Quantitative Methodology for Family Science

4

Alan C. Acock and Isaac Washburn

Quantitative methods are widely used in studying families because there are so many research questions they can address. What variables predict the person you select as a life partner? Why do some families flourish while others dissolve? Why do some families persist when all members have a miserable relationship with one another? Does cohabiting prior to marriage improve or hurt marital outcomes? Is marriage more beneficial for women or for men? These are a few of the questions we investigate using quantitative methods.

Quantitative analyses and statistics emerged as pivotal scientific methods when causal determinism was rejected in quantum mechanics and replaced by a probabilistic view of the world in the early twentieth century (Liboff, 2002). Since then, statistical analyses have played an increasing role in family scholarship. These methods fit family scholarship nicely, as few aspects of family life are deterministic, a probabilistic methodology is essential. We cannot determine who will have a successful life, but we can identify family processes that enhance the probability of achieving this success.

There are two broad classes of quantitative methods, namely, those based on surveys or observations and those involving some type of experimental or quasi-experimental design. Adopting an experimental design allows a stronger causal argument than would be possible with a survey, but it is difficult to utilize experimental designs to answer many family-related questions. By far, survey analysis is the method most commonly used in quantitative research published within the major family journals.

We can further subdivide both survey and experimental research based on whether the data are collected at a single time (cross-sectional) or over time (longitudinal), either for a panel of the same people or using a separate cohort of people each time. A panel design has many advantages for a causal analysis because a cross-sectional study can only demonstrate that variables covary, while a stronger causal case can be made with a panel design. A review of the major family journals shows a recent transition from cross-sectional designs to an increased use of longitudinal designs.

In this chapter, we examine both survey and experimental methods, covering topics of how data are collected, how variables are measured, and how statistical analysis is utilized. We attempt to set a high standard for what the best practices are, recognizing that current research often falls short of these best practices. We do this in the hope that the next generation of researchers will work to conform to if not exceed these standards.

A.C. Acock, PhD (✉)
Department of HDFS, College of Health and Human Sciences, Oregon State University, 322 Milam Hall, Corvallis, OR, USA
e-mail: alan.acock@oregonstate.edu

I. Washburn, PhD
Oregon Social Learning Center, Eugene, OR, USA
e-mail: ijnewtonmm@hotmail.com

G.W. Peterson and K.R. Bush (eds.), *Handbook of Marriage and the Family*,
DOI 10.1007/978-1-4614-3987-5_4, © Springer Science+Business Media New York 2013

Data Collection Practices

Sample Type

Surveys that sample some members of a target population are a widely used way of obtaining data for quantitative studies. Some surveys focus on a narrowly defined research question; others include a wide range of research questions. An example of the first type would involve interviewing couples to measure maternal and paternal depression just prior to the birth of their first child and then each month after the birth until the child is 1-year old.

The second type of sample is a broad range survey. A research team with funding from the National Institute of Health, for example, may conduct a broad based survey of parents with a 12-year-old child at the start of the study, repeating interviews annually until the child is 25. This research team has a general focus, perhaps how families influence the transition from adolescence to adulthood, but their survey instruments include items on far-reaching topics. Hundreds of subsequent researchers may utilize these data to cover a wide variety of topics, many of which were unimagined by the research team that designed the survey.

The small, local survey, because of its narrow focus, typically offers good measures of the key concepts. For example, maternal depression might be measured using a well-developed 20–40-item scale. The large, national survey, because it has a comprehensive suite of hundreds of concepts to measure, uses a very short list of questions for each variable. Sometimes there may be only a single question serving as an indicator of a complex concept. Often there are only 1–5 items rather than a well-developed scale to measure complex concepts such as marital satisfaction.

The strength of the narrowly focused survey is its precise measurement of key variables. Is this important? When we have poor measurement of independent variables and hence a lot of measurement error, we get biased results. Acock (1989) showed that when there was a lot of measurement error, the effects of a variable in a complex model will generally be underestimated even to the point of reversing the direction of the true relationship.

Is the small, narrowly focused survey best because of superior measurement? Perhaps not, the large scale and comprehensive national surveys allow us to generalize our findings to the broadest possible population. If we limit our study to a small sample of college students or to people who live in a small university town, we cannot generalize beyond these communities. What we find for a particular group of college students may not be true for young adults who chose not to attend college.

National surveys typically provide high quality probability samples that use complex sample designs. These surveys may sample clusters of observations or oversample certain groups such as minorities. These differences from a simple random sample require special adjustments that are not available with all statistical software programs. This is particularly problematic when a researcher is analyzing a subsample such as couples who are recently married selected from a national sample of all married couples (West, Berglund, & Heeringa, 2008). A complex sample design that can be used to generalize to all married couples may not allow generalization to subgroups such as recently married couples without complex weighting procedures that are developed just for this subsample.

This often leaves the family scholar in a quandary. Select a national survey and you have very limited measures of key concepts and often very limited ability to predict outcome variables. Conduct a highly focused, small-scale survey and you have trouble generalizing your findings. A review of major family journals shows that the large scale, national surveys are responsible for more publications than the smaller samples that are highly focused.

Non-probability Sample

Is a non-probability sample okay? Many studies in fields such as medicine rely on convenience samples of patients who volunteer to participate in a randomized clinical trial. The convenience sample is problematic if underrepresented groups respond differently to the treatment. Early medical studies often were limited to men because they were thought to be "less complicated" biologically than women. The unanticipated consequence was that research of special interest to women was not conducted.

Today, much medical research still relies on volunteer samples. This works when the processes being studied do not depend on economic, racial, regional, or cultural variation among participants. A pill that lowers blood pressure is assumed to work for Baptist as well as for atheists; for people who are married as well as for people who are cohabiting. The probabilistic effects are thought of as universal.

This rarely works for family research. The effect of close supervision of adolescents differs between cultures where it is normative and cultures where it is not. History, culture, class, race/ethnicity, and religion closely bound family life. Our study sample delimits what generalizations can be justified. Suppose you want to predict factors that influence a woman's marital satisfaction and you have a sample of married couples that have at least one child who is 12 years old. Many of these couples will have been married for at least 12 years and some for much longer. Wives who had a very low level of marital satisfaction are likely to have divorced before the study. Also, the factors that are important to a woman who has been married for 15 years and has an adolescent child may be dramatically different than the factors that are important to younger couples or childless couples.

A final problem with highly focused surveys is that they often result in a restricted range on many variables. Statistical analysis relies on variation to predict outcomes. Any artificial restriction on that variation results in underestimation of the effects of a variable. A sample drawn from a small college town might result in most parents having at least some college. A survey of mothers who are identified as being in a high-risk family may include many that have less than a high school degree. Both surveys will have truncated variation on education and, as a consequence, education will have a smaller effect on any outcome variable than it would in a representative sample that reflected the full range of parental education.

Complex Designs

Most national probability samples are what we call complex sample designs. They may involve elements of stratification and clustering. If you are doing a national survey that involves face-to-face interviews, you may first randomly sample a group of 50 counties proportional to size. Sampling proportional to size simply means that a county with a large population has a higher probability of being selected than a county with a small population. As a second stage, you may randomly sample 20 blocks within each county, again proportional to size. As a third stage, you might randomly select five people from each block. This results in a sample of $50 \times 20 \times 5 = 5{,}000$ people. This is a probability sample, but it is not a simple random sample. Why use this complex sample design? The cost savings are huge. You only need to have an interview staff in 50 counties and, within the counties, only 20 blocks. When doing interviews within a block, if a person is not available, the interviewer has four other people close by who might be available.

A researcher using a national survey needs to incorporate the design features into the analysis. Few family scholars have done this and thus findings can be quite biased. The people who live on the same block (cluster) are going to be more homogeneous than people selected at random from the United States. More than likely, neighbors have similar education levels, income, and ethnicity. They also share a common culture to a greater extent than people selected randomly from across the United States. It is important to adjust for the dependencies caused by the clustering. The effects of not weighting are equally problematic when we are trying to estimate a population central tendency. If the groups we oversampled have below average income, without weighting we will underestimate the true average income. Weighting can be ignored in certain situations.

If the relationship between variables is the main focus then weighting might be left out. This may be okay if the relationships are the same for your subgroup as they are for all other subgroups. With standard statistical software such as Stata (2009), SAS, and even SPSS with an add-on module, adjusting for complex sample designs is available for a variety of statistical procedures. To date, however, most publications have failed to incorporate the sampling design in the analysis.

Presenting Sampling Methods

In presenting the results of our analyses, it is important to describe the method of sampling. Using a national secondary data source is no excuse to ignore the sample design. Readers of an article should have a clear understanding of who was sampled, where they were sampled, how they were sampled, and why they were sampled. Often a table of demographic information that compares your sample with population characteristics is sufficient and recommended. This simple addition, if editors were to require it, would improve the quality of articles printed in our field.

We will only briefly mention the variety of data collection methodologies used in family research. These include face-to-face interviews, mailed questionnaires, telephone interviews, and Internet-based surveys. Each of these has advantages and disadvantages. Face-to-face interviews have reactivity between the interviewer and participants where characteristics of the interviewer may bias responses (see, e.g., Williams, 1994). Telephone interviews reduce this bias but must keep possible response options quite simple and the increasing number of people with only cell phones also presents a problem as it is illegal to solicit cell phones. All approaches have problems with individuals refusing to participate, and this is especially serious with Internet surveys. Individuals who refuse to be interviewed may be very different from those who agree to. This will introduce bias in the findings. Any information about refusals is helpful.

A final issue in sampling is the amount of incentive each participant is provided. Participants deserve some compensation for their participation, however the amount of compensation may be too little or too large. A study focusing on families whose incomes are below 200% of the poverty level that offers a $100 incentive to participate could be paying too much. Participants may feel they have no choice, rendering the incentive tantamount to a bribe. This compromises the meaning of voluntary participation. Researchers should disclose information about the compensation when presenting their results.

