OBSERVATIONAL MEASUREMENT *of* BEHAVIOR

PAUL YODER

FRANK SYMONS

SPRINGER PUBLISHING COMPANY

Observational Measurement of Behavior

Paul Yoder, PhD, is a professor in the Department of Special Education of Vanderbilt University. As the past director of the Observational and Quantitative Methods core for the Kennedy Center of Vanderbilt University, he has consulted for numerous single-subject and group-design researchers for over 20 years. He has conducted methodological studies and written methodological and measurement articles and chapters for both single-subject and group-design literature. Among these are three simulation studies relevant to sequential analysis. He currently teaches a recurring course on observational measurement for doctoral level students at Vanderbilt and is frequently asked to provide workshops on various aspects of observational measurement, including sequential analysis. He has been and continues to be an active user of observational measurement in his research on early communication in children with disabilities.

Frank Symons, PhD, is an associate professor in the Department of Educational Psychology at the University of Minnesota. He is the current Director of Observational Methods Lab at the University of Minnesota. He has published methodological articles on (a) visual analysis of observational data and functional analysis, (b) field testing observational technology versus paper and pencil methods, and (c) sequential and observational analysis of medication effects for severe behavior disorders. Additionally, Symons coedited a previous book on direct observational research methods and their application for research on individuals with intellectual and developmental disabilities and has written multiple chapters related to direct observational research methods. He teaches doctoral seminars in single-subject experimental research design, observational methods, and research methods in educational research in the Department of Educational Psychology at the University of Minnesota.

Observational Measurement of Behavior

PAUL YODER, PhD
FRANK SYMONS, PhD

SPRINGER PUBLISHING COMPANY
NEW YORK

Springer Publishing Company, LLC
11 West 42nd Street
New York, NY 10036
www.springerpub.com

Acquisitions Editor: Philip Laughlin
Project Manager: Molly Morrison
Cover Design: Mimi Flow
Composition: Newgen

E-book ISBN: 978–0–8261–3798–2

09 10 11 12/ 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Yoder, Paul J.
Observational measurement of behavior / Paul J. Yoder, Frank J. Symons.
p. ; cm.
Includes bibliographical references.
ISBN 978–0–8261–3797–5
1. Behavioral assessment. I. Symons, Frank James, 1967– II. Title.
BF176.5.Y63 2009
150.28'7—dc22 2009043107

Printed in the United States of America by Hamilton Printing.

The author and the publisher of this Work have made every effort to use sources believed to be reliable to provide information that is accurate and compatible with the standards generally accepted at the time of publication. Because medical science is continually advancing, our knowledge base continues to expand. Therefore, as new information becomes available, changes in procedures become necessary. We recommend that the reader always consult current research and specific institutional policies before performing any clinical procedure. The author and publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance on, the information contained in this book. The publisher has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Contents

This page intentionally left blank

Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 9

CHAPTER 10

Tables

CHAPTER 2

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CHAPTER 9

CHAPTER 10

Foreword

Undergraduate textbooks in the social sciences fly high and fast over the fields we researchers spend our days carefully cultivating. The typical text emphasizes self-reports and other-reports (interviews and questionnaires), perhaps some psychophysiological measures, and case studies and other qualitative approaches. Systematic observational methods are too often mentioned only in passing. This book, which would be of value to serious undergraduates, graduate students, and practicing researchers alike—goes far to right the balance.

Observational measurement is presented as an important (and sometimes neglected) yardstick for those of us who want to study behavior in a quantifiable, replicable, and scientific way. This book emphasizes careful attention to measurement from the first chapter, which nicely locates systematic observation within its conceptual measurement domain, to the last, which returns to a thorough discussion of the foundational concept of validation. Throughout, the authors are not only concerned with the techniques and mechanics of observational methods. They take pains to explain conceptual underpinnings and to place techniques within the larger research enterprise.

To illustrate their points the authors use many examples from their own and others' research. In most other texts, discussions of research methods are specialized, emphasizing either group-design or single-subject studies, but rarely both. Here, reflecting the authors' long involvement in research on children with disabilities, their examples come from both camps. All readers will find the many examples illuminating, but readers involved with single-subject studies will enjoy the attention given to designs they recognize, here firmly placed in the context of general measurement concerns.

In sum, this book will be useful to students and researchers at all levels who want to deepen their understanding of concepts and techniques in the observational measurement of behavior. Researchers with diverse disciplines and interests in the social sciences have used observational measurement—from ethologists and animal behaviorists to developmental, social, and educational psychologists. But this book may be especially appealing to those concerned with typical and atypical development of infants and children, whether their research goals are primarily theoretical or more practical and clinical.

Yoder and Symons bring decades of work to bear, and it shows. The topics one might expect are here: developing coding schemes and designing coding manuals, determining sampling methods and metrics for observational variables, training observers and assessing their agreement, and performing sequential analysis on observational data. Yet the whole is presented with broad scholarship and conceptual depth. A unique strength of this book is its attention to conceptual underpinning and its strong emphasis on fundamental psychometric concerns, from measurement theory to validity. Yoder and Symons have explicated the technical issues of observational measurement well and have placed the whole enterprise in the context of doing science, where it certainly belongs. If this book has the influence it should, authors of undergraduate texts will surely take notice of new activity in the field.

Roger Bakeman, PhD
Professor Emeritus
Georgia State University
Atlanta, Georgia

Preface

Researchers use many approaches to collect, summarize, and communicate their observations of behavior. We could have tried to address all of these at a superficial level in a moderate-sized book or we could have addressed all of them comprehensively in a thick, expensive book. Instead, we address a subset of these approaches at a comprehensive level to cover a set of approaches we consider important and frequently misused while keeping the book a reasonable length for a semester-long course. Open-ended approaches to observational measurement will not be covered in this book. They are covered well in other sources (e.g., see Denzin & Lincoln, 2005). Instead, we focus on a type of observational measurement that is particularly well suited to addressing highly specified, falsifiable research questions.

THE SCOPE OF THIS BOOK

The set of measurement principles we address are particularly well suited to falsifying hypotheses using a quantitative approach to the scientific method. Such an approach requires that we define in detail our methods of observing prior to beginning the study so that the results of our studies are replicable. We call this approach to observational measurement *systematic observation* (Suen & Ary, 1989). In systematic observation, we decide the following *before* observing:

(a) the key behaviors we are going to mark (i.e., a coding manual; see chapter 3),

(b) the context of measurement (i.e., the procedure and setting; see chapter 1 and chapter 2),

(c) whether the session will be observed live or from a recorded medium (i.e., session recording method; see chapter 4),

(d) the method of sampling the behavior from the observation session (i.e., behavior sampling; see chapter 4),

(e) the method by which the observer indicates that an instance of the behavior has occurred (i.e., behavior recording method; see chapter 4), and

(f) the metric used to represent the levels of the behavior (e.g., number vs. proportion vs. duration, etc.; see chapters 5–7).

Each of the terms and operations will be defined in detail in this book. Importantly, decision-making guidelines will be provided to help the reader select among the most common options. When possible, empirical evidence will be used as the basis for the guidelines. When data is not available, logical arguments will be given as rationale for the guidelines. When known, the ramifications of each decision will be provided.

A particular type of metric, indices of sequential associations between two behaviors observed in the same session, will be treated in greater detail than other metrics because of its complexity, frequent misuse, and potential value for many who use observational measurement. This topic is often referred to as the *sequential analysis of behavior* (Bakeman & Gottman, 1997). Simulation studies have provided empirical guidance for the many decisions the investigator must make when using indices of sequential association. These decision-making guidelines and their empirical support are presented in this book (chapters 6 and 7).

Interobserver agreement and reliability are among the most discussed and disagreed upon topics in the field of observational measurement of behavior. Chapters 8 and 9 attempt to address these complex topics in an honest, straightforward manner and make sometimes bold recommendations.

Chapter 10 addresses the very important topic of validity of observational variables. This topic may appear unnecessary from one measurement perspective because accuracy of observation may be all that perspective cares about. However, careful consideration to what we decide to count as an incidence of the behavior of interest is clearly necessary for all measurement perspectives. Additionally, other measurement traditions require much attention to this topic. The primary validation methods are covered.

ABOUT THE WEBSITE

The website that accompanies this text (available at www.springerpub.com/yoder/supplements) includes many experiential exercises that will help readers understand and apply the techniques discussed in this book. Discussions in various chapters refer to the electronic files.

- In chapter 2, in the discussion of a sample generalizability study, sample data are provided in Excel files so that readers can run the provided SPSS syntax and compute the *g* coefficient and the person variance on the sample data. The g calculators make transparent how results from analyses can be used in planning the number of observation sessions or raters required in an observational study when measuring generalized characteristics.
- In chapter 4, electronic files will help readers attempt to code a sample observation session. Files include a demonstration version of ProcoderDV, a software program that assists in coding, and a media file of a session to code.
- In chapter 5, raw data and a statistical syntax file for a fictitious proportion simulation study are provided electronically so readers can confirm the results of the simulation as presented in the text. An Excel file containing the arcsine transformation formula is also provided.
- In chapter 7, the website includes a timed-event data file that will be used for time-window analysis. Additionally, a demonstration version of MOOSES, a software program that assists in sequential analysis, is provided.
- In chapter 9, readers who would like to understand the relationship between chance agreement and kappa are provided formulae in an Excel spreadsheet. Data are also provided for an exercise demonstrating the effect between-person variance has on the intraclass correlation coefficient.

REFERENCES

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Denzin, N. K., & Lincoln, Y. S. (2005). *The handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Erlbaum.

This page intentionally left blank

Acknowledgments

The driving force behind directly observing behavior is discovery to make a difference. I am thankful for the opportunities families provide me and I am grateful for the inspiration from many mentors and colleagues along the way, including Travis Thompson, Paul Yoder, Jon Tapp, Frank Epling, David Pierce, Steve Holborn, Jim Bodfish, Don Bailey, Mark Wolery, Joe Wehby, Bill MacLean, Steve Warren, Patrick Rivard, Scott McConnell, Joe Reichle, LeAnne Johnson, Jennifer McComas, and John Hoch, all of whom excel at modeling endless enthusiasm, creative curiosity, and insightful intellect. I would like to express my good fortune and extreme gratitude to Paul for extending the opportunity to work with him on this project. I am also grateful for the continued support of my program (Special Education) and department (Educational Psychology) at the University of Minnesota for the Observational Methods Lab. Finally, I would be remiss without acknowledging the patience and passion of my family—Stacy, Stewart, and Elisabeth along with Festus, Trudie, and Rose—for all of their behavior that I get to constantly observe.

—*Frank Symons*

Observing behavior allows us to discover with unusual clarity of thinking and communication that which we consider important. The children and families who participate in our research trust that we will use their time wisely to address shared questions. I am grateful for their patience with our imperfect efforts to deserve that trust. I am grateful to Roger Bakeman for excellent writing and continual contributions to modern observational measurement knowledge. I am grateful to Vanderbilt University and the University of Canterbury which provided the sabbatical support during the writing of this book. Mostly, I am grateful to my wife, Deborah, who provides me with loving support of my goals.

—*Paul Yoder*

This page intentionally left blank

1 Introduction and Measurement Contexts

OVERVIEW

The purpose of this chapter is to review a number of underlying issues that, although not always explicitly articulated in a given research report, are critical to understanding the logic behind and the strategies used in the different research approaches to quantify behavior using systematic direct observation. In this chapter, we promote hypothesis-driven research as a general approach to improve the scientific rigor and the interpretability of any given study. We then move to a discussion on measurement issues that concern distinctions between behavior as context dependent and behavior as a sign of a generalized characteristic. This distinction is then used as the basis for considering foundational measurement issues related to operationalism and operational definitions, and ultimately, the interpretative framework for a given study and its findings. Wherever appropriate, we draw distinctions between philosophical and design traditions to help readers understand the different ways investigators may think about what they are measuring and why. Three key concepts (influential variables, "structuredness," and ecological validity) are introduced, defined, and discussed in relation to direct observation research methodology. Finally, two measurement decisions (whether to measure in a structured procedure and whether to derive a variable score by averaging sessions scores across many sessions) are described and a rationale is provided for each.

SYSTEMATIC OBSERVATION

The set of measurement principles we address is particularly well suited to falsifying hypotheses using a quantitative approach to the scientific method. Such an approach requires that we define in detail our methods of observing prior to beginning the study so that the results of our studies are replicable. We call this approach to observational measurement *systematic observation* (Suen & Ary, 1989).

Systematic observation is an alternative to *self-report* (i.e., asking the participants what they do) or *other report* (i.e., behavioral ratings or reports completed by asking others who draw from cumulative experience with the participant's behavior) methods of measurement. There are situations in which observational measurement may be more scientifically valid than self-report and other report. First, the observations allow detailed descriptions of behavior and its social and nonsocial context. For example, we may be interested in the antecedents or consequences of a particular type of social behavior. Some social exchanges may occur without conscious knowledge by the participant or others who know the participant. Because the exchanges in which the antecedent-behavior or behavior-consequence sequences occur may be fast moving, asking participants and others to "note and report" on such exchanges may be unsuccessful in capturing the phenomenon of interest.

Second, observations are often more valid than self-report when the participant is preverbal or limited in his or her verbal or cognitive ability to be aware of or report on the phenomenon. For example, nonverbal participants cannot report on their interest in communicating for social reasons, but we can directly observe the frequency with which a participant uses communication that is presumed to have only socially rewarding consequences (e.g., declaratives).

Third, other reports of participant behavior (e.g., parental checklists for the child's behavior) may reflect the characteristics of the reporter (e.g., socioeconomic status) as well as characteristics of the participant (Najman et al., 2000; Yoder, Warren, & Biggar, 1997). The influence of reporter characteristics may explain, in part, why it is commonly found that different reporters on the same child often disagree in their responses (Smith, 2000).

COUNT CODING SYSTEMS

This book focuses on a class of systematic observational measurement called *count coding systems*. Count coding systems are designed to lead the observer to count the number of instances and/or duration of instances of the key behaviors. All variable metrics (e.g., rates, proportions, indices of sequential associations, latencies) are derivatives of number or duration of key behaviors or time between key behaviors. In systematic observational measurement, the primary alternatives to count coding systems are checklists and rating scales. The latter are covered in detail in other sources (Cairns, 1979; Primavera, Allison, & Alfonso, 1997).

Rating scales involve the observer rating on a Likert-like scale, his or her global judgment about the quality or quantity of a particular class of behaviors. For example, after observing a parent and child interacting for 20 min, the observer might rate the parent on "parental responsivity" by indicating where, on a 5- or 7-point scale, the parent fell. The behavioral anchors of "almost all of the time" and "almost never" might be assigned to the end points of the scale for each item. In contrast, a checklist requires that the observer indicate the presence or absence of a particular behavior during the key observation period. In this example, the observer might indicate whether the parent displayed any instances of "responsivity" during the session.

Count coding systems generally provide a larger range of potential scores and more steps between values than do rating scales or checklists. Such measurement properties provide a potentially more sensitive measure of the key variables than do rating scales and checklists. Additionally, count coding systems do not require that the observer "calibrate" his or her concept of what is meant by each of the values on the Likert-like scale. This is particularly useful when it is unclear what the optimal levels of the object of measurement look like or when observers' concepts of "optimal" differ. However, it must be said that count coding systems tend to require more time to implement than rating scales and checklists. Therefore, the gain in precision comes with a cost in resources (e.g., personnel time, training, etc.). Putting it all together, we will refer to the approach covered in this book as "systematic observational count measurement."

IMPORTANCE OF FALSIFIABLE RESEARCH QUESTIONS OR HYPOTHESES

To implement well the type of observational measurement that we discuss, it is important that the investigator formulates, *prior to collecting data*, a very specific and falsifiable statement of the prediction. The syntax used, whether it is a statement or a question, is not important. It is important that the statement specifies (a) the dependent and independent variables, (b) the investigator's expectations of an association or a difference, and (c) the investigator's expectations regarding direction of the association or difference (e.g., a positive association or that the experimental group [or phase] is greater than the contrast).

The more specific the research question or hypothesis, the more guidance it will provide for designing the measurement system used to assess the independent and/or dependent variables. Creating such falsifiable research questions is important because findings that confirm very specific predictions are more likely to replicate than findings that confirm vaguely stated predictions. This is not magic. When extant data and theory that support such specificity is sufficiently developed to generate confirmation, it suggests a field that is relatively mature. Falsifiable predictions are much easier to disconfirm than they are to confirm. This is a simplification of the positivist philosophy of science. This book assumes that readers understand and are able to formulate falsifiable predictions in the form of hypotheses or research questions.

BEHAVIOR AS "BEHAVIOR" VERSUS BEHAVIOR AS A SIGN OR INDICANT OF A CONSTRUCT

Investigators may differ either implicitly or explicitly on whether or not they believe what they are measuring represents a tendency to behave that continues to exist outside the measurement context and measurement period. If the investigator is interested only in what occurs during the observation session, he or she is interested in the target behavior for its own sake. For example, a participant raising his hand before speaking in a class is a behavior that is important for his own sake. Therefore, it may not be measured as a sign of some larger concept or psychological characteristic (e.g., compliance, self-regulation). It is therefore clear that measuring behavior for its own sake can be important because it may help to solve problems that may influence more enduring and generalized behavior change.

In contrast, other investigators are interested in measuring the number or duration of behaviors as signs of psychological characteristics called "constructs" (Cronbach & Meehl, 1955). Investigators taking this perspective readily accept the notion that the "real" object of measurement is something that cannot be seen directly but must be inferred from observables. The general public accepts this approach in other domains. For example, the change in mercury level in a mercury-based thermometer is not the same entity known as "temperature." The rising or falling of mercury is only a sign of temperature change. Similarly, behaviors may be seen as a reflection of the constructs that generate them. Constructs have been divided into states (i.e., temporary behavior levels that are highly unstable over time, and context) versus skills/characteristics (i.e., more stable behavior levels over short periods of time and contexts that are designed to assess the same construct). In the past, the latter used to be called "traits." However, the term "characteristic" will be used in this book instead of "trait" because the latter has the connotation that it is genetically caused and relatively immalleable. These are unnecessary assumptions when thinking about measuring stable characteristics. The distinction between "behavior as behavior" and "behavior as an indicant of a generalized characteristic" is, in part, related to operationalism.

