Handbook of Adult Development and Learning (2006), Ed. C. Hoare, Oxford University Press

Chapter 3

Research Design and Methodological Issues for Adult Development and Learning

Grace I. L. Caskie and Sherry L. Willis

Both adult development and learning are established areas of research in their own right. Each has its own traditions and practices with respect to the type of research that is conducted. Given the relative youth of research conducted at the intersection of adult development and learning, investigators have the unique opportunity to carve out a new tradition of research but also face new challenges in blending together these two related yet previously separate disciplines. Decisions about the questions that should be asked and the theories that should be pursued need to be made by individual researchers. Issues such as these are not included in the scope of this chapter. Rather, the aim here is to provide an overview of the key issues involved in designing and evaluating various types of research endeavors that may be useful for those working in the combined areas of adult development and learning. Readers are referred at various points to publications with greater detail on a particular topic, some of which are entire texts devoted to the explanation of just one topic. The first section of the chapter describes several study types, including experimental, quasi-experimental, descriptive, cross-sectional, longitudinal, and sequential designs. The second section focuses on sampling issues and how the type of sample can influence the generalizability of research findings. The third section discusses measurement issues, including validity, reliability, appropriateness of a measure, scales of measurement, the measurement of change, and scale development. The final section of the chapter briefly addresses the issue of statistical significance versus clinical significance, with respect to evaluating the importance of one's research findings.

STUDY TYPES

Experimental Studies

In an experimental study, participants are randomly assigned to membership in one of two or more groups

that will be compared. For example, in the ACTIVE clinical trial on cognitive training of older adults (Ball et al., 2002), the participants were randomly assigned to one of three cognitive training groups or a control group. By assigning participants randomly to an experimental condition (or a control group), variables other than the independent variable (i.e., the variable on which participants are grouped or that is used to predict the outcome variable) that may influence the behavior or construct being studied should be equally distributed among the experimental and control groups. Thus, through the process of random assignment, experimental designs allow a researcher to control for possible confounding factors that would decrease a study's internal validity.

A word should also be said about the concern of artificiality in experimental studies. Although experimental studies are designed to approximate real situations, controlling for too many variables may limit the external validity of the study. For example, if a research study only included men (e.g., National Longitudinal Surveys of Older Men; Center for Human Resource Research, 1992) to control for the effect of gender on the results of the study, it may be difficult to generalize the study's findings to women. Having too many experimental controls can also decrease the generalizability of a study. A greater number of controls leads to greater artificiality in the experimental setting, making it more difficult to generalize behavior in the lab to a real world setting.

Willis (2001) discussed five types of experimental designs that are particularly useful for conducting behavioral interventions with adults. These designs vary in terms of the comparison or control group used in relation to the experimental group that receives the intervention. First, the *no-treatment control group design* uses a comparison group that receives the same pre- and postintervention assessment battery but otherwise receives no intervention or contact. This type of design is important for providing initial evidence of an intervention effect and an estimate of any gain in score that may result simply from being tested more than once (i.e., the practice effect). The second type of experimental design, the *nonspecific control group design*, provides a placebo treatment to the control group. The placebo treatment must not include any factors considered critical to the intervention. For psychological and behavioral research, it may be difficult to design a placebo treatment that has no effect at all on the outcome measures and that will maintain the blindness of participants to their treatment condition. Third, the *component control group design* is a between-group design in which various components of the intervention are implemented separately and in combination. An essential requirement, therefore, of using this design is the ability to identify and implement independently the various components of an intervention. Comparison of the different groups can help identify the most effective intervention component or combination of components. Fourth, the *parametric design* involves systematically varying the level of one experimental factor while holding all else constant. For example, the number of training sessions may be increased by one for several experimental groups being compared to determine the optimal number of training sessions needed (Hofland, Willis, & Baltes, 1981). Finally, the *comparative design* involves comparing two or more distinctly different intervention approaches to the same problem to determine which may be more effective. Although this might seem at first to be a very useful design, concerns exist with using this type of comparison. Because interventions may differ on a number of factors, it can be difficult to determine the most salient component for producing change and may result in few factors being held constant across intervention groups. Additionally, a number of practical issues exist with the use of this design, including the necessity of administering multiple interventions at the same site (to avoid confounding site with the type of intervention) and of having trainers who are equally proficient at conducting all interventions.

Quasi-Experimental Studies

Many studies exist for which random assignment to groups is entirely possible and appropriate. However, sometimes a researcher wants to compare groups to which individuals cannot be randomly assigned because the characteristics of interest are preexisting, such as gender, age, or educational level. These studies are called *quasi-experimental studies* (Campbell & Stanley, 1953; Cook & Campbell, 1979). Cross-sectional studies of age differences and longitudinal studies of age changes, described in greater detail shortly, are examples of quasi-experimental studies because individuals cannot be assigned to an age level. In practice, most studies include multiple independent variables or grouping variables, which may include a mixture of randomly assigned variables and

classificatory (i.e., not randomly assigned) variables that define the groups that will be compared. For example, Park, Smith, Morrell, Puglisi, and Dudley (1990) compared the recall of young and old adults who were randomly assigned to one of six experimental conditions (two verbal integration conditions crossed with three cueing conditions). This study is quasi-experimental because it included experimental conditions (verbal integration and cueing) and a classificatory variable (age group). In contrast to the experimental design, causal conclusions cannot be made in a quasi-experimental design regarding those variables to which individuals were not randomly assigned. When preexisting groups such as young and old individuals or males and females are compared, it is unknown whether differences are truly due to the group difference (i.e., age or gender) or to other factors that may covary with the independent variable. For example, different age groups represent different cohort experiences, including education, which can influence learning and development.

Confounding factors are variables that decrease the internal validity of the study—the ability to conclude that differences in the dependent variable (i.e., the behavior or construct that we are trying to explain or for which we want to examine group differences) were caused by the experimental condition rather than other extraneous variables. For example, suppose we conducted a study on the effect of child-rearing instruction on the quality of parent–child interactions. If all of the participants assigned to instruction on child-rearing techniques were mothers and the control group consisted entirely of fathers, assignment to experimental condition would have been completely confounded with gender of the participant, and it would be impossible to untangle whether any differences observed in the outcome measure (e.g., improvement in parent–child interactions) are due to receiving the instruction or to the fact that all who received the instruction were women. If random assignment had been used, the men and women in the sample would have been equally distributed into the experimental group and the control group. Confounded variables such as in this hypothetical example may not be as obvious in real studies, yet may be as insidious as confounding group with site or location, such as if all older adults come from an urban environment and all younger adults are from suburban college campuses. Alternative explanations for group differences should always be considered when random assignment has not been used.

Thus, the well-known admonition that "correlation does not necessarily imply causation" is misleading in that the ability to make causal conclusions is actually dependent on whether random assignment was used than the type of statistical measure used to describe the relationship. Correlation is simply a statistic that describes the linear relationship between the independent variable and the dependent variable, whether or not individuals were randomly assigned to the levels of the independent variable. If participants were randomly assigned to the number of training sessions (e.g., as in a parametric design) and a strong positive correlation were found between number of training sessions and number of correct responses at the end of the memory training, one would have a basis for concluding that more training sessions caused better memory performance. However, if we correlated number of adult education courses reported by each participant with memory performance, causal conclusions could not be made because there are other variables (e.g., socioeconomic status) that could be related to these variables that explain their relationship.

Descriptive Studies

In descriptive studies, individuals are not randomly assigned to any experimental conditions and may not be assigned to groups at all. Two basic methods for conducting a descriptive study are naturalistic observation and case study. The methods used in descriptive studies are more likely to be considered qualitative rather than quantitative, although they may also be incorporated into a more quantitatively oriented study.

