject designs*. The most common repeated-measures design is the *crossover study*, which is a



longitudinal study in which every subject in every group receives a series of different, counterbalanced treatments. That is to say, each subject in each group receives the same treatments, but in a different order.

Repeated-measures designs are practical because they can reduce the logistical problems of conducting research in classrooms, and they can be done with limited sample sizes.

Example of a Repeated-Measures Crossover/Counterbalanced Design

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861–876.

This study investigated the effect of open-book and closed-book tests on long-term retention. A random sample of 36 subjects was recruited for the study from a human subjects pool at a university. Each subject experienced four treatment and two control conditions.

Treatment condition = Open-book

Open-book condition 1 = Read passage first and was then tested while rereading passage

Open-book condition 2 = Was tested while reading the text for the first time (simultaneous answering)

Treatment condition = Closed-book

Closed-book condition 1 = Read passage and was tested with book closed

Closed-book condition 2 = Read passage, was tested, and provided feedback on answers

Control condition

Level 1 = Read passage and was not tested

Level 2 = Was tested

The sample was divided into six groups, with six subjects in each group. Each subject in each group received all treatment and control conditions in alternating sequence, as represented in Figure 9.6.

The researchers predicted "that taking a test would enhance long-term retention more than studying the passage once . . . and that providing feedback after an initial test would produce a positive effect on long-term retention" (p. 863). The DV was a short-answer test matched with the reading passages. There were two measurements of achievement; one occurred immediately after the treatment, and the second took place 1 week later. The test questions were graded with a 3-point holistic rubric. "Two raters scored 10% of the tests and the Pearson product moment correlation between their scores was $r = 0.98$" (p. 865). A one-way ANOVA was used as a measurement check, and there was consistency in scoring.