Operationalism is a historical movement in psychology in which observable behaviors are used to define how constructs (i.e., concepts created to explain a phenomenon) are measured. Although the history of operationalism is beyond the scope of this book, it has a long history in the behavioral and social sciences with many different proponents and opponents (Rogers, 1989). In the following section, we present two interpretations of operationalism (semantic vs. methodological) that are relevant in our discussion concerning behavior as behavior versus behavior as an indicant of a construct.

TWO INTERPRETATIONS OF OPERATIONALISM

Operationalism emerged as an important idea in psychology in the 1930s. This approach to defining concepts attempted to reduce the subjectivism prevalent in the psychology of the time, and its impact continues today. Contemporary accounts of operationalism still value the notion of using observable behaviors to define how concepts will be measured. The disagreements regarding operationalism are over whether one considers the operations by which one studies a phenomenon as *synonymous* with the concept of interest. We will present two interpretations of the operationalism movement in psychology: the semantic and the methodological interpretations (Feest, 2005).

The semantic interpretation asserts that the meaning of a concept can be *exhaustively* defined by stating particular observable manifestations of a concept. Viewed from a natural science perspective, psychological measurement is analogous to physical measurement, with a low level of inference regarding psychological characteristics that may cause the behaviors (Johnston & Pennypacker, 1993). The semantic interpretation has received four general criticisms (Rogers, 1989). First, it may be used to justify an unproductive belief that it is scientifically useful to measure without defining the object of our measurement (e.g., intelligence is what intelligence tests measure). Second, there is a need to make statements about people even though it is impossible to exhaustively rephrase each concept in terms of observables. Critics making this objection claim that concepts can be, at best, confirmed by observables, not exhaustively defined by them. Third, the semantic interpretation of operationalism can result in an uncritical and uninformed introduction of an unnecessarily large number of new concepts, each of which is completely dependent on the context in which it is measured. Approaching all observations in this way would fly in the face of the widely accepted notion that the same concept can apply in multiple contexts. It is again worth noting that there are divergent views on the problems of operationalism including the position that operationalism was never meant to be interpreted in this manner (Feest, 2005).

In contrast, from a methodological operationalism perspective, operational definitions of concepts are partial and temporary specifications used to study the real concept of importance. For example, in classical stimulus–response learning theory, Tolman's operationalization of "hunger" was "time from last feeding." He never thought that time from last feeding was synonymous with hunger. Nor did he deny the existence of the subjective feeling of hunger. He simply thought that the subjective feeling of hunger was a poor measure of the concept of "hunger" because it is easily confused with other needs (e.g., boredom, need to dull psychological pain via food, etc.). In this sense, time from last feeding was a reflection of the real concept of interest: hunger. Those professionals who adopt this view of operationalism will be more open to the concept of behavior as a reflection of a construct, but open themselves up to criticism that they are not necessarily measuring what they think they are measuring. The latter issues will be covered throughout the book with a culminating summary in chapter 10.

DISTINCTION BETWEEN CONTEXT-DEPENDENT BEHAVIOR AND GENERALIZED TENDENCIES TO BEHAVE

Because (a) behaviors that are considered important for their own sake and (b) behaviors that are thought to be reflections of states are *not* expected to inform us of what occurs outside of the measurement context, we lump these together in this book. We will call these *context-dependent* behaviors. Context-dependent behaviors do not require the same degree of complex consideration as measuring behaviors that are considered reflections of generalized constructs.

In contrast, measured behaviors that are thought to reflect generalized constructs are thought to represent stable (in the group design sense of the word) skills or characteristics. We refer to these as *generalized characteristics*. Generalized tendencies or characteristics allow individual differences among people to exist across multiple contexts and over time (Cronbach & Meehl, 1955). If we are measuring a generalized characteristic, people with different scores on an observational variable will generally hold their rankings within the study sample if measured a brief time later or in a different measurement context that is also designed to evoke the key behavior. When studied from a group design perspective, individual differences in level or change on the behavior are stable over time and context (i.e., high positive correlation among the rankings of the variable measured in different contexts or at different times). When referring to stability over contexts, we mean stable across contexts that realistically evoke the key behaviors, and not just any possible context. We would not expect stability in measures of aggression from the playground to the movie theater. The movie theater probably inhibits signs of aggression, while the playground may elicit them. Finally, we recognize that the term “stability” is used to mean something very different in single-subject research (i.e., a flat trend or unchanging variability). It is the group design meaning of the term “stability” that we intend to convey here.

Measuring generalized characteristics by observing key behaviors requires an inference that what we are observing reflects an ability or skill we cannot directly observe. Therefore, measuring generalized characteristics requires more attention to how we define our observational measure, select a measurement strategy, and interpret studies purporting to examine these than measuring context-dependent behavior.

It is important to note that the same behavior or set of behaviors can be measured as (a) a context-dependent behavior in one study and (b) a generalized characteristic in another study. For example, "sitting" may be measured as a context-dependent behavior when an intervention study shows that prompting and reinforcing a child for staying on a pillow helps the child do so during times the pillow, prompts, and rewards are present. We conclude this is treating sitting as a context-dependent behavior because the pillow is never withdrawn to test for generality when the pillow is not present. When measuring sitting as a behavior in the middle of the generality continuum, the pillow is faded (i.e., systematically removed) and sitting is measured in the setting where the pillow used to be present (e.g., same classroom, same activity, same peers). Finally, sitting could be measured in a way that places it at the far end of the generality continuum by fading the pillow and by measuring sitting in another classroom and in another group activity.

Behavior change that is readily reversible is considered more context dependent than behavior change that takes a very long time to decay after support conditions are no longer present. For example, if removing the pillow in the above example results in the child no longer sitting (i.e., his behavior returns to pretreatment levels of sitting), then such a reversal suggests that the sitting behavior was context dependent. On the other hand, assume you know a "good writer" who writes daily. If he is prevented from writing for a month, upon returning to writing, he will still be able to write as quickly and efficiently as if he were writing daily. However, if he is prevented from writing for decades, upon returning to writing, it will take him longer to write and his sentence structure will not be as efficient as it was when he was writing daily. Such a pattern describes a generalized characteristic that is not readily reversible.

RATIONALE FOR IDENTIFYING HOW WE ARE CONCEPTUALIZING OUR OBJECT OF MEASUREMENT

The distinction between context-dependent behaviors and generalized characteristics is not trivial or just an academic distinction. Being aware and consistent in how we conceptualize our object of measurement helps us make measurement-related choices that are consistent with our conceptualization. In this chapter, we will introduce the guideline that when we are intending to measure a generalized characteristic, we should (a) measure it in a structured procedure and/or (b) measure it in many sessions and average the session scores together to derive our dependent-variable score at the participant level. We do so in this first chapter because understanding these concepts helps to understand why direct observational measurement is not everyone's measurement method of choice. Before discussing the rationale for these two recommendations, we must introduce four other concepts: influential variables of a measurement context, the notion of "structuredness" of a measurement context, the notion of "ecological validity" of a measurement context, and the tension between structuredness and ecological validity.

Influential Variables of a Measurement Context

Among the many elements of a measurement context that may affect the occurrence of key behaviors are (a) the location or setting, (b) the activities, (c) the materials, (d) the instructions to the participants (if any), and (e) the people involved (e.g., administrator, peers, etc.). It is important for the investigator to consciously decide prior to collecting data which variables within the measurement context should be kept constant across sessions or participants and which should be left to vary across sessions. The variables that are likely to affect the occurrence of the key behaviors are called *influential variables*. Variables that are not likely to affect the occurrence of the key behaviors are called *noninfluential variables*. Only the former class of variables needs to be considered when selecting or designing measurement contexts.

Structuredness

The degree to which we keep influential variables constant across sessions or participants is the degree of structure our measurement context possesses. One may wish to control influential variables in the measurement context when measuring generalized characteristics. Individual differences or changes in scores over time within a person are assumed to reflect something about the participants, not the differences in influential variables in the measurement contexts.

One may also want to structure the measurement context because having many instances of key behaviors in at least some participants across some phases of the design is generally more desirable than observing only a few instances in all participants and design phases for several reasons. First, variability among participants or sessions (i.e., at least some participants or sessions yield "high" scores) is necessary for variables to be stable over time and context. Such stability is the hallmark of a generalized characteristic. Second, in single-subject designs addressing a dependent variable that occurs in the treatment sessions, one must use a procedure that evokes the key behaviors to demonstrate a change during the treatment phase. This type of procedure often "structures" the session in some way. Third, in AB design variants, such as a multiple-baseline design, the more immediate the change is after the onset of the treatment phase, the more confident the judges tend to be in inferring a functional relation between independent and dependent variables (Kazdin, 1981; Lieberman, Yoder, Reichow, & Wolery, in press).

Ecological Validity

The extent to which measurement contexts resemble or take place in naturally occurring (unmanipulated) and frequently experienced contexts has been called "ecological validity" (Brooks & Baumeister, 1977). There is a legitimate societal need to know the extent to which participants use key behaviors in uncontrolled conditions that the participant frequently experiences (Brooks & Baumeister, 1977). When selecting a measurement context for a generalized characteristic, there is a tension between selecting a structured context and selecting an ecologically valid one. To understand this tension, it is necessary to introduce the concept of representativeness.

Representativeness

One definition of "representative" is "typical" (*Shorter Oxford English Dictionary*, 2002). The Oxford dictionary definition of "typical" that most closely matches the intended meaning for the present context is "usual" or "familiar through frequent or regular repetition." Neither definition is scientifically useful because it is not clear how one would test the typicality of a score or the familiarity of a context. A more scientifically useful definition of representativeness is stability (in the group design sense of the word) across contexts that evoke the behaviors that are signs of the generalized characteristic. Research questions testing stability over contexts are more falsifiable than research questions testing familiarity or frequency of exposure. A research question involving stability over contexts might be, “Do the rankings of participants’ word use in communication samples with an examiner have a high and positive (e.g., above .70) correlation with the participants’ word use in communication samples with their mothers?” It is not clear how one would phrase an analogous falsifiable research question that tested the “typicality” of a score, either in a group design or in a single-subject design.

Testing representativeness (i.e., stable across contexts) is more easily falsifiable using a group design than using a single-subject design. A similar question phrased in a single-subject design might be, “Are the number of words used with the examiner within the range of the number of words used with the mother?” In this single-subject design question, we are attempting to confirm a null hypothesis. Finding evidence that might support a “no difference” hypothesis through chance is easier than finding noteworthy differences or associations. Therefore, stability across context as an operational definition of representativeness is more scientifically useful in a group design measurement context than in a single-subject context.

Tension Between Structuredness and Ecological Validity

Because ecologically valid contexts are often unstructured, it is extremely important that investigators avoid the reasoning that naturally occurring measurement contexts increase the probability that the observational variable scores from such contexts are more typical or representative than the scores from structured measurement contexts (Schmuckler, 2001). It may not be clear why unstructured measurement contexts often produce *less* stable scores across context than do structured ones to all readers, but they often do. When we reasonably expect stability (in the group design sense of the word) over contexts, we do so because we expect the individual differences on the observational variable from the two contexts to primarily reflect individual differences on the same generalized characteristic. By definition, unstructured measurement contexts produce variability among sessions or participants in part because they do not control many of the variables that influence the scores. In contrast, because structured sessions do control influential variables, scores from these procedures are less likely to be influenced by variables other than what we want to measure.

Figure 1.1 illustrates the covariation between (a) demonstrations of generality and reversibility, structuredness of the measurement

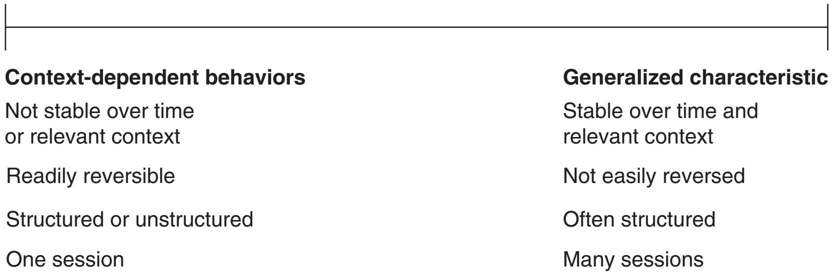

Context-dependent behaviors
Not stable over time or relevant context
Readily reversible
Structured or unstructured
One session
Generalized characteristic
Stable over time and relevant context
Not easily reversed
Often structured
Many sessions

Figure 1.1 Continuum of degree to which a measured entity is context and time dependent.

context, and number of sessions across which session scores are aggregated for each participant's score and (b) the extent to which the behavior is considered a context-dependent behavior versus a generalized characteristic.

RECOMMENDATIONS FOR MEASURING GENERALIZED CHARACTERISTICS FROM OBSERVATIONS

When reading published articles, one way to identify the extent to which an object of measurement is being studied as a generalized characteristic is to note the extent to which the measurement procedure is structured. Unless scores are averaged or summed across several sessions (a rare event), behaviors measured in unstructured procedures are more reasonably thought of as measures of potentially context-bound behaviors than measures of generalized characteristics. Similarly, if the investigator wishes to measure a generalized characteristic and there is no cultural value against measuring the behaviors reflecting the characteristic in a structured context, then selecting a structured measurement context for the yet-to-be-conducted study is more efficient (i.e., takes fewer sessions) than measuring the behavior in an unstructured context. However, if one uses structured procedures, one must restrict one's generalization of what one is measuring to similar contexts as those used.

If values or the rationale for the study requires that a generalized characteristic be measured in the natural environment, then it is extremely likely that the averaging or the summing across many sessions will result in a more stable estimate of each participant's score than would observing a single unstructured observation. Conceptually, this practice is understandable from the perspective of domain sampling. It is useful to conceptualize the entire set of measurement contexts that evoke the key behaviors as the universe, and the mean of the observational variable score from all of these contexts as the most representative score. Since we cannot exhaustively sample every evocative measurement context, we must sample (i.e., observe in) many of the contexts from this universe. The more representative this sample is of the universe, the more probable that the mean sample score will approximate the mean universe score. The degree to which the sample and the universe mean are similar is influenced by the selection process and the number of observed contexts. Practically speaking, we cannot randomly select our measurement contexts. Instead, investigators who need to derive a single score that is "representative" of all potentially valid contexts (an extremely demanding challenge) usually systematically sample different measurement contexts in an attempt to include a variety of measurement contexts from the universe of evocative measurement context for our generalized characteristic of interest. Additionally, all things being equal, the larger the sample, the closer our sample mean will be to the universe mean. It is often more realistic to restrict the *types* of measurement contexts across which we expect our measure of generalized characteristics to be stable. In this context, we can average across many unstructured sessions all of which are a certain type (e.g., circle time in a preschool class) and restrict our generalizations to similar contexts.

Practically, generalized characteristics vary in the extent to which their scores vary among contexts. Chapter 2 will describe a method, decision studies, for empirically determining how many sessions are needed to derive a stable estimate of the generalized characteristics.

POTENTIAL DISADVANTAGES OF SYSTEMATIC OBSERVATIONAL COUNT MEASUREMENT

Now that we have covered the reason why single observations are often inadequate to reliably measure generalized characteristics (i.e., the single observation may produce a variable score that is a poor estimate of the universe mean score), it should be clear why other reports (e.g., parent reports) have legitimate appeal as alternatives to systematic observation. Specifically, other reports about the participant's behavioral tendencies potentially draw on a wide range of experiences with the participant. If the reporter is able to synthesize across his or her experience with the participant while keeping his or her biases from influencing his or her report, then other reports have much potential for producing valid estimates of generalized characteristics. Because sampling many observational sessions and averaging scores to produce a single estimate is expensive, and thus rare, many investigators prefer other reports over systematic observation when measuring generalized characteristics.

On the other hand, if parents or other reporters are not able to keep their biases from influencing the report of the participant's behavior, then using the average of many observation sessions may produce a more valid estimate of the generalized characteristic than other reports or self-reports. Additionally, systematic observation will almost always be a more valid way to report on context-dependent behaviors than is other report of the participant's behavior. If one is not aware of the distinction between context-dependent behaviors and generalized characteristics, one might mistakenly overgeneralize and believe that systematic observation is *always* more valid than other report. Ultimately, the relative validity of other report versus systematic observational measures of generalized characteristics is an empirical question. Additionally, these empirical comparisons of relative validity will need to occur for each combination of population and generalized characteristics. This is arguably impractical. Therefore, for the foreseeable future, investigator's preferences will surely affect the selection of systematic observation versus other report when measuring generalized characteristics. Others have written about the advantages and disadvantages of systematic observation versus other report methods of measuring generalized characteristics (Jacobson, 1985). One approach to this ongoing debate is to measure a generalized characteristic using multiple methods (e.g., both other report and observational measurement) and aggregate them if they are correlated or look for convergence of findings (Cook & Campbell, 1979).

RECOMMENDATIONS

In this chapter, we defined what we mean by systematic direct observation and we discussed the distinction between measuring a context-dependent behavior versus a generalized characteristic. This distinction is very important for proper framing and interpretation of a study and for many measurement decisions. Two of these measurement decisions are the degree of structure and the number of the measurement contexts one needs to average across to derive a participant's variable score. When generalized characteristics are the object of measurement, measurement contexts should be structured and/or scores from many observational sessions need to be averaged to derive the participant's variable score.

REFERENCES

Brooks, P., & Baumeister, A. (1977). A plea for consideration of ecological validity in the experimental psychology of mental retardation. *American Journal of Mental Deficiency, 81*, 406–416.