Experimental and Quasi-Experimental Designs

Experimental and quasi-experimental studies are rarely used in family studies. One reason for this is that it is hard to manipulate the types of independent variables family scholars study. Evaluation research is one area where experimental and quasi-experimental designs are being used. The work done by Patterson and colleagues at the Oregon Social Learning Center are good examples of this (Marion, Forgatch, Patterson, Degarmo, & Beldavs, 2009; Patterson, Chamberlain, & Reid, 1982). The most important application of these design techniques is in studies that involve an intervention such as an approach to counseling. A true experiment requires randomization of participants to the different conditions. A waitlist is often used in a randomized trial where those in the control group are offered the program content after the data for the study is collected.

Randomization and Random Sampling

Random sampling deals with how your sample was selected from the population and deals with external validity or the generalizability of your sample. Although random samples are rare, we often have a complex sample design where each participant has a known probability of being selected. These are called probability samples rather than random samples. Random assignment (or randomization) refers to using a random process to assign participants to each condition and deals with internal validity. Separately, each technique does something different. A study of the relationship between infidelity and divorce in a national probability sample may be generalizable to the nation (external validity). The problem is that infidelity cannot ethically be randomized (internal validity), and so other traits of individuals that cheat on a spouse might be responsible for both the divorce and the infidelity. A second study of a voluntary sample where couples were randomly assigned to receive a positive or negative topic to discuss and then the couples' emotional closeness was measured has the benefit of randomization. However, the findings cannot be generalized to a large population. In the case that neither randomization nor random samplings are used, both causality and generalizability are problematic.

More on Experiments and Quasi-Experiments

A quasi-experiment lacks randomization. If participants have not been randomized then it is possible that pretest differences in variables other than the variable of interest could be the cause of relationships among variables. It is important to remember that randomization does not protect against all threats of internal validity. Participants in the control group may develop a correct or incorrect understanding of the purpose of the study and change behavior based on the understanding. For example, control group participants in a study of marital communication skills may look online for ideas (Shadish, Cook, & Campbell, 2002).

An important way to avoid this issue is by measuring how well your control and experimental groups actually remained control and experimental groups. School interventions are a good example because of the alternative programs available to schools. During the course of an intervention, control schools might implement a different but overlapping program. For example, a new drug and alcohol prevention program is tested in a series of schools and the control schools start the D.A.R.E. program half way through the study. The design is no longer treatment and control, but treatment A vs. partial treatment B. Measures of the level of fidelity of the implementation need to be included for the experimental condition, but may also be needed for the control group if any contamination is possible.

A concern of any experiment or quasi-experiment is the duration of effects. Once a treatment has ended, how long do the effects of the treatment last? Do they persist for 6 months, 2 years, or a lifetime? The test of the longevity of affects needs to be worked into the design from the beginning of the study. A pretest and posttest require a follow-up test.

Power

Whether using a survey or an experimental approach, the statistical power of your analysis is important. Family researchers pay attention to type-one error, relying on results being statistically significant at the 0.05 level. Very few published articles address the issue of power, which is an important complement of statistical significance. Statistical power refers to the ability to obtain a statistically significant result when the true result is a difference or relationship that the researcher considers substantively significant. If our sample is too small, we lack power to show a result is statistically significant even when there is an important real difference. By contrast, if our sample is very large, we have power to demonstrate a result is statistically significant even when it is substantively trivial.

Consider a comparison of the time fathers and mothers spend each evening playing with their children. First, we need to decide how much of a difference is substantively important. We will say a difference of one half of a standard deviation is important. This is generally regarded as a medium effect whereas 0.2 standard deviations would be a small effect and 0.7 standard deviations would be a large effect (Cohen, 1988). How many observations do we need to have a power of 0.80? Figure 4.1 illustrates this showing that for a power of 0.80 we need about 128 people. Assuming we select the same number of women as men we would need 64 women and 64 men. If we did this study over and over again, we would expect about 80% of the studies would have a statistically significant difference at the 0.05 level, when there was a 0.5 standard deviation difference between the women and men.

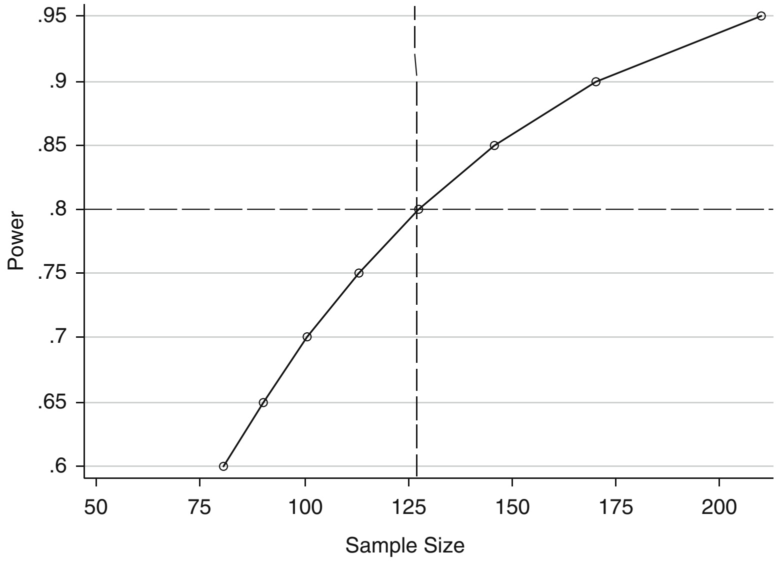


Fig. 4.1 Relationship between sample size and power

What if you can only afford to study 40 women and 40 men? Your power would be just –0.60. In other words, you face a 40% risk of finding an insignificant result when there really is a moderately strong difference. Would you be willing to go through all the work to do such a study when the risk of failure was this great? If such a study finds no significance gender difference, there is little confidence this was because there really was no difference. Many funding agencies insist on a power analysis before they will fund a project. There is little justification to fund an underpowered study that has a high risk of failure. Conversely, the funding agency may be reluctant to fund an extremely large and costly study when they feel a study with a much smaller sample would have ample power. Some funding sources would like to see a power of at least 0.80 and others would want a power of 0.90.

Can you have too much power? Not really, so long as you are sensitive to potential misinterpretations of statistical significance. Imagine you are doing the same comparison of time women and men spend playing with their children, but you are using a large national survey that included 6,500 women and 6,500 men. You find the difference is statistically significant at the 0.05 level. Does this mean the difference is substantively significant—is it important? We don't know from this information, because with a sample this large you have a power of over 80% to detect a statistically significant difference when the actual difference is trivial—0.05 of a standard deviation. If the standard deviation of time spent with the child were 10 minutes, this would be a difference between women and men of just 30 seconds would be statistically significant.

The solution to the potential problem is to interpret the size of the statistically significant difference you observe in terms of its substantive importance. If the observed difference is trivial then the statistically significant result is also trivial. What is statistically significant is that the finding is unimportant. Some large studies misinterpret results as important when they are statistically significant without first making a substantive interpretation of the effect size.

Imagine having an intervention to increase emotional support that fathers give to expecting mothers. The mean for your control group that does not receive any intervention is 50 with an SD = 10. The mean for your treatment group is 51 and it also has an SD = 10. The effect size using Cohen's *d* is 0.1. Figure 4.2 shows

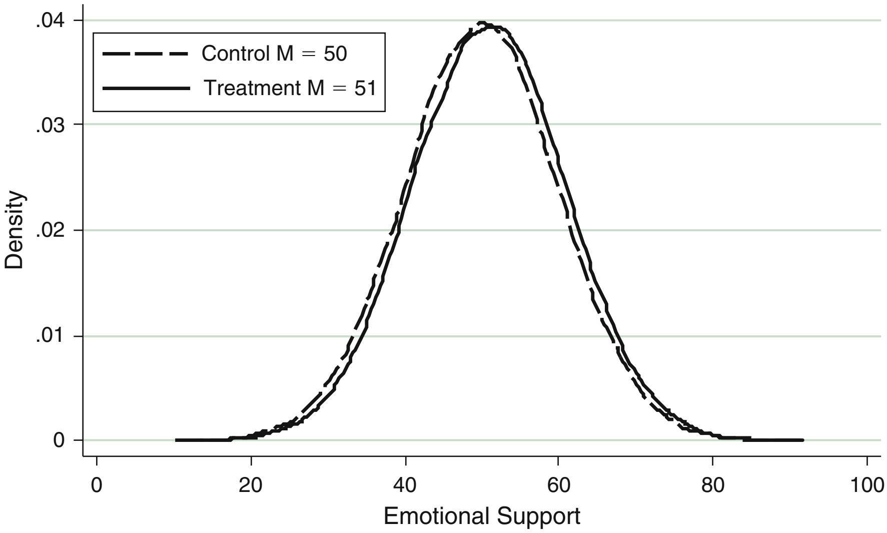


Fig. 4.2 A small effect

hypothetical data that is approximately normal for a sample of 5,000 in the control condition and 5,000 in the treatment condition. With such a large sample, the difference is statistically significant, $t(9{,}998) = -5.60$, $p < 0.001$. An inspection of the two distributions in Fig. 4.2 tells us that the result is extremely significant statistically, and with a $d = 0.1$, that we can be confident that the effect is very weak. Indeed, in this hypothetical data, 45% of the participants in the control group scored above the mean score for the treatment group.

