Prologue

What's New in the Second Edition?

Wow, it's hard to believe that this second edition is upon us! I'm excited to present it to you. I have added a number of new chapters and have expanded on quite a few topics that I had covered in the first edition. In order to make sure that a couple of the new chapters are as accurate as possible, I enlisted the help of a couple of my friends and colleagues. I added a new chapter on Bayesian longitudinal modeling in collaboration with Mauricio Garnier. Mauricio is a Bayesian Buff and teaches the Bayesian SEM course in the annual Stats Camp in June (see *www.statscamp.org*). I also added a chapter on mixture modeling wherein I enlisted the help of Whitney Moore, who is a mixture Maven. Third, I added a chapter on within-person change models, particularly the recently introduced random intercept cross-lagged panel model (RI-CLPM). Here, I enlisted the assistance of Danny Osborne, who is a within-person change Wizard. With Buffs, Wizards, and Mavens like them, the new chapters add considerably to the depth and breadth of what's covered.

Anyone who has seen me give a talk on missing data knows just how much I love missing data. I decided to add missing data as a focus area of scholarship the moment John Graham blew my mind with a talk on modern treatments for addressing the ubiquitous nature of unplanned missing data. And there have been some significant advances in the missing data literature. So much so, that I have inserted a new Chapter 3 dedicated to the topic of missing data, particularly as it applies to longitudinal data.

One of the new sections included that I'm very excited about is the discussion of how to test models of experimental effects. This conversation riffs off the recent criticisms of the null hypothesis testing traditions that still permeate our psyche and our literatures.

A PERSONAL INTRODUCTION AND WHAT TO EXPECT

How Statistics Came into My Life

For many years, folks encouraged me to write a book on structural equation modeling (SEM). I'd reply that there are lots of books already out there, especially when it comes to the basics of SEM. Ah, they'd answer back, but there aren't any that cover it quite the way you do, especially in the context of longitudinal SEM. Mind you, "covering it quite the way I do" does not mean my way is more erudite than those of others, but it is unique and, I hope, somewhat informative and entertaining. I was an English literature major as an undergraduate at the University of California, Riverside. I came to the challenge of learning statistics with trepidation. When I realized that statistics is where logic and common sense intersect, I learned what Bill Bukowski later described as the point at which the poetry of mathematics becomes the elegant prose of statistical reasoning. I discovered that statistics isn't math but a system of principles to guide my research.

Although I was an English literature major, I was also interested in psychology, and in my senior year I thought I would give the intro to statistics course a try. Larry Herringer, the teaching assistant for the course, was patient and worked very hard to explain the concepts in basic terms. He spoke to me. I learned the material well enough to be invited to help Larry with his dissertation research. A few months later, Larry introduced me to a young assistant professor who was interested in recruiting a graduate student to work with him. I spent a few hours talking with Keith Widaman and Larry about what a PhD program in developmental psychology under the mentorship of Keith would be like. I know that I didn't think it all through because I was enchanted. What was it? Serendipity or a fool's errand? The deadline to apply for the PhD training program was the next day. I applied.

After a few weeks I heard from the graduate admission committee that I could not be accepted to the graduate training program because I had not taken the Graduate Record Examinations (GREs), which were not necessary for the master's program in English to which I had been accepted at a nearby state college. The admission committee at UC Riverside gave me a conditional acceptance if I would take the GREs and post a combined score that was above their minimum threshold. A session of the exam was scheduled for 3 weeks after I heard this news. I thought OK, what the heck. I felt comfortable with the verbal portion of the GREs. I was petrified of the quantitative portion. I had avoided math courses through my entire undergraduate training (in fact, because of various transfers I made, I somehow got more credit than I should have for the one intro to algebra course that I did take). My roommate at the time (now an accomplished neurologist, Bret D. Lindsay, MD) volunteered considerable amounts of his time to tutor me in math in order to somehow help me make up for a lifetime of math avoidance.

I took the GREs and waited for the snail mail report on whether I had achieved the requisite combined score and would be in the PhD program at UC Riverside or

would attend nearby California State University to get a master's degree in English. On May 31st I had to notify Cal State of my decision. I had not received my GRE scores. Again, I acted irrationally. I declined the Cal State offer and waited further. On June 1st the mail arrived with official word: I was above threshold and, much to my surprise, my quantitative score was actually better than my verbal score. I immediately went to Larry's office. We celebrated by splitting a bottle of beer that Larry had stowed away for a special event. I then informed Keith.

Keith was patient with me for the duration of my graduate training. As I was Keith's graduate student, many of the other students presumed that I knew what I was doing, and I became an unofficial stats consultant. Putting myself in such a position forced me to understand statistics more deeply and to find ways to communicate it better. In 1991, when I was packing my things to take my first academic position as a research scientist at the Max Planck Institute in Berlin, Germany, I found the empty bottle of beer that Larry and I had shared. On it are Larry's initials and the date of June 1st. The beer was a bottle of Berliner Pils—serendipity again? My 7 years at the Max Planck Institute were incomparable learning years (*Herzlichen Dank an meine Kolleginnen und Kollegen*). The empty bottle sits on a shelf in my office as a reminder of the journey.

My Approach to the Book

If you don't like math or were never much for equations, don't panic (and don't forget your towel). Throughout this book I do cover the math, but I don't rely on it. Plus, I make sure to walk through the equations so you can get a clear sense of what the elements of the equations mean (hopefully you'll soon get comfortable with them along the way). I have never enjoyed math for math's sake. As mentioned, things changed when I took my first statistics course and realized that the numbers and the equations meant something important to me. I presume you are reading this book because you have an inherent desire to learn about how to do longitudinal modeling to answer some interesting questions. The numbers that are presented will mean something to you. If you are still wary of the numbers, you'll find plenty of other ways to understand the concepts. For example, I use a lot of analogies, metaphors, and similes—I'm a veritable metaphor machine. You'll hear about everything from clouds to plants, toothpaste to irrigation systems, and pressure cookers to flocks of birds. Each of these ideas is used to help clarify a core SEM concept. (See my essay on metaphor in science referenced in the Recommended Readings at the end of this prologue.)

I try to channel my lectures in this book (and I advise the authors who contribute volumes to the Methodology in the Social Sciences series that I edit for The Guilford Press to take a similar approach). You should hear my voice as if you were sitting in my class, attending one of my lectures, or experiencing one of my Stats Camp seminars. The tone is light and informal—I am talking to you. The material, however,

is not without substance. I have learned over the years that I can reach all levels of audience if I make sure that I start each topic or main theme at the very beginning. As I work up through the material, I focus very carefully on linking each rung of the ladder as I ascend to the top of where the field has arisen. I cycle through each topic from its foundations to its peak. This way, there's a comfort zone for understanding regardless of your prior experience or background.

I offer my preferences, recommendations, and opinions based on over 30 years of applied experience. My goal is to show you *how* you can think about statistics in general and longitudinal SEM in particular, not *what* to think. I try to avoid presenting things as proscriptive and prescriptive rules but instead emphasize principles, choices, and justifications. My limited warranty is that anything I espouse may not hold up in the future. At least at the time of this writing, smart and knowledgeable scholars have reviewed the material I present. These smart and knowledgeable reviewers have also voiced opinions on some topics that are different from mine. I have tried to indicate when others may have a different view.

Key Features of the Book

One convention that I employ is an equation box. Each equation is set apart from the text, and each element of the equation is defined in the note to the equation. This convention allows me to focus the text on telling you the meaning of things without bogging down the flow with having the equation in the text and using that distracting "where" clause. I also include a glossary of key terms and ideas introduced in each chapter. The glossary definitions are meant as a refresher of the ideas for those who are already familiar with much of this material, as well as reinforcement for those who are just learning this material for the first time. I recommend that you scan through the glossary terms and definitions before you read the content of each chapter. At the end of each chapter, I also highlight a few key readings with a short annotation. There is a separate list at the end of the book for all the references that are cited, as well as key references that may not have been cited but are important works that have either shaped my thinking or have had an impact on practices in the field.

I maintain a set of support pages including online resources for this book and scripts for every example that I present. (Go to *www.guilford.com/little-materials* for up-to-date directions on where to find these resources.) These scripts are written in various software packages, including LISREL, Mplus, and R (lavaan package). Other scripts, examples, and resources are also available at *www.guilford.com/little-materials*. On these pages I also try to post other relevant materials. You'll find guides to LISREL, Mplus, and R (lavaan), what to report, how to read the output, and so on. If you find that a guide is not clear, contact the authors of the guide and request clarification—an update with clarification included will then be posted. I will also be adding new material to augment the existing material presented in this volume.

In addition to my role as Professor at Texas Tech University, I founded and teach regularly in my "Stats Camps." I also created a nonprofit Stats Camp Foundation to operate and maintain Stats Camp events for many years to come. In the Stats Camp events, we offer a broad selection of seminars related to advanced quantitative methods. Traditionally, I teach an SEM foundations course during the first week of Stats Camp, and then I teach longitudinal SEM during the second week. These summer stats camps are 5-day courses that have lots of time allocated for personal consultation and time to work on your own data. We have had countless campers thank us for helping with completing a dissertation or a complex model for publication. We also offer a new Analysis Retreat model as part of the Stats Camp family. These retreats are meant to allow you to unplug, come to a cool location, and work with our team of experts to ensure your modeling is done with precision and best practice in mind. We help guide your decision making as you progress to a final product (publication, dissertation) and help you troubleshoot any issues you might encounter along the way.

Each of the figures that I created for this book is also available in its original form on the support pages. I used Adobe Illustrator to create most of the figures; however, some of the new figures are products of other software. Note that I hold the copyright on the figures (mainly so that I can easily use them in other material). I don't mind if you use or modify them, but please acknowledge the original source.

Overview of the Book

When I teach SEM, I use the metaphor of a knowledge tree to introduce how I organize and present the material. The tree's trunk contains all the parts that are essential to build an effective SEM model. After the trunk is well established, it is relatively easy for me to present different techniques as branches from the core trunk. Therefore, and as you can see from the Contents, I spend a considerable amount of time focusing on many of the foundational issues related to longitudinal SEM. The design and measurement issues that I describe in Chapters 1 and 2, for example, provide some insights that I have gained as a developmental researcher—the proverbial "I wish I had known then what I know now" kinds of insights. These chapters also cover essential preparatory steps that must be done well in order to fit a good SEM model. Chapter 3 is the new chapter on missing data in the context of longitudinal research. Properly treating the inevitable missing data that occur is also an essential preparatory step!

In Chapters 4 and 5, I detail the foundational material associated with SEM in general. As I mention in those chapters, I present that material because I think it is very important that persons reading this book are on the same page when it comes to how I describe the foundational elements of SEM. Plus, I can emphasize brand new developments in many of these foundational elements (e.g., effects-coded method of

identification), as well as my intuitions on some topics where definitive guidance is still lacking.

