SUGGESTED READINGS

Girden, E. R. (1992). *ANOVA: Repeated measures*. Newbury Park, CA: Sage.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Howell, D. D. (2009). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth.

Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (2nd ed.). Newbury Park, CA: Sage.

Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Thousand Oaks, CA: Sage.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Kirk, R. E. (1994). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.

Stevens, J. P. (2007). *Intermediate statistics: A modern approach*. Hillsdale, NJ: Lawrence Erlbaum.

Turner, J. R., & Thayer, J. F. (2001). *Introduction to analysis of variance: Design, analysis & interpretation*. Thousand Oaks, CA: Sage.

Chapter 12

Quasi-Experimental Studies

The major characteristics of a true experimental design, like those considered in the preceding chapters, include random assignment of participants to the various experimental and control conditions of the study. This procedure makes it more likely that groups initially are equivalent and therefore eliminates alternate explanations of results. In many instances, random assignment is not possible or not permitted by an administrator when, for example, special programs in institutions, medical settings, or mental health clinics need to be evaluated. Participants, such as those being treated after attempting suicide, abused children, drug addicts, and so forth, would not be amenable to random assignment. For these reasons, experimenters sometimes are forced to work with intact groups. The net effect is that the various groups cannot be assumed to be equivalent before treatment. Therefore, the studies are not considered to be true experiments but instead are referred to as seemingly experimental or **quasi-experimental designs**. A variety of such designs are available. Keep in mind that we are dealing with intact groups, not groups that have been randomly assigned to different conditions. Therefore, the important issue is to work with these possibly nonequivalent groups in such a way that we may still be able to draw cause and effect conclusions.

There are two basic categories of these designs. Repeated measures types obtain a number of measures of the dependent variable before and after the intervention (independent variable) on the same group of participants. Any change in behavior after intervention is less likely to be contaminated by the usual confounds. This is the rationale behind the **interrupted time-series study**. It is suitable for behaviors that can stabilize but still have room for improvement or

change in trend. These might include changes in accidents involving a driver who was under the influence of alcohol or drugs before and after legislation was introduced regarding drunk drivers, changes in rate of recovery for certain surgeries before and after hospital policy changed, or changes in some aspect of scholastic achievement before and after a critical curriculum change was introduced. To minimize some of the threats to internal validity (history, maturation, selectivity, pretest sensitization), a control group is added, if feasible. This group will not receive the treatment, but it is another intact group, not one that was formed by random assignment. The design is known as **multiple time-series study**.

Separate groups' quasi-experimental designs include experimental and control groups. Participants in each condition, however, are not randomly assigned; they are intact groups (e.g., separate classes of students, patients in different mental health clinics, workers in different parts of a factory). Therefore, groups cannot be assumed to be equivalent with respect to variables that might affect performance, and the designs are known as **nonequivalent control group designs**. First, there is the **posttest-only nonequivalent control group design**. In this case, posttreatment scores of an experimental group are compared with an untreated control group. We are talking about intact groups, so there is no basis for assuming group equivalency or for assuming that posttest differences might have occurred anyway, even without treatment. Another alternative is the **nonequivalent pretest–posttest control group design**, which is used so frequently in medical research and involves between-groups comparisons. It resembles the pretest–posttest design, which includes a control group that only gets the tests but differs from it in that the groups come from two different sources and hence are not considered equivalent at the beginning. They are preexisting rather than random groups. A third type of nonequivalent control group employs **cohorts**, groups of individuals who are similar in age, group membership, or both and, presumably, experience. Thus, second-year students in a Psy.D. program are considered to be cohorts. Performance in an Intermediate Statistics course might be compared before and after the introduction of a computer program, using the new group of students compared with those second-year students of 1 or 2 years ago. Should there be an improvement, it would be difficult to interpret. Admission policies may have become more stringent and resulted in brighter students in the new group, the faculty member may have improved teaching skills, the textbook may have been changed to a clearer one, and the experimental students may not be as phobic about statistics!

Although the factor of initial nonequivalence on the dependent variable cannot be controlled experimentally, it is sometimes possible to achieve some degree of equivalence statistically, by means of **analysis of covariance (ANCOVA)**. With this technique, an initial measure (pretest) of the dependent variable, or a variable that is linearly correlated with it, is made. This is the covariate or concomitant variable. Then, that part of variability of the posttreatment dependent variable associated with the covariate (r^2 for the entire set of posttreatment and covariate scores) is removed. The final analysis of variance (ANOVA) is performed on the remaining part of variability to determine how much is due to treatment and how much to experimental or random error. And the group means that are compared have been adjusted linearly.

As in any statistical procedure, there are certain assumptions made before the final analysis is performed. First, the relationship between the covariate and dependent variable is assumed to be linear, because this is the assumption underlying regression analysis, the basis for the adjustment. Second, the most glaring assumption is that the relationship between the covariate and dependent variable is about the same in all groups (i.e., all within-group slopes are equal). Therefore, there should be some indication that a test of "homogeneity of regression slopes" has been performed. If they are not equal, ANCOVA is inappropriate, and another statistical test would be performed. A final assumption relates to measures of the covariate. If these are other than standardized test scores, they are subject to errors of measurement, can affect treatment mean differences, and can produce misleading results. In checking for accuracy of the statistical test, df = number of groups – 1 for the numerator and total N – number of groups – 1 for the error term.

It is worth noting that there is considerable controversy concerning the use of ANCOVA as a statistical control procedure when experimental control is not possible. First, unless the covariate is measured before treatment is introduced, you cannot be certain that it was not affected by treatment (unless it is something like age, level of education, etc.). Therefore, removing its effect from total variance may also remove some of the treatment effect. Second, the intent of ANCOVA is to equalize groups on that variable by removing its effect. But such equalization may not be realistic. It would be similar to using initial weight as a covariate to compare weight gain by fathers and young sons after a special diet. Realistically, they are not equal in weight. There are other arguments, but the most potent is that ANCOVA does not produce equivalent groups from intact groups. Although these groups will be equivalent with respect to the covariate(s), there will always be other variables that differentiate intact groups, which otherwise would be distributed by random assignment. The researcher can only hope that critical ones that can just as easily account for the results have been isolated. Caution factors associated with quasi-experimental designs are found in Box 12.1.