Great progress has been made in estimating statistical power. A free software program called G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) does this for many types of analysis. For more complex analyses and especially those with missing data, it is helpful to estimate power using a simulation. In a simulation, the researcher provides a model with estimates that would be considered important. Then a hypothetical population is generated based on these parameter estimates. The next step is to draw many random samples, 1,000 or more, from this hypothetical population. The proportion of these samples that have statistically significant results is the estimate of power (see Davey & Savla, 2009).

Multicollinearity

Many variables are often needed to predict an outcome about family life. Many family and individual variables could be at play, including parental conflict, income, mother's education, and emotional support of the child. When pairs or combinations of predictors are highly correlated, it is difficult to separate their individual effects. Sometimes they cancel each other out and neither is significant. Sometimes one of them is highly significant and the other one, that is almost as important, become insignificant. Multicollinearity is evaluated using the variance inflation factor (vif), a reflection on how much the standard errors are inflated due to the multicollinearity. Family researchers often use a statistic called tolerance that is 1/vif. Tolerance is the variance in each predictor that is not explainable by the other predictors. If your set of predictors explains 95% of the variance in a single predictor, then only 5% of its variance is available to explain the outcome variable. When our variables contain substantial measurement error, that 5% may be mostly measurement error.

The meaning of multicollinearity is perhaps clearer in Fig. 4.3. On the left, Panel A, we have

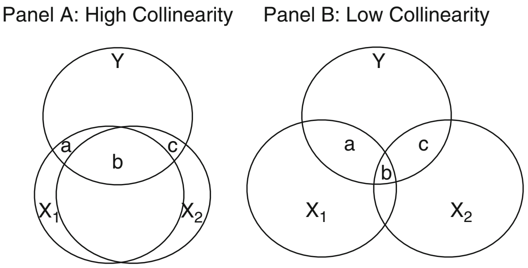


Fig. 4.3 Multicollinearity

two predictors X_1 and X_2 that are highly correlated. Perhaps these are the marital satisfaction of the wife and her husband. We are predicting an outcome variable, Y. Most of what X_1 explains is also explained by X_2 (labeled b in the figure). X_1 only gets credit for what it contributes uniquely, the sliver labeled a. There is no way to decide how to allocate the shared or joint effect area, b, so multiple regression simply does not try to allocate it. The figure on the right, Panel B, shows what happens when there is not a lot of correlation between X_1 and X_2. In that case, the shared explanatory power becomes minimal and what can be allocated to X_1, labeled a, is much larger. When there is a lot of multicollinearity, a rule of thumb is a vif should be less than 10 or the tolerance should be greater than 0.10, you may have a substantial R^2 even though each predictor seems to have very little direct effect.

Solutions to multicollinearity are sometimes available. When two variables, say X_1 and X_2 enter into an interaction (X_1X_2) term, the individual variables will be highly correlated with their product. Many researchers center their variables prior to generating the interaction term to mitigate this problem, where $x_1 = X_1 - M_{X1}$ and $x_2 = X_2 - M_{X2}$. Sometimes when there are several highly correlated variables, a composite score can be used or the variables can be included in a factor model. Multicollinearity is not always a serious problem. If you have a set of control variables and are not interested in their individual effects but simply need to control for them, multicollinearity within the set of controls (but not between them and the primary independent variables) is not a serious problem.

Attrition

The increase in longitudinal analysis in family studies means that attrition is an enormous problem. Attrition occurs when participants who start the study, drop out. The reasons for their dropping out are rarely random. People who find an intervention disagreeable are more likely to drop out. People who have personal problems such as being clinically depressed are more likely to drop out. People in a control condition are often more likely to drop out. Without considering attrition, the findings can be highly biased.

An author is obligated to provide detailed information on attrition. How many people were present at the start? How many of these people were present at each subsequent wave of measurement? There should be data on relevant background variables for all initial participants. It is then possible to compare the people who drop out to those who stay for the length of the study. Do more men drop out? Do people who are less motivated drop out? If there is a control group, are members in this group more likely to drop out? By identifying the variables for which there are substantial differences, it is possible to include scores on those variables as covariates. In epidemiology and medical research an approach called intent to treat is becoming standard. This strategy is based on the effect for people you intend to treat regardless of whether they dropped out for any reason (see Lachin, 2000).

One major mistake with attrition is the use of listwise or casewise deletion that occurs in panel studies where people are measured at 3, 4, 5, or more waves. For example, we might measure adolescents at the start of each grade from grade 7 to grade 12. If a student were measured every year except the ninth grade, the student would be dropped because of incomplete data. Eliminating such a student destroys the data you have for the student for five of the sixth years. Excluding that student gives you less information to estimate the statistics for each of those waves and will likely bias your results.

Missing Values

Missing values occur when participants in a survey or experiment simply do not answer questions. In longitudinal studies, missing values can also occur when participants are not available to answer any questions at one or more waves of data collection. It is not unusual for 30–50% of participants to have some values missing. There are many reasons why the values may be missing. You have people who accidently skip an item. On the day that a questionnaire is administered in a school, you may have 10% of the children absent. Some questions involve sensitive information (frequency of oral sex, for example) and participants may refuse to answer.

Types of Missing Values

There are three types of missing values. Missing completely at random (MCAR) means just that. You might be doing a survey of preadolescents and each child gets a random sample of 70% of the items. This might be done to keep the questionnaire short enough to match the attention span of preadolescents (Graham, Hofer, & MacKinnon, 1996). Missing at random (MAR) means that missing values are explained by the other variables in your survey. Some variables might not be of substantive interest, but are important in explaining the missingness. These other variables, called auxiliary variables, provide the mechanism for predicating missingness. Such variables include gender, age, race/ethnicity, and cognitive ability. Once you control for these auxiliary variables and other variables in your study, the remaining missing values are MAR. There is no test for the MAR assumption because there could always be some variable you did not include that helps explain missingness. At the same time, if you have a wide variety of variables and auxiliary variables in your dataset, the MAR assumption is reasonable. The third type is missing not at random (MNAR). This happens when you fail to meet the MAR condition. When this happens you should make every possible effort to find auxiliary variables that allow you to treat the data as MAR (Molenberghs, Beuchkens, Sotto, & Kenward, 2008).

Only studies that involve planned missingness are likely to meet the MCAR assumption. The more common situation is to have the MAR assumption be reasonable. Until recently, family researchers have failed to properly report and analyze data that meet the MAR assumption. Listwise deletion, where an observation is dropped if the observation has any missing data, overlooks a wealth of information about the observation and results in ignoring 30–50% of all participants. At best, this is a great loss of statistical power and will return biased results if the missing values are not MCAR.

Multiple Imputation and Full Maximum Likelihood Estimation

Landmark work by Rubin and Little (Little & Rubin, 1987; Rubin, 1987) provided an integrated treatment of multiple imputation as a solution to missing values. Since then, statistical software programs (e.g., SAS, Stata, Mplus, and SPSS) have provided a simplified way of doing this. Although the software solutions vary somewhat in how they impute missing values, all of the solutions provide much better results than earlier approaches. Multiple imputation assumes MAR. It is important to remember that these methods do not make new information. The parameter estimates and their standard errors assume the missing values are consistent with the observed data. They also incorporate a random error for each imputed value to insure that the uncertainty of the imputation process is incorporated. Multiple imputation should not be confused with single imputation procedures that yield biased estimates of the standard errors.

An alternative approach is offered by the structural equation modeling (SEM) programs such as LISREL, EQS, Amos, and Mplus. These statistical packages use a full information maximum likelihood approach that also uses all available information in the dataset. These approaches produce results that are similar to those of multiple imputation (Acock, 2005).

There are certain pitfalls when doing multiple imputation. First, you want to include all variables in that are in your analysis as well as selected auxiliary variables when doing the imputation. Including additional auxiliary variables that either predict the value a person has or predict who will or will not answer the item helps. Adding up to five auxiliary variables that help predict the score and five auxiliary variables that help explain missingness is reasonable. A second concern is that you do not want to push multiple imputation too far. If you have a massive amount of unplanned missing values, it is unlikely that you can justify the MAR assumption.

Some researchers are hesitant to impute missing values on the dependent variable. This is a mistake because the assumptions of the multiple imputation process assume the full covariance matrix is being analyzed. Leaving out any variable violates this assumption. There is a rapidly developing literature on working with missing values. Schafer (1997) provides an accessible book length treatment of multiple imputation. Schafer and Graham (2002) provide guidelines, as do Graham (2009) and Peugh and Enders (2004). Guidelines specific to family research are provided by Acock (2005).

Measurement

To many family scholars, measurement seems to be an unimportant issue. They use a secondary dataset within which they search for items that measure the concepts of interest. They may generate a new scale consisting of 3 or 4 items to measure a complex concept. Since they are analyzing secondary data, it is impossible to do any pretesting. The methods sections of many articles in leading family studies journals pay little attention to measurement. They may report that they used a scale that some other research showed was reliable on a different population. Whether a scale was reliable on some other population is sometimes important, but this is not demonstrating that it is reliable or valid for your population. Rarely is there any evidence of the validity of the measure and even more rarely is such evidence shown for the study population. The current state of measurement in family studies is inadequate. If there is a normative standard in family research, it is probably to simply report the alpha reliability.

Most statistical procedures such as multiple regression assume perfect measurement in the predictor variables. When there is just one predictor the bias is simple; the greater the measurement error, the lower the correlation. With several predictors, this bias is more complicated. If one predictor has a little measurement error and another predictor has a lot of measurement error, the effect of the predictor with less error may be exaggerated and the effect for the predictor with more measurement error may be underestimated (Acock, 1989). When multicollinearity is a problem, measurement error is especially problematic and much of the independent variance in a variable may be based on errors in its measurement.