Cairns, R. (1979). *Analysis of social interactions: Methods, issues, and illustrations*. Hillsdale, NJ: Erlbaum.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about? *Journal of History of the Behavioral Sciences, 41*, 131–149.

Jacobson, N. S. (1985). Uses and abuses of observational measures. *Behavioral Assessment, 7*, 323–330.

Johnston, J. M., & Pennypacker, H. S. (1993). *Readings for strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kazdin, A. (1981). Drawing valid inferences from case studies. *Journal of Consulting & Clinical Psychology, 49*, 183–192.

Lieberman, R., Yoder, P., Reichow, B., & Wolery, M. (in press). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly.*

Najman, J., Williams, G., Nikles, J., Spence, S., Bor, W., O'Callaghan, M., et al. (2000). Mothers' mental illness and child behavior problems: Cause–effect association or observation bias. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 592–602.

Primavera, L., Allison, D. B., & Alfonso, V. C. (1997). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–90). Mahwah, NJ: Erlbaum.

Rogers, T. B. (1989). Operationism in psychology: A discussion of contextual antecedents and an historical interpretation of its longevity. *Journal of the History of the Behavioral Sciences, 25*, 139–153.

Schmuckler, M. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*, 419–436.

Shorter Oxford English dictionary (5th ed., Vol. 2). (2002). Oxford, UK: Oxford University Press.

Smith, S. (2000). Making sense of multiple informants of child and adolescent psychopathology. *Journal of Psychoeducational Assessment, 25*, 1139–1149.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Yoder, P., Warren, S., & Biggar, H. (1997). Stability of maternal reports of lexical comprehension in very young children with developmental delays. *American Journal of Speech-Language Pathology, 6*, 59–64.

2 Improving Measurement of Generalized Characteristics Through Direct Observation and Generalizability Theory

OVERVIEW

This chapter quantitatively defines two critical measurement concepts—*true score* and *measurement error*—through the generalizability theory and provides a rationale for their consideration and use in research relying on direct observation methods. We then present decision studies as a method to plan future studies that intend to measure generalized characteristics through observation. We complete the chapter with a discussion about issues specific to single-subject design in relation to the study of individual differences and generalized characteristics. Before beginning our discussion on the application of the generalizability theory for direct observation count data, it is useful to introduce two terms (idemnotic and vaganotic) that identify different sets of key assumptions underlying two different measurement concepts and their respective approaches.

TWO CONCEPTS OF MEASUREMENT

An *idemnotic* concept of measurement requires that the phenomenon of interest is (a) measured along a continuum, (b) has an absolute and often preexisting possible minimum, and (c) uses units or steps that are established independent of variability in the phenomenon being measured

(Johnston & Pennypacker, 1993). Most scales in the physical sciences use an idemnotic scale (e.g., weight). The absolute minimum weight on earth is zero. The steps in the scale are kilograms or pounds. Of critical importance is the point that population variability among participants is irrelevant to the meaning of idemnotic scales. In this sense, it is a good match with single-subject research design. Therefore, an idemnotic conceptualization of measurement underlies the measurement of most dependent variables in single-subject experimental research. However, it should be noted that much fidelity of treatment or fidelity of procedure measurement in group or single-subject research also uses an idemnotic concept of measurement. In the latter case, we frequently want to see nearly uniform (e.g., 100%) fidelity, not variability among participants or sessions.

The *vaganotic* concept of measurement is the dominant implicit or explicit concept in studies by investigators who are most interested in individual differences. It is the measurement approach used to conceptualize the assessment of the predictors and the dependent variables in most group designs. The meaning of high and low is relative to a group (Johnston & Pennypacker, 1993). The groups can either be the sample of participants in the study or another reference group (e.g., a standardization sample in a norm-referenced test). However, because most observational measures are not norm-referenced tests, vaganotic observational measurement almost always means that an individual's score is interpreted in reference to other participants in the study sample.

Observational variables such as number or duration of behavior can be conceptualized using either idemnotic or vaganotic approaches. Both conceptualizations of observational measurement use physical aspects of behavior to assess their objects of measurement. However, we have asserted in chapter 1 that when we are measuring generalized characteristics, we assume that we are measuring levels of a characteristic that occurs in other contexts, too. That is, we implicitly or explicitly assume that individual differences on our measure of a generalized characteristic are stable in contexts that occur outside of our measurement context (i.e., it is "representative"). Here, we are using "stable" in the group design sense of the term. From this perspective, the concept of representativeness is a group research design and a vaganotic measurement concept. With this in mind, readers are asked to take on a group design perspective when reading this chapter because the purpose of group design is to explain variance among participants (either individuals or aspects of distributions of individuals such as means). Similarly, the purpose of this chapter is to provide guidance on improving our ability to directly measure generalized characteristics. Some of the issues relevant for single-subject research design and generalizability theory are discussed at the end of this chapter.

GENERALIZABILITY THEORY AS A MEASUREMENT THEORY FOR VAGANOTIC MEASURES

Classical measurement theory was created to communicate systematically about a vaganotic concept of measurement (Crocker & Algina, 1986). This theory states that an observed score = true score + measurement error. Similarly, reliability = true score variance/observed score variance. These abstract concepts take on a concrete meaning when interpreted through the perspective of generalizability theory (Cronbach, 1972; Shavelson & Webb, 1991).

To understand generalizability theory conceptually, it helps to imagine that we have observed the key behavior of interest in four different contexts each for 10 participants. Classical measurement theory calls the observational variable score (e.g., count of a key behavior) from any one context and participant, an "observed score." Theoretically, the "true score" is the mean of the observed scores from all valid measurement contexts for the generalized characteristic of interest. In generalizability theory, we *estimate* a participant's true score by averaging all the *available* observed scores for that participant. It is assumed that the observed scores that are averaged all come from measurement contexts that are designed to elicit the key behavior of interest. To help understand how accurate the central tendency of a distribution of observed scores is in estimating the true score, it is useful to consider an old study by Galton. Galton asked a crowd at a county fair to estimate the weight of an ox. While no one guessed the correct weight (including experts), the mean of all guesses was within 1% of the exact weight (Galton, 1907). This study illustrates the principle that each observed score from a single measurement context will probably be a relatively poor estimate of the true score. However, averaging across all observed scores serves to cancel out the inconsistent aspects of the measurement situation, leaving a better, more reliable estimate of the true score than any one observed score provides. A classic measurement truism is "Given enough sows' ears, we can indeed make a silk purse" (Green, 1978).

In practice, our estimate of "true" between-person variance comes from among-person variance in a set of estimated true scores from a reliability sample (Shavelson & Webb, 1991). Generalizability theory calls this true score variance or person variance. In the abstract, it is this variance that group researchers care to explain.

Conceptually, measurement error is the portion of the observed score variance that is *not* due to true score variance. At an individual level, measurement error within a participant is the average deviation of observed scores around the estimated true score for that participant. At the group level, measurement error is reflected in different rankings of the participants on the dependent variable depending on the measurement context or observer that generated the observed score (Shavelson & Webb, 1991). Once again, it is useful to remember that the true score measurement of the generalized characteristic at the group level is above and beyond any one measurement context or observer. When a single number is used to quantify measurement error at the group level, we call this error variance.

EXAMPLE: GENERALIZABILITY (G) STUDY WITH MULTIPLE SESSIONS AS A SINGLE FACET

As a simple example of a study that provides estimates of participant variance and error variance, we provide a spreadsheet in Table 2.1. This spreadsheet contains 40 estimates of "number of requesting acts" from 10 participants whose number of requesting acts is measured in 4 interactions (i.e., observation sessions). The data in this example are "fully crossed" (i.e., all participants are observed in 4 conversations). Generalizability (G) studies can be conducted using partially crossed designs, but fully crossed designs provide more information and are thus emphasized here (Shavelson & Webb, 1991). These data were arranged in repeated measures format. Because there are four sessions per participant, each participant has four rows devoted to him or her. The columns in the spreadsheet are labeled by the factors in the analysis of variance (ANOVA) design and by the dependent variable label. In G studies in psychology or education, participants are almost always one of the factors. The values in the participant or person column are the ID numbers. Factors in the design that represent "error" are called "facets." In this example, we are only studying differences in the dependent variable due to a single source of error (i.e., using different sessions). Therefore, this example is a single-faceted study. The values in the session column are 1–4, representing the four observation sessions. Finally, there is a column

Table 2.1

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES SPREADSHEET FOR 1-FACETED GENERALIZABILITY STUDY

PERSON	SESSION	DEPENDENT VARIABLE	PERSON	SESSION	DEPENDENT VARIABLE
1.00	1.00	5.00	6.00	1.00	5.00
1.00	2.00	4.00	6.00	2.00	6.00
1.00	3.00	3.00	6.00	3.00	4.00
1.00	4.00	3.00	6.00	4.00	6.00
2.00	1.00	2.00	7.00	1.00	7.00
2.00	2.00	4.00	7.00	2.00	6.00
2.00	3.00	3.00	7.00	3.00	5.00
2.00	4.00	4.00	7.00	4.00	6.00
3.00	1.00	6.00	8.00	1.00	6.00
3.00	2.00	7.00	8.00	2.00	6.00
3.00	3.00	5.00	8.00	3.00	3.00
3.00	4.00	4.00	8.00	4.00	4.00
4.00	1.00	5.00	9.00	1.00	2.00
4.00	2.00	5.00	9.00	2.00	3.00
4.00	3.00	3.00	9.00	3.00	2.00
4.00	4.00	4.00	9.00	4.00	2.00
5.00	1.00	3.00	10.00	1.00	3.00
5.00	2.00	4.00	10.00	2.00	4.00
5.00	3.00	1.00	10.00	3.00	2.00
5.00	4.00	4.00	10.00	4.00	4.00

for the dependent variable scores. G studies can and often do have multiple dependent variable columns. The example data are also contained in an Excel file named "1-faceted g study data for SPSS in Excel" that is on this book's website (www.springerpub.com/yoder/supplements) so that readers can run the provided Statistical Package for the Social Sciences (SPSS) syntax on the example data.

Table 2.2 presents the SPSS syntax and part of the output from the results of the ANOVA applied to the data in Table 2.1. Using a repeated measures data arrangement format allows us to use the default settings of a univariate ANOVA and to treat participant and session as fixed factors. The results provide the needed mean squares (*MS*) and numbers (*N*) that will be used to compute the variance estimates and reliability (i.e., generalizability or *g*) coefficient. The *g* coefficient is a type of intraclass correlation coefficient.

In Table 2.2, we have highlighted the *MS* and degrees of freedom (*df*) that are needed to compute the *g* coefficient. It is critical to note that only the person (i.e., ID number) and interaction term involving the person factor are relevant to computing these variance estimates and *g* coefficient. The reason for this is that G studies in psychology and education are usually interested in sources of influence on the ranking of participants' dependent variable scores only. A small person × session interaction *MS* relative to the variance estimate for the person factor means that the ranking of the participants on the dependent variable did not vary much among the observation sessions used to measure the dependent variable.

To compute the *g* coefficient and the person variance for this example, we have provided an Excel file named "1-faceted g calculator" on the website (www.springerpub.com/yoder/supplements). Readers may wish to input the indicated *MS* and *N* (*N* for session is degrees of freedom for session + 1) from the SPSS output in Table 2.2 into the relevant cells in the Excel spreadsheet. The 1-faceted g calculator should indicate that the variance estimate for the person is 1.35, while the variance estimate for the person × session interaction term (i.e., the same as the *MS* for this term) is only .63. The formula in the Excel spreadsheet for person variance indicates that it is the average difference between person variance and error variance (i.e., [person variance—error variance]/number of sessions). The formula in the Excel spreadsheet for the g coefficient indicates that g is the person variance on the dependent variable/total variance in the reliability sample on the dependent variable. The observed *g* coefficient for 1 observer and 1 session is .68. Conceptually, this *g* coefficient

Table 2.2

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES SYNTAX AND OUTPUT OF ANALYSIS OF VARIANCE

```
UNIANOVA
dv BY person session
/METHOD = SSTYPE (3)
/INTERCEPT = INCLUDE
/CRITERIA = ALPHA (.05)
/DESIGN = person session person*session
```

ANOVA Results

	TESTS OF BETWEEN-SUBJECTS EFFECTS				
SOURCE	TYPE III SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	*F*	SIG.
Corrected model	88.375	39	2.266	.	.
Intercept	680.625	1	680.625	.	.
Person	54.125	9	6.014	.	.
Session	17.275	3	5.758	.	.
Person * Session	16.975	27	0.629	.	.
Error	0.000	0	.		
Total	769.000	40			
Corrected total	88.375	39			

dv: dependent variable.

means that 68% of the measured variance in the dependent variable (across all 40 cases) is due to true score variance.

CONSEQUENCES OF A LOW *G* COEFFICIENT

The lower the *g* coefficient, the more influence variables other than the generalized characteristic of interest have on the observed scores. A group correlational study that attempts to account for variance in participants' aggression during recess play provides an example of the issues (Stoolmiller, Eddy, & Reid, 2000). This large study found that 70% and 20% of the variance in aggression scores was influenced by variables that varied among observation sessions and observers, respectively. This means that only 10% of the variance in any one estimate of the aggression scores was due to factors within the students. Therefore, only 10% of the variance in the any one set of observed aggression scores could be accounted for by student-level predictors such as their achievement level or their socioeconomic status (SES). Obviously, this made it extremely unlikely that such predictors could be identified if only one observer in one observation session was used to estimate aggression. In other words, the consequences of using a variable with a low g coefficient is an increased probability of Type II error, assuming that multiple significance testing is controlled for or avoided. In fact, low reliability usually causes an increase in Type II error. This occurs because measurement error is usually randomly distributed around the true score (Thompson & Vacha-Haase, 2000).

One way of increasing the variability due to participants when a single estimate of the generalized characteristic is unreliable is to average the scores from many sessions and use the average score as the participants' dependent variable score. Therefore, we need a way to indicate how many session scores need to be averaged to derive reliable estimates of the generalized characteristic. Decision (D) studies are designed to provide this guidance.

DECISION STUDIES

Decision studies allow us to posit different scenarios (e.g., number of sessions and/or number of observers) to estimate how many sessions and/or observers we need to achieve a criterion level of g coefficient. Just as many group investigators use power analyses to plan future studies, one can conduct a D study to plan future group-based observation studies. Benchmarks for an acceptably reliable measure vary between .6 and .8, depending on the area of study (Bakeman, McArthur, Quera, & Robinson, 1997).

For this example, we selected .8 for our criterion level g coefficient. Using the computations provided in the 1-facet g calculator spreadsheet, we see that we need to average across at least two interaction sessions (i.e., sessions) to derive a "reliable" (above .8) estimate of the number of requesting acts.

It should be noted that one *can* conduct a D study with only two sessions. However, providing more and a wider sample of types of relevant sessions (all of which elicit the key behaviors) provides a more accurate estimate of how many sessions one needs to average across to derive reliable estimates of the generalized characteristic. Similarly, it is even more informative to conduct at least a two-faceted G and D study with sessions and observations as facets than it is to conduct a planning study with sessions as the only facet. Next, we turn to a published example of such a study.

McWILLIAM AND WARE AS AN EXAMPLE OF A TWO-FACETED DECISION STUDY

In this exemplary study, the investigators wanted to know how many observers and sessions were needed to derive reliable (above a g coefficient of .8) estimates of nine types of "engagement" (McWilliam & Ware, 1994). In the G study part of the analysis, 47 participants were observed in 4 classroom sessions by 3 observers. The design was fully crossed. Therefore, there were 564 cases (i.e., 47 * 4 * 3). Every 10 s, observers indicated whether a target child was engaged or not. If engaged, the observer indicated which of the three types and which of the five levels of engagement characterized the observed instance. The data were summarized in nonmutually exclusive categories such that nine dependent variables were derived: three types, five levels, and one unengaged category. The metric of the variable was the proportion of 10-s interval that the observer scored a particular type, level, or presence of engagement. Nine separate ANOVAs were conducted, one for each dependent variable.

The results of the D studies for a single observer (the number of observers most of us use to generate our primary data) indicated that it would take forty 15-min sessions to derive a reliable estimate of engagement with materials (i.e., contextually and developmentally appropriate actions on an object). In contrast, it took five 15-min sessions to derive a reliable estimate of encoded engagement (i.e., rule-governed action or interaction). This is evidence that when assessed from unstructured classroom settings in preschoolers with disabilities, engagement with materials is much more influenced by variables that vary among classroom sessions than is a generalized tendency to exhibit encoded engagement. The general lesson here is that some variables are more influenced by contextual variables than others. For our purposes, G and D studies provide an empirical approach to quantifying this reality in relation to planning and executing studies using direct observation.

Another lesson from this study is that some variables cannot be practically measured in particular participants in unstructured environments. For example, using 10 sessions and 6 observers as the practical limit of what most investigators could afford, their data indicate that differentiated engagement (i.e., nonrepetitive behavior directed to a person or object) cannot be practically measured to a reliable degree in preschoolers with disabilities when measured in unstructured environments.

A final lesson from this D study is that if observing more sessions is not possible given limited resources, we may be able to derive reliable estimates of the generalized characteristic by averaging estimates across multiple observers and fewer sessions. For example, McWilliam and Ware show that although one could derive a reliable estimate of encoded engagement with 5 sessions and 1 observer, one could also derive a reliable estimate of encoded engagement with only 3 sessions if 4 observers coded all sessions, and the investigator averaged across all 12 estimates to derive the estimate of the encoded engagement for a particular participant.

