2017

Research Methods in Occupational Health Psychology

Irvin Sam Schonfeld
CUNY Graduate Center

Chu-Hsiang Chang
Michigan State University

TWO

Research Methods in Occupational Health Psychology

KEY CONCEPTS AND FINDINGS COVERED IN CHAPTER 2

A goal of occupational health psychology (OHP) researchers is to generate and organize knowledge bearing on the relationship between work-related psychosocial factors and the health of workers. Like researchers in other branches of psychology, OHP researchers elaborate theories, develop hypotheses, and devise ways to test those hypotheses. The testing of hypotheses can give rise to new knowledge. That knowledge can help in the development of interventions to improve occupational health.

One task of OHP researchers is to elaborate scientific theories. A scientific theory is a logically consistent model that describes and explains relationships among constructs. Constructs are higher-order abstractions, which are discussed more fully later in the chapter. An example of a construct from another branch of psychology is intelligence. Examples of prominent constructs in OHP include decision latitude and psychological distress. A scientific theory in OHP describes and explains relations among the constructs that are the foci of OHP. Theories in OHP explain relationships among constructs such as decision latitude at work, psychological job demands, psychological distress, and so forth. It is unlikely that a single theory can address all the relationships that are the subject of OHP research; a theory, however, should address the interrelationships among some of those constructs.

Karl Popper (1963) developed the idea that a scientific theory must be able to generate hypotheses. According to Popper, a hypothesis is a conjecture that is falsifiable. In other words, a hypothesis is a statement that, when gauged against observations a scientist assembles, can be shown to be false. Alternatively, a hypothesis can be shown to be consistent with the assembled observations. Popper, in fact, criticized psychoanalytic theory, asserting that it is not a *scientific* theory because it does not generate these falsifiable statements called hypotheses.[1] An example of a theory generating a statement that is falsifiable comes from Karasek (1979). His theory was built on the idea that the constructs decision latitude and psychological job demands, dimensions that characterize work roles, contribute to (another construct) distress in workers. Decision latitude refers to the amount of autonomy a worker has to decide on the means to achieve work-related demands. Psychological workload refers to aspects of a job such as the complexity of work-related tasks. A hypothesis that follows from Karasek's theory is that compared with other workers, workers having jobs that combine little latitude with heavy demands will experience higher levels of psychological distress. Research on decision latitude and psychological workload will be discussed at length in Chapters 3 and 4.

Science can depart from the Popperian view. Hypotheses can derive from accumulated observations, that is, induction, and not necessarily from a theory (Ng, 1991). In other words, induction can help build a foundation of observations that leads to hypothesis generation (and theory development). Durkheim's research on suicide was inductive. His research connecting suicide risk to the business cycle (see Chapter 1) was built on "social facts" or accumulated observations rather than hypotheses that derived

---

[1] Popper's critique of Freudian psychology is instructive. If a Freudian encounters a man attempting to hurt a child, the Freudian would explain that man's behavior in terms of repression stemming from some aspect of the Oedipal complex. A Freudian who learns of a man who sacrifices his life to save a child would explain the man's self-sacrifice in terms of sublimation. According to Popper, every configuration of human behavior is a verification of a preconceived set of ideas. No human behavior contradicts psychoanalytic ideas. What the Freudian fails to do is "go out on a limb" to use psychoanalytic ideas to predict behavioral differences among individuals that will occur in the future. In other words, what the Freudian fails to do is to come up with testable hypotheses.

from a theory. One can observe departures from the Popperian ideal in contemporary research. Spector and Zhou (2014) took an inductive approach to studying gender differences in counterproductive workplace behaviors (harms employees inflict on coworkers and workplaces). Occasionally, a researcher taking an inductive approach to a problem will employ a work-around when dealing with Popperian journal reviewers by "inventing a theory leading to hypotheses," as if the hypotheses were generated beforehand (P. Spector, personal communication, March 2014).

Susser (1979) wrote that "the most cogent test of a hypothesis . . . is to attempt disproof" (p. 54). Before attempting disproof through hypothesis testing, a distinction must be made between conceptual hypotheses and operational hypotheses (Kleinbaum, Kupper, & Morgenstern, 1982). A conceptual hypothesis reflects ideas, for example, that the abstractions (constructs) decision latitude and psychological job demands influence another construct, psychological distress. To conduct empirical research, one needs operational hypotheses. To create operational hypotheses that are the real-world analogues of conceptual hypotheses, constructs have to be operationalized, that is, reflected in real-world measures. Operationalizations are imperfect, but they can capture enough of what constructs represent to allow investigators to conduct research. Decision latitude may be represented by averaging a worker's responses to a small number of specially written, highly focused questionnaire items that ask workers to estimate how much freedom they have to make decisions about the tasks they ordinarily perform on the job. Or one may operationalize decision latitude by averaging independent experts' judgments regarding the amount of latitude workers with certain jobs are allowed. Only after a researcher decides how to operationalize the constructs targeted for research can a study be planned, data collected, and operational hypotheses tested.

A reader may react to the hypothesis just described by alternatively hypothesizing that job demands and latitude play no role in the development of psychological distress *even if* research reveals that workers in high-demands–low-latitude jobs experience high levels of distress. The Karasek hypothesis (the combination of high demands and low latitude leads to distress) may have a rival hypothesis that better explains the relationship between work and psychological distress.

Researchers observe that high-demand–low-latitude jobs tend to be low paying. Perhaps it isn't that high-demand–low-latitude jobs drive distress. Rather, it may be that the earnings attached to jobs play a decisive role in the development of distress. Compared with workers in higher latitude positions, workers with little latitude tend to be employed in lower paying jobs. Job-related economic and social disadvantages could be the drivers of psychological distress. OHP researchers concern themselves with rival hypotheses, and make it a practice to evaluate alternative hypotheses (Platt, 1964).

When a theory generates hypotheses researchers find to be consistent with observations, the theory gains prestige in the research community. Conversely, a theory that generates hypotheses found to be inconsistent with the relevant observations loses prestige. Theories themselves are neither proven nor disproven. They are insulated from direct testing or falsification. A hypothesis is the currency that is tested. As hypotheses go, the theories behind them gain or lose favor.

RESEARCH DESIGNS

There are a variety of research designs. Specific circumstances pertaining to each line of research narrow the choices for OHP investigators. These designs include the

experiment, quasi-experiment, cross-sectional study, various longitudinal study designs, and so forth. In this section, a number of research designs used in OHP research are explored. The research designs are presented with a minimum of statistical exposition, although there will be some, recognizing that this section is devoted to research designs themselves and not to statistics. Apart from a description of each research design, there is an example of the design as it has been implemented in practice.

Experiment

As in the biomedical sciences, in the psychological sciences the experiment is a vehicle used in assessing cause–effect relations (in biomedical research, an experiment designed to test the efficacy of a treatment for a disease or disorder is called a "randomized controlled trial" or "clinical trial"). An experiment has at least two key features that should be underlined.[2] One feature is that the effects of two or more rival treatments or interventions are compared. Ordinarily, the experimental group comprises research participants (also known as "research subjects") who are assigned to a special treatment or a control condition. In OHP research, the special treatment can be a new, potentially health-improving way of managing one or more units within an organization. The treatment could be "special" because it is a new intervention or a modified version of an existing treatment. The control condition could be a condition in which participants are exposed to no treatment, a waiting list control condition, or an existing treatment. An OHP investigator can compare the efficacy of multiple rival treatment conditions that reflect different modifications of the standard way the targeted work organization operates and a treatment that is equivalent to the standard way the organization operates.

The second feature of an experiment involves allocating experimental units to the rival treatments. The term "experimental unit" often refers to research participants, but it can also refer to clusters of individuals such as work units. The experimental units must be randomly assigned to the rival treatments. Random assignment means that every participant in a participant pool (e.g., workers in an organization) has the same probability of being assigned to any of the rival treatments. The advantage of random assignment over other methods of assigning participants is that, on average, the participants in the rival treatment groups will be similar on most background characteristics. These background characteristics include factors that were measured at the outset of the experiment and factors that went unmeasured. There are many unmeasured background characteristics. Random assignment ensures that the groups will, on average, be similar on those background characteristics (measured and unmeasured). The groups will thus differ in one way—exposure to one of the two or more rival treatments. The groups will not differ on most other characteristics. Thus, the rival treatments are *not* likely to be *confounded* with other factors that could potentially explain the effect of the treatments on the participants.

Randomization works best when there are large numbers of participants to allocate to the rival treatment groups. For instance, it is more advantageous to allocate 200 participants per treatment group than 10 participants per group. Consider the statistic

---

[2] These features are not necessarily features of experiments in the physical sciences (e.g., chemistry or physics). The experiment as described here applies to biomedical and psychological research.

known as the "standard error of the mean." The standard error is the standard deviation of a sampling distribution. Without being overly technical, the standard error of the mean reflects how much treatment means would vary if the researcher would repeatedly draw same-size random samples from the population of interest (e.g., workers in the manufacturing sector), expose each of those samples to a particular treatment, and graph the means of all those samples. Compared with means drawn from large samples (where standard errors tend to be small), means obtained from small samples would vary to a much greater extent (the standard error would be large). Put another way, as sample size increases, the standard error of the mean decreases. The result is that means obtained from large samples would be more reliable—more stable—than means obtained from small samples. In general, it is a good practice to conduct experiments using large samples because the estimates of treatment effects are more stable.

Flaxman and Bond's (2010) research on stress management training is an example of an OHP-related experiment. They randomly assigned London government workers to two rival treatments, stress management training and a waiting list control group. A waiting list control group is usually scheduled to be exposed to the experimental treatment after the treatment has been completed in the first group, which was the case in this study. When the first group has completed the treatment but the members of the waiting list control group have not yet begun the treatment, the two groups are compared on the dependent variables. One dependent variable in Flaxman and Bond's study was psychological distress, which they operationalized by mean scores on a self-reported, psychological symptom scale. Flaxman and Bond found that at the end of the 6-month period the study was in the field, members of the experimental group experienced significantly less distress than members of the waiting list control group (who had not yet begun the promised treatment). Flaxman and Bond found that among the workers in either group who experienced the highest levels of psychological distress at the study's outset, those in the experimental group improved by a "clinically significant degree."