When a researcher uses naturalistic observation, he or she has no direct contact with the individuals under study. The researcher simply observes the individuals in a natural environment (e.g., a senior citizen center, library, park, or grocery store) and then records information about behaviors that are demonstrated. For example, in Baltes et al. (1980), interactions between nursing home residents and staff members were observed and coded in terms of whether they supported dependence behaviors in the resident. Amato (1988) also conducted a naturalistic observation study in which the behaviors of men caring for young

children in public places were observed. Although naturalistic observation could be used to observe learning in adults, certain types of adult development may be more difficult to tap into given the lack of direct contact with the individuals under study. For example, one might expect that adults in a computer literacy course would learn new skills, which may in turn lead to changes in their self-perceptions and self-efficacy about their computer skills. Computer literacy might be indirectly measured by observing the number of errors made, but self-perception (an internal process) may be more difficult to assess accurately through observation alone.

In contrast to the descriptive method of naturalistic observation, case studies involve extensive direct contact between the researcher and the individuals under study. Typically, research using the method of case study involves only a few individuals to make the intense detail-gathering process a more feasible task. In addition to interviews and direct observation, the case study method may include examination of medical records or psychological measures. In addition to case studies of individuals, case studies can also be conducted of a process or a situation (e.g., Cervone, 2004). Detailed information about how to conduct a case study can be found in Stake (1995).

Examples of case study research also tend to be more qualitative than quantitative. Gillem, Cohn, and Throne (2001) described the case studies of identity development in two biracial individuals. The two subjects were studied with the use of a semi-structured interview schedule primarily composed of open-ended questions about their current identity, influences of family members on their identity development, and their experiences as biracial individuals. Honos-Webb, Stiles, and Greenberg (2003) reported on a case study of a woman who had completed psychotherapy. Measures of depression and self-esteem were collected, and transcripts of therapeutic sessions were also analyzed and rated. Finally, Levinson's (1978) book on the process of adult development for males included data collected from multiple interviews of 40 men about their lives over the period of late adolescence to their late forties. Levinson also presented in great detail the case studies of four men ("James," "William," "Paul," and "John"). Levinson's work demonstrates the incorporation of interviews and administration of psychological measures that may occur with the case study approach; specifically, both the men and their wives were interviewed, and tests such as the Thematic Apperception Test were also given to the men.

Cross-Sectional Studies

The focus of a cross-sectional study is to gather information about age differences. Participants are grouped by age, and then all groups are assessed at one point in time. For example, numerous studies have examined age differences in memory and other cognitive training (Craik & McDowd, 1998; Park et al., 1990), personality characteristics (Costa et al., 1986), or self-efficacy and attributions for performance in various domains (Lachman & Jelalian, 1984; Lachman & McArthur, 1986). When a researcher's interest is in knowing how various age groups differ on some construct or behavior at a particular point in time (e.g., voting behavior, attitudes about abortion), a cross-sectional study is very appropriate. Cross-sectional studies of age differences can also inform a researcher about possible trends for age changes and about the age range necessary to study a particular developmental process when planning a longitudinal study (Schaie, 1996b).

It is important to keep in mind, however, that the influences of age and cohort are confounded in cross-sectional studies. Cohort membership is typically defined by the year, or the range of years (e.g., the Baby Boomers), in which a group of participants was born. By definition, then, the age groups used in a cross-sectional study must be drawn from different birth cohorts. Thus, a cross-sectional study cannot fully determine whether any observed differences between the age groups are due to increased age (i.e., maturation) or to cohort differences. For example, if older adults were found to outperform younger adults on a timed test of simple mathematics problems, it may appear that mathematic ability increases with age. However, an alternate explanation is that the individuals in younger cohorts may have less experience working problems by hand rather than by hand-held calculator. In other words, the experience of learning to do simple math problems with (or without) a calculator may be cohort-specific. In addition, it should be noted that the influence of the time of measurement (i.e., period effects) on the participants' performance cannot be examined in a cross-sectional study because the data are collected at only one point in time.

When conducting a cross-sectional study, a potential problem is that of equating the age groups on relevant demographics that may influence the relationship between age and the dependent variable. For example, in comparing young and old individuals, it may be difficult to find groups with equivalent educational levels or experiences. Furthermore, there is the possibility of the bias of survivorship in the selection of older individuals for inclusion into a study.

Longitudinal Studies

The purpose of a longitudinal study is to examine age changes in the same group of individuals over time. The multiple assessments gained through longitudinal data collections are essential to understanding change in a construct or behavior over time as well as the variability and predictors of such change over time (Alwin & Campbell, 2001). Investigators have sometimes compared samples of individuals at different ages (i.e., a cross-sectional study) and concluded that differences found on the dependent variable could be attributed to chronological age. However, this type of conclusion must be treated with caution because cross-sectional and longitudinal studies do not always show the same age trends. For example, research on the adult development of mental abilities has shown wide discrepancies between cross-sectional and longitudinal data collected on the same subject population over a wide age range. For some dependent variables, substantial age differences obtained in cross-sectional data were not replicated in longitudinal data, whereas for other dependent variables, longitudinal age changes reflected more profound decrement than was shown in the comparable cross-sectional age difference patterns (Schaie, 2004; Schaie & Strother, 1968).

The most basic design for assessing changes over time due to age is a single-cohort longitudinal study. With this design, a single group of individuals (of similar age) are observed at two or more occasions in time. For example, Helson and Moane (1987) examined the personalities of a sample of women during their senior year of college and followed up twice with this cohort, once in their mid- to late twenties and finally in their early to mid-forties. A slightly more complex single-cohort longitudinal study, the Nordic Research on Aging project, included the study of 75-year-old individuals from three countries—Denmark, Sweden, and Finland—who were later reexamined at age 80 (Heikkinen, Berg, Schroll, Steen, & Viidik, 1997). With this second example, we again have a single cohort of individuals (those persons aged 75 years), albeit from three locations, which was assessed twice to obtain longitudinal information. Schaie and Hofer (2001) stated that longitudinal studies in adult development have been of three types: (1) studies that were begun to understand childhood development but with assessments continued into adulthood; (2) studies of young adulthood with continued assessments into midlife or later; and (3) studies specifically designed to assess the adulthood period with representative samples. An extensive overview of longitudinal studies of adult development can be found in Schaie and Hofer (2001).

Because all of the individuals in a single-cohort longitudinal study share the same cohort membership and the set of similar experiences that accompanies membership, this type of study cannot inform the researcher about any cohort differences in the construct being studied. Also, the single-cohort longitudinal design confounds age changes in the dependent variable with time-of-measurement (i.e., period) effects occurring over the calendar time during which change is assessed. The confound of age changes and time-of-measurement effects means that the researcher cannot be certain that the observed behavioral change is due to the individuals' increase in age rather than that something has changed about the environment between the different times of measurement. For example, suppose that in 2000 we had asked a group of 50-year-olds their opinions about the likelihood of a terrorist attack on the United States within the next year. If we reassessed this cohort in 2003 (at age 53) and found that the perceived likelihood of a terrorist attack on the United States had increased, at least two explanations are possible for the observed increase: (1) as people age from 50 years to 53 years, they become more anxious; or (2) specific events occurred during the three years between assessments to change people's perceptions. In this case, one might reasonably assume that the change in perceptions was more likely due to the occurrence of the terrorist attacks on September 11, 2001, than to any sort of maturational change that occurred during that three-year period.

The influence of the time of measurement can also influence the meaning of a particular construct or the implication for a finding in a longitudinal study. Caspi, Elder, and Bem (1987) found that a sample of

ill-tempered boys born around 1928 tended to maintain this personality style into adulthood. As a result, these men experienced significantly poorer outcomes as adults, which was at least partially influenced by changes that had occurred in society by the time they reached adulthood. Specifically, Caspi and colleagues pointed out that a greater emphasis had been placed on interpersonal skills as a prerequisite for success in the workplace during this time period. In earlier time periods, where job success focused on the ability to perform physical labor, the implication of having a confrontational personality style may have had less impact, and outcomes for these men might have been better. Questions that are asked in a longitudinal study may also have different implications at different times in history. For example, for male adolescents in the 1920s and 1930s, having a mother who worked outside the home was a typical indicator of poverty status (Hayward & Gorman, 2004), whereas it would have a very different meaning for most adolescents today.