Chapter 6 brings the foundations material to its full "tree" height. The longitudinal CFA model that I describe in this chapter is the measurement model for any quality longitudinal SEM model. I then turn to the basics of a longitudinal panel model in Chapter 7 and extend the panel model to the multiple-group case in Chapter 8. I also introduce dynamic P-technique data and how to fit a multiple-group panel model to such data.

Chapter 9 is the new contribution with Danny Osborne where we explore a couple of hybrid models for modeling within-person change, with a particular focus on the random intercept cross-lagged panel model. Mediation and moderation in panel models is covered in Chapter 10. Then, I bring in multilevel models, including growth curves in Chapter 11. Chapter 12 is also a new contribution with Whitney Moore where we discuss various finite mixture models (e.g., latent profile/class models) to look for unknown heterogeneity (i.e., unique subgroups) in the sample. Chapter 13 presents the foundations of the Bayesian SEM approach, which is the third new contribution, with Mauricio Garnier. I round out the book with Chapter 14, which details a jambalaya of models that can be fit to longitudinal data.

DATASETS AND MEASURES USED

The primary examples that I use in the various chapters are summarized here (a few others are introduced in the respective chapters, particularly the new chapters). I'm extremely grateful to my colleagues and friends for providing me access to their data and allowing me to put this material on the support pages for this book at *www.guilford.com/little-materials*. For all the datasets, missing data were handled using some form of imputation. The internal consistencies of all measures were around .80, so I don't report their specific values here. A few examples come from published papers and are cited when I present them, and so I don't detail them here.

My Dataset with the Inventory of Felt Emotion and Energy in Life (I FEEL) Measure

Participants were 1,146 sixth- through ninth-grade students (50% boys, 50% girls) from an urban school district in the northeastern United States. The sixth-grade students attended nine different elementary schools, seventh- and eighth-grade students attended a single middle school, and ninth-grade students were enrolled in high school. At the first two measurements (fall 1999 and spring 2000), all students were enrolled in the sixth (n = 382), seventh (n = 378), or eighth grade (n = 386). The third measurement occurred in fall 2000. Approximately 70% of the sample was European American, 15% African American, 6% Hispanic American, and 9% from

another ethnic background. Socioeconomic status (SES) ranged from lower to upper middle class.

Students who had written parental consent to participate and who provided their own assent (overall around 80% of eligible students) were administered a series of questionnaires over several sessions within their classrooms by trained research assistants. Teachers remained in the classrooms but worked at their desks. Students were assured that school staff would not see individual students' responses to the questionnaires. The broader data collection effort included self-report and peer-report measures of aggression, victimization, and aspects of self-regulation; the survey order was counterbalanced.

The I FEEL

The I FEEL (Little, Wanner, & Ryan, 1997) measures 14 dimensions of internalizing symptoms (i.e., positive emotion, negative emotion, positive energy, negative energy, connectedness, loneliness, positive self-evaluation, negative self-evaluation, calmness, anxious arousal, fearful arousal, hostile arousal, somatic symptoms, physiological symptoms) and asks respondents to report about their own experiences during the prior 2 weeks. The negatively valenced subscales (e.g., negative mood, loneliness) were based on existing depression, loneliness, and anxiety instruments after categorizing the existing scales (e.g., Children's Depression Inventory, Positive and Negative Affect Schedule) into subfacets (e.g., self-evaluation, social difficulties). To provide a balanced measure of mood, positively valenced dimensions (e.g., positive self-evaluation, connectedness) were created to complement each negatively valenced subscale (except for the somatic and physiological symptoms dimensions). Each subscale contains 6 items (total of 84 items), and each item is measured on a 4-point scale (*not at all true*, *somewhat true*, *mostly true*, *completely true*). Higher scores on each subscale correspond to the name of the subscale (e.g., on the Negative Emotion subscale, higher scores indicate greater levels of negative emotion).

Gallagher and Johnson's MIDUS Example

The Midlife in the United States (MIDUS) national survey was initiated in 1994 by the MacArthur Midlife Research Network in order to explore the behavioral, psychological, and social factors associated with healthy aging. The initial MIDUS sample consisted of a nationally representative sample of 7,108 individuals who were recruited using random digit dialing procedures and who completed a phone interview. Of this initial sample, 6,329 individuals (89%) completed an additional battery of self-report questionnaires, including measures of negative affect and neuroticism. The MIDUS2 survey was initiated in 2004 as a longitudinal follow-up to the MIDUS1 sample. For that analysis, the researchers selected all the participants who had at least partly completed the Negative Affect and Neuroticism scales for both

MIDUS1 and MIDUS2 and who had reported an age in MIDUS1. These inclusion criteria left 3,871 individuals (2,143 females, 1,728 males). The age of these individuals ranged from 25 to 74 (mean = 47.32, standard deviation = 12.40) at Time 1 and from 35 to 84 (mean = 56.25, standard deviation = 12.34) at Time 2.

Neuroticism

Neuroticism was measured using a four-item scale in both the MIDUS1 and MIDUS2 surveys. This scale asked participants the extent to which four adjectives (*moody*, *worrying*, *nervous*, and *calm*) described them on a 4-point Likert scale with response options ranging from *not at all* to *a lot*. A mean score was computed across the four items after reverse-scoring the items so that higher scores on the scale indicated higher levels of neuroticism.

Negative Affect

Negative affect was measured using six items in both the MIDUS1 and MIDUS2 surveys. This scale asked participants "During the past 30 days, how much of the time did you feel . . . ?" and participants responded using a 5-point Likert scale with response options ranging from *all of the time* to *none of the time*. Example items include "hopeless" and "so sad nothing could cheer you up." A mean score was computed across the six items after reverse-coding items so that higher scores indicated higher levels of negative affect.

Dorothy Espelage's Bullying and Victimization Examples

Participants included 1,132 students in fifth through seventh grades from four public middle schools in a Midwestern state. Ages ranged from 11 to 15 years, with a mean of 12.6 years, in the first wave of data collection. Students included 49.1% (n = 556) female and 50.9% (n = 576) male, with a racial distribution of 56.5% (n = 640) African American, 26.1% (n = 295) European American, 11% (n = 124) other or biracial, 3.8% (n = 43) Hispanic, 1.5% (n = 17) Asian, and 1.1% (n = 13) American Indian or Alaskan Native. Data were collected over five waves (spring 2008, fall 2008, spring 2009, fall 2009, and spring 2010) and included three cohorts.

Peer Victimization

Victimization by peers was assessed using the University of Illinois Victimization Scale (UIVS; Espelage & Holt, 2001). Students were asked how often the following things had happened to them in the past 30 days: "Other students called me names"; "Other students made fun of me"; "Other students picked on me"; and "I got hit and pushed by other students." Response options were *never*, *1–2 times*, *3–4*

times, *5–6 times*, and *7 or more times*. Higher scores indicate more self-reported victimization.

Substance Use

Alcohol and drug use was assessed with an eight-item scale (Farrell, Kung, White, & Valois, 2000) that asked students to report how many times in the past year they had used alcohol and/or drugs. The scale consisted of items such as "smoked cigarettes," "drank liquor," and "used inhalants." Responses were recorded on a 5-point Likert scale with options ranging from 1 (*never*) to 5 (*10 or more times*).

Family Conflict

The Family Conflict and Hostility Scale (Thornberry, Krohn, Lizotte, Smith, & Tobin, 2003) was used to measure the level of perceived conflict and hostility in the family environment. The scale contained three items from a larger survey designed for the Rochester Youth Development Study. Respondents indicated on a 4-point Likert scale how often hostile situations had occurred in their families in the past 30 days. Responses range from 1 (*often*) through 4 (*never*). In addition, a Sibling Aggression Perpetration Scale was created and included five items that assessed the aggression between siblings. Items were created to be parallel to items from the University of Illinois Bully Scale (UIBS).

Family Closeness

The Parental Supervision subscale from the Seattle Social Development Project (Arthur, Hawkins, Pollard, Catalano, & Baglioni, 2002) was used to measure respondents' perceptions of established familial rules and perceived parental awareness regarding schoolwork and attendance, peer relationships, alcohol or drug use, and weapon possession. The subscale included eight items measured on a 4-point Likert scale ranging from 1 (*never*) to 4 (*always*). Example items included "My family has clear rules about alcohol and drug use" and "My parents ask if I've gotten my homework done."

Bullying

Bullying was measured using the eight-item UIBS (Espelage & Holt, 2001), which includes teasing, social exclusion, name calling, and rumor spreading. This scale was developed based on student interviews, a review of the literature on bullying measures, and extensive factor analytic procedures (Espelage, Bosworth, & Simon, 2000; Espelage et al., 2003). Students indicated how often in the past 30 days they had engaged in each behavior (e.g., "I teased other students"; "I upset other students

for the fun of it"). Response options were *never*, *1 or 2 times*, *3 or 4 times*, *5 or 6 times*, and *7 or more times*.

Homophobic Teasing

Homophobic teasing was measured using the Homophobic Content Agent–Target scale (HCAT; Poteat & Espelage, 2005). The HCAT scale was used to assess homophobic name-calling perpetration. This perpetration scale contains five items and measures how many times in the past 30 days a child has called other students homophobic epithets. Students read the following sentence: "Some kids call each other names: *homo*, *gay*, *lesbo*, *fag*, or *dyke*. How many times in the last 30 days did YOU say these words to . . . ?" Students then rated how often they said these words to five different types of people, such as a friend, someone they did not like, or someone they thought was gay. Response options were *never*, *1 or 2 times*, *3 or 4 times*, *5 or 6 times*, or *7 or more times*. Higher scores indicate higher homophobic name-calling perpetration.

OVERDUE GRATITUDE

In any endeavor as large and as long in the making as this book, the number of persons to whom one is indebted is huge. My graduate mentor and friend, Keith Widaman, is responsible for all the good ideas I write about herein. The bad ideas are my mistakes. He taught me everything I know about statistics and SEM (but not everything he knows). My friend and colleague Noel Card has been instrumental in helping me hone many of these ideas over the years. Noel had been a coinstructor in the annual Summer Stats Camps that we conduct every June (see *www.statscamp.org* for more information) and regularly offers seminars on his own. He has been a long-standing collaborator. Noel also served as a reader and reviewer of the current material and stepped in to write the Foreword. My friend and colleague Kris Preacher has also patiently rectified my thinking on some the topics, and maybe I have influenced his thinking on a few; but either way, the discussions, comments, and assistance that Kris provided are invaluable to me (he's just plain good folk!). My first KU graduate student, and now colleague and friend, James Selig, has been an instrumental part of this journey, and he returns once in a while to teach in Stats Camp. James also provided invaluable feedback on the entire contents. My ex-wife, Patricia Hawley, was also instrumental in the evolution of my career as is my wife, April. During the development of the first edition, my colleagues in CRMDA at KU kindly provided feedback and support along the way: Pascal Deboeck, Chantelle Dowsett, Sadaaki Fukui, David Johnson, Paul Johnson, Jaehoon (Jason) Lee, Alex Schoemann, John Geldhof, Carol Woods, Mijke Rhemtulla and Wei Wu. Lightening my administrative burden, Shalynn Howard, Jeff Friedrich, and Jo Eis Barton

have allowed me those few precious extra minutes in a day to write and work on this. The many students in the quant program at KU, in the undergraduate minor at KU, and in the CRMDA are too many to name, but their contributions have been immeasurable and essential. After I moved to TTU, many folks worked with me to refine and expand on many of the extant topics that I covered in the first edition: Kyle Lang and Charlie Rioux, who did post-docs with me and coauthored a number of methods works as well as myriad students—Britt Gorrall, Lola Odejimi, Zack Stickley, Esteban Montenegro. I also need to express gratitude to Daniel Bontempo and Allison Tracy, with whom I have worked closely over the past years as part of a team conducting an evaluation of the Dating Matters© teen dating violence prevention intervention supported by the CDC.