As mentioned earlier, the term "experimental units" applies not only to individuals; the term can also refer to clusters of individuals. In Flaxman and Bond's experiment, individual workers were the experimental units that were randomized into experimental and control groups. But sometimes it is not practical to randomize individual workers into rival treatments. Instead, OHP researchers randomize larger units. For example, organizational units can be randomly assigned to rival treatments. In a study that involved 16 Los Angeles schools, Siegel, Prelip, Erausquin, and Kim (2010) randomly assigned each school to an experimental intervention or a control condition. Employees at the eight experimental schools were provided a health-promotion intervention (e.g., to encourage healthy eating, walking). The eight control schools received a stipend but not the intervention. Compared with employees in the control schools, employees in the intervention schools showed statistically significant reductions in body mass 2 years after the experiment began.

Quasi-Experiment

The quasi-experiment is similar to an experiment; both are used to compare the effects of rival treatments on dependent variables. The quasi-experiment, however, differs from the experiment in an important way. In the experiment, research units are randomly allocated to rival treatment groups. By contrast, in the quasi-experiment, intact or preexisting groups are exposed to the rival treatments or a treatment and a

no-treatment control condition (e.g., a stress-reduction intervention, such as a yoga class, is introduced in one administrative office but not in an administrative office elsewhere in the same city). There is no random assignment to the rival treatment groups in a quasi-experiment.

The absence of random assignment of participants to rival treatments places a limitation on quasi-experiments. In a quasi-experiment, a researcher cannot assume that the participants in the rival treatments are similar on most background characteristics. In response to this limitation, researchers who conduct quasi-experiments often assess the study participants on many background factors before the individuals are exposed to treatments. The researchers conduct statistical tests to ascertain whether the groups differ on any of the *measured* background factors. If a researcher finds that the treatment groups differ on a background factor that could potentially explain the hypothesized effects of the treatment, the researcher can implement statistical adjustments to equate the treatment groups. A fundamental problem with the quasi-experiment, however, is that one or more *unmeasured* background factors may explain potential differences in the outcome variables, differences that ostensibly appear to have emerged as a result of the impact of the treatment. The best way to control the impact of unmeasured background characteristics on treatment outcomes is to conduct an experiment with random assignment. However, sometimes there are obstacles to randomly assigning workers to different groups (e.g., manager resistance), making a quasi-experiment the most viable option for investigators.

An example of an OHP-related quasi-experiment is Bond and Bunce's (2001) study of the impact of work reorganization in a sample of British government employees. The intervention consisted of employee-driven action research, in which employees collaboratively researched work-related problems, and then developed and implemented research-informed organizational changes. The purpose of these changes was to increase employee control over work processes that would reduce stress-related problems. In order to reduce the chance of cross-unit contamination (i.e., members of the experimental group disclosing to control workers elements of the treatment, potentially precipitating change in the control group), Bond and Bunce employed a wait-list control group consisting of a work group located in a different building. The researchers took steps to select a control group comprising workers who were similar in age, gender, and education to the workers in the experimental group. Bunce and Bond found that the workers in the experimental group showed better mental health and lower absence rates than workers in the control group.

Internal Validity of Experiments and Quasi-Experiments

In discussing their study's limitations, Bond and Bunce (2001) wrote that "we have inevitably had to use a quasi-experimental design and are, therefore, left open to various threats to internal validity" (p. 300). Internal validity, which is distinct from scale validity (a topic discussed later), refers to the extent to which a study's design allows researchers to draw cause–effect conclusions. One prominent threat to a quasi-experiment's internal validity is that some unmeasured factor, and not the treatments, may have affected the dependent variables. The beauty of a true experiment is that random allocation of participants into rival treatment groups evens out potential background differences among the members of the rival groups. Ordinarily, the true experiment, in comparison with the quasi-experiment, has greater internal validity.

Cross-Sectional Study

The most commonly employed research design in OHP—and probably all of psychology—is the cross-sectional study. In a cross-sectional study, the researcher obtains measures on a sample at one point in time. The cross-sectional study is ill-equipped to answer questions about cause and effect. If cross-sectional research finds that two factors are related, that finding does not ensure that one factor caused the other. Because the two factors were assessed at the same point in time, and a cause must antedate the effect (temporal precedence of the cause), the cross-sectional study can rarely establish temporal precedence of one factor over the other.

Schonfeld (1990) conducted a cross-sectional study of coping in teachers. He found that two types of coping were related to lower levels of distress. One type of coping, positive comparisons, involves such psychological strategies as comparing oneself favorably to another teacher. Another type, direct action, involves such behavioral strategies as making pronounced efforts to turn a failing student around. Higher levels of each type of coping were related to lower levels of distress. Because coping and distress were measured at the same point in time, it is not clear which came first. It is possible that distress influenced coping patterns or that one or more unmeasured third factors gave rise to both psychological distress and the coping patterns. Thus, a cross-sectional relation between distress and coping is compatible with any of three conditions: (a) coping affects distress, (b) distress affects coping, and (c) third factors give rise to both distress and coping. Although cross-sectional research cannot ordinarily help us draw cause–effect conclusions, such research can be helpful because researchers can learn that certain factors are related. The pattern of correlations found in a cross-sectional study can offer clues investigators would like to follow up using other study designs.

Case-Control Study

A case-control study is often conducted for the purpose of identifying factors that are associated with a disorder. A case-control study involves at least two groups of individuals. The members of one group, or the cases, have a disorder to be studied. The members of the second group, or the controls, are free of the disorder. The study design, however, is more aptly termed a "case-comparison study" (MacMahon & Pugh, 1970). The term "control group" as it is used here is not an apt term because the group is not a control group in the way we understand the term as it applies to an *experiment*.

Information on the backgrounds and histories of the group members is ascertained through interviews, questionnaires, medical exams, and record checks to determine whether one or more factors are more common in the backgrounds of members of the cases or controls. If the researcher intends to identify factors that are *specific* to a particular disorder, he or she would include a third group that has a disorder other than the disorder that is the focal concern of the investigators. Some factors, such as low socioeconomic status, are related to a great many disorders. If a researcher plans to identify risk factors that are specific to a disorder such as schizophrenia, it would be helpful to include a second control group with a different mental disorder in addition to a disorder-free control group.

Link, Dohrenwend, and Skodol (1986) conducted a case-control study to identify risk factors specific to schizophrenia. The study comprised three groups: a case group consisting of schizophrenic patients; two control groups, one comprising individuals suffering from depression, and the other, community residents with no evidence of

psychopathology. Link et al. found that compared with the depressed and healthy controls, the schizophrenic participants were significantly more likely to have had first-time jobs (which were likely to antedate their first schizophrenic episodes) that exposed the cases to noisome work characteristics. Noisome work characteristics not only refer to frequent exposure to loud noise, but also include adverse air quality, high levels of heat, cold, or humidity, or hazards. The schizophrenic and depressed participants did not differ in the average occupational prestige of their initial jobs or in their levels of education.

Because case-control research is retrospective, individuals suffering from a disorder may attempt to find the meaning of their current condition (Tennant, Bebbington, & Hurry, 1981). In this "effort after meaning," the cases may magnify some experiences, diminish others, and fail to recall yet others, misleading investigators trying to identify risk factors for the disorder being studied. Link et al. (1986) surmounted this difficulty. The researchers linked the participants' first full-time jobs to independent, objective data that characterized the working conditions associated with the jobs. Those objective data were derived from the Dictionary of Occupational Titles (DOT), a document that was periodically published by the U.S. Department of Labor (now replaced by an online database). To create the DOT, analysts evaluated a large number of characteristics of thousands of U.S. jobs. Link et al. (1986) used DOT evaluations, which, of course, were obtained independently of the participants in the case-control study, to characterize the participants' jobs. Thus, the noisome job characteristics were ascertained independently of the personalities of the study participants.

Longitudinal Study

The defining characteristic of a longitudinal study is that data are obtained from the same sample at two or more points in time (sometimes called a "panel study"). Because it is not ethical to conduct experiments in which investigators knowingly assign workers to better and worse job conditions, longitudinal studies, by assessing existing working conditions and worker health at different points in time, can be valuable in testing hypotheses that certain working conditions are related to future worker health. In conducting longitudinal research, it is often helpful to assess the health problem and the hypothesized risk factors for the problem at every data collection point. For example, if an OHP researcher is studying risk factors (e.g., heavy psychological workload) for depression, it is often helpful to assess the risk factors and depression at each wave of data collection.

Some longitudinal research does not follow this pattern. In a 1-year longitudinal study of Canadian teachers, Burke and Greenglass (1995) used multiple linear regression (MLR) to predict burnout (time 2) from work stressors assessed 1 year earlier (time 1). Although the authors found that time 1 workload predicted emotional exhaustion (a core component of burnout) at time 2, the analysis was problematic in a way that is instructive. The design of the study does *not* take into account the stability of exhaustion between times 1 and 2. By statistically adjusting for time 1 exhaustion, the investigator adjusts for the stability of exhaustion over time, and as a result, the regression of time 2 exhaustion on time 1 exhaustion and work stressors will help ascertain how much exhaustion increases or decreases from its baseline, if at all, as a function of work stressors (Kelloway & Francis, 2013). Because Burke and Greenglass did not statistically control for time 1 exhaustion, the study was problematic. The analysis resembles cross-sectional analyses in which it is not clear which factor developed first. Kasl (1983) called such a study a "phoney" longitudinal study.

A more effective regression analysis would examine the impact of time 1 workload on time 2 emotional exhaustion, while statistically controlling for time 1 exhaustion (among workers who remained in the same job for the length of the study). In this way, the OHP investigator could learn whether workload predicts future exhaustion adjusting for any confounding of workload and exhaustion at time 1 as well as the relation of time 1 exhaustion to exhaustion 1 year later. It would be equally important to conduct an analysis in which time 1 exhaustion predicts time 2 workload, while controlling for time 1 workload. Such an analysis evaluates the *reverse* causal hypothesis that emotional exhaustion contributes to higher workload (e.g., burned-out teachers getting into a rut in which they work with little respite).