PRACTICE USING A G CALCULATOR ON DATA FROM A TWO-FACETED G AND D STUDY

On the website (www.springerpub.com/yoder/supplements), we provide a spreadsheet called “2-faceted g study data for SPSS in Excel.” These data can be imported by SPSS or other statistical programs. One can then use SPSS’s “general linear model,” “univariate,” and select “ID,” “session,” and “observer” as fixed factors and “DV” as the dependent variable. Once run, the output should provide results identical to those in Table 2.3. The variance estimates for the measurement error (i.e., the sum of the interaction terms’ variance estimates that involve person or ID) can be computed using the Excel spreadsheet provided on the website called “2-faceted g calculator.” This calculator has been explained in detail in another published work (Taylor, Yoder, & McWilliam, 2006).

Table 2.3

RESULTS OF PRACTICE 2-FACETED G STUDY

SPSS ANOVA output

	TESTS OF BETWEEN-SUBJECTS EFFECTS		
SOURCE	TYPE III SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
Corrected model	88.375(a)	39	2.266
Intercept	680.625	1	680.625
Session	5.625	1	5.625
ID	54.125	9	6.014
Observer	11.025	1	11.025
Session * ID	6.125	9	.681
Session * Observer	.625	1	.625
ID * Observer	6.725	9	.747
Session * ID * Observer	4.125	9	.458
Error	.000	0	.
Total	769.000	40	
Corrected total	88.375	39	

Input into g calculator

MSID	MSIDXOB	MS3WAY	MSIDXSESSION	NSESSION	NOBS	NID
6.014	0.747	0.458	0.681	2	2	10

Results of g calculator: the variance estimates

VARID	VARIDXOBS	VARIDXSESSION	VARERROR	G
1.261	0.1445	0.1115	0.714	0.63848101

(*Continued*)

Table 2.3

RESULTS OF PRACTICE 2-FACETED G STUDY (Continued)

Results of g calculator: D study results

G	hypNsession	hypNobs
0.63848101	1	1
0.74604348	2	1
0.79043042	3	1
0.81466527	4	1
0.75339806	1	2
0.83870968	2	2
0.87160878	3	2
0.88904556	4	2

Table 2.4 provides the formula for these variance estimates. It should be noted that the variance estimate for the three-way interaction between person × observer × session is the *MS* for that term. Table 2.3 provides the results of the 2-faceted g calculator when the appropriate *MS* and *N* values are entered from the SPSS ANOVA output.

The relative contribution of the various sources of influence can be illustrated in a pie chart (Figure 2.1). As is often the case in 2-faceted g studies, the largest source of error is the three-way interaction between person × observer × session. Conceptually, when the three-way interaction variance estimate is large, three primary explanations exist. First, it may indicate that the consistency with which observers rank participants on the dependent variable varies by session. Second, it may indicate that the consistency with which sessions rank participants on the dependent variable varies by observer. Third, it may mean that influential facets are not included in the design (Shavelson & Webb, 1991). The interaction between person and session quantifies the extent to which rankings among participants on the dependent variable vary depending on the session from which the observed score is derived. The interaction

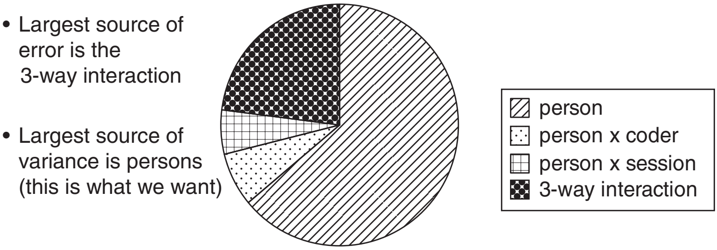
Table 2.4

FORMULAE FOR VARIANCE ESTIMATES FROM A FULLY CROSSED 2-FACETED ANOVA

SOURCE OF VARIATION	EQUATION FOR CALCULATION OF VARIANCE COMPONENT FROM MEAN SQUARES (*MS*) AND NUMBERS (*N*)
Persons (p)	(MSp—MSps—Mspo + MSpso)/(Nr × Ns)
Person × Observer (po)	(Mspo—MSpso)/Ns
Person × Session (ps)	(MSps—MSpso)/Nr
Person × Observer × Session (pso)	MSpso

Source: Shavelson and Webb (1991).

between person and observer quantifies the extent to which ranking among participants on the dependent variable vary depending on the observer. The total error variance (i.e., VARerror on the 2-faceted g calculator) is the sum of the variance estimates for the interaction terms involving persons. In this example, the person variance is much larger than the total error variance. The formula in the 2-faceted g calculator makes it clear that the formula for the g coefficient is person variance/total variance. Total variance is person variance + total error variance. Therefore, a g coefficient of .64 means that 64% of the variance in the



- Largest source of error is the 3-way interaction
- Largest source of variance is persons (this is what we want)

Figure 2.1 Variance estimates for the example 2-faceted g study.

reliability sample's dependent variable scores is due to true score variance among participants even if one only uses one observer and one session.

In terms of the D study, if one selects .8 as the criterion *g* coefficient, one would predict that averaging across four sessions from one observer would be sufficient to derive a reliable estimate of the generalized characteristic. If this is too expensive, then one could average across two sessions which are coded by two observers. Anything more than this could be argued to be an unnecessary expense (i.e., result in diminishing return). Anything less could be argued to be a waste of time (i.e., elevated probability of Type II error).

ACCURACY OF D STUDY PROJECTIONS

One of the criticisms of D studies has been that we rarely meet the assumptions of the analysis (Kane, 2002). One assumption is that the sampled observers and sessions are representative of the universe of observers and sessions for a particular generalized characteristic. The representativeness of the observers and sessions sampled vis-à-vis the universe of possible observers and sessions is influenced by the number of observers and sessions we sample (i.e., sample size) and the method by which we sample them. Neither of these is optimal in typical D studies.

Despite failing to meet the above set of assumptions, it is worth pointing out that we routinely accept the value of power analysis when planning studies despite its reliance on unrealistic assumptions. Power analysis assumes perfect reliability of measures and known effect sizes (Cohen, 1988). We continue to use power analysis because they provide information that is better than the alternative: planning without any knowledge of the required sample size to detect probable effect sizes. Similarly, we assert here that D studies provide guidance that is better than the alternative: using single unstructured measurement contexts to measure generalized characteristics. The current practice of using single sessions to measure generalized characteristics can be shown to be a poor use of resources for many characteristics, populations, and measurement contexts. However, it should be noted that D studies are never a substitute for estimating the reliability of the data used to test a research question.

IMPLICATIONS OF THE LESSONS OF G AND D STUDIES FOR SINGLE-SUBJECT RESEARCH

The recommendation that we need to use either structured measurement contexts or aggregate scores across multiple sessions to derive reliable estimates of generalized characteristics does not always match the logic or purpose of single-subject research designs. Applied to single-subject research, the recommendation to average across several sessions would take the form of averaging the scores across several (e.g., 3) sessions within the same design phase and graph the average as the data point for that temporal unit (e.g., a week). One would obviously have to observe more sessions in each phase to derive sufficient data points per phase. This suggestion may not be feasible (or acceptable to some investigators) when using single-subject research designs because within participant variability is part of the data that is used to infer a functional relation between the independent and dependent variables (Kennedy, 2005).

On the other hand, the recommendation that one use structured measurement contexts to derive single session estimates of generalized characteristics may be amenable to single-subject research designs depending on the purpose of the study or project. For example, it has become an industry standard for the study of severe problem behavior to use structured sessions as analogs within multielement designs to isolate and test possible contingent relations between antecedents (discriminative stimuli), consequences (putative reinforcers), and severe problem behavior (e.g., self-injury). Before further discussing why structured sessions are particularly useful to many single-subject designs when studying generalized characteristics, it is useful to discuss some background information first.

One internally valid single-subject design is the withdrawal design (Kennedy, 2005). A withdrawal design demonstrates that the dependent variable changes in the expected direction when the independent variable is applied and then returns toward baseline levels when the independent variable is withdrawn. However, many, if not most, generalized characteristics are not readily reversible. Therefore, withdrawal designs are not applicable for many generalized characteristics. In some cases, an adapted alternating treatments design can be used to address questions regarding a functional relation between a treatment and a generalized characteristic. For example, one might produce two sets of equally learnable vocabulary words and teach only one of these. If the taught words are learned and generalized faster than the untaught words, then one might reasonably conclude that the participant acquired a generalized understanding of the taught words through the treatment.

Unfortunately, the value of this design is limited to situations in which the investigator can persuasively argue or show that the goal sets are equally learnable to the study participant(s). Because of the limitations of withdrawal and adapted alternating treatments designs in studying generalized characteristics, variants of the AB (baseline phase treatment phase) design, such as multiple-baseline across participants, are particularly common when studying such dependent variables (Kennedy, 2005; Lieberman, Yoder, Reichow, & Wolery, in press). Multiple-baseline across participants designs use stable, staggered baselines and replicated changes in level, trend, or variability only during the treatment phase to infer a functional (i.e., causal) relation between the independent and dependent variables (Kennedy, 2005).

Using structured measurement contexts might be particularly useful for multiple-baseline across participant designs. Many measures of generalized characteristics, especially if measured in unstructured measurement contexts such as many naturally occurring settings, are logically likely to demonstrate a gradual change that begins many sessions after the onset of the treatment phase. This is particularly true in populations that learn slowly (Zeaman & House, 1977). Experts (i.e., reviewers on the editorial boards that regularly use multiple-baseline designs in their research and regularly review such studies) tend to agree with each other about whether a functional relation exists much less often when dependent variables change many sessions after the onset of the treatment than when dependent variables change immediately after the onset of the treatment phase (Lieberman et al., in press). Because structured sessions control for sources of error variance, any influence of the independent variable on the dependent variable is more likely to be detected rapidly in structured sessions than in unstructured sessions. However, some investigators will value generalization to naturally occurring and coincidentally unstructured measurement contexts more than to structured sessions.

A DILEMMA

Putting all these factors together creates a dilemma. On one hand, group design logic is generally better suited to testing research questions about treatment effects on generalized characteristics. On the other hand, clinicians and teachers often care more about treatment effects on individuals than they do about treatment effects on groups. Single-subject design is the only way we can infer treatment effects on generalized characteristics at the individual level. Many applied researchers value measuring whether newly learned skills generalize to unstructured natural contexts more than to structured ones. Demonstrating a functional relation for generalized characteristics measured in unstructured natural contexts is difficult, particularly in the type of single-subject designs most suited to measuring effects on generalized characteristics (i.e., multiple baselines across participants). There is currently no consensus on how to resolve this dilemma.

RECOMMENDATIONS

We have stated that a group design approach to measurement allows us to think about representativeness, an attribute of generalized characteristics, in a scientifically testable way (i.e., based on stability over contexts). In classical measurement theory, a measurement theory that is most easily applied to the group design concept of measurement, an observed score is composed of a true score plus measurement error. The greater the proportion of the observed score that is true score, the more stable across contexts the observed score will be.

Generalizability theory provides a way to quantify true score and measurement error using the *MS* and *N* values from ANOVA. From these ANOVA results, we can estimate the variance estimates that can be used to compute a reliability coefficient called the g coefficient. Conceptually, a g coefficient is the proportion of between-participant variance in observed scores that is true score. Decision studies, which use the results of generalizability studies, demonstrate that the more session scores one averages across, the more reliable the estimate of the generalized characteristic. Decision studies also provide a way to predict how many sessions one needs to average scores across to derive an estimate of the generalized characteristic with criterion level reliability (e.g., above .8).

REFERENCES

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*(4), 357–370.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt.

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Galton, F. (1907). One vote, one value. *Nature, 75,* 450–451.

Green, B. (1978). In defense of measurement. *American Psychologist, 33,* 664–670.

Johnston, J. M., & Pennypacker, H. S. (1993). *Readings for strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kane, M. (2002). Inferences about variance components and reliability—generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement, 39,* 165–181.

Kennedy, C. (2005). *Single-case designs for educational research*. Boston: Allyn and Bacon.

Lieberman, R., Yoder, P., Reichow, B., & Wolery, M. (in press). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly.*

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18*(1), 34–47.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Stoolmiller, M., Eddy, J., & Reid, J. (2000). Detecting and describing preventive intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Consulting and Clinical Psychology, 68,* 296–306.

Taylor, C., Yoder, P., & McWilliam, R. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28,* 139–153.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174–195.

Zeaman, D., & House, B. (1977). Zeaman and House's attention theory. In C. D. Mercer & M. E. Snell (Eds.), *Learning theory research in mental retardation: Implications for teaching* (pp. 94–141). Columbus, OH: Merrill.

3 Designing or Adapting Coding Manuals

OVERVIEW

In this chapter, we cover the definition of a coding manual, the important role of the research question for defining the scope of the coding manual, and the steps involved in creating a coding manual. The steps involved in creating a coding manual can be divided into (a) conceptually defining the context-dependent behavior or the generalized characteristic of interest, (b) deciding the level of detail at which to measure the context-dependent behavior or the generalized characteristic, (c) deciding whether to define the lowest level categories by physically based versus socially based definitions, (d) defining the lowest level categories, (e) defining segmenting rules, and (f) defining start and stop coding rules. We also indicate the potential value of flowcharts when the coding process is complex. After discussing how complex modern coding manuals can be, we suggest that the old advice, that coding manuals should be sufficiently short to be included in method sections, is unrealistic in many current studies. We close the chapter with a summary of principles for designing a coding manual.

SELECTING, ADAPTING, OR CREATING A CODING MANUAL

Definition of a Coding Manual

A coding manual is a set of rules, definitions, examples, and near nonexamples that guide the observers in counting and/or indicating the duration of the behaviors of interest. The coding manual consists of at least start and stop coding rules and definitions and examples of categories. We suggest that including close nonexamples of categories is also very useful. In the case of a particular type of behavior sampling (i.e., event sampling), a set of rules used to define the onset and offset of events (i.e., segmenting rules) is usually needed. In the case of a complex coding manual, a flowchart providing an overview of the coding process as a series of yes/no decisions is valuable.

Relation of the Coding Manual to the Research Questions and Predictions

If the coding manual is not selected or created on the basis of observational variables from each falsifiable research question, there is a risk of creating a disconnect between what the investigator wants to know and how he or she is measuring the variables (Bakeman & Gottman, 1997). The level of distinction or detail at which the observational variables are stated in the research questions do not have to be as fine-grained as those distinguished in the coding manual. That is, the context-dependent behavior or generalized characteristic may be stated as the implied independent or dependent variable or as the predictors in the research questions.

However, if there are lower level categories, these should be the types of the context-dependent behavior or generalized characteristic stated in the research question or prediction. For example, one might pose a research question as “Does teacher praising of student prosocial behavior result in less frequent student antisocial behavior?” The question does not specify the types of antisocial behavior that will be addressed. The literature indicates that there are certain types of antisocial behavior that are suppressed when prosocial behavior increases, but other types of antisocial behavior that are independent of prosocial behavior. The lower level categories should make this distinction.

Whenever an existing coding manual has been shown to be scientifically useful for the purpose and in the population for which the investigator intends to address the research question, it makes sense to use or adapt an existing coding manual (Primavera, Allison, & Alfonso, 1997). Importantly, the research question should not be changed to match existing coding manuals. Often, existing coding manuals need to be simplified, modified, or expanded to address an investigator's exact research question. To the extent that existing coding manuals need modification to meet the investigator's needs, the process may resemble that used to design a coding manual.

Recommended Steps for Modifying or Designing Coding Manuals

Conceptually Defining the Context-Dependent Behavior or the Generalized Characteristic

This step is conducted to clarify the meaning of the generalized characteristic or context-dependent behavior that theory, data, or logic suggest is scientifically useful. Because these definitions necessarily vary by topic, only guidelines can be provided here for selecting among competing definitions.

The criteria by which an informed investigator selects conceptual definitions for context-dependent behaviors and generalized characteristics are based on methods of content validation, sensitivity to change, or construct validation. More information on these topics will be covered in chapter 10. Brief coverage of these topics will be provided in this chapter.

Observational measurement for all designs needs to meet the criteria for content validation as a method of conceptually defining a behavior class (Haynes & O'Brien, 1999; Primavera et al., 1997). One way to define content validation is by majority opinion of experts; another is through consensus of a professional group.

Researchers concerned with treatment efficacy should select a conceptual definition for their independent variables that is empirically or theoretically consistent with prior reports demonstrating the efficacy of the independent variable or theoretical considerations in which the independent variable is considered to be a critical treatment component. Similarly, the way that dependent variables are conceptualized should have a strong empirical and/or theoretical support in relation to their sensitivity to change during the time frame needed by the research design.

Group designs that employ measures of generalized characteristics call for construct validation as a method of judging the relative scientific value of competing conceptual definitions for variables (Haynes & O'Brien, 1999). Briefly, construct validation is a cumulative process by which empirical studies test whether particular conceptual definitions yield variables that perform as expected by theory and logic. Conceptually related variables (e.g., aggression, anger) should be empirically associated with one another whereas conceptually unrelated variables (e.g., aggression, happiness) should not. Further, our conceptual distinctions should discriminate groups of known characteristics in predictable ways (Cronbach & Meehl, 1955; Primavera et al., 1997).

Deciding the Level of Detail at Which the Behaviors Should Be Distinguished

Often, context-dependent behaviors or generalized characteristics can be subdivided into subordinate categories or types. The issue here is whether subordinate categories are needed, and if so, how many distinctions are scientifically useful. Addressing the issue of category distinctions will also lead to decisions about which subordinate categories are grouped under a single superordinate category.

There is a balance between being so specific that the category has little social importance or communicative efficiency versus being so vague that the class no longer allows falsification of the prediction, guides clinical practice, or allows detection of change due to treatments (Haynes & O'Brien, 1999). The decision to lump or divide is not one that can be made in the abstract in a way that holds uniformly for all investigators because each substantive area will have different theories and empirical literature that guide this decision. However, two principles can be stated that warn against making "too many" distinctions. There is replicated evidence that the more discriminations observers have to make about an event, the lower the interobserver agreement (Jones, Reid, & Patterson, 1975; Taplin & Reid, 1973). Additionally, the lowest level or most specific subordinate categories must be defined at a high-enough level that in at least some sessions (in single-subject research) or some participants (in group research) there are "many instances" of the category. The definition of many instances depends on the size of the association, group difference, or single-subject effect. The general principle is that when the obtained range of scores is very small, it is very difficult to detect expected associations or effects. A guideline from

Yoder's personal research is that using two to five subordinate categories within a single superordinate category tends to serve research purposes better than using more than five subordinate categories.