The Problem with Alpha

There are two versions of alpha. One of these is unstandardized (estimated using the variances and covariances). The other is standardized (estimated using the correlations of the items). For the standardized version, the formula for alpha is:

$$\alpha = \frac{k\bar{r}}{(k-1)\bar{r}+1}$$

where k is the number of items and $\bar{r}$ is the mean correlation of the k items. Note that two parameters determine the value of alpha, k and $\bar{r}$. The larger the average correlation among the items the larger alpha and the more items the larger alpha. Figure 4.4 shows a graph of this relationship between k, $\bar{r}$, and α.

With an average correlation of just 0.3, alpha reaches the desired value of 0.8 with just 9 items. Going from 10 items to 50 items does not result in much increase in alpha, but even with an average correlation of 0.5, a 3-item scale will not reach an alpha of 0.8. Alpha, a measure of internal consistency, is the most widely reported measure of reliability. If you have 50 items that have a mean correlation of 0.10, your alpha will

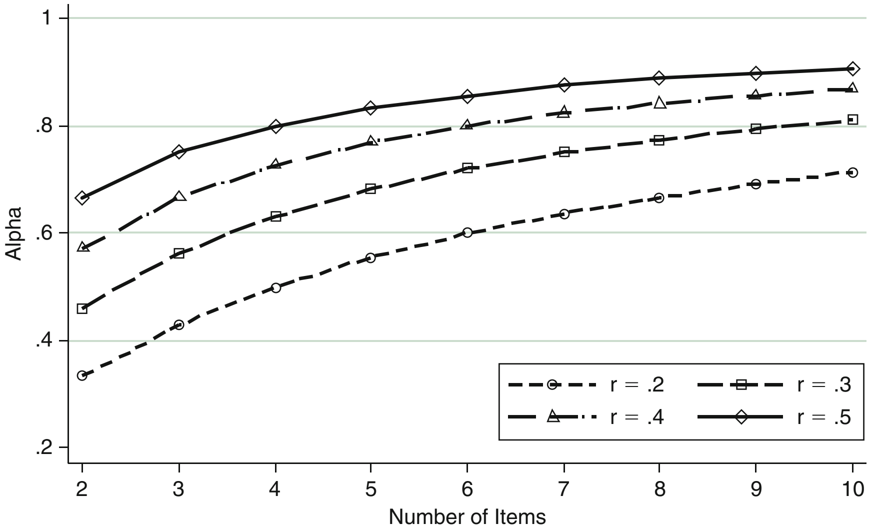


Fig. 4.4 Relationship between k, r

be good, i.e., over 0.80. This is because you have so many items even though they do not have much in common. Remember that when you square 0.10 you get 0.01 meaning that 1% of the variance in a pair of items is shared and 99% is not. These items, in spite the high alpha do not have much in common. When a researcher reports a high alpha for a scale consisting of 50 items, this does not automatically mean the scale is good. It is reliable, but pairs of items might not share much in common.

We can select items to maximize alpha by adding a couple items that will be highly correlated. The most efficient way to do this is to add items that have similar means and standard deviations to the items we already have. If we have three items with a mean of about 3.0 and a standard deviation of about 1.5 on a 1–5 scale, we need to add items with similar means and standard deviations. Essentially, we increase alpha by adding items that are largely redundant. In so doing, we increase our alpha to an acceptable standard, say 0.80, but still do a bad job of measuring real differences between people. Even when we add items, the items we add tend to differentiate people who are in the middle of the distribution from each other rather than people who are at the ends of the distribution. This results in a serious truncation of variance where our measured variance is much less than the true variance of the concept. Truncated variance leads to small correlations and insignificant findings.

Consider a measure of satisfaction with the sexual relationship you have with your partner. The underlying continuum from very dissatisfied to very satisfied is shown in Fig. 4.5. Say we have a sample of just five people, Jim, Juan, Sonya, Maria, and Dre. Let's imagine we knew their true score on the scale, represented by the location of their names along the continuum.

Jim's true score is 0, Dre's is just under 2, Juan is between 3 and 4, Maria is a bit over 4, and Sonya is a 6. Our five items, labeled A, B, C, D, and E, are arranged by their degree of difficulty. The easiest item to endorse by agreeing or strongly agreeing is item A. Since Juan, Maria, and Sonya have higher true scores than Jim or Dre, this item distinguishes them as more satisfied with their sexual relationship. Notice, we have no item that helps us show that Dre is more or less satisfied than Jim. Three of our five items, B, C, and D, let us distinguish very closely between Juan and Maria. These items give us largely redundant information. If we replaced items B, C, or D with an item that was easier, say between 0 and 1, we could use it to discern between Jim,

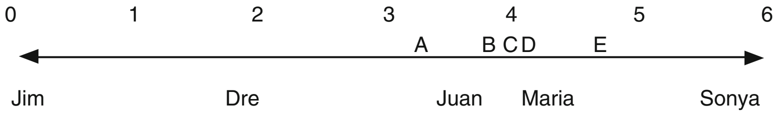


Fig. 4.5 Relationship between true score and degree of item difficulty

who is miserable, and Dre, who is somewhat dissatisfied. Without such an item, Jim and Dre will have the same response pattern and our scale will miss an important difference.

A common example given by measurement experts who use Rasch Modeling is measuring how high a person can jump. If each person is given five jumps, but the height of the obstacles to jump over are all between 3.5 and 4 ft, how useful is our measure? It is great if we want to distinguish between people who can jump 3.5 ft and those who can jump slightly more. It is terrible if we want to measure how high competitive athletes can jump or people with some serious limitation on their ability to jump. We need to have a series of items that lets us differentiate people across the full spectrum of the concept we are measuring.

Another byproduct of our reliance on alpha is that many of our scales result in highly skewed distributions. Scales that do a poor job of differentiating people at the ends of the continuum tend to have a big lump near the top (or bottom) of the distribution. We see this with marital satisfaction where a large number of people strongly agree that they are satisfied. Surely some of the people selecting the strongly agree response are more satisfied than others. By putting a premium on additional items that provide redundant information, and hence are highly correlated with existing items, we fail to ask the type of questions needed to differentiate people at the extremes.

Dimensions of a Measure

Is there a single underlying continuum on which we can locate people such as the one illustrated in Fig. 4.5? Consider marital satisfaction. Can a single score represent such a complex concept? Perhaps, but only in the most general sense can multiple dimensions be represented by a single score. Graduate students taking the GRE get a score on Quantitative ability and a score on Verbal ability as well as a total score. Suppose Juanita has a 780 on the quantitative section and a 350 on the verbal and Rick has a score of 350 on the quantitative and 780 on the verbal for the same total score. If you were selecting for a math program, would you say they were equally qualified because they both had a total score of 1,130?

We have the same problem when we use a scale to measure marital satisfaction. Juanita may be highly satisfied with the sexual aspect of her marriage, somewhat satisfied with the parenting role, somewhat dissatisfied with the division of labor for housework, and completely dissatisfied with the emotional support she receives. Rick may have a very different pattern, but both of them may have the same total composite score. They have very different marriages and very different relationships, but we would measure them as having the same.

Many measurement experts argue that if you have more than one dimension you should have more than one scale (Furlow, Ross, & Gangé, 2009; Kirisci, Hsu, & Yu, 2001). Just like the math GRE may be more important than the verbal GRE for predicting performance in a math program, different dimensions of marital satisfaction may have different consequences. The total score glosses over these distinctions and leads to weak predictive power.

Others argue that creating many small measures is not the answer (Cheng, Wang, & Ho, 2009). In particular, it is argued that bandwidth (the amount measured) and fidelity (the accuracy of measurement) are often conflicting. If we have large bandwidth (with multidimensional measures), then we sacrifice fidelity. If we have good fidelity (a measure of a single dimension), we lose out on bandwidth.

If you feel there is a higher, general dimension of marital satisfaction and that specific dimensions such as financial security reflect the general dimension, you might want to do a second order confirmatory factor analysis to measure marital

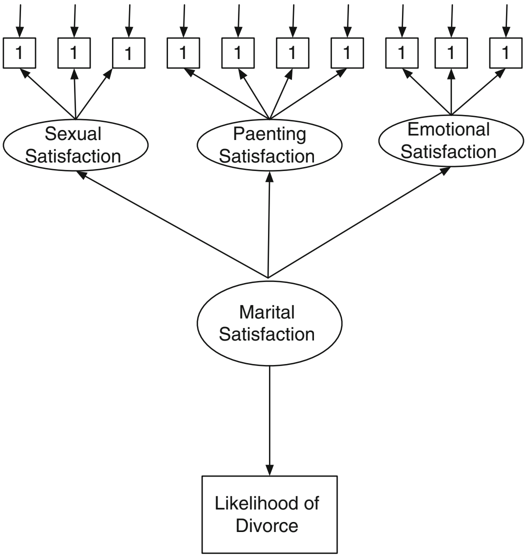


Fig. 4.6 Second order confirmatory factor analysis

satisfaction. This is illustrated in Fig. 4.6 where marital satisfaction leads to the likelihood of divorce. Those who argue that you need a single dimension would use multiple regression of the specific dimensions of marital satisfaction to predict the likelihood of divorce. If you believe that a second order factor exists, then either a multidimensional Rasch model or Item Response Theory (IRT), or a SEM model would be your best avenue for controlling for this. These approaches are rarely reported in family research.

Family studies are at the stage of development where substantial improvements in the strength of findings are possible by focusing measurement on a single dimension. Both factor analysis and principal component analysis allow us to see if a set of items is measuring a single dimension or multiple dimensions. If researchers do such analysis, it is rarely reported in journal articles. When there are two or more dimensions, the best measurement may be to have separate scales, each representing a different dimension.