Kivimäki, Elovainio, Vahtera, and Ferrie's (2003) 2-year longitudinal study of personnel working in 10 hospitals was methodologically sounder than the study conducted by Burke and Greenglass. Kivimäki et al. assessed the relation of time 1 organizational justice (e.g., fairness of workplace procedures) to time 2 health indicators (e.g., certified sickness absence), statistically controlling for the health indicators at time 1. Organizational justice still predicted health indicators at time 2. In other words, lower levels of justice were related to increased health problems, controlling for health problems at time 1. The researchers also found that health indicators at time 1 were *not* related to organizational justice at time 2, suggesting that the direction of the effect is from justice to health and not the reverse. Kivimäki et al. were able to establish the temporal precedence of a risk factor over health outcomes. Temporal precedence is an important piece of evidence, although not the only evidence needed, supporting the hypothesis that low levels of organizational justice contribute to ill health in workers.

In addition to controlling for the time 1 health outcomes in research on the impact of time 1 working conditions on time 2 health outcomes, other quality-control concerns bear on longitudinal studies. One concern is timing. How much time should elapse between time 1 and time 2? Should there be a 10-year lag between measurement periods? Probably not, because: (a) the nature of the job of an individual who remains in the same position can change over time; (b) people often change jobs given a sufficiently long follow-up period. What about a 6-month lag? It depends. The timing of data collection in a longitudinal study depends on the nature of the jobs being studied and the investigators' preliminary knowledge of the impact of risk factors on health outcomes (Kasl, 1983; Kelloway & Francis, 2013).

Cohort Studies

A particular type of longitudinal study has been employed in epidemiological research, namely, the "prospective cohort study," which is often more simply called a "prospective study." A cohort is an identifiable group of people. They may be people born at a certain time and in a certain place. They may be British civil servants aged 35 to 55 working in 20 departments in London in 1985, as in a study described by Kuper and Marmot (2003). In studies that follow a fixed cohort, which largely reflects the types of cohorts seen in OHP research, "no entries are permitted into the study after the onset of follow-up" (Kleinbaum et al., 1982, p. 56).

In a prospective cohort study, the research participants are initially assessed for the presence or absence of the risk factors or exposure variables. For example, Kuper and Marmot initially ascertained who in their sample was exposed to jobs with low

decision latitude and high demands, a combination that has been hypothesized to be harmful to coronary health. A key feature of the prospective study is that participants who are found to have the target disorder during an initial screening are excluded from the longitudinal follow-up. The study strategy is designed to ensure that exposure to the risk factor occurs *before* the disorder develops in any of the participants (the temporal precedence of the cause). Researchers test hypotheses regarding whether newly incident cases of the target disorder that emerge over the course of the follow-up period are more likely to develop in initially healthy participants who were exposed to putative risk factors than in initially healthy participants who were not exposed. Kuper and Marmot found that over approximately 11 years, heart disease was more likely to develop in civil servants who were exposed to low-latitude–high-demands jobs than in civil servants who were not exposed. Kuper and Marmot also statistically adjusted for differences between the exposed and the nonexposed groups on confounding variables such as age, salary grade, and so forth.[3]

Meta-Analysis

The next two sections describe different approaches to meta-analyses, methods that have gained considerable attention in the research community. Meta-analyses are methods for obtaining results by combining numerous data sources. The two-stage meta-analysis is older than the one-stage meta-analysis. When reading older research reports, two-stage meta-analyses are often simply referred to as "meta-analyses."

Two-Stage Meta-Analysis

Because meta-analyses pull together diverse studies with their many different samples, they provide a foundation for making generalizations that is broader than the conclusions OHP researchers can draw from results involving a specific sample. Compared with the statistical power to detect effects in any single study, statistical power is greater in a meta-analysis. A meta-analysis works differently from the research designs the reader already encountered in the chapter. The "participants" in a two-stage meta-analytic study, that is, the traditional meta-analysis, are individual studies. Like most OHP studies, a meta-analysis begins with a hypothesis. Typically, the meta-analyst attempts to identify every study that sheds light on the hypothesis of interest (e.g., compared with control interventions or a waiting list control, a certain type of intervention leads to better worker health). To identify relevant studies in journal articles, book chapters, dissertations, and theses, the analyst searches databases such as PsycINFO and PubMed. The analyst also reviews the reference lists of reports already identified (Box 2.1).

---

[3] The "retrospective cohort study" is another type of cohort study, although rarer in the OHP literature. To conduct such a study, one needs to identify archival data that can be reassembled for the purpose of constructing a retrospective cohort. The structure of a retrospective cohort study is similar to that of the prospective cohort study except that in a retrospective study all the events have already taken place. Initially healthy workers are identified at a point in time. Some were exposed to a risk factor, and some were not. All the workers are followed over time using existing records to identify the occurrence of a target health event. In a classic retrospective cohort study (Case, Hosker, Mcdonald, & Pearson, 1954), workers in the chemical dye industry and control sectors were followed using documentary records. Dye workers were found to be at extremely high risk for bladder cancer.

BOX 2.1 Databases

OH psychologists use a number of databases when identifying studies for a meta-analysis as well as for locating previous research that is relevant to an intervention they are initiating or new research they are planning. One database is PsycINFO. PsycINFO, which is a resource provided by the American Psychological Association, covers the world literature in psychology as well as a large fraction of the literature in allied disciplines (e.g., psychiatry, education, sociology). Most university libraries and many public libraries provide access to PsycINFO.

The reader can learn more about PsycINFO at the following website: www.apa.org/pubs/databases/psycinfo/index.aspx

PubMed is another database OH psychologists consult for meta-analyses and other purposes. PubMed covers the world literature in the life sciences. PubMed is sustained by the United States National Library of Medicine of the National Institutes of Health, and can be directly accessed by anyone with a computer and the Internet.

The website is: www.ncbi.nlm.nih.gov/pubmed

A database that largely parallels PubMed is MEDLINE, which can be found online in university and public libraries.

EMBASE is a biomedical database that can supplement PubMed and MEDLINE. EMBASE covers the proceedings of many conferences.

It should also be noted that sometimes the same publication is indexed in two or more databases. For example, the *Scandinavian Journal of Work, Environment & Health* is indexed in PsycINFO, PubMed, MEDLINE, and EMBASE. However, when conducting an OHP-related literature search, it is often helpful to search multiple databases to get the best coverage. Even if a journal is covered in more than one database, there are omissions in one database that are covered in another. By searching multiple databases, the chance that a relevant paper will be missed is reduced.

The analyst broadcasts alerts on Internet-based "listservs" that ask listserv participants for relevant studies they conducted or know about, including unpublished studies on the topic of interest. Some studies do not get published because investigators with null results are discouraged from publishing. It is important to conduct a broad search for research reports, even those with null or unanticipated negative findings. Unanticipated findings can include results in which the health of the experimental group became worse than that of the control group; however, these studies should still be considered for use in the meta-analysis. Doctoral dissertations and master's theses are scoured because dissertation and thesis writers with negative findings have less incentive than writers with positive findings to publish results (Box 2.2).

BOX 2.2 Listservs

OH psychologists communicate with each other by in-person meetings, e-mail, and telephone. Another method of communication that has become popular in OHP is the listserv. The American Psychological Association (APA) hosts a listserv relevant to OHP, and the European Academy of Occupational Health Psychology (EA-OHP) hosts one.

(continued)

Box 2.2 Listservs (*continued*)

The listservs provide a forum for discussions, and members can use them to broadcast requests for information and receive replies. Conference announcements and requests for papers are also broadcast on the listservs. Anyone can join a listserv.

To join the APA's OHP listserv, visit this website:

http://lists.apa.org/cgi-bin/wa.exe?A0=OHPLIST

To join the APA's other listservs, please write to the following e-mail address: listserv@lists.apa.org

To join the EA-OHP listserv, go to www.jiscmail.ac.uk/cgi-bin/webadmin?A0=EA-OHP

The meta-analyst develops criteria to judge the quality of each of the identified studies. The analyst uses these criteria to decide what studies to include in the meta-analysis. The decision to include or exclude a study is based solely on the quality of the study, and not a study's results. For a meta-analysis to assess the efficacy of an intervention, the analyst may decide that he or she will include only randomized experiments, and exclude quasi-experiments and intervention studies that lack a control group. In a meta-analysis of the impact of a psychosocial job factor such as decision latitude on a later health outcome, the analyst may include only (a) longitudinal studies for which time 1 measures of the health outcome are controlled and (b) prospective cohort studies. The analyst often excludes cross-sectional studies or conducts a separate meta-analysis of the methodologically weaker studies.

From the results section of each study, the meta-analyst extracts important findings. One kind of finding is the number of standard deviations the mean of an experimental group is above or below the mean of a control group. The statistic is called "Cohen's *d*." It is a measure of effect size, that is, how much better (or worse) the experimental group is compared with the control group or how much worse a group of workers exposed to a risk factor is compared with a group not exposed. Cohen's *d* is used when the dependent variable is measured on a continuous scale. The Pearson correlation coefficient, *r*, between the predictor and the health outcome can also serve a meta-analysis. Other statistics that can be extracted for the purpose of a meta-analysis are the odds ratio (*OR*) and the adjusted *OR*. These latter statistics are used when a dependent variable is a disease endpoint, which is binary (a participant was either diagnosed with the target disorder or was not). The adjusted *OR* reflects the influence of the risk factor on the disease endpoint, statistically controlling for other factors.[4] By converting the result of every study to be included in a meta-analysis into Cohen's *d*s or *OR*s, the meta-analyst has taken studies that may have assessed risk factors and dependent variables with a variety of different measures, and made those measures comparable by creating a common metric (Cohen's *d* or an *OR*). Although effect sizes can be classified as large, medium, and small, it

---

[4] Without being overly technical, an *OR* of 1.50 means that the risk of a disorder is approximately 50% greater in individuals exposed to a factor than in nonexposed individuals; an *OR* of 2.00 means that the risk of a disorder is approximately twice (or 100% greater) in exposed individuals than in the nonexposed. An *OR* of 0.50 means that the risk of a disorder in those exposed to a factor is half of that in the nonexposed; instead of a risk factor, we have a protective factor. An adjusted *OR* of 1.50 means that the risk of a disorder in the exposed group is approximately 50% greater, statistically adjusting for the other predictive factors.

should be noted that small effect sizes do not necessarily reflect minor or unimportant effects (Rosnow & Rosenthal, 1989), particularly when the effects bear on health or mortality (see Chapter 4).