Box 12.1 Caution Factors With Quasi-Experimental Studies

- History did not produce posttreatment performance.
- Maturation did not produce posttreatment performance.
- Pretest experience did not produce posttreatment performance.
- Changes in the measuring instrument did not produce posttreatment performance.
- A group was not specifically selected because it would react to treatment.
- Posttest performance is not due to regression toward the mean.
- Posttest performance is not due to interaction between select group and history or maturation.
- Posttreatment performance is not due to selective loss of participants.
- Nonequivalent control group is carefully matched with the experimental group.
- Covariate and other test measures are reliable and valid.
- Testing conditions are uniform.
- The experimenter is blind with regard to purpose of the study.
- The statistical assumptions of ANCOVA are met.

We will evaluate two studies, the first together and the second by you alone. Each introduces nonequivalent groups, but they are introduced for different reasons: one uses the group as an experimental group, and the second uses it as a control group.

STUDY EXAMPLE 12.1: "CARE-BASED MORAL REASONING IN NORWEGIAN AND CANADIAN EARLY ADOLESCENTS: A CROSS-NATIONAL COMPARISON"

The present study compares moral development in early adolescents from two countries: Norway and Canada. The Norway data are contemporary; the Canadian data were gathered at an earlier time. Both sets of data are compared with gender as a second factor.

The Study

Skoe, E. E. A., Hansen, K. L., Mørch, W.-T., Bakke, I., Hoffmann, T., Larsen, B., et al. (1999). Care-based moral reasoning in Norwegian and Canadian early adolescents: A cross-national comparison. *Journal of Early Adolescence, 19*(2), 280–291. Copyright © 1999 by Sage.

The theoretical and empirical foundations of psychology are rooted deeply in Western culture, especially that of the United States. . . . The area of moral reasoning is no exception. Most theories and research have been North American. However, development always occurs in a cultural context . . . and socialization varies both across cultures and across countries within cultures. Although prominent theories of moral development . . . stress social experience, little research has examined cultural variation, especially in the case of the Gilligan . . . theory.

Gilligan . . . has proposed that an "ethic of justice" characterizes the moral voice of men, and "ethic of care" characterizes women. . . . However, whereas some studies . . . have found gender differences in moral orientation (care/justice), others have not. . . . Both genders resort more to the care orientation than to justice when discussing relational real-life dilemmas, and both tend to use justice more than care when discussing nonrelational dilemmas. . . . Women's tendency to report more relational conflicts results in their showing more care responses than do men. . . .

Extensive research on gender differences . . . has tended to obscure a second important implication of the Gilligan . . . work: the concept that care reasoning, like justice reasoning, follows specific developmental pathways and varies individually. The Skoe and Marcia . . . Ethic of Care Interview (ECI) measures the following five hierarchical levels of development in the care ethic, in accordance with the Gilligan . . . theory: exclusive self-concern, questioning of self-concern as a sole criterion, primarily other-concern, questioning of other-concern as a sole criterion, and balanced self-and-other concern. The ECI uses a real-life dilemma in addition to standardized dilemmas involving care issues. Recent research . . . validates a care-based approach to moral development. . . . The ECI has been found to be related positively to age, ego development, the Kohlberg justice stages, androgyny, empathy, and prosocial behavior. . . .

No gender differences have been observed on the ECI among late adolescents and young adults in North America or in Norway. . . . However, in a study of Canadian early adolescents . . . girls scored higher than did boys on the ECI, and more girls scored at ECI Level 2 (conventions of goodness, caring for others). These gender differences, occurring in early adolescence but not in young adulthood, are consistent with North American research in related areas such as justice-based moral development . . . ego development . . . and prosocial moral reasoning. . . . To date, no research has investigated levels of care in Norwegian children.

. . . The early adolescent gender differences commonly reported in North America are not universal necessarily. . . . To the extent that national social contexts vary, developmental patterns also might differ. . . . Values vary cross-nationally, reflecting different mental programs that are developed in the family and

reinforced in schools, organizations, and the media. One such program concerns gender role expectations.

The present study compared care-based moral reasoning and real-life moral dilemma content in Norwegian male and female early adolescents with data obtained previously from Canadians of the same age (Skoe & Gooden, 1993). There is reason to believe that gender role socialization differs in Scandinavian and North American contexts . . . American gender concepts traditionally [are] more stereotyped than were those of Sweden. Swedish women were assigned more instrumental traits (e.g., ambitious, hardworking, willing to take a stand); American women were ascribed more expressive ones (e.g., compassionate, caring, eager to soothe hurt feelings). Only the Swedish women were described as liberated.

▶ *(Note that many American woman, myself included, would look at these findings and wonder which American country was surveyed. The study was conducted in 1988, but the title of the book refers to American youths.)*

. . . Recent research has shown that Scandinavian girls and boys alike value self-sufficiency and independence. . . . Therefore, fewer gender differences in care-based moral reasoning would be expected [in] Norway than in Canada.

. . . In Norway, as compared to North America, boys and girls both tend to demonstrate traditionally masculine instrumental traits of self-reliance and independence. . . . Unlike Americans . . . Norwegian adolescent girls and boys have been found not to differ significantly on masculinity measures. . . . Thus, Norwegian norms for feminine behavior might emphasize individuation more, and interpersonal connection less, than might Canadian norms. . . .