In general, investigators should make distinctions only when their research questions or predictions call for the distinction. One instance in which distinctions should be made is when theory or empirical data indicate that a subordinate category has antecedents or consequences that are functionally different from others (Haynes & O'Brien, 1999; Johnston & Pennypacker, 1993). Another way to decide whether it is important to distinguish potential subordinate categories is based on whether one subordinate category is developmentally easier, functionally simpler, or differentially predictive of later socially important outcomes than other subordinate categories in the same superordinate class.

Physically Based Definitions, Socially Based Definitions, or Both?

To some readers, it may seem strange to justify the use of conceptual definitions of the categories we wish to measure. However, in chapter 1, we indicated that an early interpretation of operationalism, the semantic interpretation, seeks to remove subjective interpretations of what is observed. Definitions of behaviors that rely only on detection of the presence or absence of stated behaviors are called "physically based" categories (Bakeman & Gottman, 1997). These emphasize only the physical dimension of the behavior, not the presumed function or higher order inferences about what the form of behavior may reflect in terms of psychological processes (e.g., attention). This type of category is very narrowly defined and is often composed of an exhaustive list of the behaviors that the investigator considers the measurable examples of the category.

For example, an investigator wishing to measure communication may list three behaviors that are examples of "intentional communication" and consider this the complete list of behaviors he will code (e.g., point to object, reach to an object, and give object to person). These examples might be selected based on what others have coded, what was expected to occur in the observation sessions, and what could be easily defined on the basis of observable behaviors (i.e., operational definitions).

In contrast, socially based coding manuals differ from physically based coding manuals, in that the former tends to have categories with more exemplars or behavioral forms and requires observers to make a judgment regarding whether the behavior in question has a particular function or meets a series of conceptual criteria. The conceptual definitions emphasize theoretically important distinctions and attributes of examples that discriminate them from near nonexamples, and examples given are illustrative rather than exhaustive. Nonexamples are given to help observers define the boundaries of the concept. The theoretical rationale for the inclusion of particular key phrases is given to observers to train them to make informed decisions about rare, but legitimate, examples of the concept. For example, the conceptual definition of nonverbal "intentional communication" might be the label given to a class of interest. A conceptual definition might be "gestures, nonword vocalizations, or expressions of emotion combined with evidence of coordinated attention to object and person used to convey a message to another person." Many examples of intentional communication and many near nonexamples provide observers with an idea of the class of behaviors they are meant to code. Details about appropriate examples and nonexamples will be given in the next section on defining the lowest level categories.

The scientific value of a coding manual is judged by the extent it accomplishes the goals for which it was created. Using this criterion, there is no empirical evidence that physically based categories are more valuable than socially based categories. In fact, it can be convincingly argued that too much reliance on what has been already measured or what is definable only through a small set of observable behaviors can reduce the social importance or the degree to which the category represents the generalized characteristic of interest (i.e., the content validity). One can argue that we cannot exhaustively list all possible exemplars of many categories. Attempting to do so can result in an unnecessarily long and complicated coding manual. By teaching observers the critical concepts underlying the needed distinctions, we are equipping observers to make informed decisions regarding the potential inclusion of frequent and rare examples of the category. Doing so may enable us to measure the generalized characteristic at a level that meets the "grandma rule" (i.e., grandmother understands the general concepts and can immediately see the importance of the concept).

Defining the Lowest Level Categories

We recommend using elements of physically based and socially based coding approaches to define the lowest level categories. Observers are likely to benefit from both conceptual and operational definitions of the categories. There is professional consensus regarding the value of operational definitions (Rogers, 1989). Operational definitions use only words with observable referents to define concepts. Conceptual definitions provide the framework observers need to judge whether marginal examples fit the rationale for inclusion in the category.

It is recommended here that authors of coding manuals (a) define their categories in conceptual terms, (b) underline key ambiguous terms, (c) use conceptual and operational definitions to further define the underlined ambiguous terms, (d) provide nonexhaustive prototypical (i.e., very good or frequently occurring) examples of the category, and (e) provide near nonexamples to define the boundary of the category. For example, our conceptual definition of intentional communication might be as follows: "Gestures, nonword vocalizations, or expressions of emotion combined with evidence of coordinated attention to object and person."

In the coding manual, all underlined phrases would be conceptually and operationally defined. As an example, the phrase "coordinated attention to object and person" can be conceptually defined as any behavior that shows *attention to the object or event* about which the child is communicating that occurs within 3 s of behavior that shows *attention to the person* to whom the child is communicating. "Attention to object or event" might be operationally defined as the child looking at, actively touching with the hand, or talking or signing about at the object or event of interest. "Attention to person" might be operationally defined as the child looking at the other's face, touching any part of the other person with the hand, immediately imitating the other's action or vocalization, or immediately and accurately responding to the other's question. These are considered first-level operational definitions. If interobserver agreement in using the category definition is poor or discrepancy discussions among observers indicate that other terms are ambiguous, further breaking down of terms used in the operational definition could occur (i.e., second- and third-level operationalizations) until an acceptable point-by-point agreement is achieved. Alternatively, more examples and near nonexamples could be used to firm up the observers' conceptual definitions of terms. Obviously, the deeper such operationalizations occur, the more difficult the coding manual will be to follow. Disagreements among observers can be created from having so much specification that the content validity of the category is lost. Therefore, there is a balance between operational specification and content validity.

Prototypical examples of the category should be provided to help the observer relate the operational definitions at a level of analysis that is more commonly used by educated consumers of the research. It is recommended that between three and five prototypical examples that illustrate different aspects of the conceptual definition be used. For example, "giving an object to a person to ask it to be opened" could be provided to convey that a nonverbal and nonvocal example for a requesting purpose is acceptable. An example like "vocalizing 'ah' and pointing to an airplane and looking at the mother" could be provided to convey that a non-word vocalization for a declarative purpose is acceptable. Finally, "smiling while holding a hammer and looking to Mom" could be provided to convey that a smile is one way to express emotion and that the child's behavior must show attention to the adult to be coded. It is important to state in the manual that these examples are illustrative, not exhaustive.

Near nonexamples of the category are provided to help the observer define the boundaries of the concept. Near nonexamples are often superficially similar in form or topography to true examples, but differ from true examples in an important way. Often the distinguishing attribute is pointed out for the readers. For example, a near nonexample of intentional communication is as follows: "The child is looking at a hamster. The adult says the child's name. The child begins to shift her gaze from the hamster to the adult's face after the adult has said the child's name." This is not an example of intentional communication because the gaze shift from object to person was caused by the adult calling the child's name, not the child's wish to attempt to share her interest in the hamster with the adult.

Sources of Conceptual and Operational Definitions

There are three primary sources for conceptual and operational definitions: (a) the scientific literature, (b) the existing coding manuals attained from the authors of past studies, and (c) our own qualitative studies of "expert" knowledge (i.e., pilot studies). In quantitative studies, the most common source is the extant empirical literature. If the subordinate category has been sufficiently studied, there are often conceptual and operational definitions available in methods section or appendices of extant empirical articles. If the definition given in articles is not sufficiently specific for observers to attain criterion levels of interobserver agreement, which will be discussed more in chapters 8 and 9, then it is perfectly acceptable for readers to request coding manuals from authors via e-mail or phone.

If the subordinate category is relatively new or understudied at the time the coding manual is being designed, then one may need to conduct his or her own qualitative study to "discover" and generate conceptual and operational definitions that enable criterion level interobserver agreement. Because a qualitative approach is relatively uncommon in quantitative texts, we attempt to strike a middle ground in the level of detail at which we describe this interesting process by providing one example. Eliciting knowledge from experts is considered one of the most challenging aspects of understanding and modeling "expert systems" (i.e., software that is designed to simulate the implicit knowledge of human experts; Hoffman, Shadbolt, Burton, & Klein, 1995). The complex processes involved in expert systems and eliciting knowledge from experts is beyond the scope of this chapter but it is important to discuss briefly because it represents one of the more important but difficult areas for both novice and expert investigators in the development of direct observational measurement systems.

As an example, a fictitious doctoral student wanted to derive definitions for varying "levels of saliency" of parental behavior intended to direct the attention of their children with autism. The parental behaviors (i.e., cues) used to direct the children's attention were gestures, words, actions, and noises of toys created when the parent activated the toy. The doctoral student hypothesized that more salient cues would be more successful in directing children's attention than would less salient cues. She could not find an operational or sufficiently elaborated conceptual definition of "salience" in the extant literature to guide reliable judgments regarding parental attentional cues in a free-play session with the parents' children. However, she was confident that "she knew a salient cue when she saw one" and that other people who were familiar with young children with autism would be able to identify salient cues, too. She wanted a content-valid definition of saliency so she thought it was important to rely on more than her own judgment of what saliency meant. She needed a panel of "experts."

The first step was to identify the "experts." The definition of expert varies depending on the content area (Hoffman et al., 1995). However, it is probably useful to select people with both explicit (those who teach others to do the skill) and implicit (those who practice the art or skill being studied) knowledge of the context-dependent behavior or generalized characteristic of interest. The doctoral student selected faculty members responsible for teaching college students who were training to be early childhood educators of young children with autism. These faculty members had also spent at least 100 hr interacting with children with autism. Four faculty members were selected to allow evidence of convergence without relying on a sole expert.

Although there are many types of materials one can select to elicit information, one of the most efficient ways to elicit information is to identify "test" or "tough" cases (Hoffman et al., 1995). However, such materials tend to elicit an uncomfortable feeling of being evaluated (Hoffman et al., 1995). In contrast, "familiar" materials tend to elicit rapid expert judgments and do not elicit anxiety from experts (Hoffman et al., 1995). Therefore, it has been suggested that a combination of familiar and test materials may be useful in eliciting cooperation and information (Hoffman et al., 1995). In our example, the doctoral student edited twelve 3–10 s video clips of parents directing their children's attention. The children's responses to the parental behavior were carefully removed. The doctoral student located and edited two clear (i.e., familiar) examples, each of "high," "medium," and "low" saliency categories, based on her own intuitive definitions. The doctoral student selected three more "test" examples that she judged to be "somewhere between high and medium saliency." Finally, three more test examples were selected that the doctoral student considered "somewhere between medium and low saliency."

Asking experts to sort materials into categories is one informative method for eliciting information (Hoffman et al., 1995). In our example, the doctoral student independently asked each expert to sort the video clips into one of three categories that varied by saliency level. No time limit was given and experts could change their minds. This continued until each expert said that he or she was finished with the sorting.

The relative value of individual versus group interviews is unknown. However, empirical studies indicate that when experts meet as a group, they tend to rapidly find and argue over the small number of points on which they disagree (Hoffman et al., 1995). It is consensus that we seek. Therefore, we prefer that experts be interviewed independently.

The literature on eliciting expert knowledge indicates that structured interviews are more efficient than unstructured ones (Hoffman et al., 1995). In our example, the doctoral student examined the sorted materials (a) to identify how the test cases were sorted and (b) to identify any familiar cases that were classified in a category that was unexpected. Generic probes were asked about these items (Hoffman et al., 1995). The questions and answers were audio recorded for later analysis. For example, assume clip 2 is a test case for classification to either high or medium and clip 4 is a familiar case for high saliency. The doctoral student might ask “Why did you classify clip number 2 as highly salient?” to elicit rules that might define high saliency. It should be noted that both the clip and the category were indicated in the interviewer’s question. This aids later analysis.

Further imagine that clip 4 was unexpectantly classified as medium saliency. The doctoral student might say “I noticed you classified clip 2 as medium and clip 4 as high” to record on the tape how the clips were classified. Then the doctoral student might ask the expert, “What is different about clip 2 (i.e., the ‘familiar’ clip that was expected to be classified as high but was classified as medium) than clip 4 (i.e., the test clip that was categorized as high)?” The latter is meant to elicit particularly useful rules that might reveal “conditional rules” (i.e., those that are used under some conditions but not others). If such conditional rules are uncovered, then the doctoral student might test her understanding of the condition of the rule by asking “What if (the condition) were not present, would (the rule) apply?” This type of structured interviewing would be continued until the interviewer felt she understood the experts’ rationale for their sorting.

After interviewing all experts and transcribing (or marking wave files of statement via a computer software such as NVivo; Bazeley & Richards, 2000), commonalities among the experts’ responses to questions are analyzed. This is sometimes called “theme analysis” or “category analysis” (Bazeley, 2007). In our example, the doctoral student identified (a) the number of sensory modalities and (b) the number of behaviors as two themes that at least three of the four experts used to justify her choices. By “behaviors,” we mean actions such as “gesture to,” “moves on,” “talks about,” and “operates” referent objects.

To “test” the accuracy with which the abstracted themes classified the 12 video clips, clips were assigned 3 through 1 to correspond with high through low classification, respectively. Average numerical scores were derived across experts to estimate “expert” classification of each of the 12 video clips. Using the two themes of number of actions and number of sensory modalities, and the average expert sorting of the 12 video clips, the doctoral student derived the following operational definitions for the three levels of saliency. These definitions classified the 12 video clips in accordance with the average expert sorting. High saliency was defined as using at least three behaviors to draw the child’s attention to an object that appealed to at least two sensory modalities. Medium saliency was defined as using two behaviors to draw the child’s attention to an object that appealed to two sensory modalities. Low saliency was defined as one behavior to draw the child's attention to an object that appealed to at most two sensory modalities. This complex definition classified all 12 clips reliably by 2 observers. Ultimately, the test of such a process is whether a study using a systematic observational measurement system relying on its derived definitions results in confirming hypotheses regarding saliency and the child's correct responses to parents' attentional directives.

Defining Segmenting Rules

Once it is clear what the lowest level of distinction will be, the onset (and if needed the offset) of instances of categories may need to be defined. If the number or duration of events will be the metric of interest, then segmenting rules are necessary. Interval coding or time sampling behavior sampling does not require segmenting rules for reasons indicated in the next chapter. Briefly, interval coding means that one indicates whether at least one instance of the key behavior has occurred within a fixed interval of time (e.g., 10 seconds).

One can see the need for segmenting rules when events occur close in time. For example, assume we are measuring the number of communication acts. The child reaches for a ball and looks at an adult. Then one second later, the child says "ah" while looking at the ball and looks at the adult. Finally, another second later, the child says "ball" and looks at the adult. Segmenting rules are needed to determine whether this cluster of child behaviors is 1 versus 3 communication acts. In-seat behavior is an example of a behavior for which duration is important. In addition to precise definitions of onset, definitions of offset are necessary to differentiate examples of in-seat behavior from near nonexamples of offset. For example, is the momentary lifting of both buttock cheeks from contact with the chair seat sufficient to end in-seat behavior or does such contact need to cease for more than one second?

Segmenting rules almost always involve a certain amount of arbitrariness. For example, we may decide to treat the potentially three clusters of behavior in our example above as one act because the onset of the first behavior occurred within 3 s of the onset of the last behavior. The use of a temporal criterion, the use of "onset" instead of "offset" as the boundaries for the temporal criterion, and the decision to call the reach and gaze, the vocalization and gaze, and word and gaze, all "parts" of the same act are all questionable, but defensible. Our empirical and theoretical knowledge about most social phenomenon of interest is almost never sufficiently specific to guide this level of decision making. But, these decisions must be made; hence, there is some degree of arbitrariness. Usually, the best we can do is to make sure that our decisions are defensible and consistently applied. If they are carried out consistently, the potential lack of content validity that may occur from their existence will be worth the almost certain gain in reliability due to reduced interobserver disagreement on segmenting.

Defining When to Start and Stop Coding

Having clearly defined the lowest level coding distinction and with guidance about how to segment potential events, the next step is to decide when observers should begin and end their coding of an observation session. This section will include two types of start–stop signals. First, there are those at the beginning and end of the observation session. Because of the inconsistency of behavior among many observational sessions, it is useful to make explicit the signal for beginning and ending coding.

One ill-advised way to do this is to ask the adult administering the procedure to say when to start and stop coding. Alternatively, a start signal might be when the clock of the media file turns from 0 to 1 s. This requires that the administrator or cameraperson be consistent in giving the verbal signal or in beginning the clock relative to the onset of the observation procedure. The problem with these approaches is that it shifts the responsibility for providing the signal consistently to the administrator or cameraperson. Investigators rarely check on the consistency of the timing for such signals.

Instead, it is better to use a behavior that the participant does or one that the examiner does in the course of conducting the procedure to mark the beginning and ending of coding. For example, one might indicate that coding begins when the adult first speaks about the objects in the session or first speaks to the child. Similarly, stop signals might be when the examiner removes all toys from the table. Whatever the signals, they need to be included in the recording of every session and occur at times that do not exclude many codeable acts.

Another use of start–stop signals is to define the duration of the codeable sections of the observation session. When such start and stop coding rules are provided, they are done so as to potentially reduce the measurement error due to unexpected events or events that are known to inhibit or interfere with the occurrence of key behaviors in a way that does not reflect the phenomenon of interest. For example, the participant may be off-screen if the session is being recorded for later coding. Or there may be something unexpected occurring that inhibits the occurrence of a codeable behavior (e.g., a fire alarm). Or the participant may be engaging in behavior that is incompatible with the coded behavior (e.g., crying hard might be considered incompatible with intentionally communicating). Or the participant may not be providing an opportunity for the key behavior (e.g., if the key behavior is parent talk about the child's focus of attention and the child is not attending to anything).