Once the meta-analyst has extracted the relevant statistics from the results sections of the highest quality research reports, the analyst averages the relevant effect sizes regardless of whether they were statistically significant or not. Lack of statistical significance reflects effect size within the constraints of sample size. It is important to average effect sizes over many high-quality studies. The analyst may average every Cohen's *d* bearing on a dependent variable such as psychological distress. For example, for every randomized experiment in which (a) an experimental intervention was pitted against a control condition and (b) a continuous measure of distress was the dependent variable, Cohen's *d* is extracted, and all the *d*s are averaged (Cohen, 1992). All other things being equal, the results of larger studies are more reliable than the results of smaller studies. Study findings based on larger samples are weighted more heavily than study findings based on smaller samples.[5] In a meta-analysis in which they converted the findings from each of 55 randomized treatment-control contrasts to Cohen's *d*s, Richardson and Rothstein (2008) found that job-related stress management programs had, in terms of mental health, an average effect size of a little more than half a standard deviation when compared with control conditions. Larger, more beneficial, effects were found for cognitive behavioral interventions in comparison with the effects of other interventions. In a meta-analysis that averaged adjusted *OR*s from five longitudinal studies, Stansfeld and Candy (2006) found that low workplace social support was related to a 30% higher risk of later depression or severe psychological distress.

One-Stage Meta-Analysis

The aforementioned approach to meta-analysis is a two-stage approach. In the first stage, individual data from any one study are analyzed at the level of the individual study, typically by the original study team. The study's results are found in the results sections of journal articles, dissertations, and so forth. The meta-analyst extracts the relevant findings from the results sections of the studies that meet preestablished quality criteria. Then the meta-analyst converts the relevant findings to *d*s or *OR*s, which, in turn, become the input into the second stage of the two-stage meta-analysis, the averaging of the results of the original research reports.

A different kind of meta-analysis, a one-stage approach, has begun to gain adherents. In the one-stage meta-analysis, researchers obtain the raw data from all contributing studies, and analyze the data at the level of each individual participant rather than at the study level (Stewart et al., 2012). One-stage studies require cooperation from the original study teams, whereas in two-stage studies, the data needed are often found in existing results sections and ordinarily don't require cooperation from the original investigators.

Studies that contribute to one-stage meta-analyses, like studies that contribute to two-stage meta-analyses, can use different measures of key variables. In two-stage meta-analyses, study results are converted into a common metric like Cohen's *d* or adjusted *OR*s. By contrast, in one-stage meta-analyses, a different kind

---

[5] The weighting scheme differs depending on whether the meta-analysis uses a fixed effects or random effects model (Borenstein, Hedges, & Rothstein, 2007).

of "harmonization" must take place because the data are merged at the level of the individual participants. If workplace autonomy, an important workplace psychosocial factor in OHP, is measured differently in the original studies to be used in a one-stage meta-analysis, researchers may decide to operationally define individuals in a semiconsistent manner from study to study. For example, the meta-analysts could operationally define individuals with scores above a study sample's mean on an autonomy scale score as high in autonomy, and individuals below that mean as low in autonomy, and then repeat the algorithm for each contributing study even if contributing studies use different scales composed in different languages. In a one-stage meta-analysis involving more than 56,000 individuals from six different longitudinal studies, Fransson et al. (2012) found that among physically active workers, those who held high-workload–low-latitude jobs, compared with other physically active workers exposed to more favorable job conditions, were about 20% more likely to become inactive 2 to 9 years later.

One- and two-stage meta-analyses tend to yield similar results. One-stage meta-analyses, however, provide more statistical power in the context of assessing interactions (Stewart et al., 2012).

Final Comment on Meta-Analyses

Meta-analyses are not without shortcomings. Research on the quality of two-stage meta-analyses of biomedical treatment studies suggests that results are not always in agreement with large ($n > 1,000$), well-controlled clinical trials that were run subsequent to the meta-analyses with similar foci (LeLorier, Grégoire, Benhaddad, Lapierre, & Derderian, 1997). Large, well-controlled clinical trials are the gold standard in biomedical research on clinical interventions. One-stage meta-analyses are subject to information loss owing to procrustean harmonization procedures that require the chopping of different scales at different midpoints, depending on the sample and the measures used.

Other Research Designs in OHP

The next sections examine other research designs that have earned considerable attention. These include the diary study, the natural experiment, and the interrupted time-series. Another category is a broad family of qualitative research methods.

Diary Studies

As the name suggests, diary studies often involve the collection of data on a sample of workers every day over a period of time, such as 1 or 2 weeks. Some diary studies (e.g., Marco, Neale, Schwartz, Shiffman, & Stone, 1999) are run with the help of electronic devices at selected moments during a day or over 1 or more days. Diaries can also involve paper-and-pencil questionnaires (Green, Rafaeli, Bolger, Shrout, & Reis, 2006), telephone interviews (Almeida, Wethington, & Kessler, 2002), and specially designed websites (Schonfeld & Feinman, 2012). The advantage of diary methods is that they help the researcher examine mental states and events occurring at the workplace almost as those events occur in real time, mitigating problems of memory decay. Statistical methods are now available to help minimize subject loss if participants contribute data during some but not all data-collection periods, a circumstance that would be a problem for MLR analyses (Raudenbush & Bryk, 2001; Schonfeld & Rindskopf, 2007).

One particular kind of diary study involves the sampling of representative moments throughout a day or workday in real time. The method is called "ecological momentary assessment" (EMA). The individual reports on momentary states while completing the assessment *in situ.* Typically in EMA, research participants carry electronic devices that signal them, often at randomly[6] generated times, to report their experiences (e.g., affective states, stressors). The accuracy of the characterization of a person depends on having a representative sample of momentary states "much in the way that representative sampling of participants is seen as essential for valid inferences from a sample to a population of people" (Stone & Shiffman, 2002, p. 237).

An innovation in research design involves coordinating EMA and more standard longitudinal designs in which waves of data collection are separated by months. During a 2-year, six-wave longitudinal study of stressors affecting teachers, McIntyre et al. (2016) evaluated the feasibility of teachers completing an iPod-based diary on a single day or on 2 or 3 consecutive days during the fall, winter, and spring. The teachers completed the diaries up to seven times per day, each time after a class period ended. McIntyre et al. found very good compliance, item completion, and user-friendliness. The findings suggest that this kind of research design could be useful in efforts to look closely at stress processes as they play out during the work day as well as provide a look at the longer-term picture. Moreover, the close look can provide clues to identifying transactions that can be improved upon in terms of both reducing teacher stress (or stress in other types of jobs) and improving student achievement.

Natural Experiment

The idea of a natural experiment was introduced in the first chapter, with a description of Parkes's (1982) research on two groups of student nurses who, as part of their training, were randomly rotated through medical and surgical wards in different orders. A natural experiment simulates a true experiment in which participants are randomly assigned to treatment and control conditions. In a natural experiment, "nature" or social conditions outside the investigator's control assign, in an implicitly random manner, individuals to alternative treatments. Without the implicit randomization, a researcher would not be able to draw a causal inference, and the research would reduce to an observational study (DiNardo, 2008). Sometimes, studies are mislabeled natural experiments (e.g., Kompier, Aust, van den Berg, & Siegrist, 2000). For example, they lack a comparison group or, when they have a comparison group, they lack implicit randomization.

Hearst, Newman, and Hulley (1986) described a natural experiment that capitalized on the lottery the U.S. military used to identify men who would be eligible for the draft during the Vietnam War. Compared with men who on account of the lottery were ineligible for the draft, draft-eligible men, over the course of 10 postwar years, were more likely to die by suicide and vehicular accidents, adjusting for wartime deaths. Because about only a quarter of the draft-eligible men served in the military and just 9% of the draft-exempt men served, Hearst et al. (1986) used those figures to calibrate the actual risk associated with military service. The researchers supplemented the analysis with a case-control study that also linked service to subsequent suicide and vehicular death.

---

[6] In the study described next, which involves teachers, participants, for practical reasons, cannot complete an electronic diary at random times during a school day.

Interrupted Time-Series

A time-series involves multiple observations over time. The observations could be of (a) the same individuals or units (e.g., organizational units) or (b) different but similar individuals (e.g., different individuals who come from the same population). What makes a time-series an *interrupted* time-series is that at a point in time within the sequence of observations, a treatment or environmental event occurs. Researchers assess whether the observations obtained after the event differ from the observations obtained before it (Cook & Campbell, 1979).

Many time-series studies extend long enough to encompass more than one environmental event. For example, in an individual-level interrupted time-series study, Eden (1982) followed 39 nursing students for 10 months over five observational periods—three objectively low-stress periods interlarded with two objectively high-stress periods. He found that anxiety, blood pressure, and pulse rate rose during the high-stress periods and fell during the low-stress periods. Although the study was uncontrolled, it is likely that the stressors were causally related to the outcomes. Given that the high-stress periods were sandwiched between the low-stress periods, it is improbable that history and maturational effects, the most likely confounders, could explain the zig and zag of the outcome measures.

Interrupted time-series designs more commonly employ aggregate-level, population-based data. For example, Norström and Grönqvist (2015) assembled data on unemployment rates and suicide rates from 30 countries. The researchers found that, except in Scandinavian countries, male suicide rates over a 52-year period rose or fell with the rise or fall of a country's unemployment rate.[7] In general, the effect of unemployment was stronger in countries with less generous protections for the jobless. A weakness of interrupted time-series studies like Norström and Grönqvist's is that by employing aggregate-level data, one ordinarily cannot discern whether a suicide victim was himself unemployed (see the section of Chapter 3 entitled "Two Pathways for Research on Unemployment"). On the other hand, a country's or region's unemployment rate can serve as a proxy for general economic conditions. Other aggregate-level interrupted time-series studies show that periods of economic recession are associated with elevations in the country's or region's suicide rate (Oyesanya, Lopez-Morinigo, & Dutta, 2015).