. . . In the present study, fewer gender differences were predicted in Norway than in Canada and stronger national differences for girls than for boys. Biological and social changes focus early adolescents' attention on gender role behavior in new ways. Norway and Canada might differ even more in gender expectations for females than for males. If so, Canadian and Norwegian girls might differ more dramatically than might boys from both countries in the gender role content that is important in early adolescence. Thus, it was expected that Norwegian and Canadian girls would differ on care-reasoning level and on self-reported real-life dilemma content, whereas no such differences would be found for boys. Specifically, because female role expectations and socialization practices traditionally stress interpersonal connection, it was predicted that more Canadian than Norwegian girls would score at ECI Level 2 (caring for others) and generate relational real-life dilemmas.

▶ *(Keep in mind that data from Canada were collected about 7 years earlier than data in Norway.)*

1. What was the rationale for the study?

Research on moral reasoning has been conducted mainly in North America. But development occurs within a culture, and socialization varies between cultures and in countries within cultures. Its development in different cultures hardly has been studied, particularly regarding Gilligan's theory. Gilligan proposed that male moral orientation is dominated by justice ethics, whereas a female's orientation is dominated by care ethics. Research support is ambiguous, but orientation may depend on the type of conflict generating the orientation. Relational dilemmas generate a care orientation; nonrelational dilemmas generate a justice orientation.

An implication of Gilligan's theory that has not received much attention is variations in the development of care reasoning. The ECI measures five levels of its development, from self-concern to a balance between concern for the self and concern for others. A study of North American and Norwegian late adolescents revealed no gender differences in ECI. But in a study of Canadian early adolescents, girls scored higher than boys on the interview, and more girls scored at the second level of development. No study investigated Norwegian children.

2. What was the purpose of the study?

This study compared moral reasoning and real-life moral dilemma content of Norwegian early adolescents with data obtained earlier on Canadian early adolescents. Because of differences in gender role expectations between countries, fewer gender differences in moral reasoning were expected among Norwegian than Canadian children. But because gender expectations are emphasized more for females, greater between-country differences were predicted for females. In particular, more Canadian than Norwegian females were predicted to score at the second level of development (caring for others) and to generate more relational real-life dilemmas.

Method

Participants

The Norwegian sample consisted of 79 (45 girls and 34 boys) fifth- and sixth-grade early adolescents, with a median age of 12.0 years, from a small coastal city in northern Norway. The Canadian sample, also used in a previous study . . . consisted of 46 sixth grade students (23 girls and 23 boys), with a median age of 11.8 years, from a Canadian East Coast town. All participants were Caucasian and volunteered to participate. The Canadians, who started school 1 year earlier, had more education (= 6.0, $SD = 0.0$) than did the Norwegians (= 5.4, $SD = 0.50$), $F(1,123) = 63.10$, $p < .001$, indicating a need to control for grade level in analyses.

▶ *(Note that the Norwegian location is described as a "city," whereas the Canadian location is described as a "town," a unit that is generally smaller. Moreover, median ages are reported, and this is usually the statistic of choice with the presence of extreme scores. Median for Norwegian children is 12 years, and they were from fifth and sixth grades. Canadian children's age is 11.8, and they were from sixth grade. In addition, a standard deviation (SD) of 0.00 is highly unusual but in this case indicates that all the Canadian children had attended school for exactly 6 years, whereas the Norwegian children had attended school for 5 or 6 years, with more attending for 5 years because the mean is below 5.5. What is important is that if Canadian children had more education, they may have had extra gender role training as well, if this is conducted in the school. If this is true, grade level per se is not an appropriate covariate. Finally, nothing is reported about economic level of the children nor their family status—for example, marital status of the parents.)*

3. Who were the participants?

The Norwegian sample included 45 girls and 34 boys, all volunteers and all Caucasian, from fifth and sixth grades, who lived in a small coastal city. The Canadian sample, tested in 1993, included 23 girls and 23 boys, also Caucasians, who volunteered, from sixth grade, who lived in a coastal town. Median ages were 12 and 11.8 years, and Canadian children had more education than the Norwegian children.

4. What are some questions regarding group equivalence, except for country of origin?

The Norwegian children were exposed to social and cultural practices that occurred 7 years later than those to which Canadian children were exposed in their country. Also there may have been bigger differences in ages between the two groups than are reflected in the medians. Moreover, Canadian children may have experienced more training in gender role expectations because of their extra year of education. Finally, economic levels may not have been equivalent, and this might affect the training the children received.

Procedure

Permission to conduct the study was obtained from the school administration. Subsequently, consent was obtained both from the children and their parents. A female researcher first individually administered the real-life dilemma and then an

interpersonal dilemma from the children's version of the ECI . . . chosen for its cross-national appropriateness: "Siri/Per has been invited by her/his friend, Kari/Jon, for dinner on Friday. The next day, another friend, Vigdis/Nils, invites Siri/Per to a big birthday party on the same Friday. What do you think Siri/Per should do?" The real-life dilemma was elicited first to avoid biasing real-life dilemma choice by providing an example beforehand. Participants described a personal moral conflict and then standard probes were used, such as "What were the conflicts for you in that situation?" and "Did you think it was the right thing to do?" The standardized dilemma was read aloud to the participants while they read along. Probes such as "What do you think Siri/Per should do?" and "What would you do if you were in the same situation? Why?" were used to examine dilemma reasoning.

▶ *(Note that noncounterbalancing was deliberate, but justification would have been stronger with experimental support of the assumption that the self-report would have been influenced by the presented dilemma. Note, too, that a female tested both males and females.)*

The audiotaped interviews were scored according to the *Ethics of Care Interview Manual*. . . . Level 1 is survival (caring for self). The aim is basically to protect self; there is little care for others. Level 1.5 is transition from survival to responsibility. Although aware of the needs of others, the person still will favor self-interest in relationships. Level 2 involves conventions of goodness (caring for others), characterized by a concept of responsibility. *Good* is equated with self-sacrificial care for others, and *right* is externally defined, often at the expense of self-assertion. Level 2.5 concerns transition from a conventional to a reflective care perspective and from goodness to truth and honesty in relationships. Commitment to others still is important, but there is more flexibility, thoughtfulness, and struggle with the dilemmas than at the earlier levels. Level 3 fully realizes the ethic of care (caring both for self and others). This perspective focuses on the dynamics of relationships and dispels the tension between selfishness and responsibility through a harmonization of the needs of self and others. . . .