In all cases, the total length of codeable time is measured so that the total codeable time might be used to "prorate" the number or duration of key behaviors. Prorating in this way is based on an assumption that more codeable time is related to more frequent or longer total duration of the key behaviors. This testing this assumption and its ramifications will be covered in chapter 5. At this point, readers are asked to note that if observers do not indicate stop and start coding times within the session, then variance in the amount of codeable time among sessions or participants cannot be analyzed or used to prorate the number or duration by codeable session duration.

Examples of start–stop rules that are used in the middle of sessions are usually different from those used at the beginnings and ends of sessions. For example, we may decide that the unexpected, inhibiting, or interfering events must occur for a criterion length of time to be sufficiently problematic to end the codeable section of the session. Similarly, we may decide that the participant needs to return on-screen long enough to allow key behaviors to be coded to restart coding. Again, such stop and start criteria are usually best if they are participant or examiner/tester behaviors rather than determined by the examiner's or cameraperson's explicit signal to start or stop coding. Although many of these criteria may seem arbitrary, they are extremely useful for improving interobserver agreement.

THE POTENTIAL VALUE OF FLOWCHARTS

Although not necessary for simple coding manuals, a flowchart that illustrates a predefined sequence of yes/no decisions can increase interobserver agreement. A flowchart is particularly useful when the coding manual requires several decisions before marking the code for a behavior. Without a flowchart, one source of measurement error is the observer forgetting to ask one of the essential questions before coding. Therefore, they are particularly useful for observers who have not yet learned the coding system well or for reminding experienced observers of the sequence and particular decisions that must be made to reliably code a behavior. In general, flowcharts ask a series of yes/no questions to help the observer decide if a codeable behavior has occurred (i.e., unitize), then ask another series of yes/no questions to help the observer classify the relevant behavior (i.e., classifying), and so on. An example of a flowchart is provided in Figure 3.1. In this example, the underlined terms would be conceptually and operationally defined in the coding manual. Appropriate examples and near nonexamples would be given for underlined phrases when necessary to achieve criterion level interobserver agreement.

DO CODING MANUALS NEED TO BE SUFFICIENTLY SHORT TO BE INCLUDED IN METHODS SECTIONS?

It is reasonable to question whether coding manuals produced with the above considerations in mind can really fit into the page limits of many journals. The answer is "no." Additionally, we should not write our methods sections as if a simplification of our coding manual is what was used to generate the data reported in articles. Instead, it is important to recognize that current observational research can be sufficiently complex to require complex coding manuals. Such manuals are simply too long to be included in the methods section. Some journal editors will not allow long manuals to be put into appendices. However, we can, and should, provide conceptual definitions and, when necessary, first-level operational definitions in the methods section with a notation that the complete coding manual is available from the author. Such public examination of the real coding manual is a necessary part of accountability to professional peers and a part of testing whether our coding manual is sufficiently complete to allow replication of research findings.

RECOMMENDATIONS

In this chapter, we recommend that investigators find, adapt, or develop a coding manual to measure the observational variables stated or implied in their research questions. If the investigator needs to develop

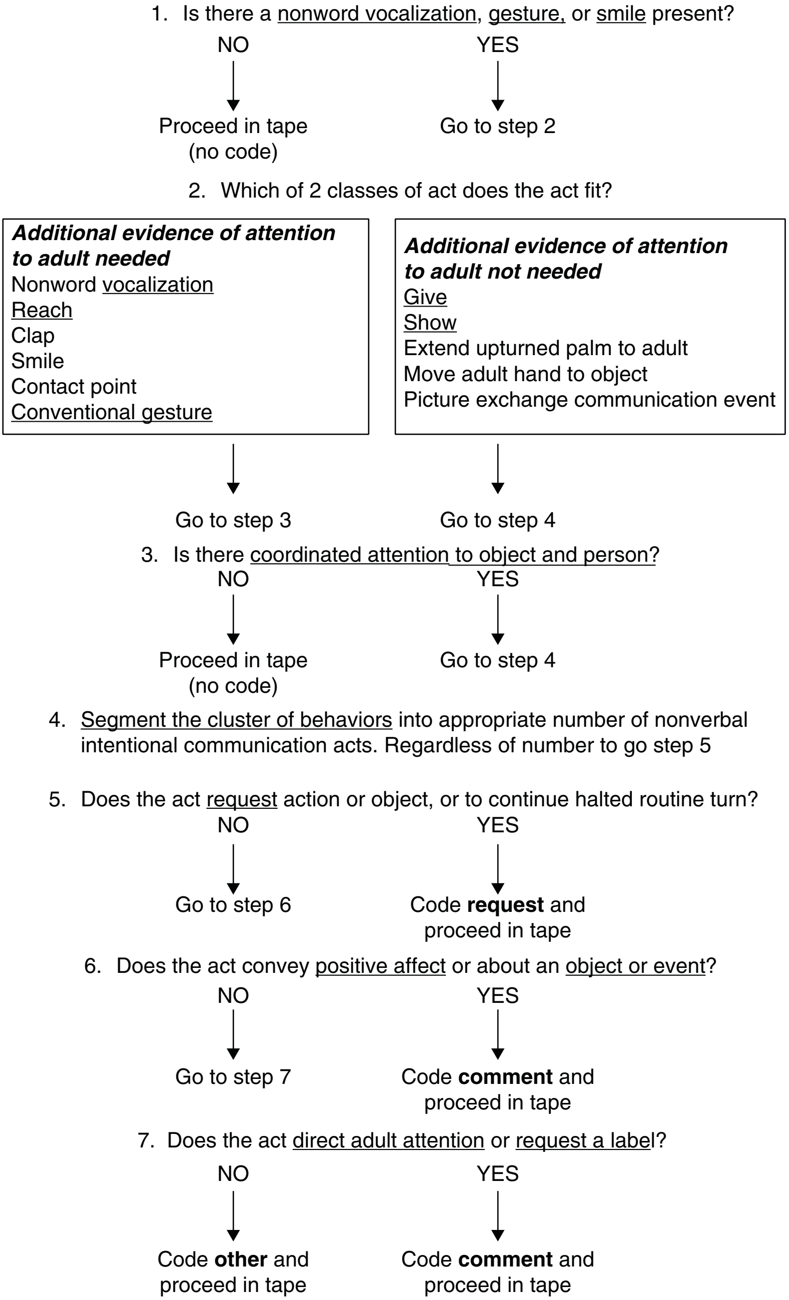

1. Is there a nonword vocalization, gesture, or smile present?
NO
YES
Proceed in tape (no code)
Go to step 2
2. Which of 2 classes of act does the act fit?
Additional evidence of attention to adult needed
Nonword vocalization
Reach
Clap
Smile
Contact point
Conventional gesture
Additional evidence of attention to adult not needed
Give
Show
Extend upturned palm to adult
Move adult hand to object
Picture exchange communication event
Go to step 3
Go to step 4
3. Is there coordinated attention to object and person?
NO
YES
Proceed in tape (no code)
Go to step 4
4. Segment the cluster of behaviors into appropriate number of nonverbal intentional communication acts. Regardless of number to go step 5
5. Does the act request action or object, or to continue halted routine turn?
NO
YES
Go to step 6
Code request and proceed in tape
6. Does the act convey positive affect or about an object or event?
NO
YES
Go to step 7
Code comment and proceed in tape
7. Does the act direct adult attention or request a label?
NO
YES
Code other and proceed in tape
Code comment and proceed in tape

Figure 3.1 Illustration of a flowchart for a coding manual on intentional nonverbal communication.

or adapt a coding manual, we recommend that a clear definition of the context-dependent behavior or generalized characteristic be developed. Often there is need to distinguish among different types of context-dependent behavior or generalized characteristics. We recommend that the phenomenon of interest be subdivided only to the extent that past data or theory justifies it. Once the lowest level categories have been selected, we recommend using both conceptual definitions, which require social judgments, and operational definitions, which require physical detection, to define the lowest level categories. When event sampling is used to derive number, duration, or a proportion is needed, developing a set of rules to define the onset and offset of the behavior is important. Indicating the rules for starting and stopping coding at the margins of the session as well as during the session are also frequently useful. Finally, complex coding is aided by use of a flowchart that sequences and specifies decisions in the form of yes/no questions. A coding manual is only a part of the overall measurement system. It influences and is influenced by behavior sampling and session and behavior recording methods: the topics of the next chapter.

REFERENCES

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Bazeley, P. (2007). *Qualitative data analysis with NVivo*. London: Sage.

Bazeley, P., & Richards, L. (2000). *The NVivo qualitative project book*. London: Sage.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Haynes, S. N., & O'Brien, W. H. (1999). *Principles and practice of behavioral assessment*. New York: Kluwer.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62*, 129–158.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Jones, R. R., Reid, J. B., & Patterson, G. R. (1975). Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 3, pp. 42–95). San Francisco: Jossey-Bass.

Primavera, L., Allison, D. B., & Alfonso, V. C. (1997). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–90). Mahwah, NJ: Erlbaum.

Rogers, T. B. (1989). Operationism in psychology: A discussion of contextual antecedents and an historical interpretation of its longevity. *Journal of the History of the Behavioral Sciences, 25*, 139–153.

Taplin, P. S., & Reid, J. B. (1973). Effects of instructional set and experimental influences on observer reliability. *Child Development, 44*, 547–554.

This page intentionally left blank

4 Sampling Methods

OVERVIEW

In this chapter, we define a "measurement system." We discussed measurement context in chapters 1 and 2 and the coding manual in chapter 3. In this chapter, we discuss the rest of the elements of the measurement system. We begin by presenting the general options for behavior sampling. Next, we briefly discuss the general options for participant sampling. Then we provide a brief review of the empirical evidence regarding the very important issue of participant reactivity to being observed. We address the issues regarding whether to code live or tape a session for later coding. Then we discuss how the decision to use paper and pencil versus computer programs to record coding decisions can affect the rest of the measurement system. We provide an exercise that is designed to introduce readers to coding and to explore four behavior sampling methods of the same observation session. Finally, we provide recommendations regarding these considerations.

THE ELEMENTS OF A MEASUREMENT SYSTEM

We call the process by which we measure our independent and dependent variables and predictors a "system" because it has many elements that influence each other. In addition to the measurement context and a coding manual, a measurement system is composed of (a) a behavior sampling method, (b) a participant sampling method, (c) a session recording method, and (d) a coding decision recording method.

BEHAVIOR SAMPLING

Because observational measurement is expensive in terms of observer time and demanding in terms of observer attention, skill, and judgment, there are several options to divide the observation session in order to reduce the cost. These options make compromises regarding the amount of information and the type of information that is noted from the observation. The term “behavior sampling” is used to refer to these different options.

Behavior sampling methods are categorized by whether and how they divide the observation session. There are superordinate categories including “continuous,” “intermittent,” and “interval” sampling. Each is briefly introduced here and described in more detail below.

In continuous behavior sampling, there are no divisions made in the observation session and the entire observation session is coded. This is the most expensive, yet most complete way to code.

In intermittent behavior sampling, periodic intervals are observed and all instances of key behaviors occurring in that interval are coded. For example, we might observe 10 min out of every hour of an observation session per participant (Primavera, Allison, & Alfonso, 1997; Repp, Roberts, Slack, Repp, & Berkler, 1976).

In interval sampling or interval coding, the entire observation session is divided into a fixed duration of temporally defined intervals (e.g., 10 s), and the presence or absence (not number) of key behaviors in each interval is coded. Alternate versions of interval coding involve (a) coding the number of instances of the key behavior in each interval (relatively uncommon) and (b) observing for an interval (e.g., 10 s) and recording for a following, sometimes briefer, interval (e.g., 5 s). Each general sampling method described above (continuous, intermittent, interval) varies in accuracy depending on the different decisions made and each is described in more detail in the following sections.

Continuous Behavior Sampling

There are two types of continuous behavior sampling methods: *timed event* and *event*. Continuous timed event sampling requires indicating the time of occurrence of the onset (and sometimes, offset) of each instance of a key behavior. By *onset*, we mean the beginning of a behavior. When duration is to be estimated, the offset of each key behavior incidence is also recorded. By *offset*, we mean the end of a behavior. In chapter 3, we discussed making sure that one includes definitions for onset and offset in the segmenting rules for timed event coding. Unless these definitions are precise, it will be difficult to achieve criterion level of interobserver agreement for onset/offset times.

Usually, timed event coding is most accurately implemented when using a computer that is designed to automatically record the time of onsets and offsets for an event occurrence (i.e., observational software). By counting the number of seconds between onset and offset, the observational software computes the duration of each instance of the key behavior. Total or average duration of events can then be derived.

The second type of continuous behavior sampling is called *continuous event sampling*, which has also been called the *tally method.* This method requires counting or tallying the number of instances of each key behavior that occurs during the observation session. In its simplest form, the observer tallies the number of instances under each key behavior label without recording the time of onset. Offset is irrelevant to event sampling. Because offset is not noted, event sampling can be used to quantify number, but not duration. It is worth noting for future reference that "number" is also equivalent to the number of onsets of the key behavior. Although accurately recording the number of instances of a behavior does not necessarily require a precise identification of the time of onset (e.g., counting the number of tantrums a child has), clear thinking about "number" is aided by thinking in terms of onset (defining when a tantrum begins will help to differentiate between separate instances of tantrums and lead to a more accurate estimate of number of tantrums).

Intermittent Behavior Sampling

Intermittent behavior sampling can take the form of intermittent timed event and intermittent event sampling. It is among the cheapest methods of behavior sampling because the entire observation session is not coded but, rather designated time intervals within the session are observed to count key behaviors (e.g., 10 min out of 60 min). Like continuous event sampling, intermittent event sampling involves recording each occurrence of key behaviors that occur during the observed period. If intermittent event sampling is timed, the onset and/or offset of the event are recorded as well. This method was used in the 1970s, in part, because live coding was the most common coding method (perhaps because of concerns about reactivity to being recorded or expense and lack of widespread availability of recording equipment). However, intermittent sampling, whether it is event or timed event, is rarely used currently (Primavera et al., 1997). Its infrequent use is likely due to the greater availability of recording and computing technology and the greater known accuracy of another time-saving method: interval sampling (Repp et al., 1976).

Interval Sampling

There are three primary types of interval sampling methods: whole, momentary, and partial. Interval sampling has been and continues to be a commonly used method across a variety of disciplines. Thirty-four percent of all articles published in the journal *Child Development* in the 1980s used interval sampling (Mann, Have, Plunkett, & Meisels, 1991). Twenty-one percent of all observational studies published in the *Journal of Applied Behavior Analysis* between 1967 and 1977 used interval sampling methods (Kelly, 1977). There is a body of empirical work examining and evaluating the relative accuracy of interval sampling methods (Rojahn & Kanoy, 1985).

Whole interval coding involves coding the behavior "present" in an interval if the key behavior occurs during the entire interval. For example, to evaluate whether a student remains seated during whole class instruction, a 5-min observation session may be divided into 15-s time intervals, in which the focal student is required to maintain a complete sitting posture in the chair at his or her desk for the entire 15-s interval to be scored. Among the interval sampling family, whole interval coding is consistently found to be the least accurate method for estimating either duration or number (Powell, Martindale, Kulp, Martindale, & Bauman, 1977; Primavera et al., 1997; Rojahn & Kanoy, 1985). A related interval sampling method that is sometimes reported is when a key behavior is required for one-half or more of the duration of the interval to be coded. Although the logic does not perfectly fit into any of the three types of behavior sampling, the use of half the duration of the interval as a criterion duration for the event to be coded is closest to whole interval coding but the rule has functionally reduced the duration of the interval by half.

Momentary interval coding is used when the observer marks a behavior as present if and only if the behavior occurs at the boundary of the interval (e.g., the end of the interval). This means observers can only note what occurs when the interval boundary occurs. At least eight studies have tested the relative accuracy of the interval sampling methods to estimate duration and have found consistent results (Rojahn & Kanoy, 1985). Momentary interval sampling is the most accurate method for estimating relative duration (the number of intervals with 1 s/total intervals), particularly when interval duration is brief (e.g., 10 s), as compared to the known duration. One can show the mathematical relationship between the shortest interoccurrence interval and the optimal interval duration (Powell et al., 1977; Suen & Ary, 1989). However, a simple way to address the fact that momentary interval duration estimates are most accurate when the interval duration is brief is to use the shortest interval duration one can afford.

Partial interval coding means that the observer marks one and only one occurrence of a key behavior when the behavior occurs *anytime* during the interval. Five studies have evaluated the relative accuracy of the interval sampling methods to estimate number (Rojahn & Kanoy, 1985). Partial interval coding underestimates number when more than one event occurs during an interval. Repp et al. (1976) examined the relative accuracy of estimating number with partial versus momentary interval coding by examining the proportion of 10-s interval that had more than one event under six different key behavior conditions: (a) frequent (10/min) and clustered, (b) frequent and spread-out evenly, (c) moderate rate (1/min) and clustered, (d) moderate rate and spread-out evenly, (e) infrequent (0.1/min) and clustered, and (f) infrequent and spread-out evenly. Under all conditions, the partial interval coding was more accurate in estimating the number than momentary interval coding. The clustered versus spread-out condition, as studied, did not appear to affect the proportion of intervals with more than one event. However, under frequent (as defined by 10 events/min) occurrence conditions, more than 60% of the 10-s interval contained more than one event, resulting in gross underestimation of number in any interval sampling method.

A more complete view is that the accuracy of partial interval coding in estimating number is a function of a complex interaction between (a) interval duration (shorter intervals are better [in relation to accuracy]), (b) rate of key behavior occurrence (relatively low-rate events are better), (c) pattern of occurrence (spread-out events are better), and (d) average duration of the key behavior (shorter events are better) (Rojahn & Kanoy, 1985). These findings indicate that the currently available mathematical methods for estimating the error from partial interval coding when estimating number are generally too simplistic or too complex, depending on the situation (Hartmann & Wood, 1983). If one wants to estimate the number using partial interval sampling, one suggestion has been to select the interval duration based on the known characteristics of the key behavior by referring to empirically derived tables and figures in the Rojahn and Kanoy study. The problem with this suggestion is that characteristics of behaviors change over time or among participants. Importantly, almost all of the error in estimating number in the Rojahn and Kanoy study involved underestimating number. If underestimation of number occurs more in a particular single-subject design phase (e.g., baseline), group design condition (e.g., control) or levels of predictor, then using partial interval coding to estimate number can result in an increased probability of Type I error. If underestimation of number is distributed equally across design phases, groups, of levels of predictors, the probability of Type II error is increased.