Qualitative Research Methods

Qualitative research is not a unitary concept. Qualitative research embraces a variety of methods. Some qualitative OHP researchers interview workers to ask about stressful work experiences and ways in which workers manage or cope with a stressor (e.g., Dewe, 1989; O'Driscoll & Cooper, 1994). Sometimes, workers complete questionnaires in which they freely write responses to open-ended questions about work (e.g., Schonfeld & Santiago, 1994). Other investigators use focus groups (e.g., Kidd, Scharf, & Veazie, 1996), which are group interviews. Some researchers place themselves in locations in which they can observe workers firsthand (Kainan, 1994). Other researchers employ participant observation (Palmer, 1983): the researcher obtains the job he or she intends to study, and learns about the job from the inside.

Glaser and Strauss (1967) propounded a bottom-up approach that has influenced investigators who conduct qualitative research. Glaser and Strauss advanced the

---

[7] Female suicide rates rose or fell with the unemployment rate only in Eastern European countries.

view that researchers should let theoretically interesting categories and hypotheses emerge from qualitative data; it is important *not* to approach qualitative data with preconceptions about what the data should reveal. Content analysis is an empirical methodology that helps organize and make sense of qualitative data. The term "content analysis" refers to a method of analysis, with its own procedures, that helps a researcher obtain insights into textual and symbolic phenomena (e.g., writing, speech, images), including the manifest and latent (shared and unshared) meanings of those phenomena (Krippendorff, 2004).

Although qualitative methods have limitations including the inability to test hypotheses (Schonfeld & Farrell, 2010), they have a number of advantages. For example, qualitative research can help with hypothesis generation, the identifying of as-yet undiscovered stressors and coping strategies, and help with understanding difficult-to-interpret quantitative findings (Schonfeld & Mazzola, 2013). With regard to the latter, Büssing and Glaser (1999) used qualitative data to help understand a paradoxical finding in a quasi-experiment involving nurses. In comparison with control nurses in traditional wards, experimental nurses worked in innovative wards in which job stressors such as time pressure were reduced because they were given greater responsibility but for fewer patients. The experimental nurses, however, experienced higher levels of emotional exhaustion. Qualitative data revealed that compared with nurses in traditional wards who had more piecemeal patient contact, the nurses in the experimental group had less opportunity to withdraw from difficult patients and thus experienced greater interactional stress.

Some OHP investigators collect both quantitative and qualitative data in a single study and integrate them into one set of data analyses.[8] An advantage of this strategy is that researchers can obtain a more complete picture of the stress process (Mazzola, Schonfeld, & Spector, 2011). Elfering et al. (2005) used qualitative methods to identify episodic work and nonwork stressors in workers at a counseling agency. The episodic stressors were assessed every day over the course of a week. Job control and chronically occurring stressors were assessed with standard, quantitatively constructed scales. Compared with chronically occurring stressors, episodically occurring stressors were more idiosyncratic, and therefore more suited to qualitative assessment. Elfering et al. (2005) found that compared with other workers, workers with more job control and lower-intensity chronic stress experienced higher levels of situational well-being in the aftermath of episodic job stressors.

MEASUREMENT

In the previous section, the types of study designs employed in OHP research were outlined. Study designs cannot be implemented in the abstract. Research cannot advance unless researchers effectively measure the factors and conditions they want to study. OHP researchers measure such entities as decision latitude, coworker support, depressive symptoms, and so forth. Before using a specific method of measuring

---

[8] This approach differs from the approach taken by Büssing and Glaser (1999), who also used quantitative and qualitative methods in the same study. Büssing and Glaser used the qualitative data to elucidate an unanticipated finding obtained in the quantitatively organized side of the study rather than combine quantitative and qualitative data in the same analyses.

a factor, OHP researchers, like researchers in other branches of psychology, require evidence that a candidate measure provides a reasonable way to proceed. Researchers can shop around for alternatives. This section briefly covers the measurement aspect of OHP research.

Reliability

Any variable measured should be measured reliably. Reliability reflects the consistency with which researchers measure a given variable. Consider a man's trip to his physician's office. The physician may want to measure the man's weight. Suppose the man steps on the scale, and she measures his weight at 150 lb (68 kg). Then 5 minutes later, the man steps on the scale again, and she measures his weight once more. She now says he weighs 125 lb (57 kg). Then after another 5-minute interval, she weighs him again. But now he weighs 165 lb (75 kg). The scale is not reliable. It does not return consistent measurements from one weighing to the next. Of course, if a year had passed between measurement occasions, it is possible that the scale would return a substantially different weight. The scale could be reliable because people lose and gain weight over time. However, because the measurement occasions in the doctor's office were close in time, the best explanation for the change in weight is that the scale is unreliable. A reliable scale would return the same weight each time the man stepped on it.

In classical test theory, the core of reliability is the "true score." The true score is an unmeasurable, underlying score thought to characterize the person or the condition being assessed. It is thought to be unchanging over short periods of time. Imagine that a sample of workers is administered a five-item job satisfaction scale. Does a worker's score on the scale map exactly onto the worker's true score[9] at that point in time? No. His true score likely contributes to the observed scale score, but a number of other factors may have also contributed to the observed score: if a coworker told him a joke; if he had an argument one morning just before completing the scale; random error.

Suppose a psychologist administers a job satisfaction scale to 100 workers. In addition, an alternate form of the scale (different items that are thought to cover the same ground) is also administered. Each worker is now assigned the mean of his or her scores on the two scales. Is that mean the worker's true score? No it isn't, but it is an improvement. Suppose the psychologist administers 10 alternate forms of the job satisfaction scale, one in the morning and one in the afternoon, every day over the course of a work week. Next each worker is assigned the mean of his or her 10 job satisfaction scores. Is that mean the worker's true score for job satisfaction? The answer is still NO. But that mean is theoretically closer to the true score.

A "reliability coefficient" (its symbol is r_{1t}.), although it looks like a correlation coefficient, is more akin to a coefficient of determination, an r^2. A reliability coefficient is an estimate of the proportion of total scale variance that is true score variance. A reliability estimate of .80 means that 80% of the variance of scale scores is estimated to be true score variance; the remaining 20% of the variance reflects measurement

---

[9] The true score underlying the observed score refers to something largely responsible for the observed score, although reliability alone does not establish what exactly a scale score represents. The section on validity describes what the creators of a scale do to help researchers feel confident that the scale is measuring what the scale is purported to measure.

error or random noise. OHP researchers rarely use scales that have reliability coeffi-
cients less than .70. As a reliability coefficient approaches 1.00 (at which point 100%
of scale variance is true score variance), more and more scale variance is reflective
of true variance. Because true variance is thought to reflect something true about
the people responding to scale items, there should be greater consistency from one
administration of the scale to the next (at least over short periods of time) and how
individuals respond from one scale item to the next. Investigators use a variety of
approaches to estimate scale reliability.

Internal Consistency Reliability

The most commonly used reliability coefficient in OHP research is the alpha coeffi-
cient, which is also known as "Cronbach's alpha." The strength of the alpha coefficient
depends on item-to-item correlations. If you were to open an issue of *Work & Stress*
or the *Journal of Occupational Health Psychology*, you would likely find an article in
which a scale's coefficient alpha (sometimes seen as the Greek letter α) is reported.

According to classical test theory (Nunnally & Bernstein, 1994), how an indi-
vidual responds to scale items depends on at least two components, the individual's
true score and random error (see Viswanathan [2005] for a discussion of other test
score components such as systematic errors). True scores are additive, and sum when
the scores on individual items are added to obtain a total score. Item-level random
errors, on the other hand, are just that, random. By virtue of their being random,
the error components of items are uncorrelated with true scores and each other.
Because these errors are essentially random noise, they don't repeat in the same way
from item to item. Random errors are not additive; they average out when summing
over many items.

One consequence of the randomness of measurement error is that there is a
straightforward way to improve the reliability of a scale. Let's suppose that a researcher
would like to increase the reliability of a three-item scale designed to measure job
satisfaction. Even if the interitem correlations in the original scale are moderate,
say around .20 or .30, the researcher can improve the scale's reliability by adding
items from the same domain, that is to say, from the domain of items reflecting job
satisfaction, that have similar moderate correlations with each other and the original
items. The items' true score components will sum, but measurement error associated
with the items will not.

Alternate Forms and Test–Retest Reliability

Other kinds of reliability include test–retest reliability and alternate forms reliability
(Anastasi & Urbina, 1997). Alternate forms reliability is established by administering
two or more forms of the same test or scale within a short time frame, the instru-
ments presumably reflecting similar content but comprising different items. The
instruments should be highly correlated with each other. Alternate forms reliability
is more likely to be observed in research on psychoeducational tests than in OHP.
By contrast, test–retest reliability, which involves correlating the scores obtained
in two separate administrations of the same scale, has been employed more often
in OHP research. For example, in a measurement study conducted to evaluate the
consistency of scales designed to measure the stressfulness of schoolteachers' work
environments, Schonfeld (1996) conducted a 2-week test–retest reliability study.

Interrater (Scorer) Reliability: Continuous Measures

Interrater reliability involves having two or more *independent* scorers rate a sample of behavior. Confidence in the measure of behavior depends upon the extent to which the independent ratings are highly correlated. An example of interrater reliability borrowed from a domain outside of OHP would make the idea clear. Two professors of English are independently reading and grading the essays of 100 college students. With a solid scoring rubric and some prior practice, a high correlation between how the two professors rate the 100 essays should obtain.

One could use the Pearson correlation coefficient to assess the reliability of the two sets of ratings. Let's suppose (although this is highly unlikely) that, owing to leniency bias, Professor B awards each student's essay a grade that is exactly 10 points higher than the grade Professor A awards the essay. The Pearson would be perfect, $r = 1.00$, because the corresponding grades awarded by Professors A and B are in exactly the same relative position. This result is to be expected because the Pearson reflects the extent to which the two sets of measurements are linearly related. However, if one is specifically concerned about the exactness with which the two raters agree, that is, the extent to which the raters' ratings are replicates of each other (Bartko, 1991), then it would be helpful to use another statistic, the intraclass correlation coefficient (ICC).[10] The ICC in the essay example would be less than 1.00 because the professors' ratings are not perfect replicates of each other. Of course, applications of the ICC are not limited to professors' ratings of student essays. The ICC can apply to independent observers rating how much complexity there is in job tasks or how much autonomy workers are allowed.