Level scores for each of the two dilemmas were added, yielding a total score ranging from 2.0 through 6.0. Overall level scores then were halved and rounded to the nearest .5 level (e.g., 1.88 = level 2, 1.13 = Level 1).The correlations of care reasoning levels across the two dilemmas for the Norwegian and Canadian samples were .66 and .67, respectively (both $ps < .0001$). Two independent judges rated a random sample of 50 ECI tapes (25 girls and 25 boys). Interscorer agreements for the real-life and Siri/Per dilemmas were 94% and 96%, respectively (Cohen's kappas were .91 and .94). . . . The real-life dilemmas also were scored as either relational

(involves an ongoing significant relationship) or nonrelational (involves a person whom the participant does not know well or an issue primarily intrinsic to self). Two raters classified independently all dilemmas with 100% agreement.

▶ *(Note that no mention is made of where interviews took place, nor the time of day, nor the length of time it took to be interviewed. Moreover, the scoring range is narrow, and no justification is given for halving overall scores. Nor is it clear who the scorer was.)*

5. What was the procedure?

Participants were tested by a female research assistant. Each first described a real-life dilemma and its resolution. Then each was read an interpersonal dilemma and was asked to resolve it. Order of the two deliberately was not counterbalanced to control the effect of the standard dilemma on the one elicited. Each audiotaped interview was scored for level on the ECI. Two judges rated a random sample of tapes of 25 girls and 25 boys. Interscorer agreements for each were 94% and 96%. Finally, each real-life dilemma was classified as relational or nonrelational.

6. What were some questionable aspects of the procedure?

Only a female assistant interviewed the children, and assistants had to differ for the Canadian and Norwegian children. Testing conditions were not described at all. Standardization procedures for the interview were not reported, nor were validity or reliability coefficients reported (except for the 50 random interviews). Moreover, scoring procedure is not clear and appears to result in a very narrow range of scores.

Results

A 2 (Gender) × 2 (Norway/Canada) [ANCOVA], with ECI total score as the dependent variable and grade as a covariate, revealed a significant interaction, $F(1, 20) = 8.05$, $p < 01$. Follow-up one-way analyses showed that Canadian girls scored significantly higher than did Canadian boys, $F(1, 44) = 8.94$, $p < .005$. There were no significant gender differences in Norway, $F(1, 77) = 33$, n.s. Also, whereas Canadian girls scored significantly higher on the ECI (= 3.53, $SD = 0.53$) than did Norwegian girls ($\bar{x}$= 2.95, $SD = 0.51$), $F(1, 66) = 19.22$, $p < .0001$, Canadian ($\bar{x}$ = 3.02, $SD = 0.63$) and Norwegian ($\bar{x}$ = 3.02, $SD = 0.60$) boys did not differ significantly, $F(1, 55) = 1.00$, n.s. There was no significant effect for the covariate.

	Norway	*Canada*	
Male	3.02 (0.60;34)	3.02 (0.63;23)	3.02 (57)
Female	2.95 (0.51;45)	3.53 (0.53;23)	3.15 (68)
	2.98 (79)	3.28 (46)	3.09 (125)

$MS_{int} = 34(3.02 - 3.02 - 2.98 + 3.09)^2 + \ldots + 23(3.53 - 3.15 - 3.28 + 3.09)^2 = 2.4365$ (df=1).
$MS_e = [33(.60)^2 + \ldots + 22(.53)^2]/121 = 0.316$.
$F = 2.4365/0.316 = 7.710$.

▶ *(We have enough information to set up a table. We can do minimal checking because the adjustments for the covariate are made on between-group and interaction mean squares as well as error. But because we are unsure about the appropriateness of the covariate [including its relationship with ECI, which is supposed to be linear and the same for both countries], we can calculate an unadjusted F ratio and see how it compares with the obtained F.)*

▶ *(Interaction is significant at p < .01, so the covariate added some power to the test. Note, however, that the post hoc tests were inappropriate. Usual F tests were performed, which were wrong on two counts: The error term should have been the adjusted one from the original analysis, and the means should have been adjusted for the covariate. Nonetheless, the findings are accurate; the difference between the means for the Canadian males and females is significant.)*

A 2 (Norway/Canada) × 3 (ECI level) chi-square analysis was conducted on the participant frequencies for each ECI level (see Table 12.1). More Canadian than Norwegian girls scored at ECI Level 2, and more Norwegian than Canadian girls scored at the two lowest levels (phi = .47), $\chi^2(2, N = 68) = 15.12$, $p < .001$. In contrast, there were no significant differences among the boys.

▶ *(Note that these statements can be checked by performing chi-square tests on the data for females and males. The test for females indeed yields 15.12, as the authors report. Now, each cell can be examined to determine the extent to which observed frequency differs from its expected frequency. In fact, only one difference is significant: For Canadian girls who responded with Level 2 responses, the observed frequency is 14 and the expected frequency is (21 [row total] × 23 [column total])/68 (total girls) = 7.103. Then, z = observed frequency − expected frequency / $\sqrt{\text{expected frequency}}$ = (14 − 7.103) / $\sqrt{7.103}$ = 2.2588, p = .0048. No other difference is significant. Therefore, we can conclude only that more Canadian girls than Norwegian girls responded with Level 2 responses.)*

Table 12.1 Frequencies of Participants in the Ethic of Care Interview Levels

	Norway		Canada	
ECI Level	*%*	*n*	*%*	*N*
Girls				
1	20.0	9	4.3	1
1.5	64.4	29	30.4	8
2	15.6	7	60.9	14
Boys				
1	20.6	7	21.7	5
1.5	50.0	17	52.2	12
2	29.4	10	26.1	6

Note: ECI = Ethic of Care Interview.