The Guilford Press, under the wise and supportive leadership of Seymour Weingarten and the late Bob Matloff, has been supportive and patient in the process. Gerry Fahey did a spectacular job copyediting and William Meyer's effort to bring the pieces together is much appreciated. Most notable in this whole process, however, is the incomparable C. Deborah Laughton, editor extraordinaire, whose kind cajoling and helpful advice kept me on task and brought this to fruition.

As mentioned, I have taught this material to thousands of students over the years. Their input, questions, and comments have helped us all understand the material better. To those thousands of students who have taken SEM and related courses from me or consulted with me on matters statistical: thanks for pushing me to find clearer, simpler, and varied ways of communicating the ideas that underlie SEM. Many folks have read and commented on various drafts of the chapters. I have tried to keep a tally of them all, but I have lost track. This list is in a random order, and it is, unfortunately, incomplete: Katy Roche, Jenn Nelson, Alex Schoemann, Mrinal Rao, Kris Preacher, Noel Card, Ed Fox, John Nesselroade, Steve West, Sharon Ghazarian, James Selig, John Geldhof, Waylon Howard, and _____ (fill in your name here if I have forgotten you).

I also need to thank the wonderful set of reviewers who provided feedback on the first edition, as well as the revised and new chapters:

Leonard Burns/Psychology/Washington State University
Aaron Metzger/Psychology/West Virginia University
Kristin D. Mickelson/Psychology/Arizona State University
Douglas Baer/Sociology emeritus/University of Victoria
Ellen Hamacher/Utrecht University
Oliver Christ/Fern University in Hagen
Kevin Grimm/Psychology/Arizona State University
Ed Merkle/Psychology/University of Maryland
Sarah Depaoli/Psychology/University of California, Merced

To my family and friends, few of whom will understand what's herein but support me none the less.

APOLOGIES IN ADVANCE

With apologies to Rex Kline, I do often anthropomorphize my tables and figures, and to Brett Laursen, I also anthropomorphize variables. My tables show, my figures display, and my variables have relationships—but they don't sing or dance or anything like that. If they did, I'd be worried.

With apologies to Kris Preacher, I want to be informal in my communication. I'm a fourth-generation Montanan. Using terms like "folks" is just my style.

With apologies to the Smart Innovators in our field, I have tried to be as up-to-date as possible. And I have read a number of papers that may challenge some of my recommendations and conclusions. Where I maintain my view, I do not do so lightly; rather, I'm still not convinced that the basis for my recommendations or conclusions has been sufficiently challenged. I remain open to feedback, and I will gladly share the basis for any errata on the web page for this book (*www.guilford.com/little-materials*).

With apologies to all the persons I did not name in my acknowledgments, whose input made this book a better book than it would have been otherwise.

KEY TERMS AND IDEAS INTRODUCED IN THIS CHAPTER

Serendipity. The idea of making a fortunate discovery or finding oneself in a fortunate circumstance when the discovery or circumstance found was not what one was looking for or striving for.

Statistics. Statistics is the point at which common sense meets logic, and numbers are used to convey the ideas and the logic. More formally, statistics involves collecting (measuring), organizing (database management), analyzing (descriptively or inferentially), and interpreting (making decisions from) numerical data. Or, as Bill Bukowski (personal communication, 2008) has described it, "If math is God's poetry, then statistics is God's elegantly reasoned prose."

RECOMMENDED READINGS

Little, T. D. (2011). Conveying complex statistical concepts as metaphors. *The Score, 33*(1), 6–8.

This is an essay I wrote at the request of Dan Bernstein for KU's *Reflections from the Classroom* publication. It was reprinted in 2011 in *The Score*, the newsletter for Division 5 (Measurement, Evaluation, and Statistics) of the American Psychological Association. It conveys why I think metaphors work so well and can be invaluable as a tool for teaching advanced statistics.

Little, T. D. (2015). Methodological practice as matters of justice, justification, and the pursuit of verisimilitude. *Research in Human Development, 12,* 268–273.

I wrote this essay, to encapsulate why it is so critical to conduct research at the highest level of accuracy (which comes with complexity). It's a matter of social justice.

Little, T. D., Widaman, K. F., Levy, R., Rodgers, J. L., & Handcock, G. R. (2017). Error, error, in my model, who's the fairest of them all? *Research on Human Development, 14,* 271–286.

I had a lot of fun working with these coauthors who, collectively, have over 100 years of wisdom in conducting research and modeling data. In this regard, it's all about error management and the places where errors can occur: they begin at the beginning and seemingly never end.

Terrell, S. R. (2021). *Statistics translated: A step-by-step guide to analyzing and interpreting data* (2nd ed.). Guilford Press.

I reviewed this introductory statistics textbook for Guilford (both editions). It's the textbook I would have written: it's relaxed, speaks to you, informs you, and, if you're like me, you'll have a few LOL moments.

Adams, D. (1979). *The hitchhiker's guide to the galaxy.* Pan Books.

Here you'll learn to not panic and why forgetting your towel is a no-no.

Card, N. A. (2011). *Applied meta-analysis for social science research.* Guilford Press.

I learned all that I know about meta-analysis from Noel. He's never meta-analysis he didn't like. I like them, too. Even though my book is not about meta-analysis, I do make recommendations on reporting that should help us in our meta-analytic endeavors.

2

Design Issues in Longitudinal Studies

I focus this chapter on often neglected issues related to the foundations of longitudinal SEM (i.e., the tree trunk idea I outlined in Chapter 1). Here, I'll discuss the timing of measurements, conceptualizing time, and some innovations in measurement including more directly measuring change. In the newly created Chapter 3, I will also elaborate on modern treatments for missing data, including the idea of planned missingness. These topics are all design related, although a full appreciation of these issues likely requires some conceptual understanding of SEM, which I tried to provide in the previous chapter. The chapter order for presenting these issues is not a traditional one, for I could easily present them in a later chapter. I'm discussing them here, however, because I want novice (and even some seasoned) SEM users to appreciate these ideas before the other topics are presented, because I make reference to them along the way. Given that these issues would and should be considered up front at the design phase of a study—well before a model is actually fit to data—discussing the ideas here seems logical to me. On the other hand, I could make this chapter the closing chapter because you would then have the context of the various models to use as a frame of reference and would be able to hang these ideas on the scaffold that has already been created. My recommendation is to read this material now as Chapters 2 and 3 and read it again after going through the remaining chapters.

TIMING OF MEASUREMENTS AND CONCEPTUALIZING TIME

In the social and behavioral sciences, longitudinal studies usually index time as age in years or as the occasions of measurement, which quite often are the same (e.g., 10-year-olds measured at Time 1 are 11 a year later at Time 2 and 12 at Time 3, and so on). In this section, I discuss and highlight some alternative ways to consider time that will assist you in thinking about the design of a longitudinal study. If done

with a bit of forethought, a properly designed longitudinal study will improve your ability to capture the change process that you desire to model. For a very rich but challenging discussion of the issues that I touch on here, you can try to find a copy of Wohlwill's (1973) classic book, *The Study of Behavioral Development*. More recent (and more approachable) discussions of the importance of mapping theory to methods and methods to theory have emerged (see, e.g., Collins, 2006; Jaccard & Jacoby, 2020; Lerner, Schwartz, & Phelps, 2009; Ram & Grimm, 2007).

Most developmentalists have been schooled that "behavior is a function of age:" $B = f(\text{age})$. This way of thinking about developmental change processes is somewhat limited, however. First, age is not a causal variable and is really only a proxy of the multitude of effects that covary with maturation and experience (Wohlwill, 1973) and are partner effects in the context of influences like race/ethnicity, sexual orientation, and gender. In fact, age is probably best considered as an index of context. The context of behavior in the life of a 12-year-old Hispanic gay boy would have different connotations than the context of behavior in the life of a 16-year-old African American transgendered girl. In addition, the historical timing of measurements is often neglected as a contextual impact. The widespread impact of social media and constant connectivity has even emerged as a new field of study, what Nilam Ram has described as Screenomics.

Second, as Wohlwill notes with regard to age functions, "the particular form of this functional relationship is rarely taken seriously, let alone given explicit expression" (p. 49). In other words, a reliance on chronological age as the de facto index of change engenders a limited view of developmental change processes. A more flexible way to consider development is that "change is a function of time:" $\Delta = f(\text{time})$. Arguably, the one constant in human development is change. This alternative formulation encourages a broader consideration of the primary time dimension of a longitudinal model and moves away from the equal-interval age divisions that are typically used to index developmental change (Lerner et al., 2009). Identifying and modeling change using the right time dimension will maximize the likelihood that a study will represent accurately the developmental/change processes of interest.

Before I get too far in this discussion, I need to briefly outline the five basic types of developmental designs that are commonly used. In Figure 2.1, I have laid out a table with a hypothetical set of age cohorts along the left edge (y axis) and potential times of measurement along the horizontal dimension (x axis). The implied ages that would be assessed are in the body of the table grids. Each of the five designs can be found in this figure. In addition to the three sequential designs that are labeled in the body of the figure, the cross-sectional design and the single-cohort longitudinal design can be found. A single-cohort longitudinal study would be depicted by any given row of Figure 2.1, and a cross-sectional study would be depicted by any given column in the table.