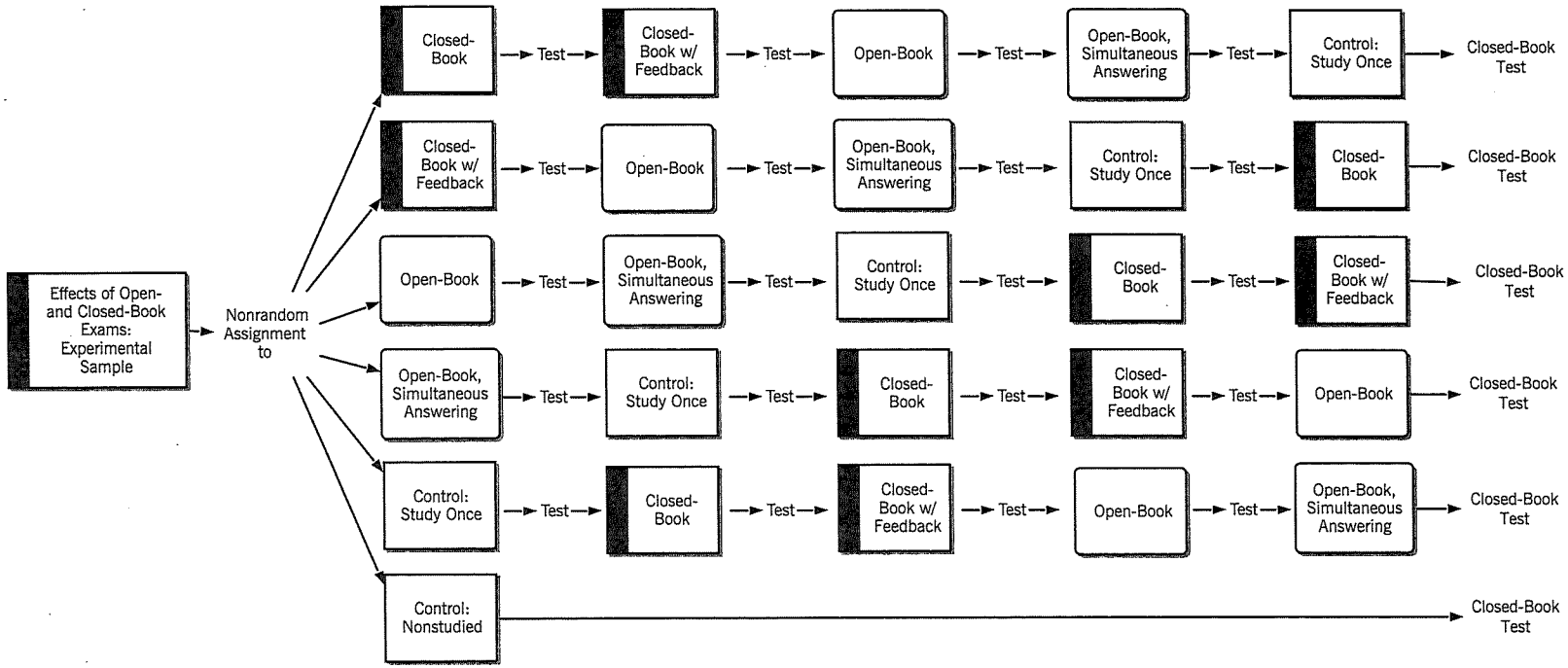


FIGURE 9.6. Design for the Agarwal, Karpicke, Kang, Roediger, and McDermott (2008) study.

The researchers used a repeated-measures ANOVA to calculate the main effect of the four test conditions on the DV of initial recall performance. The results were significant, $F(3, 105) = 11.27$. They then did multiple analyses of the four test conditions, using follow-up ANOVAs and *t*-tests to analyze the multiple main effects. On final measurement, both the open- and closed-book tests produced ES > 0.87. This repeated-measures study represents an efficient and complex approach to experimental design.

EVALUATING FACTORIAL DESIGNS

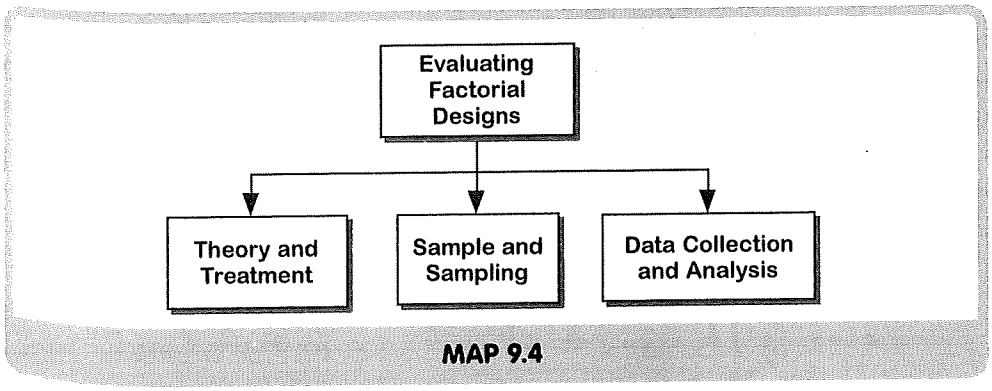
MAP 9.4

In evaluating factorial designs, a reader should apply the three categories of (1) theory and treatment; (2) sample and sampling; and (3) data collection and analysis, as first presented in the rubric for evaluating the quality of experimental designs in Chapter Eight (see Figure 8.5 there). Below is a sample evaluation of the Agarwal et al. (2008) repeated-measures study, which is organized according to these three categories.

Example: Evaluation of a Repeated-Measures Experimental Study

Theory and Treatment

The Agarwal et al. (2008) study merits a strong rating in the category of theory and treatment. The theory for the study of open- versus closed-book examinations is stated clearly and is very well supported in the literature review. There are over 40 references in the literature review, with over 40% of the citations very recent (published within the previous 5 years). The researchers state clear research questions concerning the use of testing situations to increase long-term retention of information: "We predicted that taking a test would enhance long-term retention more than studying the passage once. We also expected that providing feedback after an initial test would produce a positive effect on long-term retention" (Agarwal et al., 2008, p. 863). Overall, the IV was the use of open versus closed book in the testing conditions, but the researchers created levels of the IV to examine variations of these



MAP 9.4

conditions. The DV was a short-answer test matched with the reading passages (one each in history, science, and literature).

Sample and Sampling

The study merits only a moderate rating for sample and sampling. The researchers used a nonrandom sample of 36 subjects, who were recruited from the Human Subject Pool of the university's Department of Psychology. The repeated-measures experimental design used each individual participant as his or her own control, and each participant was involved in six treatment conditions. The advantage was that the comparison was the measure of within-subject change, and the sample did double duty as treatment and control. The disadvantage was that there might be some carryover between treatments. However, it would have been difficult to set power analysis because of the requirements for the design.