Table 12.2 presents participant frequencies for the relational and nonrelational real-life dilemma categories. A 2 (gender) × 2 (relational/nonrelational) chi-square analyses showed no significant gender differences for the Norwegians (phi = .07), $\chi^2(2, N = 46) = 8.26$, $p < .005$. Also, significantly more Canadian than Norwegian girls generated relational dilemmas, and more Norwegian than Canadian girls generated nonrelational ones, $\chi^2(1, N = 68) = 4.63$, $p < .05$. For boys, there were no significant national differences.

Table 12.2 Content Analysis of the Nature of the Relationship in Real-Life Dilemmas

	Norway		Canada	
Nature	*%*	*n*	*%*	*n*
Girls				
Relational	82.2	37	100.0	23
Nonrelational	17.8	8	0.0	0
Boys				
Relational	76.5	26	69.6	16
Nonrelational	23.5	8	30.4	7

▶ *(Here, too, statements can be checked by looking at individual cells. Note that four chi squares were calculated: for Norwegians, Canadians, females, and males. No differences are found for Norwegians. Two differences are found for Canadians: females gave fewer nonrelational responses than expected, (0 − 3.5)/ = −1.871, and males gave more nonrelational responses than expected, (7 − 3.5)/ = 1.871. This means that males gave more nonrelational responses than females, but no differences were found regarding relational responses. For females, one difference is found: Canadian females gave fewer nonrelational responses than expected. Although all Canadian females gave more relational responses, the number did not differ from what is expected by chance.)*

7. What were the results regarding use of caring solutions to the dilemmas facing boys and girls of the two countries?

First, Canadian females responded at a higher level of caring than did Norwegian females, but a similar difference was not found among males.

Second, more Canadian females responded at Level 2 of the ECI than did Norwegian females. Although the authors report that more Norwegian than Canadian females responded at the lower levels, none of these frequencies were significantly different from what was expected by chance. Third, the authors reported that Canadian females generated more relational real-life dilemmas than did Canadian males, and Norwegian females generated more nonrelational real-life dilemmas than did Canadian females (who generated none). Analysis of the data reveals that, in fact, Canadian males generated more nonrelational responses than females and that Canadian females made fewer nonrelational responses than Norwegian females.

Discussion

. . . Canada and Norway are urbanized, industrial Western countries with similar characteristics and values, but their social norms for gender behavior differ. This likely influences how young people reason morally about the needs of self and others. The present study casts doubt on the universality of developmental gender differences in care-based moral reasoning. . . . As predicted, early adolescent Canadian girls scored higher than boys on the ECI and generated more relational real-life dilemmas. In contrast, Norwegian boys and girls scored similarly on these measures. The Canadian results are consistent with the North American theory and research. . . . However, the Norwegian results do not follow that pattern. As predicted, there were significantly fewer girls at ECI Level 2 (conventions of goodness, caring for others) in Norway (15.6%) than in Canada (60.9%). Also, as predicted, significantly more Canadian girls reported relational real-life dilemmas

(100%) than did Norwegian girls (82.2%). These results are congruent with the . . . finding that American adolescents were more gender-stereotyped than were Swedish adolescents and . . . observations that in Norway, unlike in North America, girls and boys both tend to focus on self and independence.

. . . Gender stereotyping, which traditionally has encouraged females toward care for others . . . seems to remain stronger in North America than in Norway, where the current cohort of young adolescents [has] available stronger social models of female independence and participation in public life, and an environment of greater gender equality. Thus, girls in North America and Scandinavia likely perceive somewhat differing sets of characteristics as gender appropriate. The stronger Norwegian tendency to encourage girls toward self-assertion and achievement probably focuses them on their own interests, rather than on primary concern for others.

. . . Canadian girls simply might acquire Level 2 reasoning early (compared to Norwegian girls and to boys in both countries) because early adolescent sensitivity to gender expectations differentially orients them toward such reasoning at a younger age. . . .

The gender differences of early adolescence are sometimes attributed to biological factors, such as a growth spurt for girls in physical and cerebral cortical maturation . . . or to social experiences. . . . The present findings support the latter explanation. Surely, there are many major life experiences behind gender . . . and several are determined culturally. Longitudinal cross-cultural research that addresses both the ethics of justice and care is necessary to assess further developmental issues in morality and to delineate the roots of individual as well as cultural differences.

8. What did the authors conclude?

Because social norms for gender behavior differ in Norway and Canada, these differences likely influence moral reasoning about needs for the self and for others. Canadian females scored higher on the ethics scale and generated more relationship real-life dilemmas than males. Norwegian males and females scored alike. Fewer Norwegian than Canadian females responded to the dilemmas at Level 2, and more Canadian than Norwegian females responded to the real-life dilemma with relational responses. These findings support the notion that North American adolescents are more gender stereotyped than are Scandinavian females and that Norwegian males and females are more focused on independence and the self. Stereotyping of the feminine role to care more for others seems stronger in North America than in Scandinavia. Thus, gender-appropriate

responses are perceived differently by adolescents of both cultures. Canadian females may respond more at Level 2 because this orientation is taught at a younger age.

9. Are these conclusions justified?

On the basis of some statistical tests, two conclusions appear justified. Canadian females scored at a higher level than their male counterparts on the ECI, and more Canadian than Norwegian females responded to a real-life dilemma with relational examples. The remaining conclusions are negated by inappropriate interpretations of significant overall chi-squares. More important, the selection of a group that was assessed 7 years earlier makes the conclusions not justified. First, the critical factor of history cannot be ruled out as a confound. Social changes had to have occurred during this period of time, which might have put the Norwegian adolescents at an advantage and might have changed the results from what they would be, had the Canadians been measured at the same time. Second, Canadians had more education than the Norwegians, and the use of grade as a covariate (which may or may not be linearly related to ECI) did not adequately control for its effect. If training goes on at school, as it probably does, it would have been a better design to test both groups at equal grade levels and use age (which is related to ECI) as a covariate. Third, testers had to differ, and each may have had different effects on the adolescents (especially if gender affects responses). Fourth, the test itself may not fully measure level of development of moral reasoning.