Regardless of whether one is using the Pearson or the ICC in conducting research in which ratings are involved, certain other statistics need to be appreciated and published. With regard to research involving the application of the Pearson correlation coefficient (assuming normality) to evaluate the reliability of raters' ratings, an investigator needs to report in a publication the means and standard deviations (*SD*s) of each rater's ratings. Mean differences could reflect systematic differences (biases) in the raters' ratings. Differences in the *SD*s tell us whether the raters are unequally discriminating. An advantage of the ICC is that, in telling us how closely the corresponding ratings match, the coefficient is affected by rater-related differences in means and *SD*s, although that knowledge may not be obvious to a reader (D. Rindskopf, personal communication, March 14, 2014).

Interrater reliability has been employed in OHP research. Murphy (1991) examined the relation of job characteristics to cardiovascular disease disability. Murphy capitalized on the results of independent research on more than 2,400 U.S. job titles. In that research, job analysts, using observations of job incumbents and interviews of both job incumbents and supervisors, rated 194 work-related activities as specified in the Position Analysis Questionnaire. The interrater reliabilities of the analysts' ratings were satisfactory. The analysts' ratings then became building blocks for the creation of multirating measures of the occupational titles.

---

[10] The intraclass correlation is a family of statistics (Shrout & Fleiss, 1979), but in the interest of making this section relatively nontechnical, the aforementioned example was created.

Interrater Reliability: Categorical Measurement

Some measures used in OHP research are categorical. Examples of categorical measurement include diagnosing a psychiatric disorder. Categorical measurement includes content-analyzing workers' reports about job conditions (e.g., deciding that a worker's written description of a recent interaction with a supervisor reflects whether or not the supervisor was supportive). The coefficient kappa (κ) provides a useful approach to assessing the reliability of procedures employed in categorical measurement (Cohen, 1960).

Suppose, for example, two clinicians independently diagnose 100 workers for a current episode of major depression. For illustrative purposes, the clinicians do something unorthodox. The clinicians believe that the base rate (the normal rate of major depression in the general population) is 10% (it is actually lower), and assign a randomly chosen 10% of the sample a diagnosis of depression. This example is revealing because in earlier times, researchers employed percent agreement as a measure of the reliability of their categorical measurements. Table 2.1 shows what would happen.

Two clinicians assigning a diagnosis to a random 10% of the sample would agree, on average, 82% of the time. They would, on average, agree 81% of the time (0.90×0.90) on which workers are not depressed, and agree about 1% of the time (0.10×0.10) on which workers are depressed. An 82% level of agreement seems impressive; the agreement, however, is largely the result of chance, given the low base rate for the disorder. κ, which adjusts for chance agreement given the base rate (Cohen, 1960), would be .00 for the data in Table 2.1. κ typically ranges from 0, where agreement is purely chance, to 1.00, where there is perfect agreement, as in Table 2.2.[11] An acceptable κ is greater than .40, and κ is considered good when it exceeds .60.

In laying a foundation for the creation of an online diary (Schonfeld & Feinman, 2012), a preliminary study was conducted in which 74 teachers were interviewed for the purpose of identifying job stressors teachers commonly encounter. Two readers *independently* read interview transcripts, content-analyzed the text, and sorted the teachers' descriptions into stressor categories (e.g., episode of student-on-student violence). The median κ was .82 (range: .61 to 1.00). The stressors the 74 teachers identified would become the nucleus of the stressors that would be included in the online diary used in a much larger study.

Final Word on Reliability

Consider an imaginary study in which a researcher measures the height of 25 people in a New York City apartment. The researcher would expect the correlation between height and weight to be moderate, about .60. Let's suppose weight was reliably assessed at another site. What would happen if the measure of height varied in reliability? Suppose the researcher lives near an elevated subway line, with trains passing his window every minute. With subway traffic outside, his hands shake when he measures each person's height. That shaking introduces a little random noise, measurement error, into the height he records next to each person's weight. Some individuals' heights will be slightly lower than they should be, and others', slightly higher. What happens if

---

[11] κ can trend less than 0 when one rater indicates one category, whereas the other rater trends toward indicating the opposite.

TABLE 2.1	Two Clinicians Assigning a Random 10% of a Sample of Workers' Diagnoses of Major Depression		
		Clinician 2	
		Not Depressed	**Depressed**	**Row Marginals**
Clinician 1	Not Depressed	81	9	90
	Depressed	9	1	10
	Column Marginals	90	10	

TABLE 2.2	Two Clinicians in Perfect Agreement in Assigning Diagnoses of Depression to a Sample of Workers		
		Clinician 2	
		Not Depressed	**Depressed**	**Row Marginals**
Clinician 1	Not Depressed	90	0	90
	Depressed	0	10	10
	Column Marginals	90	10	

express trains rattle by, causing his hand to shake more, adding even greater variability in the measures of height? The heights he records will contain even more measurement error. Finally, suppose that his hand wobbles so much that the heights he obtains may as well be random numbers. What happens to the correlation between height and weight? As more random error is injected into the recorded heights, the correlation between height and weight weakens (or in statistical language "attenuates"). When the height variable is totally random, that is, 100% of the variability in height represents measurement error, the correlation between height and weight approaches 0. See Cohen, Cohen, West, and Aiken (2003) for a more formal description of the impact of unreliability on correlations. The aforementioned example underlines for the reader that the foundation of all OHP research (in fact, all research in psychology) is the solidity of the measurement properties of the variables being investigated.

Validity

At its most general level, validity concerns establishing that a scale measures what its creators and users profess it to measure. How do we know that a scale purported to measure the stressfulness of a work environment actually measures that construct? How do we know that a scale measuring psychological distress actually measures distress? To establish that a scale measures what the scale's creators claim it measures, three interconnected types of scale validity need to be established: content, criterion-related, and construct validity.

Content Validity

Before looking at scales used in OHP research, it is helpful to turn to psychoeducational testing because (a) it is the arena in which much work on psychometric theory began and (b) an example from that area provides an intuitive way to introduce content validity. In the area of psychoeducational testing, subject area experts establish the content validity of a test by closely examining items to ensure that they are understandable and map onto the content found in curriculum guides and other

documents bearing on educational attainment. For example, subject area experts examining an arithmetic test for fourth graders determine whether items map onto the cognitive skills and knowledge that curriculum guides from representative school districts indicate should be taught in the fourth grade.

Content validity in OHP broadly follows the path just described. In creating a scale to assess the extent to which teachers are exposed to workplace stressors, veteran teachers, that is, the experts in the area of stressors confronting teachers, could inspect the items and judge whether they represent stressors teachers ordinarily confront. In creating a depressive symptom scale, experts such as psychiatrists and clinical psychologists inspect the items to judge whether they reflect the symptoms used in clinical settings to diagnose depression.

Criterion-Related Validity

A criterion is something external to the test or scale that the test or scale is expected to predict. Criterion-related validity reflects how well, that is, accurately, the test or scale predicts that external criterion. For illustrative purposes, it is helpful to turn again to the domain of education. Consider the criterion-related validity of a test such as a U.S. college admissions exam (e.g., the ACT) that is administered to upper-level high school students. A college admissions exam's criterion-related validity would depend on how well scores on the exam predict college grades. By the same token, an OHP practitioner would expect scores on a job satisfaction scale to predict workers' contemporaneous intentions to leave their jobs or actual future quitting. A "validity coefficient" is the correlation between the scores on the scale and the behavior the scale is expected to predict. The higher the validity coefficient, the more accurately the scale predicts.

There are two kinds of criterion-related validity. Both apply to the temporal relation of the criterion to the scale in question. Concurrent validity applies to the correlation between the scale and a contemporaneous criterion, for example, the correlation between scores on a job satisfaction scale and current intentions to quit. One would expect the correlation to be positive. Predictive validity applies to the correlation between the scale and a criterion measured at some future time, for example, quitting over the next 2 years.

Construct Validity

When OHP investigators conduct research, they may be interested in such relationships as the influence of decision latitude on psychological distress. But when investigators conduct research, what they do in practice is evaluate the relation of scores on a particular 5-item decision latitude scale to scores on a particular 20-item measure of distress.

OHP researchers also want to draw conclusions at a level of abstraction that goes beyond the particular observables. They would like to draw conclusions about the relation of an abstraction called "decision latitude" to an abstraction called "psychological distress." As mentioned at the beginning of the chapter, abstractions such as decision latitude and psychological distress are constructs. Constructs, not observables, are the constituents of scientific theories (Nunnally & Bernstein, 1994). Construct validation concerns establishing that a particular observable, such as a score on a certain symptom scale, is a reasonable reflection of the construct psychological distress. As in research in other branches of psychology, the construct validity of the measures used in OHP research is of great importance (Hurrell, Nelson, & Simmons, 1998).

Scientific theories generate hypotheses. On the basis of a scientific theory, OHP investigators develop hypotheses that predict how constructs are related to each other. The constructs have to be operationalized in observable measures. Conceptual hypotheses give rise to operational hypotheses (Kleinbaum et al., 1982) that imperfectly mirror the way the constructs will behave (Cronbach & Meehl, 1955). To the extent the observables behave in ways that are consistent with the relevant operational hypotheses, OHP investigators have evidence bearing on the construct validity of their measures.

So far, the discussion of construct validity has been abstract, like a construct itself. Let's bring the discussion down to earth. Schonfeld (2001) developed a measure of the stressfulness of teachers' work environments, the Episodic Stressor Scale (ESS). Teachers indicated how frequently they encountered episodically occurring stressors (e.g., a fight between students). The ESS is supposed to reflect the stressfulness of a teacher's work environment. A hypothesis he tested was that the ESS would correlate with a second measure of workplace stressfulness, the Ongoing Stressor Scale (OSS), which assesses chronically occurring stressors (e.g., low student motivation), because the two scales presumably measure the same construct. This hypothesis reflects Campbell and Fiske's (1959) idea of convergent validity, a component of the construct validation process. Convergent validity requires that two measures of the same construct correlate substantially with each other. The ESS and OSS did, indeed, correlate substantially ($r = .65$).

Every construct in psychology can't be related to every other construct. Construct relationships have borders. A psychological theory often stipulates that a construct central to the theory is unrelated, or weakly related, to another construct. This idea derives from Campbell and Fiske's (1959) concept of discriminant validity, another component of the construct validation process. Schonfeld (2001) hypothesized that the correlation of (a) the ESS administered during the first-year teachers' first fall term with (b) depressive symptoms measured during the summer, preemployment period *before* the women entered the teaching profession would be much weaker than the correlation of the first-term ESS with the first-term OSS. That hypothesis was borne out. The fall-term ESS correlated weakly with preemployment symptoms ($r = -.01$).