Other longitudinal designs exist beyond the basic single-cohort design discussed so far in this section. Several of these designs are presented in a later section on sequential designs. The type of data collected in the design discussed is considered prospective data, wherein one begins studying a group of individuals with the intent of collecting future waves of information on this group of individuals. Longitudinal data can also be obtained retrospectively, by asking participants to recall information about earlier time periods (Alwin & Campbell, 2001). George, Hays, Flint, and Meador (2004) conducted a study of the relationship between religion and health in older adults by supplementing existing information on health in older adulthood with the collection of retrospective data on religiosity throughout the life course. Their study included a sample of community-dwelling adults aged 65 years and older on which health information had been collected prospectively from 1986 to 1996. Following the 1996 data collection, George and her colleagues also obtained life histories of the religious participation of these participants. This example demonstrates two of the problems with the use of retrospective data for studies of adult development and aging: (1) differential survival and (2) reliability of recall (Alwin & Campbell, 2001). First, religious histories were obtained from only those participants who were healthy enough and willing to complete all four waves of testing that occurred over period of a decade. Second, the recall of religious involvement at earlier time periods may be influenced by current states, including current level of religious involvement, or it may be biased by faulty recall. Yet, as George et al. (2004) pointed out, the use of retrospective data is preferable to no data at all.

Of course, despite its many advantages for the study of change over time, the longitudinal study is not without disadvantages. Collecting longitudinal data can be an expensive and time-consuming process, particularly when one desires to study the participants over an extensive period of time (i.e., over a wide age range). Also, participants are lost from longitudinal studies due to attrition over time, which may influence the results. Reasons and implications for attrition are addressed in the section on sampling included in this chapter. Longitudinal studies must also contend with the possibility of measures and research questions becoming outdated.

Time-efficient designs, such as the accelerated longitudinal design, have been proposed to shorten the amount of time necessary to study a developmental trajectory. These designs aim to reduce both the monetary and time costs of collecting longitudinal data. Under the assumption of no cohort differences, accelerated longitudinal designs link longitudinal data collected from several independent cohorts studied for overlapping age ranges (Duncan, Duncan, Stryker, Li, & Alpert, 1999; Tonry, Ohlin, & Farrington, 1991). Because no one cohort is studied, for the entire age range being investigated, distinct patterns of missing data are incorporated into the data by design. In the context of latent growth modeling, both McArdle and Hamagami (1992) and Duncan, Duncan, and Hops (1996) demonstrated that growth parameter estimates from the accelerated longitudinal design were as accurate as the corresponding true longitudinal design. However, McArdle and Hamagami also found that the standard errors for these growth parameter estimates increased as the amount of data collected per cohort decreased. This difficulty could lead to unstable parameter estimates. Large-scale simulations are needed to provide further insight into the utility of the accelerated longitudinal design.

Sequential Designs

Data acquisitions that are initially structured as either a cross-sectional or single-cohort longitudinal study can be extended into cross-sectional or longitudinal sequences (Baltes, 1968; Schaie & Baltes, 1975).

Sequential studies can also allow questions about development to be answered without involving the long time frame required for a true longitudinal design. Longitudinal sequences use the same sample of individuals from two (or more) cohorts repeatedly, whereas cross-sectional sequences use independent random samples of individuals (each observed only once) from cohorts covering the same age groups at two (or more) different points in time. For example, a longitudinal sequence might begin by studying a group of 30-year-olds in 2005, planning to retest these individuals every 5 years until they were 45 years old (i.e., retests in 2010, 2015, 2020). At the first retest point, a longitudinal sequence could then be formed by including into the study an additional cohort of individuals who were 30 years old in 2010, with the plan to assess this group also every 5 years until they turned 45. In contrast, a cross-sectional sequence might begin in 2005 with a simple cross-sectional study of individuals in three age groups: 30–34 years, 35–39 years, and 40–44 years. The data for the cross-sectional sequence would then be obtained by repeating this study at a later time point and by drawing new samples of individuals in each of the age groups that had been included in the original investigation. The critical difference between the two approaches is that the longitudinal sequence permits the evaluation of intraindividual age change and interindividual differences in rate of change, about which information cannot be obtained from cross-sectional sequences.

Schaie's "most efficient design" (Schaie, 1965, 1977, 1994) combines these cross-sectional and longitudinal sequences in a systematic way, incorporating age effects, cohort effects, and time-of-measurement effects. First, an age range of interest is defined at the time of the initial data collection and is sampled randomly at age intervals that are optimally identical with the time chosen to pass between successive measurements. The age range used must be specific to the problem under study. As an example, consider a design in which one wanted to study relationships between learning and development as individuals pass from midlife into older adulthood. In this case, to capture a range of ages in the midlife period, the age range of interest might include individuals who are 35 to 49 years old at the first time point.

Second, from this full age range of interest, samples of participants are drawn from age intervals with widths that match the amount of time expected to pass between measurements. Using the midlife example, if the researcher planned for 5 years to elapse between the first and second measurements, the samples should be drawn in 5-year age intervals within the larger age range (e.g., 35–39 years, 40–44 years, 45–49 years). Then, at the second time of measurement, participants from the first data collection are retrieved and restudied, providing short-term longitudinal studies of as many cohorts as there were age intervals at the initial data collection. Each age interval is also resampled at the second time of measurement, providing a new set of individuals to be tested within each age group. The resampling process is also shown in the example in table 3.1 where new samples of different people from each age group are added at each testing year. By Time 4 in this example, we would have collected data that covers ages from 35 to 64 years (using the 5-year testing interval that was proposed) and would have included individuals from six cohorts. The entire process can be repeated multiple times with retesting of previous subjects (adding to the longitudinal data) as well as initial testing of new samples (adding to the cross-sectional data).

The data generated by using the most efficient design are rich in that they can be analyzed with several analysis strategies (proposed by Schaie, 1965) to contrast the relative effects of age, cohort, and time of measurement on the variable being studied. Many developmentalists are most interested in the analysis of age changes and cohort changes performed in a *cohort-sequential* analysis under the assumption that time-of-

TABLE 3.1. Example of Data Collection Based on Schaie's Most Efficient Design

Age Group	*Sample*	*Time 1*	*Time 2*	*Time 3*	*Time 4*	*Cohort*
35–39 years	1	X	X	X	X	3
	2		X	X	X	4
	3			X	X	5
	4				X	6
40–44 years	1	X	X	X	X	2
	2		X	X	X	3
	3			X	X	4
	4				X	5
45–49 years	1	X	X	X	X	1
	2		X	X	X	2
	3			X	X	3
	4				X	4

Note: Each X represents a data collection for a particular sample.

measurement effects have not influenced the variable being studied. If a consistent age change is found for different cohorts, the results have greater external validity than would be provided by a single-cohort longitudinal design. For our midlife example, one of the cohort-sequential analyses that could be performed would use the Time 1 and Time 2 data from the 45- to 49-year-old age group and the Time 2 and Time 3 data from the 40- to 44-year-old age group. One practical drawback to this analysis strategy is that the analysis cannot be performed until three data collections have passed. Depending on the interval that one has proposed to use between waves of data collections, this could mean a long wait before one could begin analyzing data!

In contrast to the cohort-sequential strategy, cross-sequential and time-sequential analyses can be done earlier in the study, because both require only two data collections (as a minimum). In the *cross-sequential* analysis, cohort changes are contrasted with time-of-measurement effects, under the assumption that no age changes (or at least uniform age changes) have occurred on the variable of interest. A simple cross-sequential analysis could be conducted with the Time 1 and Time 2 data from two of the age groups in table 3.1. Finally, in a *time-sequential* analysis, age effects are contrasted with time-of-measurement effects, assuming no cohort effects on the variable being studied. The time-sequential strategy examines whether the difference between the age groups remains stable or changes over time. For example, we might want to examine whether the difference between the 35- to 39-year-old age group and the 45- to 49-year-old age group was the same or different at Time 1 and Time 2. The data from the first samples drawn from these age groups (i.e., those samples first tested at Time 1) would be contrasted with the data from the second samples drawn from these age groups (i.e., those samples first tested at Time 2). More detailed descriptions of each of these analyses, and analyses that also incorporate tests of practice effects and attrition, can be found in Schaie and Caskie (2004).