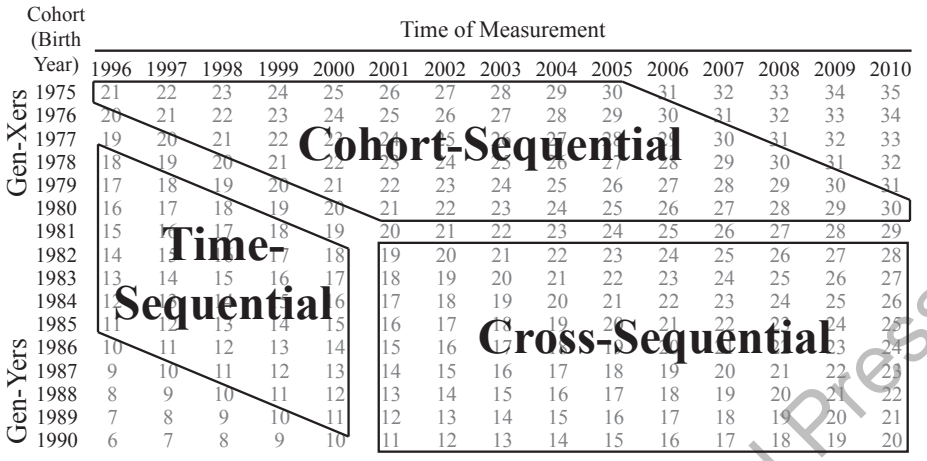

FIGURE 2.1. Traditional sequential designs. *Note.* Ages are entered in gray in the body of the table. Any given row would be an example of a longitudinal design, and any given column would be an example of a cross-sectional study. A cohort-sequential design would consist of starting a new cohort at a certain age and then following longitudinally. A cross-sequential design starts with a traditional cross-sectional study and then follows all participants longitudinally. A time-sequential design is a repeated cross-sectional design, with some participants followed longitudinally.

Cross-Sectional Design

Of the five developmental designs, only the cross-sectional design does not include a repeated assessment as part of the data collection protocol. The cross-sectional design makes age comparisons by selecting different samples of persons from different age cohorts. The cross-sectional design is quite limited in its ability to describe developmental processes. It can hint at developmental differences, but this hinting is always confounded with age-cohort differences and strongly influenced by between-group sampling variability. Cross-sectional designs are, therefore, most useful for yielding preliminary data to determine whether the measures used are appropriate for the age cohorts that would be studied in a longitudinal manner. The cross-sectional design is also useful to determine whether the internal validity relationships among the constructs are as expected and whether the age-cohort differences are in the direction of the expected age differences. If the answer to these questions is generally yes, then pursuing a longitudinal component would be warranted. If any of the answers is no, then some more planning and measurement development would need to be done before a longitudinal study would be launched. In other words, cross-sectional studies are not much more than feasibility studies for longitudinal studies.

Cross-sectional designs are well suited to addressing measurement validity issues across different age cohorts. The factorial invariance of constructs across age cohorts can be assessed and validated before engaging in a longitudinal study. I discuss factorial invariance in detail in Chapter 6; but, briefly, factorial invariance is testing whether or not the indicators of a construct measure the construct in the same

way, either across age cohorts in a cross-sectional study or across time in a longitudinal study. Assuming that factorial invariance is supported, then a number of characteristics of the constructs can be examined. Within each cohort, factorial invariance establishes the content validity of the constructs' respective indicators. When multiple constructs are measured, the concurrent criterion-related validities among the constructs can be examined. In terms of differences across the age cohorts, the differences in the mean levels of the constructs, the variances of the constructs, and the strength of the criterion associations among the constructs can each be examined. Because of the inherent confound of age, cohort, and time of measurement, however, any interpretation of the group differences as somehow reflecting age differences would not be valid.

Single-Cohort Longitudinal Design

The single-cohort longitudinal design (any given row of Figure 2.1) and the sequential designs depicted graphically in the body of Figure 2.1 each have a repeated-measures component. As such, they can be analyzed using a traditional panel model, the more dynamic growth curve model, or the alternative within-person models that Danny Osborne and I discuss in Chapter 9. The key difference between these statistical models is how variability is modeled.

The traditional panel design focuses on slices of variability that are temporally static and focuses on the individual-differences relationships in a series of sequential "snapshots." In such a snapshot, various features among the constructs are examined: namely, the content validity of the indicators, concurrent criterion validity among the constructs measured at each time point, and predictive criterion validity relations among the constructs over the interval(s) specified. In addition, changes in construct means and variances can be tested across each successive measurement occasion. The growth curve models, on the other hand, focus on the more "fluid" variability over time, with particular focus on the variability in individual trends (interindividual differences in intraindividual change). The other alternative within-person change models are also more "fluid" in how change is conceptualized and modeled. The panel model is one in which a series of vertical slices are made in a flowing developmental process, whereas the other within-person models attempt to cut a slice horizontally across the time frame encompassed by a given study.

The single-cohort longitudinal study does allow you to examine change relationships. This design is limited because age is perfectly confounded with both age cohort and time of measurement. That is, in a single-cohort longitudinal study, any observed changes may be related to age, but they are also related to any time-of-measurement effects (i.e., events that co-occur with the measurement occasion). Covid-19 is now the quintessential example of a time of measurement effect. In fact, my colleagues and I recently wrote an article that outlines numerous ways to model Covid-19 as an effect in various types of longitudinal models that were disrupted

by Covid (Rioux, Stickley, & Little, 2021). In addition, the observed effects may not generalize to other cohorts because of possible cohort effects related to the age sample that was originally chosen (e.g., baby boomers vs. Generation Xers). More concretely, let's say I selected a group of 18-year-olds to follow every year for 10 years. When I wrote the first edition in 2013, my sample of 18-year-olds would all be from the 1994 age cohort. My population of generalization, then, is youths born in 1994. The youths born in 1994 have experienced a number of events that future 18-year-olds born today likely will not experience (e.g., the 9/11 terrorist attacks, Hurricane Katrina, and of course, Covid-19). Tragic and life-changing events such as these are likely to influence a number of attitudes, beliefs, behaviors, and cognitions of a given sample of youths. In this regard, any "age" differences may be due more to the cohort from which the sample is drawn than to true age differences.

In addition to the age-cohort confound, time of measurement confounds any age differences. If data were collected in 2008, for example, the sample of participants would have experienced major events that would have co-occurred with the measurement occasion: a major economic downturn and the ability to vote in a pretty historic presidential election. At this measurement occasion, the specter of limited job or educational opportunities would have likely had an influence on many youths in the sample. Similarly, the "hope of change" that was at the core of Barack Obama's 2008 presidential campaign would have inspired or influenced many of the sampled youths at the time. Similarly, for youth measured in 2021, the political changes ushered in by the election of Donald Trump are global and likely to be long-lasting. Any estimates of the constructs' means, variances, or correlations in the sample at this measurement occasion might be a reflection of being a certain age, they might be a reflection of the events co-occurring with this measurement occasion, or they might be a reflection of the past events that the cohort has experienced. To address these confounds, methodologists have thought about ways to study age changes that are not so inextricably confounded. The three so-called "sequential" designs each attempt to remedy some of this inherent confounding.

In Figure 2.1, I have superimposed outlines of an example of the three "sequential" designs over the ages listed in the figure. Each sequential design attempts to minimize the confounding among age, cohort, or time of measurement that is inherent in any longitudinal study. To minimize this confound, developmentalists such as K. Warner Schaie (Schaie & Hertzog, 1982) and Paul Baltes (1968) discussed alternative sequencing designs that would allow one to disentangle some of these confounded effects. From these discussions, three primary designs emerged: the cross-sequential, the cohort-sequential, and the time-sequential.

Cross-Sequential Design

The cross-sequential design starts with a cross-sectional design and then follows all the participants over time. Many people confuse this design with the *cohort*-sequen-

tial design. It is *cross*-sequential because it starts with a cross-sectional design and then adds a longitudinal sequence to each cohort of the original cross-sectional sampling. In the cross-sequential design, cohort and time of measurement are the two dimensions of time that are "manipulated" and controlled by the experimenter. The cross-sequential design is perhaps the most popular longitudinal design used today, even though it is the least powerful design to use if one aims to examine a developmental function as a reflection of age. The reason for this weakness is the fact that any age differences are confounded with the interaction between cohort and time of measurement. Specifically, age differences at a younger versus older age are confounded because older cohorts (e.g., the Generation Xers or Gen Xers) would provide "age" estimates that occurred before influential time-of-measurement effects, whereas younger cohorts (e.g., the Generation Yers or Gen Yers) would provide these estimates after the time-of-measurement effects. More concretely, youths assessed at the same age but who are from two different cohorts would have confounded estimates of true age differences. For example, the older cohort, Gen Xers, would have been measured pre-Covid, and the younger cohort, Gen Yers, would be measured post-Covid. On the other hand, this design is very well suited to examining change processes over time, controlling for potential differences in cohorts.

Cohort-Sequential Design

The cohort-sequential design is like starting a longitudinal study at the same age over and over again. That is, each year, a new sample of participants of a certain age is selected and enrolled in a longitudinal study. Here, each new "cohort" is enrolled in a longitudinal sequence that covers the same age span. This design is particularly well suited to identifying age differences while controlling for cohort differences. An important limitation of the cohort-sequential design, however, is the assumption that time-of-measurement effects are trivial, because any time-of-measurement effects are confounded with the interaction of age and cohort. Potentially powerful time-of-measurement effects such as Covid can have influences across all cohorts, yet the effects would show up as a cohort-by-age interaction with this design. The problem here is that the analysis cannot disentangle whether the effect was a time-of-measurement effect or a true age-by-cohort interaction. In other words, pre- versus post-Covid effects would be confounded with the older cohort measured at older ages versus the younger cohort measured at younger ages. Although time-varying covariates to capture the time of measurement effect could possibly be included to help disentangle the potential confound.

Time-Sequential Design

The time-sequential design is probably the least used of the sequential designs, but it is particularly useful for identifying time-of-measurement effects and age effects.

The age range is kept the same and repeatedly assessed (with only some participants being repeatedly measured). With this design, the age window is critical, and repeated testing of new and continuing cohorts at different times of measurement would identify the age-related changes that are not confounded with time-of-measurement differences. That is, time-of-measurement effects can be estimated and thereby controlled when looking at age differences. In this design, the cohort effects are the fully confounded factor. That is, any cohort effect would appear as an age-by-time interaction, which would not allow one to conclude whether the effect was a cohort effect or a true age-by-time interaction. Here, an effect of Gen Xers versus Gen Yers would be confounded with younger participants measured pre-9/11 or pre-Covid versus older participants measured post-9/11 or post-Covid.

The choice of which design to use depends on what type of change function is important to examine: Which one of the three constituents (age, cohort, time of measurement) of a longitudinal design will be critically important in understanding the change phenomenon to be studied? Has other work identified which of these factors is likely to be trivial? In such situations, the choice of design becomes straightforward. If, on the other hand, you can't say which factor is worth "ignoring," then you can consider some alternatives. Some hybrid designs have also been introduced that attempt to remove the confounding among all three of the age, cohort, and time-of-measurement effects. Schaie and Hertzog (1982), for example, offered an "optimal" design that involves different random sampling schemes that effectively use aspects of both cohort-sequential and time-sequential designs to disentangle age effects, as well as cohort versus time-of-measurement differences. Similarly, the accelerated longitudinal design (discussed later in Chapter 3) can be used to estimate age effects controlling for some cohort and time-of-measurement effects.