STUDY EXAMPLE 12.2: "THE INFLUENCE OF BACKGROUND CLINICAL DATA ON INFANT PAIN ASSESSMENTS"

This is a second example of a quasi-experimental design. Here, too, data of an experimental group are compared with data gathered earlier, but the latter serve as control data, and the experimenter made sure that both groups were matched on important variables. The study attempted to isolate important variables in rating pain in infants. This study is more advanced than ones we've come across. It is a three-factor study with repeated measures on one factor. So we'll simplify things by looking at the design and then appropriate degrees of freedom for the analysis in Box 12.2.

A final point to note about this and similar studies using repeated measures is that when the circularity assumption is not met and *df*s are adjusted, that adjustment comes into play only when the calculated *F* is evaluated; it is *not used* in the calculation.

The Study

Fuller, B. F., Neu, M., & Smith, M. (1999). The influence of background clinical data on infant pain assessments. *Clinical Nursing Research, 8*(2), 179–187. Copyright © 1999 by Sage.

Box 12.2 Experimental Setup for Study Example 12.2

	Video+	*Video (J)*
Age(K)	*None Mild Moderate Severe*	*None Mild Moderate Severe (L)*
0–3	$n_{jk} = 6$	$n_{jk} = 10$
4–6		
7–9		
10–12		
Source	*df*	
Between	$(s - 1) =$	$(64 - 1) = 63$
Video	$(J - 1) =$	$(2 - 1) = 1$
Age	$(K - 1) =$	$(4 - 1) = 3$
Video × Age	$(J - 1)(K - 1) =$	$(2 - 1)(4 - 1) = 3$
Error ($within_{gp}$)	$\Sigma(n_{jk} - 1) =$	$4(6 - 1)+4(10 - 1) = 56$
Within	$K(L - 1)(n_{jk}) =$	$4(4 - 1)(6)+4(4 - 1)(10) = 192$
Pain	$(L - 1) =$	$(4 - 1) = 3$
Pain × Video	$(L - 1)(J - 1) =$	$(4 - 1)(2 - 1) = 3$
Pain × Age	$(L - 1)(K - 1) =$	$(4 - 1)(4 - 1) = 9$
Pain × Age × Video	$(L - 1)(K - 1)(J - 1) =$	$(4 - 1)(4 - 1)(2 - 1) = 9$
Error (pain × $within_{gp}$)	$(L - 1)(\Sigma(n_{jk} - 1)$	$(3)(56) = 168$

This study evaluated the influence of clinical background data on infant pain assessment performed by nurses in a clinical setting. Pain assessments of videotaped infants by nurse participants who were not provided with infant background information [were] compared with the assessments of nurse participants who viewed identical videotapes that were accompanied by written clinical background data.

Problem

In earlier studies, nurse participants assessed infant pain by viewing 3-minute infant videotapes and reading clinical background data about each infant. The background data described the infant's birth history; age; diagnosis, type of surgery, or both; time since admission or surgery; parent comments about the infant's usual behavior; and all medications, nutrition, and fluids given to the infant during the 48 hours prior to videotaping. The nurse participants in these studies reported using, as assessment cues, many details provided in this background information. . . . The importance of clinical background information in assessing pain is underscored by another study that showed that nurses who viewed only videotaped snippets (without any background knowledge of the infant) could not agree on infant pain assessment. . . . These disagreements may have occurred because the snippets were very brief, showed the infant alone, and did not include the infant's response to comfort measures or other interventions. A model of nursing assessment of infant pain . . . suggests that these assessments are based, in part, on the infant's response to comfort measures and clinical background data that inform the nurse about the likelihood of pain.

Infant age influences pain-related behaviors. Pain responses of older infants are more goal directed than are those of younger infants. . . . Pain-induced facial expressions differ between younger and older infants. . . . Such age-related differences may affect nursing assessments. Pediatric nurses report that it is easier to evaluate pain in older infants because they exhibit more behavioral cues than do younger infants. . . . If so, then younger infants may be assessed as having less pain than older infants. . . . In pain assessments based solely on videotapes, infant age differences in the richness of behavioral cues that indicate pain may have more influence on these assessments than they would were the videotapes accompanied by clinical background data.

The purpose of this study was to determine the importance of knowledge of clinical background data in nursing assessments of infant pain. Three questions related to this purpose are

1. What is the difference in levels of assessed infant pain between pediatric nurses who view infant videotapes and read infant background data and pediatric nurses who only view infant videotapes?
2. What is the influence of level of infant pain on any differences found for Question 1?
3. What is the influence of infant age on the assessment results for Question 1?

Method

Design

A quasi-experimental design compared the infant pain assessment of 24 pediatric nurses in an early study . . . who viewed videotapes of infants that were accompanied by written background information about each infant (video-information group), to the assessments of 40 pediatric nurses, recruited in 1997, who viewed the same videotaped infants without written background information (video-only group).

Variables

The dependent variable was level of infant pain as assessed by the 64 nurse participants. The three independent variables were the presence or absence of written clinical background data about each videotaped infant, age of the infant, and level of infant pain as assessed by a panel of five expert pediatric nurses. The clinical background data described the infant's birth history, age, number of previous hospitalizations and reasons for each admission; diagnosis, type of surgery, or both; time since this admission; time since surgery (if applicable); parent comments about the infant's usual behavior; and all medications, nutrition, and fluids given to the infant during the 48 hours prior to videotaping.

Participants

A convenience sample of 64 pediatric nurses with 5 or more years of pediatric nursing experience and a bachelor of science in nursing degree was used. The video-information group consisted of 24 White female nurse participants recruited in 1993 for a previous study. The video-only group consisted of 40 nurse participants recruited in 1997. One participant was male. . . . Of these participants, 36 were White, 2 were Black, 1 was East Indian, and 1 was Asian. No differences existed between the two groups in years of pediatric experience or level of education. Mean length of pediatric experience of the entire sample ($N = 64$) was 11.7 years. This experience included . . . general pediatrics to neonatal, pediatric intensive care, or both. Twenty-three percent had a master of science in nursing.