SAMPLING AND GENERALIZABILITY

The way a researcher obtains the actual group of individuals who will participate in his or her investigation has important implications for the generalizability of the study's findings. In other words, the type of sample used in a study and the external validity of the findings are interdependent. The ability to generalize a study's findings beyond the sample of individuals who participated requires that the sample is representative of the larger population of individuals to which the findings are to be applied. Certain sampling methods are much more likely to generate samples that are representative of the characteristics of the larger population of interest and, as a result, imply greater generalizability for the study.

Use of a probability sampling method, such as random sampling, ensures a representative sample because it avoids the biases involved in nonprobability samples (Babbie, 1986). In a simple random sample, every member of the population of interest has an equal chance of inclusion in the study sample. Suppose one wanted to study graduates from a particular university. A random sample of this group could be obtained by randomly selecting names of potential participants from university records. Because the likelihood of selection into a random sample was equal for all members of the population, research findings from this representative sample can be generalized to the entire population of interest from which the sample was drawn. For example, Sitlington and Frank (1990) used random sampling to obtain participants for a study of the success of learning-disabled individuals one year after completing high school. First, 2,476 individuals were randomly selected from a list of former special education students in Iowa who had graduated from high school in the years studied. Of the 2,476 graduates, 1,090 had participated in a Learning Disabilities Program as part of their special education curriculum; 911 of these individuals with disabilities were eventually included in the Sitlington and Frank study. The use of random sampling tends to be more important in applied research such as the Sitlington and Frank study than in basic research (e.g., on sensory processes), because the findings of applied research are more likely to be influenced by sample characteristics; findings are also more likely to be immediately and directly applied to some situation or environment (Stanovich, 2001). Sitlington and Frank's findings regarding their sample of learning-disabled students' lack of preparation for life after graduation could be immediately applied to other students currently in curricula designed for those with learning disabilities in Iowa high schools.

A more complex type of random sample is the stratified random sample. Stratification is most often used to ensure that the sample will include adequate

numbers of individuals from various subgroups of interest in the population. These subgroups differ on variables that are related, or which one expects are related, to the problem under study (Babbie, 1986). The Seattle Longitudinal Study (Schaie, 1996a, 2004) began in 1956 by taking random samples of individuals enrolled in a health maintenance organization from groups stratified by age (cohort) and sex. With this method, equal numbers of men and women were included in the study, and cohort sizes were also equal. Stratification may also be especially important in situations in which one must ensure inclusion of rare or hard-to-recruit groups (e.g., old-old individuals, certain ethnic minorities). For example, Klumb and Baltes (1999) used a stratified random sample to obtain equal numbers of individuals in age groups that spanned the range of 70–90+ years. It is unlikely that other sampling methods would have resulted in the inclusion of as many 90-year-old individuals as 70-year-old individuals. Rather than ensuring equal group sizes, stratification could alternatively be used to maintain a match between the proportions of various groups in the sample as those in the population. For instance, in a study of adults returning to college, it may be important to stratify the sample by age or motivation for taking courses (e.g., self-improvement versus job training courses) to maintain adequate representation of the various groups in this population that will be compared.

Yet the fact remains that true random samples can be difficult to obtain. As a result, most behavioral or social science studies use nonrandom, or nonprobability, samples. These types of samples are known more commonly as convenience samples, available group samples, or volunteer samples. Such samples are obtained by soliciting volunteers from existing groups, for example, local senior centers, churches, and university participant pools, through advertisements in the media, or, in the case of surveys, directly (e.g., from eligible passersby). Thus, these samples consist of individuals who volunteer to participate and are "available" or "convenient" for the researcher to use for that particular study. The procedure of simply including every person who responds to a survey or who volunteers to participate in testing is known as haphazard sampling (Minke & Haynes, 2003).

Generalizability is a concern with studies that use samples of volunteers because of their lack of representativeness and potential for bias. Specifically, individuals who volunteer may differ from the population to which we may want to generalize a study's findings in terms of the characteristics that self-selected them into the studies. Individuals who participate in research studies are more likely to be middle-aged (Herzog & Rodgers, 1988; Rosenthal & Rosnow, 1975; Schaie, 1959; Thornquist, Patrick, & Omenn, 1992), be better educated (Dodge, Clark, Janz, Liang, & Schork, 1993; Wagner, Grothaus, Hecht, & LaCroix, 1991), have higher incomes (Wagner et al., 1991), and have better access to transportation to and from the testing site (Dodge et al., 1993). Higher socioeconomic status may be related to a greater interest in supporting research or scientific pursuits, having more leisure time (allowing the opportunity to participate) than those who do not volunteer for a study (Dodge et al., 1993), or even the ability to understand the purpose and requirements of the research project (Wagner et al., 1991). Additionally, individuals who volunteer for certain types of studies, for example, training or prevention studies, may have greater concerns about the decline in their memory abilities (Schleser, West, & Boatwright, 1987) or health conditions (Dodge et al., 1993). Yet it is important to keep in mind that the use of available or convenient groups of participants does not necessarily invalidate that study's findings but rather points to other variables, situations, or types of samples that should also be examined (Stanovich, 2001). Replicating a study with other types of samples or in other environments or situations, can either increase the generalizability of the study's findings or indicate limitations of the research (Minke & Haynes, 2003).

Research with rare (or hard to recruit) populations is especially likely to use available group samples, partly because the sampling frame needed for a random sample can be difficult to identify (Minke & Haynes, 2003). Snowball sampling (also known as reputational sampling or sampling by referral) begins by identifying a few members of the target group, who then identify other group members who can be contacted, and so on (Kalton & Anderson, 1989). Examples of the types of populations that might be accessed by snowball sampling include minorities, particularly elderly members (Patrick, Pruchno, & Rose, 1998), the homeless (Woods-Brown, 2002), members of the gay and lesbian community (Rothblum, Factor, & Aaron, 2002), and people with disabilities (Kalton & Anderson, 1989). If a large enough number of individuals in these groups can be identified, a random sample could then be taken from those identified, but more often, all identified individuals who are willing

to participate are included in the study (Kalton & Anderson, 1989). Snowball samples may also be used when a random sample from a formally identified list (e.g., from social service agencies) is counter to the objective of the study—for example, if one wanted to study family caregivers who were not receiving any formal help—or for descriptive research where this type of sampling is used to reveal information about the process under study (e.g., identifying the most influential members of a group; Babbie, 1986). Compared to other recruitment techniques, Patrick et al. (1998) found that snowball sampling required more staff time but was highly cost-effective. Simultaneously using multiple recruitment methods (e.g., snowball, media, mailing lists, formal service organizations or support groups) to target rare populations may be the most successful and cost-effective approach (Patrick et al., 1998; Rothblum et al., 2002).

A final type of nonprobability sampling that we will discuss is purposive or judgmental sampling, which is used more commonly in descriptive or qualitative research (Babbie, 1986; Minke & Haynes, 2003). With this method, the researcher specifically targets and selects certain individuals because they display the characteristics or behaviors of interest being studied. The selection may be made after a period of extended observation (Babbie, 1986). For example, a program for increasing the social skills of adults may identify individuals who were socially isolated at a group function, or a study of individuals with conservative views may target Republican groups. Alternatively, purposive sampling may be specifically used to broaden a sample's characteristics, such as when a newly developed measure is to be tested on individuals with a wide range of ability (Babbie, 1986). Thus, participants are selected based on the purpose of the study and the judgment of the researcher. The aim of purposive sampling is to produce a sample of individuals that will be best for the question one wishes to address.

It should be noted that the use of large sampling frames or random samples does not automatically imply that studies are not subject to any restrictions in terms of generalizing their results. For example, using a random sample of college alumni may limit a study's ability to generalize its results to noncollege-educated individuals. Replication is important to establishing the generalizability of findings from studies using probability or nonprobability samples (Stanovich, 2001). Studies of adult development that use samples of people with greater income, more education, better jobs, and better health should be replicated with samples of less fortunate individuals. Thus the demographic and health characteristics of a sample should be reported so that applications or generalizations of the research can be better made.