Cohort and time-of-measurement effects are not always confounds that must be controlled by the nature of the design. Another way to think of these effects is as potential context effects that are measurable. If you include measures of these effects or measures that adequately gauge how a person responds to particular events, then you can either control for them statistically or use them to predict the amount of variance that is due to the cohort or time-of-measurement effects.

A key point of the preceding discussion is that you need to consider the overall longitudinal design that is most appropriate for the change process being modeled. As I continue to emphasize throughout this book, strong theory will guide your thinking through most of these design/statistical conundrums.

Other Validity Concerns

In addition to the cohort and time-of-measurement effects that I just described, a number of other potential validity threats are found in longitudinal studies (see Campbell, Stanley, & Gage, 1963; Schaie & Hertzog, 1982). The "classic" threats include regression to the mean, retest effects, selection effects, selective attrition, and instrumentation effects (e.g., factorial noninvariance).

Regression toward the mean is the tendency for extreme scores to move closer to the mean of the distribution at subsequent measurements. Regression to the mean is purely a phenomenon of unreliability in repeated-measures situations. The random variation is the reason that extreme scores at the first time point will tend toward the mean at the second time point. Because regression to the mean is only a function of unreliability, it is easily remedied by using latent-variable SEM. When multiple indicators of constructs are used, the variance of the construct is thereby measured without error. The random variation that is responsible for regression to the mean is removed from the measurement process by virtue of the multiple indicators. The effect of regression to the mean appears only in the manifest variables, which contain the measurement error, and not the latent variables, which are composed of 100% reliable variance.

In contrast to regression effects, retest effects are more nefarious and difficult to remedy. Retest effects occur when a measure is sensitive to repeated exposure, whether it is practice that leads to improved performance or reactivity that leads to changes in responses due to the act of being assessed. Most measures are sensitive to repeated exposures, but the impact may vary depending on the measure. Some measures will increase in mean levels, whereas some will decrease as a function of repeated exposure to the instrument. Repeated exposure can also have a homogenizing effect in that the extremes of a scale may be responded to less and less over time, thereby shrinking the variance of the variable over time. One of the best ways to estimate and correct for retest effects is to randomly assign participants to receive or not receive a given measurement occasion or use a carefully crafted multiform planned missing protocol. Such designs are referred to as intentionally missing or planned missing data designs (see Chapter 3).

Selection effects are fundamental to any study in which a sampling plan fails to provide a representative sample of the population to which one wishes to generalize. Avoiding the lure of convenience samples and pseudo-random selection will go a long way toward increasing the quality of behavioral and social science research in general and longitudinal studies in particular. A related problem of longitudinal studies is selective attrition. Selective attrition occurs when dropout from a study (attrition) is not a random process but is related to some characteristic(s) of the sample. As you'll see in Chapter 3, selective attrition is relatively easy to address using modern missing data estimation procedures if one plans ahead and measures known predictors of dropout. Note that oversampling selected groups is not the same as selective sampling. Oversampling can be converted to representative analyses by using population weights to adjust parameter estimates accordingly.

Instrumentation effects can influence longitudinal studies in a couple of ways. First, the measurement properties of the phenomenon of interest can change over time. When the measurement properties of an instrument change with age, then the measurement properties of the construct are not factorially invariant, and conclusions about changes in the constructs would not be valid. Fortunately, this type of instrumentation effect is a testable issue (for more details, see Chapters 6 and 7).

Second, instrumentation effects can influence conclusions from a longitudinal study when the measurement tool is not sensitive to change. Most measures developed in the social and behavioral sciences have been developed with so much focus on reliability that they are not very sensitive to change. Because test–retest correlations have been maximized, for example, such measures no longer contain the items or the item content that might have been sensitive to change. In fact, the test–retest model for reliability has likely resulted in long-term damage to the work of researchers who desire instruments that are reliably sensitive to change. The test–retest correlations capture only stability information, and the effort to maximize this stability in the development phase of a measure undermines the usefulness of the measure for identifying and modeling change processes. Developmentalists should consider this problematic aspect of measures when designing a study and selecting measures. I would encourage all developmental researchers to modify, adapt, or develop measures so that they are sensitive to change. We also need further work in the area of change measurement. Research on measurement can take advantage of advances in computerized testing. For example, one useful way to measure change (perhaps) would be a measure that populates itself with the responses from the first measurement occasion and then asks the respondent to indicate how much his or her current response (e.g., attitude, belief, mood, cognition, etc.) has changed from the prior occasion of measurement. Or, simply use a traditional slider-response scale at Time 1 but then at Time 2 use a slider where the mid-point is labeled "no change" and the end-points are labeled with a term that captures "less" and "more" (adapted to the question asked). The score at Time 2 would be the score at Time 1 plus the change score at Time 2. Time 3 and beyond would also be measured the same way as Time 2 and added to the prior score to get a cumulative change score (see Yu, Zhang, & Little, 2023, for an example of measuring change in this way).

Another way to measure change is to utilize a response scale that reflects a direct assessment of change. For example, in a large-scale evaluation of interest in STEM topics, we asked participants to rate their interest relative to the start of the after-school program on a sliding visual analog scale. One of the items on the science interest scale is "I am curious about science." In Panel B of Figure 2.2, I show a response option from less curious to more curious with the midpoint labeled "no change." Similarly, the retrospective-pre-post design can be used to better gauge perceived changes from pre-test to post-test. In the same after-school program evaluation, we utilized this design to also assess changes in interest. Howard (1984) introduced this design to rectify the response shift that often occurs during a traditional pre-post assessment. My colleagues and I recently re-introduced this design with empirical support for its effectiveness (Little et al., 2019) and I would encourage you to take a good look at this design for assessing perceived change (see Panel A in Figure 2.2). The basic idea is at the post-test, two questions are asked. One question asks participants to reflect and respond retrospectively to what they thought or felt at the time of the pre-test (e.g., before the intervention was initiated). Then, the second question asks the participants to rate how they think or feel now. In our

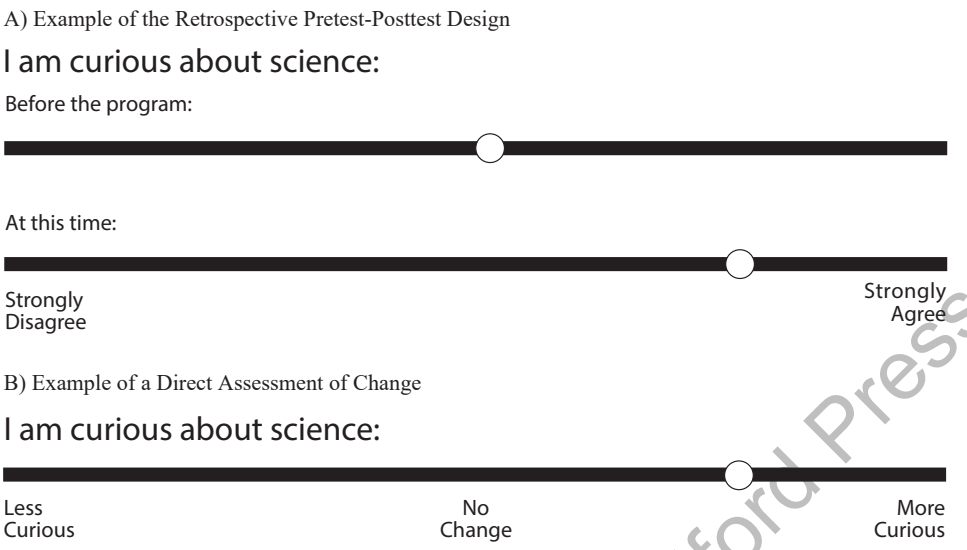

FIGURE 2.2. Two visual analog scaling methods for estimating subjective change.

paper intended to revive the design, we demonstrated that this way of assessing the intervention effects was quite sensitive to changes due to the intervention including implementation fidelity, program duration, and the like. Notably some conditions yielded no effect, which suggests that the design isn't a way to cheat and always yield positive effects.

Related to this instrumentation issue is the heterotypic versus homotypic expression of an underlying construct with age. An example of a homotypic expression would be happiness. The facial expressions of the emotion of happiness stay pretty much the same throughout the lifespan. The facial indicators of a happy expression (i.e., the various changes to the mouth, eyes, forehead) remain consistent indicators of the happy expression across all ages (i.e., the indicators are factorially invariant). An example of a heterotypic expression is aggression, particularly as it changes from toddlerhood to adolescence. Screaming and grabbing give way to name calling and punching. During early childhood, aggression is expressed with the cognitive and physical tools available to toddlers. During adolescence, aggression is expressed with the more advanced cognitive and physical tools of the adolescent. A measure of aggression that has items assessing toy taking, screaming, and kicking would probably work quite well for assessing the frequency of aggressive behavior in toddlers but would not capture the aggressive behaviors of adolescents.

If a study contains constructs that are heterotypic or uses "age-appropriate" measures for different phases of the study, then careful consideration must be given to how the construct is measured over time. When a measurement tool must change during the course of a study, the scores across the different measures must be comparable in order to talk about the same construct changing over time. With age-appropriate measures, for example, too often the younger age-appropriate measure

is swapped out completely for the one that is now appropriate for the older ages. When this kind of wholesale change takes place, all ability to map or model the changes in the construct over time are lost. There is no way to know how a score on the old measure relates to a score on the new measure. This problem can be remedied by transitioning between measures. If at least one measurement occasion exists where both the old instrument and the new instrument are given to the same group of individuals, the degree to which they measure the same thing can be tested and the measures can be calibrated. That is, how do the scores on the two instruments relate to one another? With longitudinal SEM, this process of calibration is relatively straightforward.