Rasch Modeling and Item Response Theory

Rasch modeling and IRT offer related, but competing, alternatives to scale development. Items are picked all along the continuum and thus allow us to differentiate people at both ends of the distribution as well as those in the middle of the distribution. This gives us more variance in the concept and potentially more explanatory power. Many popular scales used in family studies only differentiate within 1 SD of the mean where sales developed using Rasch modeling or IRT seek to differentiate people from 3 SD below the mean to 3 SD above the mean. These two methods assume a latent construct is being measured and estimate the item difficulties and participants score. For an accessible treatment and introduction to Rasch modeling and IRT see Bond and Fox (2007).

Validity

There are several types of validity that could be discussed in publications, but that rarely are. These include face validity, content validity, criterion validity (concurrent and predictive), and construct validity. Space does not permit giving them adequate coverage. Fortunately, they are discussed in many texts and there are two exceptionally useful treatments of them (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Shadish et al., 2002). The point we want to stress is that these are rarely mentioned when a new measure is utilized in an article. Validity goes hand in hand with reliability, and a reliable yet invalid measure is just as useless as an unreliable yet valid measure. As Family Science reaches maturity in the fields of behavioral and social sciences it is important that we present clear measures of the constructs we are studying.

Statistical Procedures

Levels of Measurement

Historically, a great deal has been written about levels of measurement since the seminal work of Stevens (1946). Nominal level refers to classification where no ordering of the classes is merited. For example, gender is a dichotomous nominal variable and marital status (married, divorced, cohabiting, widow) is an example of a polychotomous nominal variable. Where an outcome is dichotomous the most common model uses logistic regression. Where it is polychotomous the appropriate analysis is multinomial logistic regression.

Many variables we study are somewhere between being ordinal and being interval. An ordinal variable is a series of categories that are ordered such as strongly agree, agree, disagree, and strongly disagree. An interval measure is a numeric variable such as temperature Fahrenheit where the differences between values are equal, e.g., 50° is 10 more than 40° and 20° is 10 more than 10°. We can, of course, assign numerical values to an ordinal scale, e.g., strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1. However, the difference between 3 and 2, agree vs. disagree, may be much greater than the difference between 4 and 3, strongly agree vs. agree. The difference between 4 and 3 or 2 and 1 may depend partially on personality rather than a real difference of opinion. That is, some people are yea-sayer or naysayers and pick the most extreme option to any question and other people are temperamentally reluctant to pick an extreme option on any question.

There are procedures for analyzing ordinal data such as ordinal logistic regression, but these are not widely used by family researchers. More often, family researchers treat ordinal variables as if they were interval. This has been shown to produce consistent findings and regression results for interval level variables have simpler interpretations than those for ordinal regression.

The area where the most grievous problem occurs is when the outcome being predicted is a count of rare event. How often did you strike your spouse? How often did you drink more than 5 beers in the last month? Ordinary least squares and related procedures that assume normality are inappropriate. These variables often conform to a Poisson distribution where the mean and variance are equal or negative binomial distribution where the variance is greater than the mean. The appropriate procedures are Poisson regression or Negative Binomial Regression. These options are available in many statistical packages including SAS, Stata, and Mplus and can be utilized for many types of analysis including growth curves (Long & Freese, 2006).

A second option with count variables is using a POMP (Percent Of Maximum Possible) score. A POMP score is simply the count for each participant divided by the total count that is possible, or perhaps the highest count recorded. This might have the effect of creating a more normal distribution, but that is not guaranteed. Once you make the transformation, the interpretation of parameters is easier. See Cohen, Cohen, Aiken, and West (1999) for more information.

Another type of measure occurs when we have count variables that have an "excess" of zeros. Consider the question above about how often you struck your spouse in the last month. Most participants will say zero times. This would be a count of a rare event, but it is also a zero inflated count. There are two-part regressions where a logistic regression is done to predict who has done the behavior at all and a Poisson or negative binomial regression is done simultaneously for the subsample that have done the behavior at least once.

Programs such as Stata make it very easy to estimate these models for zero inflated Poisson regression and zero inflated negative binomial regression. The procedures estimate who is always zero where the behavior is not part of their repertoire and how often people do the behavior. Factors that predict the onset of marital violence, for example, might be separate from factors that predict the frequency of such violence.

Many of our measurements are censored. The distributions will be skewed with a large clump of participants at the top or bottom. Garrison Keillor in the Prairie Home Companion radio show tells us that in the mythical town of Lake Wobegon all the children are above average. When we ask parents to report on the well-being of their children we often find a distribution like the children in Lake Wobegon. If we have a 5-point scale, we will have a cluster of high scores with very few low scores. If we think the true distribution is not so skewed we could utilize censored regression to estimate the true relationships. On marital satisfaction we have many people giving the most positive response—but some of these people are likely much more satisfied than others—we just don't ask sufficiently challenging questions to make these distinctions.

A final measurement issue is how researchers sometimes collapse their data. They may have a scale that ranges from 1 to 10. They collapse the data so 1–5 is a 0 and 6–10 is a 1. They may dichotomize at some other value such as the median or the mean. When is this appropriate? When is it inappropriate? This is a reasonable thing to do if the researcher has no confidence in the variance. S(he) needs to say that a score of 1 and a score of 5 are the same—no difference since both are assigned the single value of 0. If this assumption is unreasonable, then collapsing the data will destroy meaningful variance.

Another occasion for collapsing data is when there is a checklist of behaviors such as problem behaviors and all of them are problematic. If the researcher is interested in whether there are any problem behaviors rather than the number of problem behaviors then we might collapse the data so a person checking any of the behaviors is recorded as having a problem behavior and only those who check none of the behaviors are recorded as not having a problem behavior. Of course, treating the number of problem behaviors as a Poisson model or a zero inflated Poisson model would offer more information.

Mediation and Moderation

Many questions cannot be answered by a single equation but involve two or more equations. This happens when we have one or more intervening variables that mediate the effect of our predictor on our outcome. For example, divorce has been linked with problem behavior in adolescents. However, the direct effect of divorce may be mediated by other variables including household income, involvement of non-resident parent with the child, and continued parental conflict. After a divorce the household income of the resident parent may be reduced, the non-resident parent may or may not be involved with the child, and conflict between the parents may continue or not.

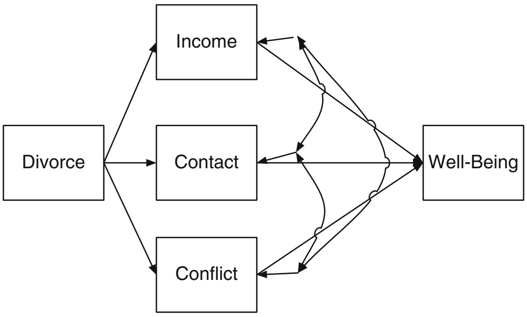


Fig. 4.7 Mediation model

The level of income, non-resident involvement, and parental conflict are more proximate causes of well-being than is the divorce itself. These variables are said to mediate the effect of the divorce. This is illustrated in Fig. 4.7 where divorce has no direct effect on well-being once we control for household income, nonresident contact, and parental conflict.

The curved lines are included because the variance in income, contact, and conflict that is not fully explained by divorce will tend to be correlated. This lack of independence needs to be incorporated in the model. We could estimate this model using seemingly unrelated regressions, but family researchers more commonly use SEM. SEM programs such as Mplus (Muthén & Muthén, 2009), LISREL (Mels, 2009), EQS (Byrne, 2006), and Amos (Arbuckle, 2009; Byrne, 2009) can do this for outcomes that are continuous, categorical, or counts. We evaluate a model such as the one in Fig. 4.7 using a few simple rules.

1. The predictor (divorce) is significantly correlated with the mediators (e.g., income).
2. The predictor (divorce) is significantly correlated with the outcome (child well-being).
3. The mediators are each significantly correlated with the outcome (child well-being).
4. The direct effect of the predictor (divorce) on the outcome (child well-being) controlling for the mediators (e.g., income) is not significant (the effect of divorce is mediated by income, contact, and conflict).
5. Or, the direct effect of the predictor (divorce) on the outcome (well-being) controlling for

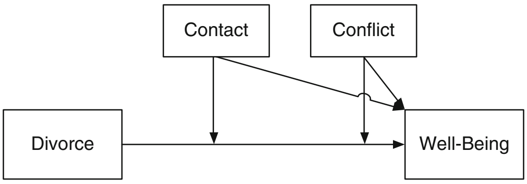


Fig. 4.8 Moderating model

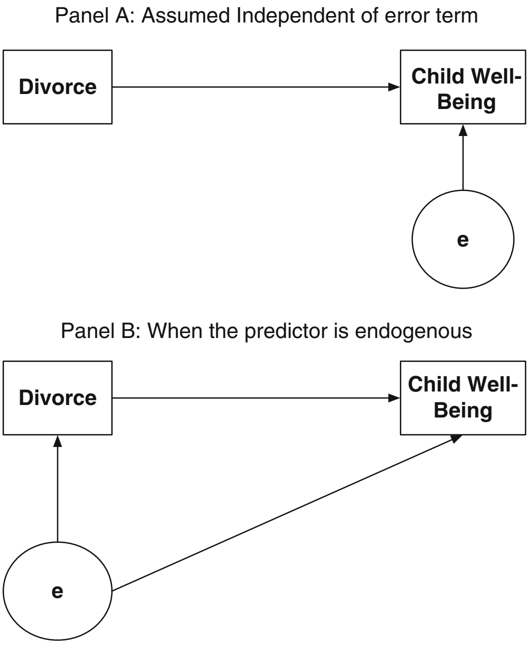
the mediators (e.g., income) is substantially smaller than the direct effect of the predictor when you do not control for the mediators (the effect of divorce is partially mediated by income, contact, and conflict).