In longitudinal research, the influence of attrition between waves of testing on the generalizability of results must also be considered. Even if a study began with a representative sample, attrition may alter the characteristics of the sample, rendering it less representative than it was previously. Participants may be lost between testing occasions for several reasons, including death or illness (especially if participants are elderly), relocation, or refusal to participate, including refusal by a caregiver (Alwin & Campbell, 2001; Johnson & Tang, 2003). Some studies have found that these types of attrition tend to bias the sample toward the middle class even more, with those remaining having better education, income, and jobs (Cooney, Schaie, & Willis, 1988; Schaie, 1988, 1996a, 2004; Sharma, Tobin, & Brant, 1986), whereas other studies have found that attrition had no impact on the representativeness of the sample (Fitzgerald, Gottschalk, & Moffitt, 1998; Johnson & Tang, 2003). Loss of participants may also be increased in studies that use longer intervals of time between waves of testing.

MEASUREMENT ISSUES

Many of the variables that developmental researchers are interested in studying cannot be observed directly. Rather, information about these constructs of interest (e.g., intellectual abilities, personality traits, learning) must be inferred by observing the participant's behavior or by other indirect means (e.g., rating scales and tests). Thus it is important to ensure that the method of measurement that one is using to assess these unobserved constructs has adequate validity and reliability, is appropriate for the population, and uses a scale of measurement that fits the question that is addressed. For the study of development, the assessment of change in a construct is also an important concept to consider, particularly in terms of identifying reliable amounts of change.

Validity

A test's validity is perhaps most important in the context of the intended purpose of the test. Thus, test

validation could be viewed as a continual process of accumulating information about the use of a measure in various contexts, rather than as something that is done only once (Suen, 1990). McDonald (1999) concluded that all of the approaches to establishing the validity of a test are interrelated and share the purpose of contributing evidence toward establishing that a particular measure is a valid assessment of the construct of interest. This section describes three approaches to the test validation process, each of which contributes different types of information toward it.

Construct Validity

The measures we employ in research need to be true reflections of the construct that is measured so that meaningful conclusions and interpretations can be made based on the information collected. Construct validity captures the idea that a test score is an accurate reflection of the construct of interest (Suen, 1990). Specifically, a measure with good construct validity is one that is related to others in the same domain and one in which convergent results are obtained with it and other measures of the same construct (Nunnally & Bernstein, 1994). The construct validity of a measure can be established by examining its similarity with other measures that are intended to measure the same construct. Using factor analysis as one possible method for establishing construct validity, Suen (1990) noted that this issue might be approached either internally (comparing items within a test) or externally (comparing the test with other tests). For example, if items in a factor analysis do not have strong loadings on the domains that they were hypothesized to measure, or if the items are spread across many domains, the construct validity of the measure is considered poor. Alternatively, a factor analysis of scores from several measures could be conducted (e.g., a multitrait-multimethod approach), in which the expectation is that the test being validated would have strong factor loadings on the same factors as other tests of that construct but weak loadings on factors represented by tests of other traits or attributes. The methods of exploratory and confirmatory factor analysis are discussed in more detail in a later section.

Content Validity

Content validity is an issue that is typically considered during the development phase of a new measure. Items that are selected for inclusion on a new measure are only a sample of the potential items that could have been used to assess the construct of interest. How well the selected items represent the entire pool of possible items determines the content validity of a measure (Nunnally & Bernstein, 1994). Determining this representativeness is typically a subjective process, based on the judgment of the individual performing the validation study. In cases in which the total set of items represents a good sample from the pool of possible items, samples of the measure's items would be expected to yield similar results. Thus, finding similar results with alternate forms of a particular measure may also help establish the content validity of a measure. Finally, the purpose of the test must be considered in establishing content validity. A particular sample of items may be a good representation of the collection of all possible items for one purpose or context but not for another purpose or context (Suen, 1990). For example, the appropriate items for assessing caregiver burden may vary based on disease progression or by whether the patient is community dwelling or institutionalized. Or a measure of personality development in childhood may not be appropriate for the study of adults.

Face validity is a concept that is related to but should not be confused with content validity. In contrast to content validity, the face validity of a measure is concerned with whether the scale appears to measure the construct it is intended to measure. Face validity is also determined after a measure has already been developed, rather than during the development stage (Nunnally & Bernstein, 1994). Issues of face validity can be important considerations in studies in which the subject matter may be sensitive (e.g., mental health) or where certain biases (e.g., the social desirability bias on measures of personality or honesty) may need to be avoided. In these cases, it may be necessary to disguise the intent of the items, resulting in reduced face validity.

Criterion-Related Validity

Criterion-related validity can be assessed by how strongly correlated the measure being validated (i.e., the predictor) is with a criterion measure (Suen, 1990). Two main types of criterion-related validity have been discussed. These types differ in terms of when the criterion is measured in relation to the predictor measure in question. With *predictive validity*,

the criterion is measured after the predictor measure; with *concurrent validity*, the criterion and predictor measures are assessed simultaneously. Good choices of criterion measures will be guided by some theoretical rationale for why the predictor measure should be related to it.

Criterion-related validity of a measure becomes especially important when scores from that measure are used to make decisions about a person (Nunnally & Bernstein, 1994). For example, certain test scores may be used to indicate cognitive impairment or to select (or exclude) individuals for a research study. For this to be a valid use of the test scores, the measure would need to have good criterion-related validity. Specifically, if the measure is used to predict a current state, such as cognitive impairment, it is implied that the measure has good concurrent validity. Or in the case where test scores are used to predict a future state (such as the ability to participate in a research study), the measure must have good predictive validity. Finally, it is also important to consider that restriction of range on either the predictor or criterion measures may attenuate their correlation and the assessment of the criterion validity of the measure. For example, if older adult learners have a more limited range of scores on a measure than a sample involving young and middle-aged adults as well as older adults, the criterion validity may appear lower than it would have in the more age-heterogeneous sample.

Reliability

The reliability of a measure can be defined as how well the measure reflects the true (unobserved) ability level of the individual being assessed; alternatively, reliability may reflect how stable measurements of a test score are over time (Nunnally & Bernstein, 1994). Typically, the second definition of reliability (i.e., stability) is of concern when researchers assess the reliability of a scale. To measure reliability in terms of stability, parallel forms of a measure are needed. The next section presents several types of reliability, which differ in terms of how the parallel forms of a measure are defined or created.

Test-Retest Reliability

One way to measure reliability is to have a group of participants take the same test twice and then find the squared correlation of the two sets of test scores. This form of reliability is called test-retest reliability. Obviously, the second administration of the measure is a parallel form of the first administration, because it is the same test. It is important to keep in mind that this estimate of reliability can be influenced by the procedure used to assess it. For example, shorter intervals between test administrations will produce higher reliability estimates than longer test intervals (Nunnally & Bernstein, 1994). Test-retest reliability estimates can also be influenced by several other factors that reduce the parallel nature of the two test administrations, such as practice effects or developmental changes. These types of effects can result in changes in the test scores at the second administration, which can reduce the reliability coefficient, even though the test itself did not change (Suen, 1990).

Alternate-Forms or Split-Half Reliability

To avoid the issues associated with test-retest reliability, another approach used to assess reliability is to construct two equivalent, alternate forms of the test. However, the effort and cost of constructing two measures rather than one may be prohibitive, and it may be difficult to ensure that the two forms are truly parallel (Suen, 1990). Instead, alternate forms can be created by dividing the scale items into two halves; this creates the parallel forms needed to assess reliability. Because the number of items from the total measure is halved in this process, reliability estimates will be attenuated relative to the total measure and must be corrected with the Spearman-Brown formula (for computational details see, e.g., Nunnally & Bernstein, 1994; Suen, 1990). One potential complication that remains with the split-half procedure is how exactly to split the test items to produce two equivalent halves. Depending on how the items are selected, various test halves can be constructed from a single test, and reliability estimates may then vary depending on how the item split was performed.