Videotaped Infants

To ensure that the assessments covered a broad spectrum of pain and infant ages, a 16-cell matrix of four ages (0–3 months, 4–6 months, 7–9 months, and 10–12 months) and four levels of pain (none, mild, moderate, severe) was filled with videotaped infants, based on the independent determinations of an "expert panel"

composed of five RN clinical nurse specialists with a BSN, and MSN and ≤ 13 years current pediatric experience. . . . The expert panel members, working independently, viewed 150 videotapes that were accompanied by notes concerning the infant's history, diagnosis, medication, and nutritional and fluid status. The experts rated the pain of each videotaped infant on a 4-point scale: *no pain*, *mild pain*, *moderate pain*, and *severe pain*. Infant videotapes were included in the matrix only when at least four of the five expert nurses agreed on the [expert] level of assessed pain (EAP). Each of the 16 matrix cells was filled with five infant videotapes. . . . Each infant was represented by a single videotape. Forty-three of the 80 infants were surgical patients. . . . Medical diagnoses included severe respiratory disorders, severe dermatitis, gastroenteritis, otitis media, and seizures. Twenty-nine infants received analgesic medication within 0 to 2 hours prior to videotaping. Parents were present during videotaping for 52 infants. Twenty-three infants were girls, 57 were boys. The videotapes showed the infant during periods of quiet and crying, the infant's behavior during a routine nursing intervention (e.g., bathing or vital signs), and the infant's response to routine comfort measures.

Procedure

All 64 participants designated which age categories of infants (i.e., 0–3, 4–6, 7–9, and 10–12 months old) were most familiar to them and were then assigned to assess the pain of 20 infants belonging to one of those age categories. Five of the 20 infants of each age category represented four pain subsets . . . : *no pain*, *mild pain*, *moderate pain*, and *severe pain*. The participants of both groups assessed infant pain on a 4-point scale in which 0 = *no pain* and 3 = *severe pain*. Six participants in the video-information group and 10 participants in the video-only group viewed the videotapes for infants of one age category. The participants of the video-information group also read written background information about each infant; the participants of the video-only group did not.

Data Analysis

For each of the 64 pediatric nurses, four mean values of the assessed pain levels were obtained. Each mean represented the level of assessed pain for the five infants that represented one of the four subsets of . . . EAP per age category. These means were compared using a three-level repeated measures [ANOVA]. The within-participants factor was EAP. The between-participants factors were infant age category and group type (video-information, video-only). For comparisons involving EAP, the results of Mauchly's Test of Sphericity indicated a rejection of the null hypothesis (i.e., the assumption of compound symmetry or equal variances and

covariances among levels of the repeated measures factor); therefore, the degrees of freedom used to determine the *F* value were first adjusted by the Greenhouse-Geisser epsilon for these comparisons. The assumption of homogeneity of variance was supported . . . so no adjustments were necessary.

Results

Mean values of pain assessed by the pediatric nurse participants for group, age, and EAP [were] shown in Table 12.3. The video-only group of nurses had lower [EAPs] (mean = 0.89) than did the video-information group (mean = 1.52), $F(1,55) = 53.00$, $p = .000$. There were no differences in the levels of assessed pain across the four age categories of infants, $F(3, 55) = 1.23$, $p = 0.31$, and no Video Group × Infant Age interaction, $F(3, 55) = 1.03$, $p = 0.31$. Mean levels of pain assessed by the pediatric nurse participants did differ across EAP, $F(2.57, 45.4) = 221.18$, $p = .000$. Individual contrasts showed that mean levels of pain differed among each of the four levels of EAP. Mean [EAPs] by the nurse participants were greatest for the EAP = severe category, next greatest for the EAP = moderate category, next greatest for the EAP = mild category, and lowest for the EAP = no pain category.

The only significant interactions were between EAP and infant age, $F(7.7, 141) = 4.1$, $p = .000$, and between EAP and group, $F(2.57, 45.4) = 8.9$, $p = .000$. The mean levels of pain assessed by the pediatric nurse participants of the video-information and video-only groups were lowest for the 10- to 12-month-old

Table 12.3 Means of Assessed Pain by the Video-Information and Video-Only Groups of Pediatric Nurses for the Five Infants Belonging to Each of the Four Groups of Expert Levels of Assessed Pain[a]

		Expert Levels of Assessed Infant Pain			
Age	*Group*	*None*	*Mild*	*Moderate*	*Severe*
0–3 months	Video-information	0.56	1.63	1.67	1.90
	Video only	0.24	0.65	0.72	1.12
4–6 months	Video-information	0.63	1.40	1.83	2.23
	Video only	0.36	0.72	1.13	1.67
7–9 months	Video-information	0.63	1.43	1.80	2.50
	Video only	0.29	0.73	1.00	1.25
10–12 months	Video-information	0.60	1.1	1.93	2.30
	Video only	0.43	0.65	1.50	1.7

a. Data for infants of all age categories are combined.

infants belonging to the EAP = mild category but highest for the 10- to 12-month-old infants belonging to the EAP = moderate category, in comparison to assessed levels of pain for the younger age groups. Nurse participants in the video-only group rated the 0- to 3-month-old infants as having lower levels of pain in all EAP categories than they rated older infants. This differs from the nurses in the video-information group who rated 0- to 3-month-old infants in the EAP = mild category as having a higher level of assessed pain than infants in other age groups.