Figure 2.3 shows a couple of ways in which measures across time can be calibrated and linked if there is at least one occasion of measurement at which both measures are administered. In Panel A of Figure 2.3, the indicators of the two different

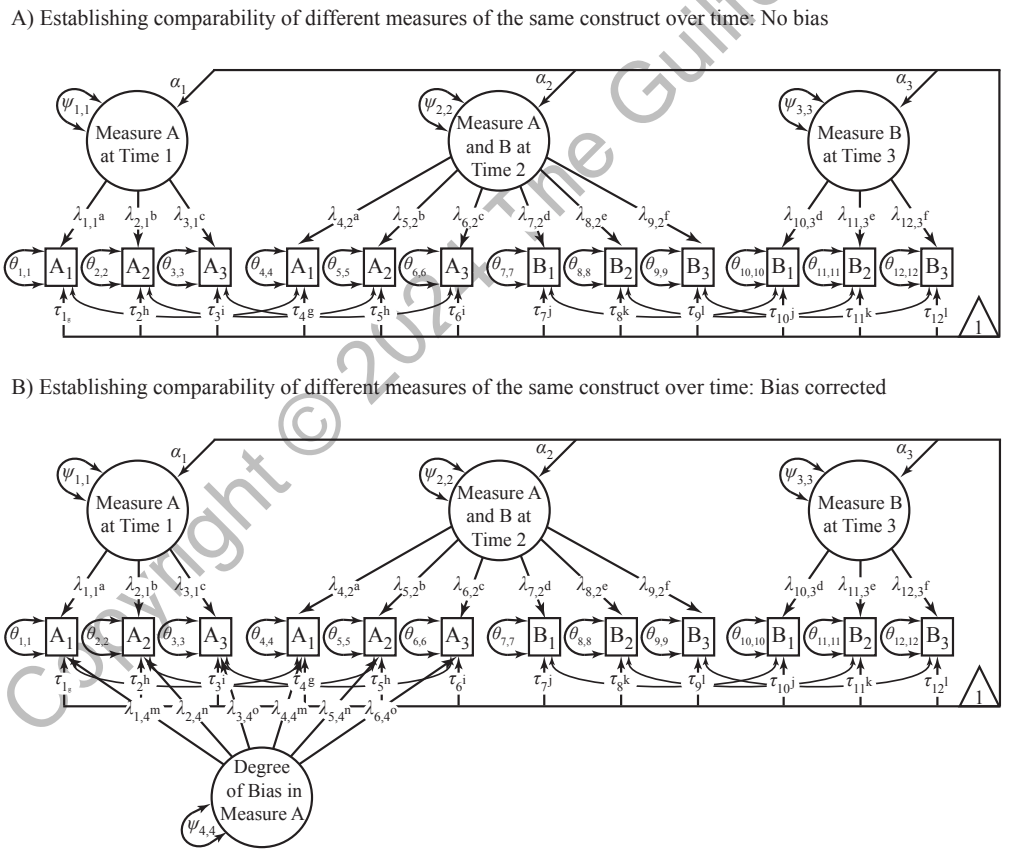

FIGURE 2.3. Two models for establishing comparability of different measures of the same construct over time, with and without a "bias correction" factor. *Note.* The corresponding loadings and intercepts that are equated across time are designated by a superscripted letter (a–o). The residual variances among the corresponding indicators are allowed to associate over time.

measures load on the same construct at the overlapping time point (Time 2 in this hypothetical example). The loadings and intercepts of the indicators for Measure A are specified to be factorially invariant over time (see Chapter 6 for details on factorial invariance). The loadings and intercepts of the indicators for Measure B are also factorially invariant. If this model fits the data, then the scores on the two measures have the same latent construct meaning (i.e., the factors' scores have the same meaning), and changes in the construct over time would be accurate and comparable over time. In Panel B of Figure 2.3, I have shown a variation of this type of model but have allowed for a "bias" factor. Assuming that Measure A has some systematic bias to it (e.g., teacher reports are used for Measure A while classroom observations are used for Measure B), the bias construct corrects the scores on the teacher-reported version of the construct for differences due to the systematic bias. The model in Panel B makes the presumption that one measure is biased while the other is not. A model with a bias factor for both measures would not be identified unless some limiting constraints are placed on the model parameters.

Using a multiform planned missing data design would allow assessment of all items of both measures without burdening all participants with all items from both measures (see Chapter 3). Even if a random subsample of participants was taken to give both measures to, this subsample could provide the linking functions between the two instruments that could be applied to the whole sample. For that matter, you could derive the linking functions on an independent sample and apply them to the longitudinal study. Of course, this latter approach makes a number of assumptions about the comparability of the two samples that may not be tenable and would require using a tricky multiple-group model to provide the linking functions (I'm getting off track here . . .).

Temporal Design

Temporal design refers to the timing of the measures that you want to employ. It is an overlooked aspect of longitudinal research design, even though Gollob and Reichardt (1987) outlined three basic principles of temporal design over three decades ago. First, they stated that causes take time to exert their effects. We can modify this statement to assert that effects of change take time to unfold. Whichever way you want to assert it (i.e., as causal effects or change effects), we need some appropriate quantity of time between measurements to see it. The second principle is that the ability to detect effects (either causal or change) depends on the time interval between measurements. With the rare exception of an on–off mechanism of change, this principle is a pretty accurate reflection of most developmental change processes. Some effects are cumulative and reach an eventual asymptote, which may or may not reset. Other effects rapidly reach a peak and diminish quickly or slowly over time. Figure 2.4 depicts a number of potential patterns of unfolding (causal or change) effects. Each pattern varies in the strength of its expression over time. I

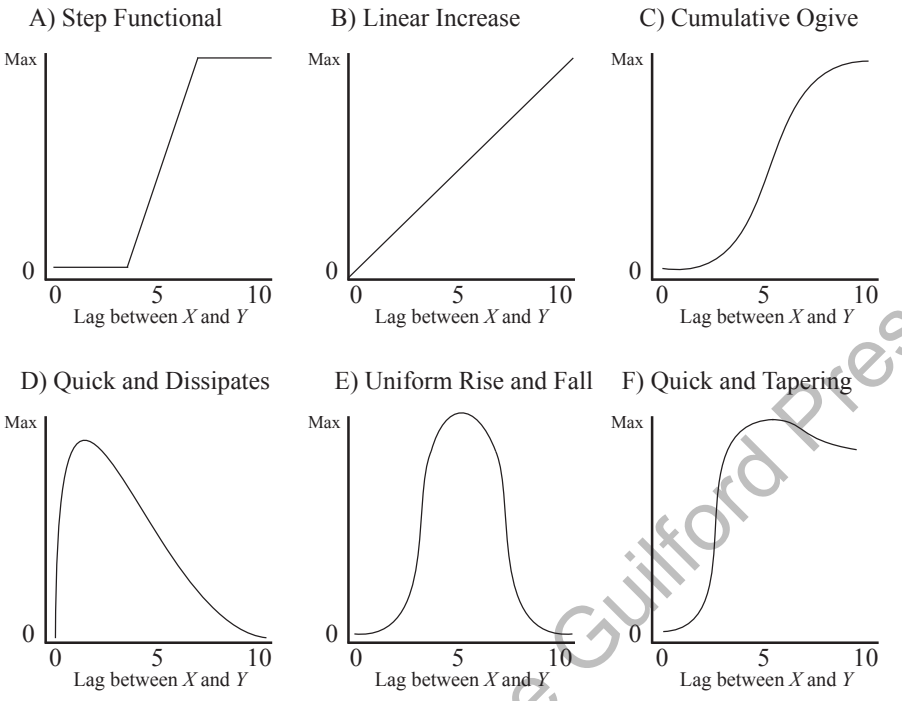

FIGURE 2.4. Some possible types of effects and their lag relations. *Note.* The time scale for the lag between *X* and *Y* can be on any scale, from milliseconds to years. The size of the effect can also vary from small to medium to large.

think it is helpful to consider the pattern of change that you might expect when you design your study so you can adequately capture the change process.

The implications of these two principles of temporal design are twofold. The first implication, which Gollob and Reichardt (1987) stated as a third principle, is that because different lags would show different magnitudes of effects, to understand the effect, assessments must occur at different lags. Measuring multiple lags would allow you to model the temporal unfolding of the effect of interest. The second implication, which can also be stated as a principle, is that because effects have an optimal temporal lag (when the effect is at maximum influence), we should design studies to measure at intervals that correspond to this optimal lag. When an optimal design approach is considered, however, it assumes that the shape of the effect is the same for all individuals, which can be a tenuous assumption. This problem is an instance of the ergodicity assumption (Molenaar, 2004). This assumption means that effects or patterns of change in a group of individuals would generalize to each individual. For many change processes, such an assumption may not be reasonable.

If the focus of the study is on the effect itself, Gollob and Reichardt's (1987) third principle dictates that a study must measure multiple lags in order to model

the unfolding of the effect. On the other hand, if you are interested in the system-of-change effects that lead to some outcome, you'd want to measure at intervals that capture the optimal lag of the effects of interest. If you measure one variable at its optimal lag and another variable at a suboptimal lag, the optimally measured variable will likely appear to be a more important change predictor than the suboptimally measured variable. This problem is similar to the unreliability problem in that a variable that is measured less reliably than another variable will often not appear to have as important an effect as the more reliably measured variable if unreliability is not corrected for. Latent-variable modeling overcomes this problem of unreliability by correcting for it when we use multiple indicators of a construct. For the optimal effect problem, a poorly timed measurement interval is more difficult to "correct for." A measurement strategy that varies lag would allow calculating the functional form of the effect over a span of occasions to identify the point of maximum effect.

In addition to ensuring the correct temporal ordering of variables, the timing of measurements must be as fast as or faster than the change process that you want to capture. That is, the amount of time between measurement occasions needs to be short enough to keep pace with the underlying process that is changing (see Chapter 1, the section titled "What Changes and How?"). Too often, occasions of measurement are determined by the convenience of the calendar (e.g., yearly intervals in the spring or biyearly in the fall and spring). If one is studying the change and development of relatively stable or trait-like constructs such as personality or intelligence, then yearly intervals of measurement are probably sufficient. If malleable constructs such as mood or self-esteem are the focus of study, the intervals of measurement must be closer together. Unfortunately, even a casual inspection of the literature indicates that the most common intervals for longitudinal studies are 1 year or longer. Such intervals can detect only slow-change trends and provide little information on the true processes that underlie the observed changes.

Lags within the Interval of Measurement

Often, a study will be executed such that a measurement occasion spans some specified period of time. For example, in a fall assessment of school-age youth, I might start collecting data in late September and finish by the end of October. Here, the occasion of measurement is actually a 6-week span. In a school-based study, this 6-week difference, from the beginning to the end of the assessment occasion, can have an impact on the measures that are collected. The potential influences include the fact that students at the end of the window will have had 6 more weeks of social interactions and educational instruction and, generally, will have undergone more developmental change. In addition, spillover becomes more likely as students are assessed in the later parts of the window. These students likely hear about the kinds of questions that are being asked and may begin to form ideas of what the study is about (for a detailed discussion of data collection lag effects in longitudinal studies,

see Selig, Preacher, & Little, 2012). Any of these factors could influence the strength of an observed effect.

Usually, if we think about this "window" at all, we think it won't have much of an impact. I have been guilty of this thinking in my own *past* work. For the "stable" constructs in my protocol, these short-term processes probably have not had much of an impact. But for the more malleable constructs (e.g., affect, esteem), these processes may have had an influence. In my older datasets, I won't know whether they did, because I did not code for the date of administration of a given protocol. Had I known then what I know now, I would have coded when a student received the protocol, and I could have created an index of time that reflects the protocol administration time. This variable could easily be included as a covariate or moderator in the models that I discuss in later chapters (see Selig, Preacher, & Little, 2012, for details of using lag as a moderator variable).