Cronbach's Alpha

As the descriptions of test-retest reliability and alternate-forms or split-half reliability have demonstrated, creating parallel forms of a test may not be a simple process. Cronbach's alpha takes what may seem like a more extreme approach to creating parallel forms of a test. With this approach, each item is

considered a separate test and the correlations of each item pair are examined. Nunnally and Bernstein (1994) noted that this intensive process became much easier with the advent of high-speed computers (especially when the number of test items is large) and is considered preferable to the other forms of reliability. The computation of Cronbach's alpha coefficient provides a conservative estimate for the average of all item-pair correlations (Suen, 1990) and can be computed by statistical programs such as SAS (SAS Institute, 1999).

Appropriateness of a Measure

The appropriateness of a measure for the type of individuals included in a study also needs to be considered. Specifically, a measure is most appropriate to use with a particular sample when it has been validated and standardized on a population similar to that being studied. For example, one would not want to use a measure that had been validated on a Caucasian sample for an African American sample, unless it could be demonstrated that no reasonable differences between ethnic groups would be expected on this construct. Questions of measure appropriateness may also relate to gender or to any other important group differences that may influence test scores. Measure appropriateness and the use of correct norms is particularly important when the test scores will be used to make decisions about individual participants (Suen, 1990).

Scales of Measurement

Variables can be measured on one of four scales of measurement: nominal, ordinal, interval, and ratio. *Nominal* variables represent unordered categories of a particular construct or trait. For example, the type of mnemonic strategy used in a memory recall task might include the two unordered categories of (a) verbal mnemonics (e.g., acrostics, acronyms, and rhymes) and (b) visual mnemonics (e.g., the method of loci). In contrast, a variable on an *ordinal* scale rank orders individuals along some continuum. Although the values on an ordinal scale are ordered, the distances between any two ranks or ordered categories do not necessarily represent equal differences between people on that variable. For example, in a study comparing three age groups, the age groups may be considered ordered categories yet may be unevenly spaced.

For both *interval* and *ratio* scales of measurement, equal distances between values on a scale represent equal differences between individuals on the dependent variable. The difference between interval and ratio scales is in the definition of the zero point. A ratio scale has a true zero point; in other words, a score of zero on a ratio scale represents a complete absence of the variable being measured. Such variables are not as common or as meaningful in studies conducted on adult development and learning as they are in the physical sciences (e.g., for height, weight, and length). With physical measurements, statements such as "Person A weighs twice as much as Person B" are valid. With measurements of learning or other aspects of development, it may not be accurate to say that a person with a score of 20 has twice as much of the construct as a person with a score of 10. For example, assume that these scores were generated from an administration of a delayed recall test. Although scores of 20 are twice as great as scores of 10, this does not imply that the memory ability of the individual with a score of 20 is twice as good as the memory ability of the individual with a score of 10. Thus the variable for delayed recall represents an interval scale of measurement. It is true that the variable in this example could be operationalized as a ratio scale by defining the variable of interest as the number of items correctly recalled. However, this interpretation may not be as conceptually meaningful as being able to make conclusions about memory ability.

Measurement of Change

When data on the same construct are collected at two or more time points, researchers typically are interested in examining the amount of change that has occurred on this construct. A well-known critique of the use of change scores is that they tend to be less reliable than measurements taken on any single occasion (e.g., Lord, 1956). However, Rogosa (1988, 1995) demonstrated that the values from which this conclusion was generated were based on an assumption that no individual differences existed in the amount of change. This is an unlikely situation for most research studies. Thus, when individual differences in the amount of change over time are allowed, reliability estimates for the change score are much better, sometimes as reliable as the test itself (Rogosa, 1988, 1995). Although having three or more time points of data provides many more analysis options than just two, the change

score should not be dismissed automatically as an unreliable or poor choice.

In addition to simply subtracting two values to create a change score, it is also possible to examine change at the individual level in terms of the amount of reliable change. We know that some individuals' scores will change over time due to random fluctuations, whereas other individuals will show meaningful change. Use of the standard error of measurement (*SEM*) can be used to define reliable or meaningful change at the individual level (see Dudek, 1979, for computational details). Change scores are compared to the range of values formed by ± 1 *SEM*, and only those outside this range are defined as having reliable change (e.g., Ball et al., 2002; Schaie, 1996a). This method then allows scores that varied only randomly to be classified as stable and those with a significant amount of change (as defined by the *SEM* value) to be classified as having had significant decline or significant improvement. For example, if the *SEM* for change in a measure had a value of 5, then change scores greater than +5 would be classified as significant improvement and change scores less than −5 would be classified as significant decline.

When considering the change over time in a battery of tests, the issue of temporal invariance must also be considered. If the relationships among the tests have changed over time, they may no longer represent the same factors or underlying constructs. Thus the demonstration of longitudinal invariance is important for making interpretable comparisons across time for factor domains (Horn, 1991). This issue is discussed in more depth in the section on confirmatory factor analysis, the method used to test measurement invariance.

SCALE DEVELOPMENT

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a method that is particularly appropriate when there is a new measure for which the underlying dimensions are unknown. EFA is often performed as a precursor to a confirmatory factor analysis, which will be described shortly. We provide a basic overview of the information necessary to determine whether this method would be appropriate to use; specific statistical details can be obtained from texts on factor analysis (e.g., Gorsuch, 1983).

With EFA, the interrelationships among a set of variables are analyzed with the intent to reduce this larger set of scores into a smaller set of common factors. Because this procedure is exploratory in nature, decisions about the "best" factor structure and the interpretation of the factors themselves are made by the individual researcher. Two common guidelines that are often used for determining the proper number of factors are the percentage of the variance explained by that set of factors and the use of scree plots. Optimally, a factor solution will account for at least 75% of the variance (Gorsuch, 1983). Scree plots may also be used to determine where the point where the addition of factors does not result in a meaningful increase in variance explained.

Another decision to be made in EFA is the type of solution to obtain. In cases where one expects factors extracted from a measure or set of scores to be uncorrelated, the orthogonal (uncorrelated) factor solution that is estimated can be used. The varimax rotation is the most popular orthogonal rotation (Suen, 1990). However, in cases where a correlation among the factors is expected or hypothesized, an oblique (correlated) factor solution, such as a promax rotation, is more appropriate to use. If an oblique solution is estimated and factor intercorrelations are low, one may return to the orthogonal factor solution.

Confirmatory Factor Analysis

In contrast to exploratory factor analysis, confirmatory factor analysis (CFA) is used when a measure has a known (or at least hypothesized) factor structure (see Bollen, 1989; Gorsuch, 1983; Jöreskog & Sörbom, 1993 for statistical and computational details). CFA uses the structural equation modeling framework and thus provides fit statistics that a researcher can use to assess directly the fit of this structure to the data that has been collected. Another benefit of the CFA framework is that multiple indicators of the same underlying construct are used to form a better estimate of an individual's true score on that construct than could be provided by a single observed measure. For example, Schaie, Dutta, and Willis (1991) determined that the cognitive battery used in the Seattle Longitudinal Study was best represented by six cognitive domains, or factors, each of which is indicated by at least three measures.

CFA can be a useful method for determining whether the relationships between observed variables and the latent constructs they represent remain

invariant across multiple groups or across time (Jöreskog, 1979). Only when factorial invariance has been demonstrated can one assume that quantitative comparisons of differences in developmental trajectories truly reflect changes in the underlying construct (see Baltes & Nesselroade, 1970, 1973). Shifts in the regression of observed variables on the latent construct, if found, would impose significant restrictions on the interpretability of age changes and age differences measured with single markers. The demonstration of factorial invariance is also important in showing that the relations between observed variables and latent constructs remain stable following the introduction of interventions that might affect such relationships (Schaie, Willis, Hertzog, & Schulenberg, 1987).

A minimum requirement of longitudinal invariance is the demonstration of configural invariance, which requires only that the indicators of the factors have the same pattern of zero and nonzero loadings across time (Horn, McArdle, & Mason, 1983; Meredith, 1993). The next level of invariance is metric invariance, or weak factorial invariance. Weak factorial invariance requires the equality of the unstandardized factor loadings across time. Meredith further proposed the level of strong factorial invariance, which additionally requires equality of the unique (error) variances and intercepts across time. Because stricter levels of invariance can be difficult to meet in many complex studies, it may only be possible to demonstrate partial measurement invariance (Byrne, Shavelson, & Muthén, 1989), where longitudinal invariance can be demonstrated for only a subset of the factors of interest across time.