Discussion

Nurses in this study assessed lower levels of pain when they assessed the infant without additional clinical data than they did when they viewed the infant and had access to clinical information about the infant. These findings underscore the importance of clinical data and clinical context in making infant pain assessments and the risk of underestimating pain when such data and context are not considered. Infant behavior by itself, however, did signal pain to nurses of this study, as indicated by the increase in levels of pain ratings in the video-only group that paralleled the increase in EAP assessments. The lower pain ratings for the 0- to 3-month-old infants compared to the older infants in the video-only group suggest that nurses may have more difficulty interpreting the pain-related behaviors of the 0- to 3-month-old infants. The 10- to 12-month-old-infants, on the other hand, seem to express their pain more effectively than younger infants, as indicated by pain ratings that were higher than those of younger infants in the moderate and severe categories of both groups. . . .

A potential limitation of this study was that the assessments of the two groups of nurses were separated by several years, posing the risk that hospital practice or medical advances might influence scoring. This seems unlikely, however, because the video scenes used in this study were identical for both groups and involved activities that have been common practice for many years. Nurses in both groups had similar pediatric experience and educational level. The only difference in the assessment situation was access to clinical information about the infants.

It is possible that parents and nurses who care for an infant for several days are able to interpret and use an individual infant's behavioral cues that indicate pain better than did the nurse participants of this study who only watched a relatively brief video scene of each infant. The video scenes viewed by the nurses also did not include "hands-on" assessment or testing of comfort measures that would be done in the clinical area. Comparisons of assessments of nurses caring for infants in the clinical setting with and without access to clinical data may be less divergent than what was found in this study and is an area for future research. . . .

CRITIQUE OF STUDY EXAMPLE 12.2

1. What was the rationale for the study?
2. What was the purpose of the study?
3. What variables were measured or manipulated?
4. Who were the participants?
5. Can the two groups be considered equivalent at the start of the study?
6. Are groups nonequivalent for all aspects of the study?
7. How were videotapes selected?
8. What was the general procedure?
9. What information is missing?
10. How were the data analyzed?
11. What steps were taken to remedy any failure to meet assumptions of the test?
12. What were the results regarding main effects of each variable?
13. What were the results regarding interactions? Are all statements accurate regarding significant interactions?
14. If you look at degrees of freedom for the error term in the *F* ratio for pain level (45.4), you'll see that it is too far removed from 168, even for an adjusted value. This suggests that the wrong error term was used (55 in the other two also are wrong; they should have been 56, but the 1 *df* won't make a difference). But you can estimate the correct error term by using an interaction that is reported significant and uses the appropriate error term, the interaction between EAP and age. The means are summarized next. Each is the weighted mean of (video+ × 6 + video × 10)/16.

	None	*Mild*	*Moderate*	*Severe*	
0–3	0.36	1.02	1.08	1.41	0.97
4–6	0.46	0.98	1.39	1.88	1.18
7–9	0.42	0.99	1.30	1.72	1.11
10–12	0.49	0.82	1.66	1.92	1.22
	0.43	0.95	1.36	1.73	1.12

Each cell mean has $n = 16$. The $SS = \Sigma 16$ (cell mean − row mean − column mean + grand mean)^2. Calculate MS_{int} and, by means of $F = MS_{int}/MS_e$ and $MS_e = MS_{int}/F$, determine the error term.

Now go back and reevaluate the *F* ratio for the main effect of pain level. The $SS_{pain} = \Sigma 16\ (Mean_{level} - Mean_{grand})^2$. Does the conclusion change?

15. The group × pain level interaction also used the wrong error term. The table is presented here:

	Video+	*Video*	
None	0.61	0.33	0.43
Mild	1.39	0.69	0.95
Moderate	1.81	1.09	1.36
Severe	2.23	1.43	1.73
	1.51	0.88	1.12

Calculate the SS_{gxp}, keeping in mind that there are 6 scores in each cell of the video+background group and 10 scores in the video-only group. Then, recalculate the *F* ratio with MS_e arrived at in Question 14. Does the conclusion change?

16. What did the authors conclude?
17. Were the conclusions justified?

For answers to these questions, see page **373**.

SUGGESTED READINGS

Best, J. W., & Kahn, J. V. (2005). *Research in education* (10th ed.). Boston: Allyn & Bacon.

Bickman, L. (2000). *Validity & social experimentation*. Thousand Oaks, CA: Sage.

Kazdin, A. E. (2002). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

Kirk, R. E. (1994). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.

Mason, E. J., & Bramble, W. J. (1997). *Research in education and the behavioral sciences: Concepts and methods*. Dubuque, IA: William C. Brown.

Neale, J. M., & Liebert, R. M. (1986). *Science and behavior: An introduction to methods of research* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Quasi-experimentation: Design & analysis issues for generalized causal inference*. Boston: Houghton Mifflin.

Wiersma, W. (2008). *Research methods in education: An introduction* (9th ed.). Boston: Allyn & Bacon.

Chapter 13

Longitudinal Studies

Longitudinal studies are typically observational or correlational designs that follow individuals over time in order to test hypotheses. There are several types of longitudinal designs, including prospective panel, retrospective panel, and repeated cross-sectional designs. In a **prospective panel design**, data are collected on the same individuals at two or more time points. In a **retrospective panel design**, the data from two or more time points in the past are reconstructed from archival information. In a **repeated cross-sectional design**, data are collected on the same set of variables at multiple time points, from individuals that are comparable, but not necessarily the same. **Longitudinal designs** are used to assess change over time and to evaluate causal relationships among variables.

One of the strengths of a well-crafted longitudinal study is the disentanglement of age, period, and cohort effects. **Age effects** refer to the impact of chronological age on the dependent variable. **Period effects** refer to the impact of a specific period of history on the dependent variable. **Cohorts** are groups of individuals who experience the same significant life events in the same period of time. Typically, cohorts are defined by when they were born (birth cohorts) and a cohort effect refers to cohort group differences on the dependent variable. Cohort effects (or at least birth cohort effects) are often conceptualized as an interaction between age and period.

An example will help illustrate these possible effects. There is growing interest in generational differences in leadership behavior. For example, older leaders tend to approach the role of leader in a more traditional and consensual manner than younger leaders, while younger leaders tend to be more task-focused and energetic (Kabacoff & Stoffey, 2001). Are these differences due to generational