It was also expected that compared with the correlation of the fall-term ESS with preemployment symptoms, the fall-term ESS would correlate more strongly with concurrent depressive symptoms and spring-term (measured 4 months later) symptoms because the construct underlying fall-term ESS scores (the extent of chaotic working conditions) would give rise to depressive symptoms. These expectations were also borne out (see Table 2.3). The fall-term ESS scores would *not* be related to preemployment depressive symptoms because the two constructs should be largely

TABLE 2.3 Correlation Matrix		1	2	3
1	Fall-term Episodic Stressor Scale			
2	Preemployment Depressive Symptoms	−.01		
3	Fall-term Depressive Symptoms	.44*	.47*	
4	Spring-term Depressive Symptoms	.31*	.46*	.55*

*$p < .001$.
Source: Excerpted from Schonfeld (2001).

independent of each other *unless* there was faulty measurement; for example, the fall-term ESS was confounded with prior symptoms/distress.

Schonfeld developed the ESS by populating it with neutrally worded self-report items in reaction to job stress research conducted in an earlier era. In the context of this study, neutral wording refers to items that asked a teacher in unemotional language to indicate how frequently he or she encountered an event such as student fighting. In an earlier research era, job incumbents were often asked how stressed or distressed they were by a work event like student fighting (e.g., Kyriacou & Pratt, 1985). Asking how disturbed a worker is by a working condition (which is fine when done in ordinary conversation) creates an instrument that simultaneously measures at least two constructs: the presumed cause, that is, the job stressors, and the presumed effect, that is, the distress the stressors are hypothesized to provoke. A consequence of this kind of scale construction is that it risks inflating the correlation between the work stressor scale and commonly used dependent variables such as measures of psychological distress. Neutrally worded self-report items, by contrast, minimize emotional language, and concentrate on pinning down how frequently each work event occurred (Frese & Zapf, 1988; Kasl, 1987; Schonfeld, 1996).

RESEARCH ETHICS

Like all professionals, OHP investigators must adhere to ethical standards. The American Psychological Association (APA), the British Psychological Society, the National Institutes of Health, and other prominent organizations have promulgated ethical standards (APA, 2010; Ethics Committee of the British Psychological Society, 2009; National Institutes of Health, 2011). Some major ethical requirements advanced by the APA are highlighted, along with the observation that the APA's ethical standards are generally consistent with those of other, comparable organizations.

The APA's ethical standards follow from a set of general principles the APA has advanced. An example of an ethical principle is the principle of beneficence and nonmaleficence. This principle requires the researcher to do no harm and safeguard the research participant's "welfare and rights." Another principle is that of respect for the rights and dignity of individuals. The principle holds that individuals have the right to "privacy, confidentiality, and self-determination."

More specific standards of conduct are derived from the principles. For example, all proposed empirical research requires approval from the researcher's home institution. The APA requires researchers to provide accurate information to their home institution's institutional review board (IRB). The IRB has the power to approve the research's launch. In the formal application to the IRB, the investigator, in describing the study's procedures, also describes the study's risks and benefits, and the steps the investigator has taken to minimize risk or harm to participants. For example, the investigator describes how he or she will safeguard the privacy of the participants.

For empirical research with human subjects, the investigator must obtain informed consent from each participant. Informed consent is often framed in a letter to the participant (one to be signed and given to the investigator and an identical copy kept by the participant). The investigator, using language that is clearly understandable to participants, explains the purpose of the research, the risks and benefits of the research, and what the investigator has done to minimize risk (e.g., how the participant's privacy will be safeguarded). The letter (in some research projects informed

consent can be communicated orally) explains that participation in the research is voluntary and that the individual can stop participating at any time with no penalty.

In some circumstances, informed consent is not needed. When an investigator uses an anonymous questionnaire that his or her IRB has approved and judged not to cause harm, informed consent can sometimes be waived. When conducting research based on archival data, informed consent is not required.

The APA has ethical standards regarding publication credit once a research effort has concluded and a group of investigators has decided to write a paper about its findings. The APA standards indicate that who becomes the first author, the second author, and so forth depends on the extent of the individual's scientific and professional contribution to the paper. The standards make it clear that having senior status (e.g., an individual who is a department chairperson) does not entitle an individual to author status. Author status depends on the individual's contribution to the project.

SUMMARY

This chapter sketches the principal designs used in OHP research. Research is often guided by theory-generated hypotheses that investigators plan to test. Sometimes, however, research is guided by the accumulation of past findings. Much research is motivated by a combination of both. It should be borne in mind that the designs found in OHP research are commonly used in other branches of psychology and in medicine. The chapter emphasizes the value of the random allocation of research participants to experimental and control conditions in research evaluating the efficacy of workplace interventions. Random assignment of participants to treatment and control groups distinguishes the experiment from the quasi-experiment.

Researchers cannot ethically conduct experiments in which they deliberately expose participants to risk factors for health-related problems. Investigators can, however, conduct longitudinal studies in which they can statistically adjust for measured confounding factors such as baseline levels of the dependent variable, the variable that is expected to be affected by the work-related exposures. A diary study is a special kind of longitudinal study, the duration of which is usually brief, but captures events at work in, or close to, real time. Investigators conduct another special kind of longitudinal study called the prospective study, a study that is limited to participants who at the time 1 baseline are free of the disorder hypothesized to emerge later as a result of work-related exposures. A natural experiment, a kind of accident of the social, if not the natural, world, mimics a true experiment. An interrupted time-series study also takes advantage of events occurring in the world provided there are a sufficient number of assessments, either at the level of the individual or at the aggregate level, before and after the occurrence of the event in question.

Other types of research designs include the cross-sectional study and the case-control study. In cross-sectional studies, variables are assessed at a single point in time. Case-control studies involve individuals with and without a disorder. Investigators obtain data on the participants' life histories in order to learn whether the cases and controls were subject to different exposures over time. Both of these designs are useful for assessing whether an exposure and a health problem are related, but are usually unsuited to drawing cause–effect conclusions. One- and two-stage meta-analyses pool data at the participant or study level. Meta-analyses provide a

basis for obtaining broad summary findings across multiple studies. Qualitative methods help us understand the work lives of individuals at a highly descriptive level. Although not an effective vehicle for hypothesis testing, qualitative methods help investigators in the context of discovery and hypothesis generation.

Measurement is an important part of OHP research. We cannot study what we cannot adequately measure. OHP researchers want to employ reliable and valid measures of the variables they investigate. Investigators want evidence that they are reliably measuring every factor pertinent to their research. In other words, researchers want evidence that a factor (e.g., decision latitude) will remain largely unchanged if measured today and tomorrow. Of course, investigators want evidence that the instruments they use are valid. In other words, researchers want evidence that an instrument they use to measure a factor (e.g., depression) is truly assessing that factor.

Research must be conducted ethically. Research should do no harm. Those who participate in a study should do so voluntarily; they must give their informed consent in order to participate in a study. Coercion is unacceptable. Participants can withdraw from a study at any time and for any reason. When publishing a paper, researchers must allocate credit fairly.

REFERENCES

Almeida, D. M., Wethington, E., & Kessler, R. C. (2002). The daily inventory of stressful events: An interview-based approach for measuring daily stressors. *Assessment, 9*, 41–55. doi:10.1177/1073191102009001006

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct.* Retrieved from www.apa.org/ethics/code/index.aspx

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin*, *17*(3), 483–489.

Bond, F. W., & Bunce, D. (2001). Job control mediates change in a work reorganization intervention for stress reduction. *Journal of Occupational Health Psychology, 6*, 290–302. doi:10.1037/1076-8998.6.4.290

Borenstein, M., Hedges, L., & Rothstein, H. (2007). *Meta-analysis: Fixed effect vs. random effects.* Retrieved from www.meta-analysis.com/downloads/Meta%20Analysis%20Fixed%20vs%20Random%20effects.pdf

Burke, R. J., & Greenglass, E. (1995). A longitudinal study of psychological burnout. *Human Relations, 48*, 187–202. doi:10.1177/001872679504800205

Büssing, A., & Glaser, J. (1999). Work stressors in nursing in the course of redesign: Implications for burnout and interactional stress. *European Journal of Work and Organizational Psychology, 8*, 401–426. doi:10.1080/135943299398249

Campbell, D. T., & Fiske, E. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016

Case, R. A., Hosker, M. E., Mcdonald, D. B., & Pearson, J. T. (1954). Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Part I. The role of aniline, benzidine, alpha-naphthylamine, and beta-naphthylamine. *British Journal of Industrial Medicine*, *11*(2), 75–104.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. doi:10.1177/001316446002000104

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, W. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavior sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues in field settings.* Boston, MA: Houghton Mifflin.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0040957

Dewe, P. J. (1989). Examining the nature of work stress: Individual evaluations of stressful experiences and coping. *Human Relations, 42*, 993–1013. doi:10.1177/001872678904201103

DiNardo, J. (2008). Natural experiments and quasi-natural experiments. In S. N. Durlaur & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). New York, NY: Palgrave Macmillan. doi:10.1057/9780230226203.1162

Eden, D. (1982). Critical job events, acute stress, and strain: A multiple interrupted time series. *Organizational Behavior & Human Performance, 30,* 312–329. doi:10.1016/0030-5073(82)90223-9

Elfering, A., Grebner, S., Semmer, N. K., Kaiser-Freiburghaus, D., Lauper-Del Ponte, S., & Witschi, I. (2005). Chronic job stressors and job control: Effects on event-related coping success and well-being. *Journal of Occupational and Organizational Psychology, 78*, 237–252. doi:10.1348/096317905X40088

Ethics Committee of the British Psychological Society. (2009). *Code of ethics and conduct.* Retrieved from www.bps.org.uk/sites/default/files/documents/code_of_ethics_and_conduct.pdf

Flaxman, P. E., & Bond, F.W. (2010). Worksite stress management training: Moderated effects and clinical significance. *Journal of Occupational Health Psychology, 15*, 347–358. doi:10.1037/a0020522