Another important application of confirmatory factor analysis is the use of this procedure to implement the Dwyer (1937) extension method. As Tucker (1971) demonstrated, it is not appropriate to use factor scores on a latent variable to estimate its regression on an observed variable. However, CFA permits the estimation of the location of some new observed variable or variables of interest within a previously known factor (latent construct) space. This situation frequently arises in aging studies; because samples are followed over long time periods, new measures and constructs are often added to a study. The extension analysis method has been used recently in the Seattle Longitudinal Study to examine the relations of a neuropsychological test battery to the established psychometric intelligence battery (Schaie, Willis, & Caskie, 2004), and of the relations of the NEO Personality Inventory (Costa & McCrae, 1992) to the Test of Behavioral Rigidity (Schaie & Parham, 1975) in Schaie et al. (2004).

EVALUATING THE MEANINGFULNESS OF RESULTS

Clinical Significance

Research findings can be statistically significant without being practically meaningful, particularly in large samples. The reverse can also be true in very small samples, where a finding may have practical significance but the sample size is so small that it is not statistically significant (Urdan, 2001). One way to assess practical or clinical significance may be to focus on whether a change in the level of performance has been observed in the practical aspects of daily life and everyday activities. For example, if a memory training study increased participants' ability to recall a list of words by an average of five words, it would be important to assess whether this increase has translated into improved performance in the older adult's everyday activities. For example, does the individual show an increased ability to recall a short grocery list? This transfer from trained psychometric abilities to more applied abilities can be a useful indicator of the clinical meaningfulness of a research result. Karlawish and Clark (2002) and McGlinchey, Atkins, and Jacobson (2002) provide information on estimates of the clinical significance of an effect.

Effect Size

Many measures of effect size exist; this section presents two of the more commonly used statistics. For linear regression analysis, the R^2 statistic is typically used (Cohen & Cohen, 1983); for analysis of variance (ANOVA), the η^2 statistic can be used. Both of these measures of effect size provide information about the amount of variance that has been explained relative to the total variance. In the context of linear regression, the R^2 value describes the amount of variance explained by the set of predictors included in the regression equation. If several blocks of predictors have been entered as is done in a hierarchical or staged regression analysis, the change in R^2 that occurs with the addition of each block can also be examined. For

ANOVA, η^2 is calculated for each effect included in the model separately, and the relative impact of each effect can be assessed. For example, if the effects of age group and training group on reasoning ability were examined in a two-way ANOVA, the proportion of variance attributed to both could be computed and compared.

Effect size is also an important consideration for power calculations. Power describes the ability of a study to detect a true difference (i.e., the long-term probability of rejecting a false null hypothesis). Cohen (1992) noted that how one chooses the correct population effect size value to be used in a power analysis is often a point of confusion for researchers. In his article, Cohen reviews his previously established conventions for a small, medium, and large effect size and notes that the meaning of these designations is dependent on the type of hypothesis test being conducted. It is particularly important for studies in new areas, such as the intersection of adult development and learning, to be sure that adequate power exists to detect hypothesized or expected population differences.

CONCLUSION

The aim of this chapter was to provide an overview of some of the important research design and methodological topics that need to be considered when proposing and conducting new research in the area of adult learning and development. More detailed treatments of these topics can be found in many of the textbooks and other sources referenced herein. When designing a research project, the choice of study type will have important implications for the types of conclusions that can be drawn. When age differences are of interest, cross-sectional data are sufficient, but longitudinal data are required to address age-related changes in a construct. Random samples ensure the generalizability of a study's results, but the difficulty of obtaining true random samples implies that researchers should be sure to address potential biases associated with the use of nonrepresentative samples when designing, conducting, and reporting research. The need for valid, reliable, and appropriate measures is implicit in all research. Finally, researchers should include information about their study findings that allow readers of the research to determine the practical significance of the work. The incorporation of these key elements into new research studies of adult development and learning will generate solid and robust research findings that can only strengthen this burgeoning field.

References

Alwin, D. F., & Campbell, R. T. (2001). Quantitative approaches: Longitudinal methods in the study of human development and aging. In R. H. Binstock & L. K. George (Eds.), *Handbook of aging and the social sciences* (pp. 22–43). San Diego, CA: Academic Press.

Amato, P. R. (1988). Who cares for children in public places? Naturalistic observation of male and female caretakers. *Journal of Marriage and the Family, 51*, 981–990.

Babbie, E. (1986). *The practice of social research* (4th ed.). Belmont, CA: Wadsworth.

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., et al. (2002). Effects of cognitive training interventions with older adults: A randomized controlled trial. *Journal of the American Medical Association, 288*, 2271–2281.

Baltes, M. M., Burgess, R. L., & Stewart, R. B. (1980). Independence and dependence in self-care behaviors in nursing home residents: An operant-observational study. *International Journal of Behavioral Development, 3*, 489–500.

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development, 11*, 145–171.

Baltes, P. B., & Nesselroade, J. R. (1970). Multivariate longitudinal and cross-sectional sequences for analyzing ontogenetic and generational change: A methodological note. *Developmental Psychology, 1*, 162–168.

Baltes, P. B., & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 219–251). New York: Academic Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research in teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Skokie, IL: Rand McNally.

Caspi, A., Elder, G. H., & Bem, D. J. (1987). Moving against the world: Life-course patterns of explosive children. *Developmental Psychology, 23*, 308–313.

Center for Human Resource Research. (1992). *NLS handbook*. Columbus, OH: Center for Human Resource Research.

Cervone, D. (2004). The architecture of personality. *Psychological Review, 111*, 183–204.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. C., & Campbell, D. T. (1979). *Quasi-experiments: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Cooney, T. M., Schaie, K. W., & Willis, S. L. (1988). The relationship between prior functioning on cognitive and personality variables and subject attrition in longitudinal research. *Journal of Gerontology: Psychological Sciences, 43*, P12–P17.

Costa, P. T. Jr., & McCrae, R. R. (1992). *Manual for the NEO*. Odessa, TX: IPAR.

Costa, P. T. Jr., McCrae, R. R., Zonderman, A. B., Barbano, H. E., Lebowitz, B., & Larson, D. M. (1986). Cross sectional studies of personality in a national sample: 2. Stability in neuroticism, extraversion, and openness. *Psychology and Aging, 1*, 144–149.

Craik, F. I. M., & McDowd, J. M. (1998). Age differences in recall and recognition. In M. P. Lawton & T. A. Salthouse (Eds.), *Essential papers on the psychology of aging* (pp. 282–295). New York: New York University Press.

Dodge, J. A., Clark, N. M., Janz, N. K., Liang, J., & Schork, M. A. (1993). Nonparticipation of older adults in a heart disease self-management project: Factors influencing involvement. *Research on Aging, 15*, 220–237.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86*, 335–337.

Duncan, S. C., Duncan, T. E., & Hops, H. (1996). Analysis of longitudinal data within accelerated longitudinal designs. *Psychological Methods, 1*, 236–248.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum.

Dwyer, P. S. (1937). The determination of the factor loadings of a given test from the known factor loadings of other tests. *Psychometrika, 2*, 173–178.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources, 33*, 251–299.

George, L. K., Hays, J. C., Flint, E. P., & Meador, K. G. (in press). Religion and health in life course perspective. In K. W. Schaie, N. Krause, & A. Booth (Eds.), *Religious influences on the health and well-being of the elderly* (pp. 246–282). New York: Springer.

Gillem, A. R., Cohn, L. R., & Throne, C. (2001). Black identity in biracial black/white people: A comparison of Jacqueline who refuses to be exclusively black and Adolphus who wishes he were. *Cultural Diversity and Ethnic Minority Psychology, 7*, 182–196.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hayward, M. D., & Gorman, B. (2004). The long arm of childhood: The influence of early-life social conditions on men's mortality. *Demography, 41*, 87–107.

Heikkinen, E., Berg, S., Schroll, M., Steen, B., & Viidik, A. (Eds). (1997). *Functional status, health, and aging: The NORA study*. Paris: Serdi.