The most common result is partial mediation where part of the effect of divorce on well-being is direct (not shown in Fig. 4.7) and part is indirect. In this example there are three indirect effects, the first of which is divorce → income → well-being. When we estimate an indirect effect, it is important to test its significance. Many articles in family journals report that there is an indirect effect without reporting its size or its level of significance (MacKinnon, Fairchild, & Fritz, 2007; Ridenour et al., 2009).

Moderation is a technical term used by many family scholars to describe what statisticians call statistical interaction. A researcher may feel that the effects of divorce are moderated by having nonresident parental contact with the child and, at the same time having minimal continued parental conflict (see Fig. 4.8). That is, when the nonresident parent stays involved with the child and there is little parental conflict, the children will have better post divorce well-being than if the nonresident parent disengages from the child or continues to fight with the resident parent. This is different from mediation in that the moderators change the relationship between divorce and well-being. The relation is stronger or weaker depending on the level of contact and conflict. The moderation effect is estimated by adding interaction terms to the regression equation. In this case we would estimate the equation:

$$\hat{Y} = a + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_1X_2 + B_5X_1X_3$$

where $\hat{Y}$ is the child's well-being, X_1 is the parents divorce status, X_2 is the level of parent child contact, and X_3 is the level of parental conflict. The two product terms represent the interaction effect. Notice that we have a main effect for X_2 and X_3 included whenever they appear in an interaction term. We have not covered centering or other important statistical issues when studies involve moderation; these are developed in Cohen, Cohen, Aiken, and West (2003).


Panel A: Assumed Independent of error term

Panel B: When the predictor is endogenous

Fig. 4.9 Endogenous regressors

A variable may be both a mediator and a moderator. A third model might combine both moderation and mediation by specifying that divorce has direct effects on contact and conflict as well as specifying contact and conflict have a direct effect on well-being, but would also have contact, and conflict moderate the direct effect of divorce on well-being.

Endogenous Regressors

Regression models make the assumption that independent variables are themselves independent of the error term as illustrated in Panel A of Fig. 4.9. In this example, the assumption is saying that everything related to child well-being other than parental divorce is unrelated to parental divorce.

In many family studies' applications this assumption is unreasonable and our regressions yield biased estimates of the effect of the predictor. That is, we do not know whether a significant effect of divorce on child well-being is because of divorce or because of common antecedent causes as shown in Panel B. What would some of these be? Poverty and a spouse that abuses the child come to mind. These variables are associated with both divorce and child well-being. We need to include all such variables in our model or identify an instrumental variable that directly causes divorce but does not directly cause child well-being. We would then estimate the model using an instrumental variable strategy (Cameron & Trivedi, 2009). Family policy research has often failed to consider the endogeneity of predictors, leading to invalid policy recommendations. There has been so much written about the adverse effect of divorce that what we think we know about divorce actually may reflect the adverse effects of other variables that lead to both divorce and the level of child well-being.

Multilevel Models

Over the last decade a new variation of regression has become increasingly important in family research. This is referred to by several names, which contributes to the confusion researchers have about the method. Statisticians usually call these mixed models. Applied researchers often use the term multilevel analysis. A popular statistical package for estimating these models is called Hierarchical Linear Modeling (HLM) and many family scholars use the name of the program as the label. Whether you want to call it mixed modeling, multilevel modeling, or HLM, it is a very important extension of traditional multiple regression.

In traditional regression we assume the observations are independent. If you have a sample of 500 mothers and measure them on ten variables, each of the mothers is independent of the other 499. If, on the other hand, you have a sample of mothers and fathers, your observations are not independent. The problem is evident when you have a variable such as household income that has exactly the same value for the mother and the father. Obviously, the two scores are not independent. Even if we have individual level variables such as age, they may not be independent across observations. If the mother is 20 years old in one family and 50 years old in a second family, we can predict that the father will be older in the second family. The mother's age is probably more similar to her husband's age than to a randomly selected man. A coefficient called the intraclass correlation is used to assess the extent of dependence. Unless this coefficient is zero (some say not statistically significant), we need to do some sort of multilevel analysis. Otherwise, we will have a miss-specified model and incorrect standard errors (Atkins, 2005).

We may have variables at several levels. Suppose we are predicting the likelihood of a 20-year old having completed high school. The first level would be individual characteristics such as parental support, supervision, career goals, grade point average, and gender. The second level could be family characteristics such as family income. A third level might be neighborhood characteristics. If 10 of our 20-year olds lived in the same very poor neighborhood with high crime rates and another ten lived in a prosperous neighborhood with low crime rates, these differences could be important. However, all ten individuals in each of these neighborhoods would have the same score on crime rate and other neighborhood variables.

As a researcher, you might feel that variables at each level are important predictors of the likelihood of graduating from high school. This approach is entirely consistent with the ecological theory behind much family research. You should not use traditional regression methods. Multilevel analysis allows you to examine how variables at each level of analysis influence the likelihood of graduation. HLM is a specialized program that is designed explicitly for doing this analysis. Major statistical packages such as SAS or Stata have comprehensive commands that work under a variety of assumptions. SPSS has a limited capability that works for continuous, interval level outcomes. Multilevel analysis can also be approached from an SEM framework and both Mplus and LISREL are widely used for this purpose. The literature on multilevel modeling is extensive. Raudenbush and Bryk (2002) provide an introduction to HLM, Singer (1998) and Little, Milliken, Stroup, Wolfinger, and Schabenberger (2006) provide useful introductions to SAS' Proc Mixed for multilevel models, and Rabe-Hesketh and Skrondal (2008) do the same for Stata.

Effect Size

Earlier we discussed the over emphasis on statistical significance and an under emphasis on substantive significance in large-scale family research. Increasingly, journal editors are insisting on reporting some measure of the size of the effect along with its statistical significance. Durlak (2009) provides a brief introduction to some of the more common measures of effect size. The U.S. Department of Education has a web resource called What Works Clearinghouse, http://ies.ed.gov/ncee/wwc/, that covers a wide variety of procedures including effect size calculations (U.S. Department of Education & Institute for Education Science, 2008).

There are many statistics called measures of effect size and these may disagree with one another (McGrath & Meyer, 2006). Cohen's *d* or Hedges *g* are often reported in experimental studies. For survey data, standardized betas are often used to measure effect size. This may or may not be appropriate. Whenever variables are measured on a meaningful scale, it is important to interpret the size of the unstandardized *B*'s. These tell us how much the dependent variable changes for a one unit change in the independent variable. A focus on unstandardized coefficients is common in other fields but underutilized by family scholars. Imagine an intervention to encourage people to buy energy efficient light bulbs. If the *B* for the intervention is 3.0 and the dependent variable is the number of energy efficient light bulbs in the household, this has a clear meaning. The intervention increased the number of energy efficient light bulbs by 3.0 per household. Is this a big effect? If you consider that the U.S. has over 100 million households, this intervention would replace 300 million inefficient light bulbs with more efficient ones. This represents an enormous savings in energy. The researcher could compute the kilowatt hours saved, the reduction in releases to the atmosphere, etc. The researcher would have little interest in whether the standardized beta weight was 0.1, 0.3, or 0.5. Standardized coefficients provide little basis for a serious cost-benefit calculation (Duncan & Magnuson, 2007).

When we have a continuous predictor or dependent variable that is measured on an arbitrary scale it is appropriate to use a standardized beta weight. This happens with scales of attitudes and beliefs. One scale of marital satisfaction might range from 1 to 10 and another from 25 to 57. A one unit change is not equivalent and because the scoring systems have arbitrary ranges, interpreting an unstandardized coefficient may be hard. The beta weight tells us how much the dependent variable changes in standard deviation units for a one standard deviation change in the independent variable. We can use the beta weights to compare the importance of several predictors when the measurement scales do not allow more meaningful consideration of the unstandardized coefficients. Family scholars have an over reliance on standardized coefficients because of the clear norms in the field specifying that under 0.2 is weak, 0.3–0.5 is moderate, and anything above 0.7 is strong. We should always recognize that a more careful interpretation of unstandardized coefficients on variables that have accepted scales is better.

There are misuses of beta weights in the family literature. One of the most common is to compare the beta weights for the same predictor in two groups. Suppose we are interested in the relationship between education and income as this differs for single and married women. The unstandardized *B*'s tell us how much income is increased for each additional year of education. If single women have a $B = 2{,}000$ and married women have a $B = 1{,}500$, then single women have a bigger payoff for education. We should not, however, compare beta weights for the same variable across groups. The β for single women might be 0.2 and the β for married women might be 0.3. This would not mean education had a bigger payoff for married women than single

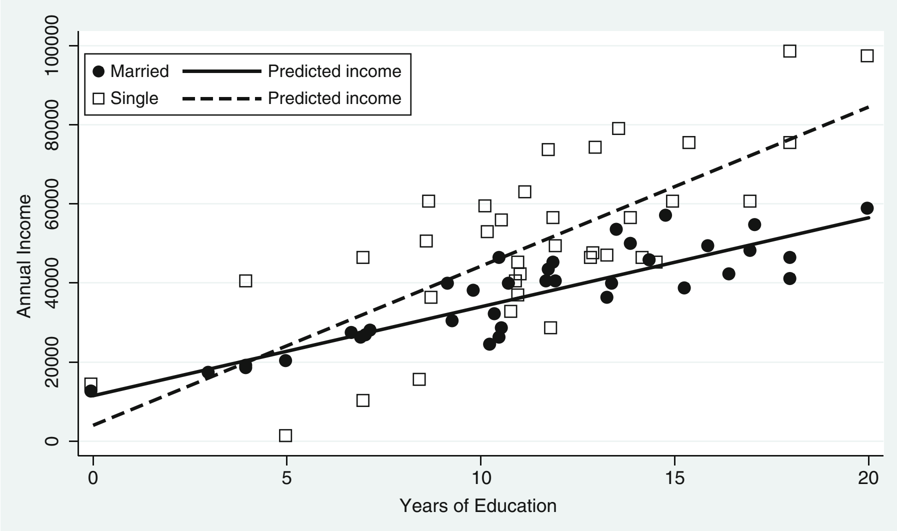


Fig. 4.10 Education and income for married and single women

women—rather we should get that information from the unstandardized *B*'s. The betas mean that if a single woman has one standard deviation more education than another single woman, she will have 0.2 of a standard deviation more income than that single woman. Notice the standard deviation for both education and income refers only to the sample of single women. If married women have a different standard deviation, then a one standard deviation change for a married woman is not comparable to a one standard deviation change for a single woman.