The more I work with researchers and study developmental changes, the more I realize how critical it is to get the timing of the measurements right. For the most part, measurements in developmental studies are selected too often on the basis of convenience than on the basis of a clear theoretical rationale. The power of an appropriately designed longitudinal study is simply underutilized in practice. It's kind of a sad state of affairs. I hope to see lots more future research that really focuses on the timing of measurements.

Episodic and Experiential Time

Aside from the traditional ways of conceptualizing time and this overlooked issue of the time lag within measurement occasions, time can be indexed to model change in a couple of other ways: episodic time and experiential time. In both of these ways of representing change as a function of time, the actual chronological age of the participants can still be included in these models as either covariates, moderators, or predictors.

Episodic time refers to the length of time during which a person experiences a particular state or context. Time before the episode and time after the episode reflect different potential change processes. Here, the index of time that we want to model is not necessarily the chronological age of the participant but where the participant is in relation to the key developmental episode that we are interested in capturing. Puberty is a classic example of a normative event that has distinct developmental repercussions regarding its timing and the change processes that occur prior to and after the pubertal event. The idea here is to organize your data so that you remap the occasions of measurement now to correspond with the "time" prior to the event and the "time" after the event. Table 2.1 provides a schematic representation of how to reorganize a longitudinal design into an episodic time design. Wohlwill (1973) provides an example of such a design using the maximum velocity of growth as the centering "event" and then modeling the growth function as time before and

TABLE 2.1. Transforming a longitudinal design into episodic time

Data collection wave crossed with episode occurrence

Pattern	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
Pattern 1	P	P + 1	P + 2	P + 3	P + 4	P + 5
Pattern 2	P - 1	P	P + 1	P + 2	P + 3	P + 4
Pattern 3	P - 2	P - 1	P	P + 1	P + 2	P + 3
Pattern 4	P - 3	P - 2	P - 1	P	P + 1	P + 2
Pattern 5	P - 4	P - 3	P - 2	P - 1	P	P + 1
Pattern 6	P - 5	P - 4	P - 3	P - 2	P - 1	P

Episodic occurrence crossed with data collection wave

	P - 5	P - 4	P - 3	P - 2	P - 1	P	P + 1	P + 2	P + 3	P + 4	P + 5
Pattern 1						W1	W2	W3	W4	W5	W6
Pattern 2					W1	W2	W3	W4	W5	W6	
Pattern 3				W1	W2	W3	W4	W5	W6		
Pattern 4			W1	W2	W3	W4	W5	W6			
Pattern 5		W1	W2	W3	W4	W5	W6				
Pattern 6	W1	W2	W3	W4	W5	W6					

Note. Multiple cohorts could also be transformed in such a manner. A dummy code to represent age cohort would be included within each pattern to account for potential cohort differences. The missing data are treated as missing at random and imputed.

time after maximum velocity. The idea here is to "group individual functions into families in terms of some parameter of the developmental function such as rate or asymptotic level, in order to arrive at meaningful relationships to other situational or behavioral variables" (p. 142). Such a reorganization of the data allows one to examine nomothetic features of the change process separately from chronological age. As mentioned, chronological age and/or cohort can still be included in such models to examine its impact on the growth functions that are modeled around the episodic centering point.

In Table 2.1, I have created a hypothetical longitudinal study of a group of adolescents and used puberty as the event. The table shows different patterns or subgroups of adolescents based on the point during the study at which they experienced the onset of puberty. If I were conducting a cohort-sequential or a cross-sequential study, I would identify the patterns in each age cohort and assign them to the new index of time, centered on the pubertal event. Even if I had censored data (i.e., kids who were past puberty prior to the study or kids who did not reach puberty during the course of the study), I can still reorganize the data according to the scheme depicted in Table 2.1. In this case, I would add a P – 6 and a P + 6 time point to the episodic time sequence in the bottom half of Table 2.1. I would then identify the Pattern 0 and Pattern 7 youths who either already had reached puberty (Pattern 0) sometime before the start of the study or did not reach it (Pattern 7) during the course of the study. The impact of censoring on the trends when including the P – 6 and P + 6 groups can be accommodated by including a dummy-coded variable for each of these latter two patterns (Pattern 0 and Pattern 7) and estimating

their effects as covariates on the parameters of the episodic time model. Similarly, if I used a cohort-sequential or a cross-sequential design and transformed it into an episodic time sequence, I could include dummy codes for cohort and chronological age as covariates or predictors in the estimated models.

By way of another example, let's say I have measured a group of 15-year-olds at 12 measurement occasions separated by 1 month. The research question is how the members in this group of youths change in their social relationships prior to their 16th birthdays versus how they change after their 16th birthdays. At the first measurement occasion, approximately 1/12 of the sample experiences his or her 16th birthday. For this subgroup, the first measurement occasion corresponds to the window in which the birthday event has occurred, and each subsequent measurement corresponds to an occasion of measurement after the event. For another 1/12 of the sample, the birthday event does not occur until the 12th measurement occasion. For this subgroup, all the measurement occasions fall prior to the event. If I added a time-sequential design or cohort-sequential design on top of these monthly measurements, I could also disentangle potential time-of-measurement effects or cohort effects.

Note that when data are reorganized around episodic events, missing data are introduced. This missing data can be readily addressed using a modern treatment for missing data (see Chapter 3). Although it is critical when creating the bins that there is covariance coverage between each adjacent column of data in order to implement a modern treatment for missing data.

A related index of time is experiential time. Grade in school is a classic example of this index of time. With this type of index of time, age within grade level can also be included to examine differences that may be related to being younger versus older within a given grade level. Relative age effects are rarely examined but when they are, the effects can be quite pronounced when grade in school is your primary index of time. Relative age occurs in school-based studies because districts and states have an arbitrary date for entrance into school. Kids whose birthday is on or just after this date would be relatively young compared to their peers whose birthdates fall later in the year even up to the day before the arbitrary date that decides entrance or not into formal schooling. Relative age would be the number of days between the arbitrary date of entry and a participant's birthdate. This new variable can be used as a predictor, moderator, or even a potential mediator.

Another example of experiential time might be the length of time in an intimate relationship. Like relative age for a given grade, age in years can be used as a predictor, moderator, or even a potential mediator of say relationship satisfaction and its changes as a function of the length of time in the intimate relationship.

Although episodic time and experiential time are related, the primary difference is the focus. Experiential time is focused on how long participants have experienced a state or process and would use chronological age as a covariate, predictor, or moderator of the change relationships being modeled. Episodic time, on the other hand,

focuses on an event or episode as a potential turning point in a larger developmental process. As with experiential time, chronological age can be included and used as a covariate, predictor, or moderator of the modeled change process.

MODELING DEVELOPMENTAL PROCESSES IN CONTEXT

Nearly every longitudinal study that has been conducted makes some sort of comment about how the modeled process is subject to contextual influences. Usually, these comments occur in the limitations section of the discussion when the authors admit that they have not measured or controlled for the impact of context on the focal developmental process. A few years ago, my colleagues and I edited a book (Little, Bovaird, & Card, 2007) that discusses the merits of and shows the methods for modeling contextual influences on developmental processes. As I stated above, age itself can be seen as a proxy of the developmental context of the organism. In the following, I summarize some of the key points that can be found in that volume.

The context in which a person develops (physically, socially, emotionally, spiritually, etc.) is multidimensional and multilayered. First and foremost, the context of this development encompasses all the circumstances in which development unfolds (i.e., its settings). The context is the set of features that influences the performance or the outcome of a developmental process. The context also defines the conditions that are relevant to an outcome. In the discussion sections of most longitudinal studies, terms such as *circumstances*, *times*, *conditions*, *situations*, and so on are used when trying to convey the layers and levels of influence. The ecology of development is also another way to think of context. Here, the ecology defines the relationship between organisms and their environment. In Figure 2.5, I display a Venn (aka Ballantine) diagram of how these ecologies can vary along a social dimension, a physical dimension, and a personal dimension.

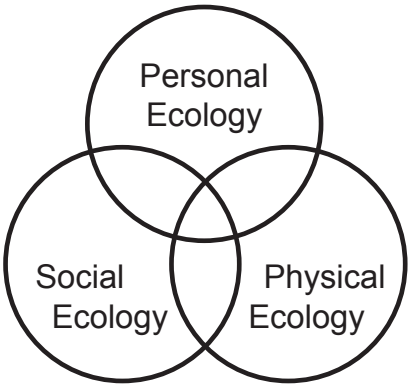

FIGURE 2.5. Ecologies of human development.

Each ecology exists in a nested manner, and each level has influences that can be measured as variables. These hierarchically nested variables can be used in an analysis model to examine their influence on the developing individual. For example, Bronfenbrenner's (1975, 1977) nested structure of the social ecology is perhaps the most famous and widely used conceptual model of context. The hierarchically nested nature of his model is depicted in Figure 2.6. The social ecology focuses on the social and cultural interactions and influences that affect the developing individual. The microsystem represents the influence of the immediate family, close friendships, and romantic partners. The mesosystem captures the next most distal level of social influences, such as peer groups, neighborhood communities, clubs, worship, and the like. Larger cultural influences are also represented at the higher levels of the nested data structures.

Keith Widaman developed a similar system of overlapping contextual influences that focuses on the physical ecology of the developing individual and not just the social ecology (Figure 2.7). The local/home ecology of the physical environment can include the in vitro environment or the immediate physical environment. At the next level, physical characteristics of the surroundings, such as neighborhood orderliness, hours of daylight, and the visible signs of community wealth, can affect the development of the individual within those contexts.

Finally, turning to Figure 2.8, I present a possible hierarchy of the personal ecology. Figure 2.8 is not meant to be a strong statement of whether the affective system is nested within the behavioral-cognitive systems or vice versa. Instead, the goal is to highlight that genes and ontogenetic expressions as well as age-related expressions of the personal ecology, are taking place within the core affective, behavioral, and cognitive systems.

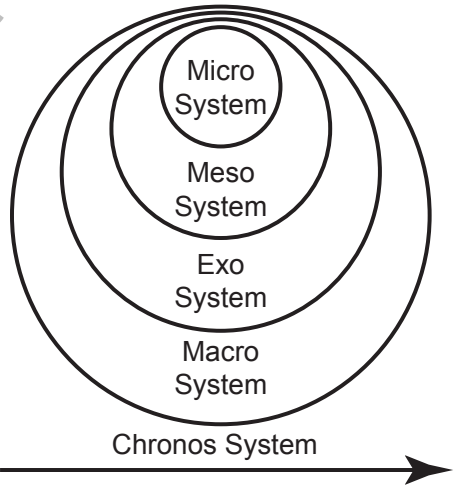

FIGURE 2.6. Bronfenbrenner's hierarchy of the social ecology.