Fransson, E., Heikkilä, K., Nyberg, S., Zins, M., Westerlund, H., Westerholm, P., & Kivimäki, M. (2012). Job strain as a risk factor for leisure-time physical inactivity: An individual-participant meta-analysis of up to 170,000 men and women: The IPD-Work Consortium. *American Journal of Epidemiology*, *176*, 1078–1089. doi:10.1093/aje/kws336

Frese, M., & Zapf, D. (1988). Methodological issues in the study of work stress: Objective vs subjective measurement of work stress and the question of longitudinal studies. In C. L. Cooper & R. Payne (Eds.), *Causes, coping and consequences of stress at work* (pp. 375–411). Oxford, England: Wiley.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Green, A. S., Rafaeli. E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods, 11*, 87–105. doi:10.1037/1082-989X.11.1.87

Hearst, N., Newman, T. B., & Hulley, S. B. (1986). Delayed effects of the military draft on mortality: A randomized natural experiment. *New England Journal of Medicine, 314*, 620–624. doi:10.1056/NEJM198603063141005

Hurrell, J. J., Jr., Nelson, D. L., & Simmons, B. L. (1998). Measuring job stressors and strains: Where we have been, where we are, and where we need to go. *Journal of Occupational Health Psychology*, *3*, 368–389. doi:10.1037/1076-8998.3.4.368

Kainan, A. (1994). Staffroom grumblings as expressed teachers' vocation. *Teaching and Teachers Education, 10*, 281–290. doi:10.1016/0742-051X(95)97310-I

Karasek, R. A. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative Science Quarterly*, *24*(2), 285–308.

Kasl, S. V. (1983) Pursuing the link between stressful life experiences and disease: A time for reappraisal. In C. L. Cooper (Ed.), *Stress research* (pp. 79–102). Chichester, England: UK: Wiley.

Kasl, S. V. (1987). Methodologies in stress and health: Past difficulties, present dilemmas, future directions. In S. V. Kasl & C. L. Cooper (Eds.), *Stress and health: Issues in research methodology* (pp. 307–318). Chichester, England: UK: Wiley.

Kelloway, E. K., & Francis, L. (2013). Longitudinal research and data analysis. In R. R. Sinclair, M. Wang, & L. E. Tetrick (Eds.), *Research methods in occupational health psychology: Measurement, design, and data analysis* (pp. 374–394). New York, NY: Routledge.

Kidd, P., Scharf, T., & Veazie, M. (1996) Linking stress and injury in the farming environment: A secondary analysis. *Health Education Quarterly, 23*, 224–237. doi:10.1177/109019819602300207

Kivimäki, M., Elovainio, M., Vahtera, J., & Ferrie, J. E. (2003). Organisational justice and health of employees: Prospective cohort study. *Occupational and Environmental Medicine, 60*(1), 27–34.

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*. Belmont, CA: Lifetime Learning.

Kompier, M. J., Aust, B., van den Berg, A., & Siegrist, J. (2000). Stress prevention in bus drivers: Evaluation of 13 natural experiments. *Journal of Occupational Health Psychology, 5*, 11–31. doi:10.1037/1076-8998.5.1.11

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks Hills, CA: Sage.

Kuper, H., & Marmot, M. (2003). Job strain, job demands, decision latitude, and risk of coronary heart disease within the Whitehall II study. *Journal of Epidemiology and Community Health*, *57*(2), 147–153.

Kyriacou, C., & Pratt, J. (1985). Teacher stress and psychoneurotic symptoms. *British Journal of Educational Psychology, 55*, 61–64. doi:10.1111/j.2044-8279.1985.tb02607.x

LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, *337*(8), 536–542.

Link, B. G., Dohrenwend, B. P., & Skodol, A. E. (1986). Socio-economic status and schizophrenia: Noisome occupational characteristics as a risk factor. *American Sociological Review, 51*, 242–258. doi:10.2307/2095519

MacMahon, B., & Pugh, T. R. (1970). *Epidemiology: Principles and methods*. Boston, MA: Little, Brown.

Marco, C. A., Neale, J. M., Schwartz, J. E., Shiffman, S., & Stone, A. A. (1999). Coping with daily events and short-term mood changes: An unexpected failure to observe effects of coping. *Journal of Consulting and Clinical Psychology, 67*, 755–764. doi:10.1037/0022-006X.67.5.755

Mazzola, J. J., Schonfeld, I. S., & Spector, P. E. (2011). What qualitative research has taught us about occupational stress. *Stress and Health, 27*, 93–110. doi:10.1002/smi.1386

McIntyre, T. M., McIntyre, S. E., Barr, C. D., Woodward, P. S., Francis, D. J., Durand, A. C., & Kamarck, T. W. (2016). Longitudinal study of the feasibility of using ecological momentary assessment to study teacher stress: Objective and self-reported measures. *Journal of Occupational Health Psychology, 21,* 403–414. doi:10.1037/a0039966

Murphy, L. R. (1991). Job dimensions associated with severe disability due to cardiovascular disease. *Journal of Clinical Epidemiology, 44*(2), 151–166.

National Institutes of Health. (2011). *NIH research ethics.* Retrieved from http://researchethics.od.nih.gov/CourseIndex.aspx

Ng, S. (1991). Does epidemiology need a new philosophy? A case study of logical inquiry in the acquired immunodeficiency syndrome epidemic. *American Journal of Epidemiology*, *133*(11), 1073–1077.

Norström, T., & Grönqvist, H. (2015). The Great Recession, unemployment and suicide. *Journal of Epidemiology and Community Health, 69*, 110–116. doi:10.1136/jech-2014-204602

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

O'Driscoll, M. P., & Cooper, C. L. (1994). Coping with work-related stress: A critique of existing measures and proposal for an alternative methodology. *Journal of Occupational and Organizational Psychology, 67*, 343–354. doi:10.1111/j.2044-8325.1994.tb00572.x

Oyesanya, M., Lopez-Morinigo, J., & Dutta, R. (2015). Systematic review of suicide in economic recession. *World Journal of Psychiatry, 5*, 243–254. doi:10.5498/wjp.v5.i2.243

Palmer, C. E. (1983). A note about paramedics' strategies for dealing with death and dying. *Journal of Occupational Psychology, 56*, 83–86. doi:10.1111/j.2044-8325.1983.tb00114.x

Parkes, K. R. (1982). Occupational stress among student nurses: A natural experiment. *Journal of Applied Psychology, 67*, 784–796. doi:10.1037/0021-9010.67.6.784

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*(3642), 347–353.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. New York, NY: Basic Books.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Richardson, K. M., & Rothstein, H. R. (2008). Effects of occupational stress management intervention programs: A meta-analysis. *Journal of Occupational Health Psychology, 13*, 69–93. doi:10.1037/1076-8998.13.1.69

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284. doi:10.1037/0003-066X.44.10.1276

Schonfeld, I. S. (1990). Coping with job-related stress: The case of teachers. *Journal of Occupational Psychology, 63*, 141–149. doi:10.1111/j.2044-8325.1990.tb00516.x

Schonfeld, I. S. (1996). Relation of negative affectivity to self-reports of job stressors and psychological outcomes. *Journal of Occupational Health Psychology, 1,* 397–412. doi:10.1037/1076-8998.1.4.397

Schonfeld, I. S. (2001). Stress in 1st-year women teachers: The context of social support and coping. *Genetic, Social, and General Psychology Monographs*, *127*(2), 133–168.

Schonfeld, I. S., & Farrell, E. (2010). Qualitative methods can enrich quantitative research on occupational stress: An example from one occupational group. In D. C. Ganster & P. L. Perrewé (Eds.), *Research in Occupational Stress and Well Being Series: New developments in theoretical and conceptual approaches to job stress* (Vol. 8, pp. 137–197). Bingley, England: Emerald.

Schonfeld, I. S., & Feinman, S. J. (2012). Difficulties of alternatively certified teachers. *Education and Urban Society, 44*, 215–246. doi:10.1177/0013124510392570.

Schonfeld, I. S., & Mazzola, J. J. (2013). Strengths and limitations of qualitative approaches to research in occupational health psychology. In R. R. Sinclair, M. Wang, & L. E. Tetrick (Eds.), *Research methods in occupational health psychology: State of the art in measurement, design, and data analysis* (pp. 268–289). New York, NY: Routledge.

Schonfeld, I. S., & Rindskopf, D. (2007). Hierarchical linear modeling in organizational research: Longitudinal data outside the context of growth modeling. *Organizational Research Methods, 18*, 417–429. doi:10.1177/1094428107300229

Schonfeld, I. S., & Santiago, E. A. (1994). Working conditions and psychological distress in first-year women teachers: Qualitative findings. In L. C. Blackman (Ed.), *What works? Synthesizing effective biomedical and psychosocial strategies for healthy families in the 21st century* (pp. 114–121). Indianapolis: University of Indiana Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. doi:10.1037/0033-2909.86.2.420

Siegel, J. M., Prelip, M. L., Erausquin, J. T., & Kim, S. A. (2010). A worksite obesity intervention: Results from a group-randomized trial. *American Journal of Public Health, 100*, 327–333. doi:10.2105/AJPH.2008.154153

Spector, P. E., & Zhou, Z. E. (2014). The moderating role of gender in relationships of stressors and personality with counterproductive work behavior. *Journal of Business and Psychology*, *29*, 669–681. doi:10.1007/s10869-013-9307-8

Stansfeld, S., & Candy, B. (2006). Psychosocial work environment and mental health: A meta-analytic review. *Scandinavian Journal of Work, Environment & Health, 32*(Special Issue 6), 443–462.

Stewart, G., Altman, D., Askie, L., Duley, L., Simmonds, M., & Stewart, L. (2012). Statistical analysis of individual participant data meta-analyses: A comparison of methods and recommendations for practice. *PLOS ONE*, 7(10), e46042. doi:10.1371/journal.pone.0046042

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine, 24*, 236–243. doi:10.1207/S15324796ABM2403_09

Susser, M. (1979). *Causal thinking in the health sciences*. New York, NY: Oxford University Press.

Tennant, C., Bebbington, P., & Hurry, J. (1981). The role of life events in depressive illness: Is there a substantial causal relation? *Psychological Medicine, 11*, 379–389. doi:10.1017/S0033291700052193

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.