This problem with standardized beta weights applies equally to correlations. Figure 4.10 shows hypothetical data of the relationship between a married woman's education and her income as well as a single woman's education and her income. The $r=0.87$ for the married women and the $r=0.70$ for the single women. Relying on the correlations, a researcher would mistakenly say that education has more effect on income for married women than it does for single women. As is clear in Fig. 4.10, the black circles for married women are closer to the solid regression line for married women, hence the correlation is high. The hollow squares for single women are more spread around; hence the correlation is lower for single women. What about the payoff of education for income? The regression line for married women is \$11,479 + \$2,251 (Education) and for single women it is \$3,985 + \$4,027 (Education). The $B=\$4{,}027$ for single women is nearly twice as high as the $B=\$2{,}251$ for married women. This means that for each additional year of education a single woman increases her expected income by \$4,027 and each additional year for a married woman increases her expected income by just \$2,251 dollars. The point is that you should not compare standardized coefficients such as *r*'s or *β*'s to compare the form of the relationship between variables in different groups. Instead, you need to compare the unstandardized coefficients such as *B*'s, means, and standard deviations.

Another problem with the reliance on standardized coefficients applies when using a binary predictor. Suppose we are estimating the effect of marriage vs. cohabiting on some outcome variable with marriage coded 1 and cohabiting coded 0. A beta weight tells us that as you go up one standard deviation on an independent variable you go up beta standard deviations on the dependent variable. You can be a 0 if you are cohabiting or a 1 if you are married. These are the only meaningful values. Some variables have an underlying continuum that has been collapsed as a 0, 1 dummy variable. It is possible to think of a latent variable for marital status that represents your propensity to be married rather than cohabiting. Often such interpretations are awkward. A far simpler solution is a semi-standardized beta weight that is standardized only on the dependent variable. A semi-standardized beta of 0.4 would mean that if you were a 1 on marital status (married) you would be 0.4 standard deviations more satisfied in your marriage than if you were a 0 on marital status. For a dummy variable a one unit change, from zero to one, usually makes more sense than a one standard deviation change. This semi-standardized beta weight was proposed by Stavig and Acock (1981) and has been implemented in Mplus (Muthén & Muthén, 2009) and Stata using commands written by Long and Freese (2006).

Growth Modeling

In the late 1990s and early 2000s there was an explosion of techniques for handling longitudinal data (Bollen & Curran, 2006; Duncan, Duncan, Stycker, Fuzhong, & Alpert, 1999; Muthén, 1991; Preacher, Wichman, MacCallum, & Briggs, 2008; Walker, Acock, Bowman, & Li, 1996). These greatly extended what could be done within a SEM framework. One of the most promising of these methods involves estimating growth curves or growth trajectories. Applied to family scholarship there are endless possible applications. Most family scholars are interested in change and what accounts for change. Growth curves provide greatly enhanced isomorphism between our dynamic theories and our quantitative methods. Growth curve analysis is designed to answer such questions.

Suppose you are interested in what happens to a wife's level of stress in the first 4 years after her husband has been diagnosed with cancer. You predict that her stress increases linearly from year to year. More complicated models are possible. Figure 4.11 shows that you have measured her stress each of the 4 years, *W*1, *W*2, *W*3, and *W*4. Because your measure of stress is imperfect, you assume there is some measurement error at each wave, e_1, e_2, e_3, and e_4.

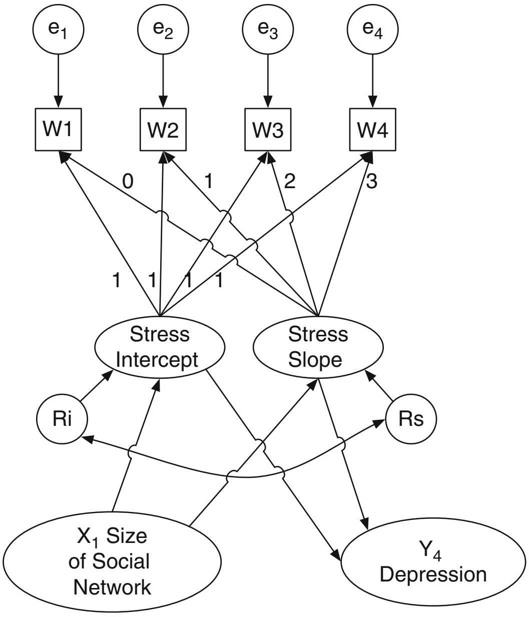


Fig. 4.11 Wife's stress level

First, you want to identify the growth trajectory for her stress. A linear growth trajectory is defined by two parameters. The intercept reflects the initial level of stress that she had at the start. The slope reflects how much her stress goes up (or down) each year. The intercept and slope represents a fixed effect that applies to everybody.

However, some mothers may start out at a high level of stress and some may start out at a low level. Similarly, some mothers may have an increase in stress (positive slope) and some mothers may have a decrease in their level of stress (negative slope). Thus, both the intercept and the slope can vary across the mothers. We call this the random effect, and R_i represents the random effect (residual variance) for the intercept and R_s represents the random effect (residual variance) for the slope. Once we identify significant random effects, the next step is to explain them using time invariant or time varying covariates. For example, the size of the woman's social network at wave 1 might explain both differences in the intercept and the slope. Because this is measured at just one time, we treat it as time invariant. It appears as X_1 Size of Social Network.

At the right side of the figure is a measure of maternal depression taken at wave 4. The figure shows a direct effect from the intercept (initial stress) to depression. Wives who have higher stress initially are more likely to have a subsequent high level of depression at wave 4. Depression appears as Y_4 Depression in the figure. The second direct effect is from the stress slope to depression. Wives who have a greater increase in stress will become more depressed than wives who are able to manage their stress more effectively. Notice, that our independent variables are the intercept and slope and not a simple score on a variable. This model would allow us to test how much change in stress influences depression.

This model can be extended in many ways. An extension would involve having a parallel growth curve. We might think of depression as a parallel growth curve rather than an outcome variable. In that case, we would have an intercept and slope for depression. We could test if the intercept on stress influenced the slope on depression and whether the intercept on depression influenced the slope on stress. Sometimes we have a growth curve for variables that are binary or counts. For example, we might do a growth curve where the intercept and slope represent how often the wife has a depressive symptom. All SEM packages can do some things with growth curves, Mplus is especially handy at doing these for categorical and count variables including extensions to zero inflated models.

Growth Mixture Models

Growth mixture models inductively identify subgroups of people who have different growth trajectories. Mixture models use quantitative methods in an inductive process. This approach grows out of latent class and latent profile analysis where subgroups are identified empirically as being homogeneous within each group and heterogeneous between. Using this inductive approach, typologies emerge from the data rather than being imposed on the data. This can be described as a person centered approach rather than a variable centered approach because it locates groups of people rather than groups of variables as is done with factor analysis (Muthén & Muthén, 2000).

Applied to growth modeling, what emerges are different groups of people who have distinct trajectories. In what has become a modern classic, Muthén and Muthén (2000) did a growth mixture model of excessive drinking for a panel of people from their late teens to when they were in their 30s. They found a normative group that peaked at about age 22 and then dropped off and a deviant group that never engaged in much excessive drinking. The third group mirrored the first group up to age 22, but then they continued at this level. Identifying groups with different directories has compelling policy applications. The appropriate intervention and its consequences vary for each of these groups. The ability to predict who will be in each group can result in much more effective interventions.

Growth mixture modeling has many potential applications to family scholarship. For example, a growth mixture model of the level of interpersonal violence in married couples might reveal distinct groupings of couples with sharply different trajectories. The trajectory of many family processes may fall into distinguishable groups. Whatever the trajectory being studied, the follow-up task is identifying time invariant and time varying covariates that predict class membership. Equally important is finding distal outcomes that vary by class membership. Perhaps four classes might emerge for parental conflict: consistently low, consistently high, decreasing, and increasing. What are the adolescent outcomes when the parents are in the consistently low or high trajectory class? Does it matter whether the parents are in the decreasing or increasing trajectory class? A tutorial on growth modeling and growth mixture modeling is available at oregonstate.edu/~acock/growth.

Survival Analysis

Survival analysis deals with whether an event happened as well as when an event happened. It is specifically designed for censored data, when only a portion of the sample experiences an event in the course of the study. It does require a time variable, so longitudinal data is needed. This use of time allows us to answer when the event is most likely to happen. Specifically, we can use a hazard function, which maps the probabilities of the event happening at a given time point. Using a logistic model, this map can then be predicated using covariates and show how it varies dependent on the covariates. The ability to predict both whether an event happened and when it happened has great significance in family studies. Whether and when a marriage ends in divorce, when a child becomes sexually active (Singer & Willett, 2003), how abuse affects later deviant behavior (Keiley & Martin, 2005) are all questions that are best suited for survival analysis. A logistic regression might give a partial answer to each of these, but would not be able to answer the when questions. For a brief introduction of survival analysis see Keiley and Martin (2005) or for a detailed treatment see Singer and Willett (2003).