How Does Interval Sampling Estimate Number and Duration?

The relative accuracy with which various types of interval sampling estimates the number or duration of a behavior matters because it affects the interval sampling method we select or even whether we select an interval sampling method at all. In general, if the behavior of interest is a short-duration behavior (e.g., under 1 s average duration), the underlying metric of interest is number. In contrast, when the behavior is thought to represent a state (e.g., attention) or behavior that has a long duration (e.g. on average 5 s), the underlying metric of interest may be duration. So, as we have stated from the beginning, it helps and is important to be clear about what your research question is and what it is you want to know in terms of your hypothesis and your knowledge of the phenomenon you study (tantrums, reading, language development, etc.). If you do not know that you implicitly wish to estimate duration, you may mistakenly select partial over momentary interval sampling. In contrast, if you are not aware that you implicitly wish to estimate number, you may not know that you will be fairly accurate in estimating number during a baseline in which the behavior of interest occurs infrequently and the interbehavior interval is long, but you may grossly underestimate number during the treatment phase when the number of the behavior is quite high and the interbehavior interval is brief. Such knowledge may lead you to select an entirely different sampling method in the first place (e.g., continuous behavior sampling method instead of an interval sampling method). In the following paragraph, we discuss briefly how interval sampling estimates number and duration.

When an interval is coded to show that a key event or behavior is present, that interval is coded as "1." When the interval is coded to show that a key event or behavior is absent, then it is coded as "0." There is no agreed-upon standard regarding how number estimates should be derived using interval data. Some researchers use the number of intervals with 1 preceded by intervals with 0 as an estimate of number (Suen & Ary, 1989). At first glance this seems sensible because number is the number of onsets and one cannot directly infer the number of onsets unless it is known that the preceding interval did not contain a key event. However, for observation sessions in which interoccurrence times are brief, such a method of estimating number may seriously underestimate true number (Rojahn & Kanoy, 1985). Alternatively, the number of intervals with 1 has also been thought to be an estimate of number (Rojahn & Kanoy, 1985). Experience leads us to prefer using the number of intervals with 1 as the estimation of number. The exercise at the end of this chapter illustrates why: using the number of intervals with 1 as the estimation of number better matches the number produced by continuous behavior sampling than does using the number of intervals with 1 preceded by intervals with 0. There is consensus that the number of intervals with 1 divided by the total number of intervals is the method of choice to represent relative duration (Powell et al., 1977; Suen & Ary, 1989).

PARTICIPANT SAMPLING

Participant sampling is the type of method used to decide which participant to code when there is more than one participant to be coded from a single observation session. The three most common types of participant sampling are as follows: focal, multiple pass, and conspicuous behavior. Multiple pass and conspicuous participant sampling are continuous behavior sampling methods. Focal sampling is an intermittent behavior sampling method.

Focal Sampling

Focal sampling involves coding one participant for a predetermined period, then coding a different participant in the group for the same period of time, and so on, until all selected participants have been coded. This can be done live without recording the session (i.e., videotaping) and also without computer software, but focal sampling reduces the length of the observed time for each participant (total observation session duration/number of participants observed). As mentioned earlier, there are many reasons why it is desirable to maximize the length of the observation time per participant when possible.

Multiple Pass Sampling

Multiple pass sampling involves selecting one participant and coding the entire session for only that participant. This process is possible when one has recorded (i.e., videotaped) the observation session for later coding. If the interaction of individuals needs to be analyzed, then one can use multiple pass sampling if one also uses a software program designed to link the codings of each individual via a time indicator that "resides" in the recorded version (e.g., videotape) of the observation session. One such computer program will be discussed in this chapter (Tapp, 2003). Multiple pass sampling is expensive, but it provides the most complete method of sampling multiple participants in the same observation session.

Conspicuous Sampling

Conspicuous sampling involves watching the entire group and noting which individual engaged in any predefined conspicuous behaviors. For example, one might note which participants engage in a sustained fight during a recess period. Interobserver agreement is likely to be acceptably high in such sampling for infrequent but salient behaviors (Suen & Ary, 1989).

REACTIVITY

One of the early criticisms of observational measurement was that the participants would act differently when watched than when they were not. A related issue is the hypothesis that people's behavior in the middle of observation sessions tends to be different from their behavior at the beginning (due to self-consciousness) or end (due to fatigue). There are studies testing both hypotheses. We bring up the issue of reactivity and studies of it in this chapter because some of the older suggestions in the literature regarding designing measurement systems are based on assumptions that reactivity is sufficiently important to sacrifice data or reliability. Despite reasonable questions about reactivity, the many studies on this issue indicate it is not nearly as serious an issue as it was first thought (Gardner, 2000). To inform our discussion, we briefly review the empirical literature from studies with human beings directly addressing the reactivity issue. There have been two primary ways to study reactivity: (a) compare the level of behavior under multiple observational conditions that vary in intrusiveness and (b) compare the level of behavior in different parts of the observation period.

Several studies have compared behavior levels under conditions of live observation (assumed to be the more intrusive) with conditions of unmanned recordings. These studies do not support the reactivity hypothesis (Fulton & Rupiper, 1962; Jacob, Tennenbaum, Seilhamer, Bargiel, & Sharon, 1994; Johnson & Bolstad, 1975; Kent, O'Leary, Dietz, & Diament, 1979). At least three studies have compared behavior levels under conditions of live observation with conditions of manned video recording. Which of these conditions is most intrusive, however, is open to interpretation. Two of the three studies found no difference in behavior between conditions (Christensen & Hazzard, 1983; Pett, Wampold, Vaughan-Cole, & East, 1992). One study examining many dependent variables found that one of their variables, aggression, was more frequent when videotaped (Pepler & Craig, 1995). If manned video recording is considered more intrusive than live recording, then this finding would support the hypothesis of reactivity. However, it should be noted that this is only one variable out of many that were tested.

Perhaps a more direct test of the reactivity hypothesis is a comparison of behavior levels during live observation versus observation conditions in which the participants did not know they were being observed. In one study, the unknown observation condition involved using a hidden recorder that was automatically activated over a 6-week period (Bernal, Gibson, William, & Pesses, 1971). No differences were found on any dependent variable tested. If reactivity does occur, it is probably more of a factor for some participants than others. For example, there is replicated evidence that fathers are more influenced by knowledge of being watched than mothers (Lewis et al., 1996; Russell, Russell, & Midwinter, 1992).

Several studies have compared the first part of the session with the middle and/or last part of a session. For the vast majority of the variables examined, there were no differences (Hughes, Carmichael, Pinkerton, & Tizard, 1979; Johnson & Bolstad, 1975; Kier, 1996). However, one study found a single variable, playing alone, to occur more in the first 10 min of the session than in the last 10 min of the same session (Kier, 1996).

In general, there is little evidence that reactivity is a large enough issue around which broad observational method policies should be made. Out of the 12 studies reviewed, all with multiple dependent variables, only 4 dependent variables were different between conditions thought to allow inferences relevant to testing reactivity. Additionally, in two of these four variables, there were condition differences in only a subset of participants. These studies lead to the conclusion that there is little empirical basis for (a) preferring live over taped observation due to reactivity concerns and (b) using only the middle section of a session due to habituation or fatigue.

LIVE CODING VERSUS RECORDING THE OBSERVATION FOR LATER CODING

Live or in situ coding occurs when the observer codes the behavior as it occurs. The observer is either present in the room with the participants or behind a one-way window during the observation session. Taping the observation for later coding can take the form of audio recording and video (with audio) recording. Audio or videotaped records can be converted to digital format, which in turn, can be used by computer software designed to control and "tag" particular segments of the tape (Bazeley & Richards, 2000; Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000; Tapp, 2003).

The primary advantage of live coding is the reduced cost of not requiring recording equipment or expertise. Another relative advantage of live observation over coding from recorded session is that sometimes the clarity of visual and audio signal is better from live observation. Finally, live observers may be able to move to a more advantageous viewing angle than a cameraperson with recording equipment. However, in our opinion, live coding does not always result in more "representative" records than taping the session for later coding, for reasons we outline below.

Not having a record of the observation session prevents a number of potentially useful approaches to coding. One cannot recode sessions. Recoding might be useful if, for example, the coding manual changes during the study because of problems with the definitions midway in a study. Recoding allows post hoc questions to be addressed. Live coding allows only single-pass sampling methods. In addition to multiple-pass participant sampling that was covered earlier, multiple passes of an observation session can be useful for complex coding systems that require the observer to focus on one dimension for one pass (e.g., all aggressive acts) and other dimensions for other passes (e.g., all prosocial acts). Some questions may require that the observer be blind to the audio aspects of the session when coding visual aspects. Multiple passes allow such blind coding. If live coding is used to implement an interval sampling method, typically every other interval is missed because they are reserved for looking at the score sheet or keyboard to record that an instance of the behavior occurred. This intermittent behavior sampling may be less accurate in estimating number or duration than continuous interval sampling (i.e., one without record intervals). "Stop-and-go" coding is not possible with live coding. Stop-and-go coding is often very helpful for complex coding manuals that require social judgment because multiple looks at the behavior may enable the observer to notice aspects of the scene that were missed the first time around. Finally, interobserver agreement checks are usually not fully independent when coding live. Independent coding for interobserver agreement checks are an important part of producing accurate data (see chapter 8).

The three primary disadvantages of taping observation session are (a) the expense of the recording equipment and the time to train personnel in using the equipment, (b) the potential loss of data due to technical dysfunction or operator error, and (c) the added time in extracting the data. However, we believe that the advantages of recording the observation session far outweigh these disadvantages. The advantages of taping and later coding are many. Such an approach allows multiple passes, stop-and-go coding, recoding many years after recording, and blind interobserver checks (i.e., the "primary" observer does not know which sessions are being checked for agreement). Additionally, duration can be more accurately coded when sessions are recorded because the precise onset and offsets are easier to determine with stop-and-go coding or slow-motion playback. If one converts the tape to a digital media file, the computer software programs can enable interval sampling without record intervals. One of the biggest advantages to recording the observation session is that it enables discrepancy discussions, which can prevent or correct observer drift, a topic covered in chapter 8.

RECORDING CODING DECISIONS

There are two primary ways to record coding decisions: paper and pencil or computer assisted. When we use paper and pencil during a live coding session, we are able to use a continuous event sampling method or an interval sampling method. If we are conducting continuous event sampling, we tally each occurrence of the event under appropriately labeled columns or boxes. If we are conducting interval sampling, we begin observing for an interval of time (e.g., 10 s) after hearing an audio signal that marks the onset of the interval. During the observation, we make mental notes of what we see. At the next audio signal, we indicate on the paper what we have seen by writing a letter or tally under the appropriately labeled column or box. With the next audio signal, we begin observing again. Alternatively, in the same recording format, the intervals can be scored as the event or behavior occurs without an "observe–record" interval combination. The choice between these two versions will be determined largely by the nature of what is being observed and observer skill.

When we use paper and pencil while playing back a taped record of the session, we *can* use any type of behavior or participant sampling method, but some are difficult to use without help from a computer. For example, we can implement timed event sampling by stopping the tape at the onset or offset and writing down the time of occurrence. The time of occurrence can be stamped on the tape or we might use a stopwatch. The reader has to only imagine stop-and-go coding to find precise onsets and offsets to understand why this method results in rapid breakage of the playback machine, fatigue in the observer, and imprecision in the duration estimates. Similarly, interval sampling with paper and pencil using a taped record is conducted as indicated for live interval sampling (i.e., observe and record intervals). Attempting to use stop-and-go coding for interval coding from a tape will usually result in slippage of audio and videotapes (i.e., the interval does not stop and start at precisely the same time each time it is played). Therefore, replaying the exact interval is not really possible without computer assistance.

When tapes of observation sessions are converted to digital files (i.e., media files), such files can be controlled by computer when used with the particular software designed for such purposes (Bazeley & Richards, 2000; Tapp, 2003). There are many examples of such software

(Hoch & Symons, 2004; Kahng & Iwata, 2000). Some of these programs are designed specifically for personal data assistants or pocket personal computers (PCs) (Tapp et al., 2006) or laptops (Tapp, 1995). These small computers allow live coding. Even when used for live coding, the computer software improves the accuracy of continuous timed event coding because onset and offset times are recorded automatically based on the computer's internal clock. Additionally, even when used to implement live coding, computer-assisted coding can improve interval sampling or event sampling because one can use touch typing to record differentiated behaviors via the use of different letters. Finding the different letters on the keyboard is a simple matter for a skilled typist. Computer records of what has been coded enables the subsequent use of analysis or counting computer programs, which reduces measurement error due to miscounting (Tapp, 1995). For example, we compared a software program designed for interval-sampled data with paper-and-pencil recording and found that the software program was more accurate and more efficient (set-up time, duration of data entry, duration of interobserver agreement calculations, and cost) than the traditional method (Tapp et al., 2006).

Using computer-assisted coding with a digital record of the observation session provides the most options for behavior and participant sampling and the most accurate implementation of these because it enables more accurate marking of onset and offset times, and training options that are not available from other methods of coding. The additional use of computer-controlled media files allows precise control of interval (i.e., virtually no slippage during replays of the same interval). This allows observers to do away with the "record" interval in interval sampling. That is, instead of observing for 10 s and recording what was observed for 5 s, the software that allows precise control of the media file allows precise stopping and starting at interval boundaries without slippage, thus eliminating the need for the record interval. This maximizes the observed time. It allows independent observers to see the same exact interval when conducting interobserver checks. It also allows precise replay of events on which observers disagree, which enables fruitful discussions regarding coding manual changes or observer training. It enables the rapid collecting of multiple examples that can be played back in succession to help an observer-in-training learn or relearn the concepts behind coding categories.

PRACTICE RECORDING SESSION

We have designed a brief exercise to (a) introduce the user to an example of a software program designed to assist coding and (b) compare the time to code, stress to code, and accuracy of four coding methods. We encourage you to try the exercise before moving forward because it will "bring to life" many of the issues we have been discussing. The exercise can be found on the book's website (www.springerpub.com/yoder/supplements). The file called "Procoder Manual for Exercise" will guide the reader through the exercise. A media file of an observation session ("smile.wmv") and a demonstration version of ProcoderDV, a software program that is designed to assist coding ("Procoder.exe"), is also on the website. Readers who intend to complete the exercise are asked to do so before reading the rest of the chapter. The sequence of opening the files on the website should be to download and set up the software ("Procoder.exe") first. When conducting the exercise, readers are asked to keep the following questions in mind:

1. What is the relative ranking of the four coding methods on time to code (1 = shortest)?
2. What is the relative ranking of the four coding methods on stress to code (1 = least)?
3. What is the cause of stress for real-time coding?
4. Does stress have anything to do with fuzzy beginnings and ends of smiles?

After completing the exercise, consider the following questions by referring to your data for all questions except for those involving timed event duration data. For timed event duration estimates and for readers who skipped the exercise (and missed all the fun), the following questions can be addressed by referring to the data provided in Table 4.1.

The data in Table 4.1 represent the results of the first author's coding and judgment on the coding of the provided media file. The length of the coded segment of the session was 2 min. The two behaviors coded were smiling (*s*) and nonsmiling (*n*). Smiling was chosen because it often has indefinite onset and offset. This is the type of behavior that can be well suited to interval sampling. In all four coding methods, the behaviors were coded as exhaustive categories (i.e., every time unit was assigned to either smiling or nonsmiling). A 10-s interval length was

Table 4.1

SUMMARY OF RESULTS OF THE BEHAVIOR SAMPLING EXERCISE

CODING METHOD	BEHAVIOR SAMPLING METHOD	RANK FOR TIME TO CODE	RANK FOR STRESS TO CODE	NUMBER OF NONSMILE PRECEDING SMILES	NUMBER OF SMILES	PERCENTAGE TIME OR INTERVALS WITH SMILE (i.e., RELATIVE DURATION)	ESTIMATE OF RATE OF SMILES
		1 = SHORTEST	1 = LEAST	NUMBER ESTIMATE #1	NUMBER ESTIMATE #2		
Real time	Timed event	3	4	13	13	53.5	6.5
Stop and go	Timed event	4	3	13	13	43.3	6.5
Partial interval	Interval sampling	2	2	0	12	100	6
Momentary interval	Interval sampling	1	1	3	7	70	3.5

selected to correspond with the method used in the Repp et al. study. A 10-s interval is common in many observational studies of humans using interval sampling. The "Practice Recording Session" questions are as follows:

1. Why were the interval sampling methods relatively easier (brief and relatively lower stress [i.e., less difficult]) when compared to timed event methods?
2. Was there much difference in the estimates of number and duration for stop-and-go versus real-time timed event coding?
3. With this simple code, was it worthwhile to do the stop-and-go timed event coding?
4. If this had been a more complex code, would it have been worthwhile to do the stop-and-go timed event coding?
5. What is the rate of smiles per minute using the timed event sampling methods?
6. How does this rate of occurrence compare with the frequent versus moderate rate conditions in the Repp et al. study on the accuracy of partial interval coding to estimate number?
7. When conducting partial interval coding, were there any intervals in which more than one smile occurred?
8. Would the number of intervals with more than one event been lower if the rate of occurrence of smiles had been less frequent (assuming they were spread out)?
9. Which is the better method for estimating number using interval sampling?
   - a The number of intervals with 1 preceded by an interval with 0.
   - b The number of intervals with 1?
10. Is the partial or momentary interval coding method better for estimating number?
11. How does this finding relate to what Repp et al. found about accuracy of estimating number?
12. Is the partial or momentary interval coding method better for estimating relative duration (percentage of time or intervals with smiles)?
13. How does this finding relate to the findings in the Powell et al. study?
14. How could you increase the accuracy of estimating relative duration using the superior interval sampling method?