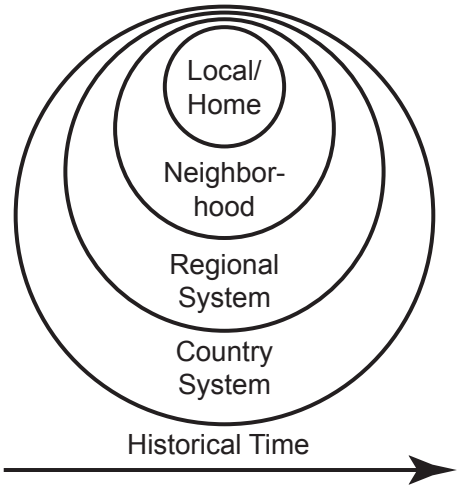

FIGURE 2.7. Widaman's hierarchy of the physical ecology.

Because contexts exist, they can be measured. The process of measuring the features and characteristics of the different levels of context often requires some innovation and careful consideration. Once a measure of a contextual influence is created or adapted, however, the nature of the contextual variable can be represented in a statistical model in a handful of ways.

Contextual variables can be entered as a direct effect that varies at the level of the individual and influences the individual directly. They can be entered as indirect (mediated) effects, whereby a contextual variable varies at the level of the individual and influences the individual through its effect on an intervening variable.

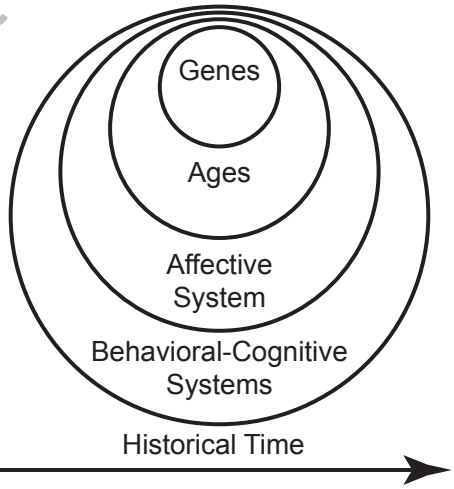

FIGURE 2.8. A possible hierarchy of the personal ecology.

An indirect effect is not necessarily a causal statement but rather an acknowledgment that the effect is distal and that its influence is now channeled through one or more proximally measured variables. Contexts can also be entered as mediating effects. Here, the distal context influences the proximal context, which in turn influences the individual. The primary difference between an indirect effect and a mediated effect is the causal hypotheses that underlie how and why the effect permeates from distal to proximal. Often this distinction is simply one of theory and additional empirical evidence supporting a strong causal conclusion. Statistically speaking, an indirect effect and a mediated effect are similar parameter estimates, but to be a mediator prior levels of the mediator and outcome need to be assessed to control for the stable variance in them. Contextual variables can be entered as moderating effects, which are interactive influences that change the strength of relationships for any of the preceding effects (see Chapter 10 for a detailed discussion of mediation and moderation).

Reciprocal effects and feedback loops are another way that contextual variables can be conceptualized. Statistically, however, such effects are entered as direct, indirect, mediating, or moderating influences. The key to demonstrating a reciprocal effect or a feedback loop is the timing and spacing of the measurement occasions. When properly designed, such effects are represented as cross-time associations that can be entered statistically as direct, indirect, mediated/mediating, or moderated/moderating influences.

The final type of statistical influence that represents the manner in which context can have an influence is via hierarchically nested effects. Nested data structures occur when the context is a larger sphere of influence that affects to some degree each of the entities contained within it. In the multilevel-modeling literature, the entities contained within the larger units are referred to as Level 1 units, and the hierarchical-level units are referred to as Level 2 units. Students (Level 1) contained in classrooms (Level 2) is the classic example. In longitudinal studies, however, the Level 1 unit is often the time of measurement that is nested within the individuals, who would then become the Level 2 units; higher units such as the classrooms would then become Level 3 units (even higher units, such as schools, would be Level 4 units, and so on; see Chapter 11 for a discussion of multilevel nested data structures as contextual variables).

Nested structures are larger units of context that can have direct, indirect, mediating, or moderating effects; or they may be mediated and/or moderated. Some key factors in modeling nested data structures include sampling enough of the larger units so that the hierarchical influence can be estimated as a random variable in the statistical model. Nested data structures can be represented as fixed effects when the number of larger units is relatively small. Here, the higher level "units" can be represented as groups in a multiple-group framework; or, if there is no evidence of moderation across the higher-level units, the units can be represented as a set of dummy-coded variables to estimate and thereby control for their influence.

In the second half of this book, I present examples of various models that contain contextual features and how they can be modeled. I also provide more detailed discussions of the steps involved. For now, my main goal is to remind researchers that context does indeed matter and that we have the analytic capacity to model their influences.

SUMMARY

In this chapter, I covered a number of foundational issues related to SEM that are closely tied to longitudinal data analyses. Perhaps the most important message to take away from this chapter is: *plan ahead.* Too often folks collect longitudinal data just for the sake of having it. Clearly, well-conceived and well-executed longitudinal studies can provide a wealth of information about change processes, growth functions, and predictors of both; however, poorly designed and haphazardly collected longitudinal data are theoretically dubious at best and empirically crappy at worst.

With careful and proper planning, a good longitudinal study would have a clear beginning, a circumscribed and efficient focus, and a clear ending. Many ongoing longitudinal studies have so many fundamental design and measurement problems that continued data collection on them is difficult to justify and mostly unwarranted. The resources that are being ill spent on haphazardly designed longitudinal studies, regardless of the theoretical merits of the project, should probably be redirected and reprioritized.

KEY TERMS AND CONCEPTS INTRODUCED IN THIS CHAPTER

Construct validity. An ongoing process of research using a particular construct. Showing that a construct has good characteristics in different contexts of samples, other constructs, age groups, and the like provides ongoing support for the utility of the construct. Any one study is a piece of the construct validity pie.

Content validity. Refers primarily to the internal relationships among the items/scores and the pool of potential items that can be selected to be indicators of a given construct. Content-valid indicators provide coverage of the domain of interest (a nonstatistical judgment) and, in a confirmatory factor analysis (CFA) framework, have strong loadings on the construct of interest and no indication of dual loadings onto other constructs or correlated residuals. That is, the indicators converge on the construct of interest and diverge or discriminate from the indicators of other constructs. The size of the loadings and the fit of the CFA model are used to inform content validity.

Context. The circumstances in which an event occurs, a setting. A context is the set of features that influence the performance or the outcome of a process. A context also defines the conditions that are relevant to an outcome. The word *context* stems from *contextus*, a putting together, and from *contexere*, to interweave, braid. Synonyms include *circumstances*, *times*, *conditions*, *situation*, *ambience*, *frame of reference*, *background*, *framework*, *relation*, and *connection*.

Criterion validity. This form of validity comes in two flavors, concurrent and predictive. In SEM models, with multiple constructs included in the model, all of the potential relationships among the constructs are potential criterion validity relationships. Traditional descriptions of criterion validity describe it as the association between a new measure and an established measure of the same general construct. This idea is narrow. A criterion is a statement of an expected association or mean difference that is supported by data. In this regard the expectation of a –.5 correlation between two constructs, for example, is a criterion validity finding. A strong statement of all the expected associations among the constructs is a broader and more rigorous definition of criterion validity.

Cross-sectional design. A design in which individuals from two or more age cohorts are assessed at only one time point.

Ecology. The ecology of human development involves examining the relationship between organisms and their environment. These ecologies can vary in a nested manner along a social dimension, a physical dimension, and a personal dimension.

Episodic time. Episodic time relates to identifying a key event or episode, such as puberty, graduation, or retirement. The assessment occasions are reorganized by centering each individual's data on the episode. Chronological age would be included in such models as a covariate or a moderator of the relationship.

Experiential time. The length of time during which individuals experience a state or influence. Chronological age would be included in such models as a covariate or moderator of the relationships.

Intensive designs. A person or group is studied on a large number of occasions. Collins (2006) defines these designs as involving at least 20 relatively closely spaced times of measurement; however, Nesselroade (e.g., Jones & Nesselroade, 1990) indicates that, when a single person is studied, there may be 100 or more times of measurement.

Panel designs. A cohort (e.g., people born in 1990) is studied at three or more times (e.g., 2000, 2001, and 2002). Collins (2006) defines these designs as involving eight or fewer times of measurement that are separated by at least 6 months.

Retrospective pretest-posttest design. Championed in the 1980's by Howard, this design was developed to circumvent the problems that traditional pre-post designs suffer, including the underappreciated response shift bias. The basic design asks respondents to make two ratings at the post-test time point: (1) rate current level on a given item and (2) rate, retrospectively, their level at the time of the pretest on the same item. This design is quite sensitive to change and differentially sensitive to different program characteristics that can impact perceived amount of change.

Sequential designs. Multiple cohorts of individuals are studied repeatedly, typically at three or more times. Attrition and retest control groups are often part of sequential designs.

Single-cohort designs. A group of individuals (members of the same birth cohort; e.g., people born in 1990) is studied repeatedly, that is, at multiple occasions (two or more times, e.g., 2000 and 2001).

Visual analog scaling. Using any form of visually enhanced response scale, usually in the form of a slider or number line in which the respondent marks a point on the visual scale to indicate his or her desired response.

RECOMMENDED READINGS

Mapping Theory with Model and Model with Theory

Jaccard, J. & Jacoby, J. (2020). Theory construction and model-building skills: A practical guide for social scientists (2nd ed.). New York: Guilford Press.

A very practice discussion of how to generate theory and modeling data to test theory.

Lerner, R. M., Schwartz, S. J., & Phelps, E. (2009). Problematics of time and timing in the longitudinal study of human development: Theoretical and methodological issues. *Human Development, 52*, 44–68.

A thoughtful discussion and alternative view on many of the issues that I present in this chapter.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605.

A detailed statistical overview of various models that can be fit to longitudinal data by the incomparable Jack McArdle—he will be missed.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Ram, N., & Grimm, K. J. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development, 31*, 303–316.

Both of these papers give outstanding rationales for matching theory, method, and model.

Wohlwill, J. F. (1973). *The study of behavioral development*. New York: Academic Press.

Wohlwill's book is a classic, although it is a dense "read." The ideas are timeless and worth spending the effort to work through.

The Retrospective Pretest-Posttest Design

Little, T. D., Chang, R., Gorrall, B. K., Waggenspack, L., Fukuda, E., Allen, P. J., & Noam, G. G. (2019). The retrospective pretest–posttest design redux: On its validity as an alternative to traditional pretest–posttest measurement. *International Journal of Behavioral Development*, 0165025419877973.

Solutions for Models to Address a Major Event

Rioux, C., Stickley, Z. L., & Little, T. D. (2021). Solutions for latent growth modeling following COVID-19-related discontinuities in change and disruptions in longitudinal data collection. *International Journal of Behavioral Development, 45*(5), 463–473.