A View of Where We Are Going and What We Need to Get There

We have reviewed a wide variety of quantitative procedures with an emphasis on the strategies and procedures that are becoming more prevalent. Quantitative research is moving toward longitudinal research that requires knowledge of SEM, growth curves, mixture models, and survival analysis. Recognition that many of the topics we study must be approached from a multi-level perspective is now widespread and the inappropriate application of traditional regression is becoming unacceptable. The best journals that deal with family research are demanding ever-higher standards for the quantitative analyses they publish.

We have also moved toward expecting a serious consideration of the assumptions of a procedure and the expectation that the most appropriate estimators be used. The choice of estimator now requires consideration of sampling characteristics. Is it a cluster sample? Are some groups underrepresented requiring weighting? Is the dependent variable a binary, categorical, or count outcome requiring special estimation techniques? Is it reasonable to assume that the independent variable is uncorrelated with the error term?

Attrition needs to be explained and appropriate adjustments incorporated into our models. The *default* use of listwise deletion is being discredited and sophisticated methods of multiple imputation and full information maximum likelihood estimation are becoming expected. There are also increasing expectations to substantively evaluate our results much more carefully than simply reporting what predictors are statistically significant. We now have an arsenal of measures of effect size and a need to provide interpretation of the strength of results. Not the least of the expectations for the substantive significance is the consideration of the duration of effects. This chapter has only introduced the topics that are important for family research. A book length discussion of some of these topics is available in Bakeman et al.'s (2006) treatment of best practices in quantitative methods.

The reliance on a single software program for quantitative analysis has become problematic. It is a mistake to say one software package is the best, because all of them are constantly expanding their capabilities. The package that is best today may be superseded by another package next year. In selecting a general software package, we have several needs. We need to have flexibility to manage large, longitudinal datasets, and manipulate the data. Programs such as SAS, Stata, and SPSS all have at least adequate capabilities. We need packages that work well with complex sample designs and both Stata and SAS are quite good at this. SPSS is far behind them at the time this is written. We need programs that offer effective ways of doing multiple imputation or full information maximum likelihood estimation and that are able to work with different types of outcome variables including binary, categorical, ordinal, and count variables. Both SAS and Stata are quite strong in these abilities and SPSS is likely to improve its own capabilities.

We probably need to have skill with specialized software such as SEM programs. Here LISREL, Mplus, and EQS have comprehensive capabilities and AMOS has some capabilities. Many family scholars who work with multilevel analysis rely on HLM, which was written explicitly for this type of analysis. Other packages have varying degrees of capability to duplicate or even extend what HLM does including MLwin, SAS, Stata, and Mplus. As with other applications, SPSS has more limited capabilities in this area.

One thing to remember about software is that anything written here applies only to these programs at the time this chapter is being prepared. Other programs such as *R*, which many family researchers find hard to use today, are developing and may soon become easier to use. *R* is unlimited in its expandability as new techniques develop. It is also advisable to have software that is dedicated to measurement including Rasch modeling and IRT. Mplus and Stata have some capabilities, but specialized programs such as Winsteps, Facets, ConQuest, RASCAL, BILOG, and MULTILOG provide far more detailed information for developing a scale.

Unfortunately, many family researchers stop enhancing their quantitative skills once they complete their graduate education. They rely on whatever was taught while in graduate school. The rapid growth in computer speed has gone hand in hand with the rapid growth in complex quantitative techniques. A complex SEM program that ran in half a day just 10 years ago will now run in seconds because of the joint improvement in computer performance and software efficiency.

Graduate programs training the next generation of family scholars vary enormously in the scope of quantitative training they require. Methodological advances are typically published by statisticians and are written for other statisticians. This results in a level of technical difficulty that is extremely challenging for applied researchers. Universities provide ever-increasing pressure to publish but fewer opportunities for professional development. How does a researcher keep up with all technical developments in quantitative methodology? There is positive news. In addition to specialized monographs, there has been a rapid increase in completely free online tutorials. It is impossible to list all of these, but one of the best web resources is the statistical portal at UCLA (statcomp.ats.ucla.edu/). This site provides researchers with videos, tutorials, and links to a wide variety of Internet resources. They have special topics for most major software packages and even provide the programming code for examples in a number of advanced methods textbooks and monographs.

It is easy to be overwhelmed by the rapid development and we need to constantly remember that the best procedure is the simplest one that is appropriate. The advantage of some of the complex methods we have summarized here is that they are appropriate because they achieve greater isomorphism between your research question, theory, and methods. For example, it is easier to assume there is no measurement error in predictors, but it is possible to estimate the measurement error and incorporate these estimates in your model. It is easier to assume a simple random sample, but it is possible to choose options that give you unbiased results with a complex sample. It is simple to use listwise deletion, but you can get less biased estimates doing multiple imputation. As a researcher you have to choose what you will do as you balance the simplicity of a method with how appropriately it meets your statistical and theoretical assumptions.

Acknowledgment We would like to thank Molly Trauten for her considerable contribution to the clarity of this chapter.

References

Acock, A. C. (1989). Measurement problems and other problems in using large data sets for secondary analysis. In K. Namboodiri & R. Corwin (Eds.), *Research in the sociology of education and socialization* (Vol. 8, pp. 201–231). Greenwich, CT: JAI Press.

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*, 1012–1028.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing*. Washington, DC: AERA.

Arbuckle, J. L. (2009). *Amos 17.0 user's guide*. Chicago: SPSS.

Atkins, D. C. (2005). Using multilevel models to analyze couple and family treatment data. *Journal of Family Psychology, 19*, 98–110.

Bakeman, R., Gottman, J. M., Brewer, D., Bub, K., Burchinal, M., Graham, F., et al. (Eds.). (2006). *Best practices in quantitative methods for developmentalists*. New York: Wiley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Byrne, B. (2006). *Structural equation modeling with EQS*. Thousand Oaks, CA: Sage.

Byrne, B. (2009). *Structural equation modeling with AMOS* (2nd ed.). New York: Routledge.

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata*. College Station, TX: Stata Press.

Cheng, Y. Y., Wang, W. C., & Ho, Y. H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth fidelity dilemma. *Educational and Psychological Measurement, 69*, 369–388.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*, 315–346.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. (2003). *Applied multiple regression/correlation analysis for behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Davey, A., & Savla, J. (2009). *Statistical power analysis with missing data: A structural equation modeling approach*. New York: Routledge.

Duncan, G. J., & Magnuson, K. (2007). Penny wise and effect size foolish. *Child Development Perspectives, 1*, 46–51.

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

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*. Advance Access published February 16. doi:10.1093/jpepsy/jsp004.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Furlow, C. F., Ross, T. R., & Gangé, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement, 33*, 441–464.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns. *Multivariate Behavioral Research, 31*, 197–218.

Keiley, M. K., & Martin, N. C. (2005). Survival analysis in family research. *Journal of Family Psychology, 19*(1), 142–156.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146–162.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Trials, 21*, 167–189.

Liboff, R. (2002). *Introductory quantum mechanics* (4th ed.). New York: Addison Wesley.

Little, R., Milliken, G., Stroup, W., Wolfinger, R., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Carry, NC: SAS.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station, TX: Stata Press.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614.

Marion, S., Forgatch, M. S., Patterson, G. R., Degarmo, D. S., & Beldavs, Z. D. (2009). Testing the Oregon delinquency model with 9-year follow-up of the Oregon Divorce Study. *Development and Psychopathology, 21*, 637–660.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11*, 386–401.

Mels, G. (2009). *LISREL for windows: Getting started guide*. Lincolnwood, IL: Scientific Software.

Molenberghs, G., Beuchkens, C., Sotto, C., & Kenward, M. G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, 70B*, 371–388.

Muthén, B. (1991). Analysis of longitudinal data using latent variable models with varying parameters. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change* (pp. 1–17). Washington, DC: American Psychological Association.

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism, Clinical and Experimental Research, 24*, 882–891.

Muthén, L., & Muthén, B. (2009). *Mplus user's guide*. Los Angeles: Stat Model.

Patterson, G. R., Chamberlain, P., & Reid, J. B. (1982). A comparative evaluation of a parent-training program. *Behavior Therapy, 13*, 638–650.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata* (2nd ed.). College Station, TX: Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Ridenour, T. A., Tarter, R. E., Reynolds, M., Mezzich, A., Kirisci, L., & Vanyukov, M. (2009). Neurobehavior disinhibition, parental substance use disorder, neighborhood quality and development of cannabis use disorder in boys. *Drug and Alcohol Dependence, 102*, 71–77.

Rubin, D. B. (1987). *Multiple imputation in nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs to generalized causal inference*. Boston: Houghton Mifflin.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323–355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Stata. (2009). *Stata reference manual*. College Station, TX: Stata Press.

Stavig, G. R., & Acock, A. C. (1981). Applying the semistandardized regression coefficient to factor, canonical, and path analysis. *Multivariate Behavioral Research, 2*, 255–258.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

U.S. Department of Education & Institute for Education Science. (2008). *WWC procedures and standards handbook,* version 2. Retrieved from http://ies.ed.gov/ncee/wwc/references/

Walker, A. J., Acock, A. C., Bowman, S., & Li, F. (1996). Amount of care given and caregiving satisfaction: A latent growth curve analysis. *Journal of Gerontology: Psychological Sciences, 51B*(3), 130–142.

West, B. T., Berglund, P., & Heeringa, S. G. (2008). A closer examination of subpopulation analysis of complex sample survey data. *The Stata Journal, 8*(3), 1–12.

Williams, G. (1994). Effects of interviewer status touch, and gender on cardiovascular reactivity. *The Journal of Social Psychology, 134*, 247–249.