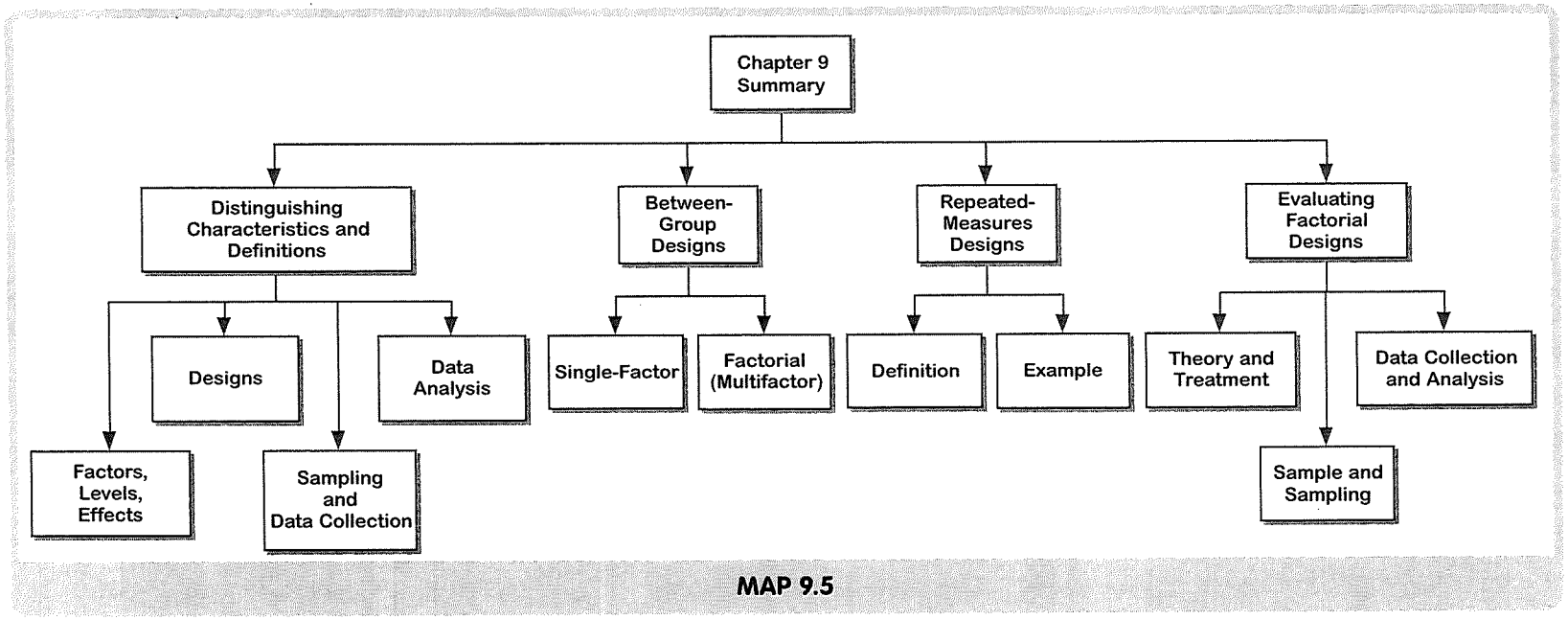
Data Collection and Analysis

The study merits a strong rating for data collection and analysis. The DV of major interest was performance on the short-answer tests given to subjects in the open- and closed-book testing conditions. The test questions required reading and grading with a 3-point holistic rubric. "Two raters scored 10% of the tests and the Pearson product moment correlation between their scores was $r = 0.98$. Given the high interrater reliability, one rater scored the remaining tests" (Agarwal et al., 2008, p. 865). The inferential tests matched the experimental design, with the use of a repeated-measures ANOVA and a series of follow-up ANOVAs and *t*-tests. Researchers performed a variety of statistical tests but were careful not to overanalyze, and they avoid overconcluding in their article. Effect sizes were published for all results, indicating the researchers' attention to the practical as well as the statistical significance of their findings.

CHAPTER SUMMARY

MAP 9.5

- ✓ Complex experimental designs can be organized into between-group and repeated-measures designs.
- ✓ These designs allow researchers to study and control multiple IVs and DVs.
- ✓ Factorial designs allow researchers to examine main effects and interactions of variables.
- ✓ Repeated-measures designs are used to study the changes of a treatment within individuals, who experience all control and treatment conditions.
- ✓ Theory and treatment, sample and sampling, and data collection and analysis are key elements in evaluating the quality and validity of experiments and quasi-experiments.


Chapter 9 Summary
Distinguishing Characteristics and Definitions
Between-Group Designs
Repeated-Measures Designs
Evaluating Factorial Designs
Designs
Data Analysis
Single-Factor
Factorial (Multifactor)
Definition
Example
Theory and Treatment
Data Collection and Analysis
Factors, Levels, Effects
Sampling and Data Collection
Sample and Sampling

MAP 9.5