Helson, R., & Moane, G. (1987). Personality change in women from college to midlife. *Journal of Personality and Social Psychology, 53*, 176–186.

Herzog, A. R., & Rodgers, W. L. (1988). Age and response rates to interview sample surveys. *Journal of Gerontology: Social Sciences, 43*, S200–S205.

Hofland, B. F., Willis, S. L., & Baltes, P. B. (1981). Fluid intelligence performance in the elderly: Intraindividual variability and conditions of assessment. *Journal of Educational Psychology, 73*, 573–586.

Honos-Webb, L., Stiles, W. B., & Greenberg, L. S. (2003). A method of rating assimilation in psychotherapy based on markers of change. *Journal of Counseling Psychology, 50*, 189–198.

Horn, J. L. (1991). Comments on issues of factorial invariance. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 114–125). Washington, DC: American Psychological Association.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1*, 179–188.

Johnson, D. R., & Tang, Z. (2003, November). *Are estimates of family processes from long-term panel studies biased by attrition? An empirical test in a twenty-year panel study*. Paper presented at the annual meeting of the National Council on Family Relations, Vancouver, British Columbia.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software.

Kalton, G., & Anderson, D. W. (1989). Sampling rare populations. In M. P. Lawton & A. R. Herzog (Eds.), *Special research methods for gerontology* (pp. 7–30). Amityville, NY: Baywood.

Karlawish, J. H. T., & Clark, C. M. (2002). Addressing the challenges of transforming laboratory advances into Alzheimer's disease treatments. *Neurobiology of Aging, 23*, 1043–1049.

Klumb, P. L., & Baltes, M. M. (1999). Time use of old and very old Berliners: Productive and consumptive activities as functions of resources. *Journal of Gerontology: Social Sciences, 54B*, S271–S278.

Lachman, M. E., & Jelalian, E. (1984). Self-efficacy and attributions for intellectual performance in young and elderly adults. *Journal of Gerontology*, 39, 577–582.

Lachman, M. E., & McArthur, L. Z. (1986). Adulthood age differences in causal attributions for cognitive, physical, and social performance. *Psychology and Aging*, 1, 127–132.

Levinson, D. J. (1978). *The seasons of a man's life*. New York: Knopf.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16, 421–437.

McArdle, J. J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research*, 18, 145–166.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significance methods: Which one to use and how useful are they? *Behavior Therapy*, 33, 529–550.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.

Minke, K. A., & Haynes, S. N. (2003). Sampling issues. In J. C. Thomas & M. Hersen (Eds.), *Understanding research in clinical and counseling psychology* (pp. 69–95). Mahwah, NJ: Lawrence Erlbaum.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Park, D. C., Smith, A. D., Morrell, R. W., Puglisi, J. T., & Dudley, W. N. (1990). Effects of contextual integration on recall of pictures by older adults. *Journal of Gerontology: Psychological Sciences*, 45, P52–P57.

Patrick, J. H., Pruchno, R. A., & Rose, M. S. (1998). Recruiting research participants: A comparison of the costs and effectiveness of five recruitment strategies. *Gerontologist*, 38, 295–302.

Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York: Springer.

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum.

Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.

Rothblum, E. D., Factor, R., & Aaron, D. J. (2002). How did you hear about the study? Or, how to reach lesbian and bisexual women of diverse ages, ethnicity, and educational attainment for research projects. *Journal of the Gay and Lesbian Medical Association*, 6, 53–59.

SAS Institute. (1999). *SAS/STAT user's guide, version 8*. Cary, NC: SAS Institute.

Schaie, K. W. (1959). Cross-sectional methods in the study of psychological aspects of aging. *Journal of Gerontology*, 14, 208–215.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin*, 64, 91–107.

Schaie, K. W. (1977). Quasi-experimental designs in the psychology of aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 39–58). New York: Van Nostrand Reinhold.

Schaie, K. W. (1988). Internal validity threats in studies of adult cognitive development. In M. L. Howe & C. J. Brainerd (Eds.), *Cognitive development in adulthood: Progress in cognitive development research* (pp. 241–272). New York: Springer-Verlag.

Schaie, K. W. (1994). Developmental designs revisited. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Theoretical issues revisited* (pp. 45–64). Hillsdale, NJ: Lawrence Erlbaum.

Schaie, K. W. (1996a). *Intellectual development in adulthood: The Seattle Longitudinal Study*. New York: Cambridge University Press.

Schaie, K. W. (1996b). Research methods in gerontology. In G. L. Maddox (Ed.), *Encyclopedia of aging* (2nd ed.) (pp. 812–815). New York: Springer.

Schaie, K. W. (2004). *Developmental influences on cognitive development: The Seattle Longitudinal Study*. New York: Oxford University Press.

Schaie, K. W., & Baltes, P. B. (1975). On sequential strategies in developmental research: Description or explanation? *Human Development*, 18, 384–390.

Schaie, K. W., & Caskie, G. I. L. (2004). Methodological issues in aging research. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 21–39). Malden, MA: Blackwell.

Schaie, K. W., Caskie, G. I. L., Revell, A. J., Willis, S. L., Kaszniak, A. W., & Teri, L. (2005). Extending neuropsychological assessments into the primary mental ability space. *Aging, Neuropsychology, and Cognition*, 12, 245–277.

Schaie, K. W., Dutta, R., & Willis, S. L. (1991). The relationship between rigidity-flexibility and cognitive abilities in adulthood. *Psychology and Aging*, 6, 371–383.

Schaie, K. W., & Hofer, S. M. (2001). Longitudinal studies in aging research. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 53–77). San Diego, CA: Academic Press.

Schaie, K. W., & Parham, I. A. (1975). *Manual for the Test of Behavioral Rigidity*. Palo Alto, CA: Consulting Psychologists Press.

Schaie, K. W., & Strother, C. R. (1968). A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin*, 70, 671–680.

Schaie, K. W., Willis, S. L., & Caskie, G. I. L. (2004). The Seattle Longitudinal Study: Relationship between personality and cognition. *Aging, Neuropsychology, and Cognition*, 11, 304–324.

Schaie, K. W., Willis, S. L., Hertzog, C., & Schulenberg, J. E. (1987). Effects of cognitive training upon primary mental ability structure. *Psychology and Aging, 2,* 233–242.

Schleser, R., West, R. L., & Boatwright, L. K. (1987). A comparison of recruiting strategies for increasing older adults' initial entry and compliance in a memory training program. *International Journal of Aging and Human Development, 24,* 55–66.

Sharma, S. K., Tobin, J. D., & Brant, L. J. (1986). Factors affecting attrition in the Baltimore Longitudinal Study of Aging. *Experimental Gerontology, 21,* 329–340.

Sitlington, P. L., & Frank, A. R. (1990). Are adolescents with learning disabilities successfully crossing the bridge into adult life? *Learning Disability Quarterly, 13,* 97–111.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Stanovich, K. E. (2001). *How to think straight about psychology* (6th ed.). Boston: Allyn & Bacon.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

Thornquist, M. D., Patrick, D. L., & Omenn, G. S. (1992). Participation and adherence among older men and women recruited to the beta-carotene and retinol efficacy trial (CARET). *Gerontologist, 31,* 593–597.

Tonry, M., Ohlin, L. E., & Farrington, D. P. (1991). *Human developmental and criminal behavior: New ways of advancing knowledge.* New York: Springer-Verlag.

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika, 36,* 427–436.

Urdan, T. C. (2001). *Statistics in plain English.* Mahwah, NJ: Lawrence Erlbaum.

Wagner, E. H., Grothaus, L. C., Hecht, J. A., & LaCroix, A. Z. (1991). Factors associated with participation in a senior health promotion program. *Gerontologist, 31,* 598–602.

Willis, S. L. (2001). Methodological issues in behavioral intervention research with the elderly. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 78–108). San Diego, CA: Academic Press.

Woods-Brown, L. Y. (2002). Ethnographic study of homeless mentally ill persons: Single adult homeless and homeless families. *Dissertation Abstracts International, 62* (10-A), 3460.