RECOMMENDATIONS

We can now complete our recommendations regarding designing an observational measurement system. As before, these recommendations assume that resources are not a limiting factor (see Tapp et al., 2006).

1 We recommend that observers code as much of the session as is possible, not just the middle section of the session. Longer sessions tend to produce more stable and sensitive variable scores than do shorter sessions. Early concerns that early and late portions of sessions would yield notably different levels of behaviors have not been supported by empirical studies.

2 We recommend using taped records of observation sessions rather than live coding. Such a practice allows stop-and-go coding, repeated viewings of key scenes, multiple pass coding, and independent agreement checks.

3 We recommend converting the taped record to a digital media file and using a computer program to assist coding. Such a practice allows precise coding of onset and offset times, eliminates slippage of intervals without skipping intervals for recording coding decisions, and eases observer training.

4 If the observers are coding discrete events (i.e., those with clear beginning and ending), we recommend using timed event behavior sampling to enable the use of discrepancy discussions (see chapter 8) that can center about acts on which there is observer disagreement.

5 If the observer is coding events with "fuzzy" onsets and/or offsets, there is a legitimate argument for using either timed event or interval sampling. If the investigator selects interval sampling, then we recommend using interval sampling with the shortest affordable interval duration. If the investigator selects interval sampling with a brief interval, we recommend using partial interval coding to estimate number and momentary interval sampling to estimate duration. For number estimation, the number of intervals with 1 s is the estimate. For duration estimation, the proportion of intervals with 1 s is the relative duration estimate.

6 If there is more than one participant to be observed during a session, we recommend using multiple passes of the media file and coding the entire session.

REFERENCES

Bazeley, P., & Richards, L. (2000). *The NVivo qualitative project book*. London: Sage.

Bernal, M. E., Gibson, D. M., William, D. E., & Pesses, D. L. (1971). A device for automatic audio tape recording. *Journal of Applied Behavior Analysis, 4*, 151–156.

Christensen, A., & Hazzard, A. (1983). Reactivity effects during naturalistic observation of families. *Behavioral Assessment, 5*, 349–362.

Fulton, W. R., & Rupiper, O. J. (1962). Observation of teaching: Direct vs. vicarious experiences. *Journal of Teacher Education, 13*, 157–164.

Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review, 3*(3), 185–198.

Hartmann, D. P., & Wood, D. D. (1983). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification* (pp. 109–138). New York: Plenum.

Hoch, J., & Symons, F. J. (2004). Computer-assisted recording and observational software programs. In A. D. Pellegrini with F. J. Symons & J. Hoch (Eds.), *Observing children in their natural worlds: A methodological primer* (pp. 214–228). Mahwah, NJ: Erlbaum.

Hughes, M., Carmichael, H., Pinkerton, G., & Tizard, B. (1979). Recording children's conversations at home and at nursery school: A technique and some methodological considerations. *Journal of Child Psychology and Psychiatry, 20*, 225–232.

Jacob T., Tennenbaum D., Seilhamer, R. A., Bargiel, K., & Sharon, T. (1994). Reactivity effects during naturalistic observation of distressed and nondistressed families. *Journal of Family Psychology, 8*, 354–363.

Johnson, S. M., & Bolstad, O. D. (1975). Reactivity to home observation: A comparison of audio recorded behavior with observers present or absent. *Journal of Applied Behavior Analysis, 8*, 181–185.

Kahng, S. W., & Iwata, B. A. (2000). Computer systems for collecting real-time observational data. In T. Thompson, D. Felce, & F. J. Symons (Eds.), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 35–46). Baltimore: Paul H. Brookes.

Kelly, M. B. (1977). A review of the observational data collection and reliability procedures reported in the *Journal of Applied Behavior Analysis*. *Journal of Applied Behavior Analysis, 10*, 97–101.

Kent, R. N., O'Leary, K. D., Dietz, A., & Diament, C. (1979). Comparison of observational recordings *in vivo*, via mirror, and via television. *Journal of Applied Behavior Analysis, 12*, 517–522.

Kier, C. (1996). How natural is "naturalistic" home observation? *Proceedings of the British Psychological Society, 4*, 79.

Lewis, C., Kier, C., Hyder, C., Prenderville, N., Pullen, J., & Stephens, A. (1996). Observer influences on fathers and mothers: An experimental manipulation of the structure and function of parent–infant conversation. *Early Development and Parenting, 5*, 57–68.

Mann, J., Have, T. T., Plunkett, J. W., & Meisels, S. J. (1991). Time sampling: A methodological critique. *Child Development, 62*, 227–241.

Noldus, L. P., Trienes, R. J., Hendriksen, A. H., Jansen, H., & Jansen, R. G. (2000). The observer video-pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, and Computers, 32*, 197–206.

Pepler, D. J., & Craig, W. M. (1995). A peek behind the fence: Naturalistic observations of aggressive children with remote audio-visual recording. *Developmental Psychology, 31*, 548–553.

Pett, M. A., Wampold, B. E., Vaughan-Cole, B., & East, T. D. (1992). Consistency of behaviors within a naturalistic setting: An examination of the impact of context and repeated observations on mother–child interactions. *Behavioral Assessment, 14*, 367–385.

Powell, J., Martindale, D., Kulp, S., Martindale, A., & Bauman, R. (1977). Time sampling and measurement error. *Journal of Applied Behavior Analysis, 10*, 325–332.

Primavera, L., Allison, D. B., & Alfonso, V. C. (1997). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–90). Mahwah, NJ: Erlbaum.

Repp, A., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of number, interval and time-sampling methods of data collection. *Journal of Applied Behavior Analysis, 9*, 501–508.

Rojahn, J., & Kanoy, R. C. (1985). Toward an empirically based parameter selection for time-sampling observation systems. *Journal of Psychopathology and Behavioral Assessment, 7*, 99–120.

Russell, A., Russell, J., & Midwinter, D. (1992). Observer influences on mothers and fathers: Self-reported influences during a home observation. *Merrill-Palmer Quarterly, 36*, 263–283.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Tapp, J. (1995). *Multiple options for observation in experimental studies*. Nashville, TN: Vanderbilt Kennedy Center.

Tapp, J. (2003). *ProcoderDV*. Nashville, TN: Vanderbilt Kennedy Center.

Tapp, J., Ticha, R., Kryzer, E., Gustafson, M., Gunnar, M. R., & Symons, F. J. (2006). Comparing observational software with paper and pencil for time-sampled data: A field test of Interval Manager (INTMAN). *Behavior Research Methods, 38*, 165–169.

This page intentionally left blank

5 Common Metrics of Observational Variables

OVERVIEW

An observational variable is composed of the object of measurement (i.e., behavior or generalized characteristic) and the metric used to quantify the object of measurement. This chapter focuses on the most common classes of metrics used in direct behavioral observation, including number, duration, and proportions. These metrics are also referred to as *nonsequential* because the sequence of events or behaviors is not reflected in the unit of measurement for the variable. Sequential metrics will be covered in chapters 6 and 7.

After defining the term *metric*, we will discuss the concept of quantifiable dimensions of behaviors, including number, duration, and space. Much of the remainder of the chapter will cover the special considerations required when one uses proportion metrics. Using one metric instead of another can change the meaning of the key variables in a research question and thus often requires restating the research question. We present the rationale behind the common practice of transforming observational variables before conducting group parametric statistical analyses. We clarify that observational variables are not categorical, as is commonly assumed, but at least ordinal. Then we discuss that using observational variables in parametric analysis is perfectly acceptable. Finally, we conclude with recommendations regarding the selection of metrics for observational variables.

DEFINITION OF METRIC

A metric is the unit of measurement that indicates the level of a quantifiable dimension about a property of behavior or generalized characteristic. There is an important distinction among the property of an object of measurement, its quantifiable dimension, and its metric (Johnston & Pennypacker, 1993). A person who is running is displaying the property of motion. Speed is a quantifiable dimension of motion. A common metric for speed is distance/time (e.g., miles/hr).

QUANTIFIABLE DIMENSIONS OF BEHAVIOR

The most commonly measured dimensions of behavior in psychology, education, and related fields are presence, time, and space. When we measure presence or absence of a behavior, we often are interested in the number (or count) of instances or occurrences of a behavior or in a proportion derived from such counts. Technically, count or number is the number of onset–offset cycles of instances of the same category of behavior. Time units are typically applied to quantify three aspects of behavior: duration, latency or reaction time, and interoccurrence or interresponse time. Duration is the time from onset of a behavior to offset of the same instance of the behavior. Latency is the time from the offset of a behavior or event to the onset of a second, different behavior (e.g., the time from the starting shot of a race to the onset of a sprinter is the start latency for a sprinter). Interoccurrence or interresponse time is the time from the offset of the first occurrence of an event to the onset of a second occurrence of the same type of event (e.g., the time between communication acts). When applied to behavior, the time units are usually seconds or minutes but can be longer.

When referring to behavior, space is typically quantified as either direction or distance (i.e., proximity). Direction involves the vector of movement (e.g., the runner was moving in a southwest direction). The metric for direction is usually degrees on a compass (e.g., he was running 17 degrees north-northwest). Distance or proximity describes how far (or near) the entity being measured is from another salient entity (e.g., John is physically close to Jim.). The metric for distance can be in either the International System of Units (SI) units (e.g., meters) or Imperial units (the English or U.S. system; e.g., yards). Almost all scientific literature uses SI units.

PROPORTION METRICS

Proportions can be created to measure behavior and its context by combining the aforementioned metrics. The most common of these is rate (i.e., number of events or occurrences of a behavior/unit of time). For example, it is sometimes meaningful to know not only that John communicated three times but also the duration of the session in which the three times were observed (e.g., 3 times/min vs. 3 times/hr). Speed (i.e., distance/time) has already been discussed as a common metric quantifying motion. Accuracy or consistency (i.e., number of correct responses or occurrences of goal behavior/number of opportunities for correct response or attempts at goal behavior) is commonly used in behavioral observations. For example, one might want to know both that John answered 4 quiz questions correctly and that 10 questions were asked (4/10 = .4). Style (i.e., number of key behaviors directed to another person/number of all coded behaviors directed to another person) is less common than the above but is used in studies of naturally occurring individual differences in how adults speak to children. For example, Betty talked about her child's focus of attention 40 times and talked to her child 100 times (40/100 = .4). There are other proportion metrics (e.g., Siller & Sigman, 2002), but in our experience the above four types are the most commonly used in observational measurement.

Proportion Metrics Change the Meaning of Observational Variables

When we use proportion metrics, the meaning of the observational variable changes from what it would be if we used number, time, or space dimensions alone. Therefore, once we have selected a proportion metric, it often increases the precision and clarity of our research question to revisit it and, if necessary, rewrite it on the basis of the specific selected metric.

Unfortunately, to increase the efficiency of communication, investigators often omit any consideration of metric when stating research questions. For example, an investigator might state her question as "Is maternal verbal responsiveness to child communication positively related to child productive vocabulary?" The omission of the metric leaves important information unspecified and makes the question less falsifiable. When reading such a question, many readers might assume that the metric involved a single dimension (e.g., number). Thus the implicit research question many readers understand might be "Is the number of maternal verbal responses to child communication positively related to child productive vocabulary?" However, because mothers cannot verbally respond unless the child communicates, the investigator might have meant "Is the proportion of child communication to which a mother verbally responds [i.e., consistency] positively related to child productive vocabulary?"

Alternatively, because mothers who talk frequently to their children may give more verbal responses to child communication, the investigator might have meant "Is the proportion of child-directed maternal talk that is a verbal response to child communication [i.e., style] positively related to child productive vocabulary?" Although these distinctions may seem trivial or too nuanced, we think they are actually quite important because the specificity of the research question or hypothesis is related directly to subsequent analyses and, ultimately, the falsifiability of the hypothesis. When the question is clear, the choice of analysis is facilitated.

When guided by the results of correlational studies, treatments often attempt to increase the value of the metric for the positive predictor of the desired outcome. The different metrics can have different ramifications for treatment or intervention approaches designed to increase the putative causal variable (e.g., maternal verbal responses). If number were the metric, then we could increase maternal verbal responsiveness by (a) teaching children to communicate more often and (b) asking mothers to notice and respond to more of their children's communication. If consistency were the metric, then we would increase maternal verbal responsiveness only by asking mothers to notice and respond to more of their children's communication. If style were the metric, then we could increase maternal responsiveness by (a) asking the mother to reduce the "other" talk she directed to her child, (b) teaching the child to communicate more often, and (c) asking the mother to notice and respond to more of her child's communication.

Empirically, the metric can influence the results of a study greatly. In a recent study, we examined the relative predictive power of three metrics of maternal talk about children's focus of attention in relation to later productive child vocabulary (McDuffie, Yoder, & Krause, 2008). The three metrics for the maternal talk variable were number, consistency, and style. The relative predictive relation to later child productive vocabulary was 0.44, 0.28, and –0.02 for number, consistency, and style, respectively. As the results show, it is important to communicate very clearly the metric of observational variables. This clear communication should be reflected in the research question by including the metric of the observational variable(s) and in the methods section by indicating explicitly the numerator and denominator of any proportion metrics used.

Scrutinizing Proportions

As discussed above and hopefully illustrated by the examples, proportions are intended to clarify the meaning of the numerator by equating or controlling a variable that affects the numerator. It has been noted, however, that when a proportion variable increases, we do not know whether it is because the numerator increased, the denominator decreased, or both (Johnston & Pennypacker, 1993). For example, if a style proportion (e.g., number of a facilitating maternal behaviors directed to a child/number of all coded child-directed behaviors) increases, we do not know whether the number of facilitating behaviors increased or whether the other child-directed behavior decreased. Thus, it may be worthwhile to consider whether a proportion metric is required, and the remainder of this section is devoted to an in-depth examination of proportion metrics.

Rate is a proportion and is often used to equate the duration of the observational session. Accuracy or consistency is often used to equate the number of opportunities to respond or the attempts at a goal behavior. Individual style is often used to equate total behaviors directed to others.

The general issue is whether (and when) it is necessary to convert a number based on count data to a proportion as a metric for subsequent analysis. Conventional wisdom in many, if not most, areas of behavioral research relying on direct observational count data is immediately to convert count data to proportions. Although we recognize the value in doing so related to the prevailing convention (discussed below), we want to draw attention to the idea that there are probably circumstances in which an automatic conversion is not necessary or desirable.

For starters, the degree to which proportion metrics are necessary is dependent on the degree to which there is a need to control for differences across subjects or sessions in some aspect of the observational context (e.g., observation time). To illustrate the issues, assume two cases. If one student talks out 5 times during 10 min of observation and another 10 times during 5 min of observation, current convention leads us automatically to convert number to rate per minute (.5 talk-outs/min and 2 talks-outs/min, respectively) as a common metric. The reason we feel the need to equate or control the denominator is that we believe it is influential on the numerator. If the property that the denominator quantified were irrelevant to interpreting differences in the numerator, then there would be no compelling reason to equate it. That is, it would be considered a noninfluential variable. When we say that the property of the denominator influences the interpretation of the numerator, we are saying there is an implicit functional or causal relation between that property and the property quantified by the numerator.

An Implicit Assumption of Proportion Metrics

The following quotation indicates an important assumption underlying proportion metrics that is not widely recognized by direct observational research conventions:

> For division by the denominator to be an appropriate method for qualifying the numerator, [there is] an implicit assumption that the correlation of the numerator with the denominator is perfectly linear, or nearly so, and that the regression line goes through the origin. Otherwise, one is either over- or underadjusting or adjusting nonuniformly over the range [of the denominator]. (Cohen & Cohen, 1984, p. 265)

In the common proportion metrics used in observational measurement, the assumption is that the numerator has a perfect *positive* relation with the denominator. Figure 5.1 illustrates this assumed relation. To understand the assumption behind proportion metrics, try replacing the word "numerator" with the real number of interest and the word "denominator" with the variable that the proportion is designed to equate or control. For example, when we compute rate of communication, we are assuming that the duration of the observation session is positively related in a linear fashion with the number of communication acts.

The notion that an assumption is made when creating a metric may be antithetical from an idemnotic measurement perspective. After all,

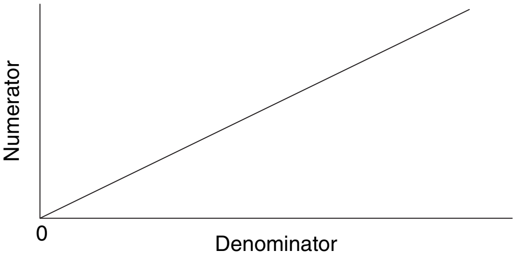


Figure 5.1 Illustration of the relation between numerator and denominator in proportion metrics.

a key rationale underlying the idemnotic measurement framework is to avoid unnecessarily complex mathematical manipulations. Such manipulations are avoided wherever possible so that the relation between the measure (including the metric) and the thing or phenomenon being measured is as transparent as possible (Haynes & O'Brien, 1999). Therefore, it is worth elaborating this point. The key point of contention is whether testing the above-stated assumption is an unnecessary complication or whether it is a necessary and warranted condition when using proportion metrics.

It is worth considering instances in which the property quantified by the denominator either does not influence the property quantified by the numerator or does so in a negative direction. Consider a situation in which one observation session took twice as long as the next because the examiner had to administer 10 probe activities regardless of how long it took to do so. It was more difficult to do so in the first than in the second session because the child was acting out in the first session but not in the second session. Further assume that child compliance with the probe activities is negatively associated with acting out. In such a case, the number of instances of the key behavior is likely to be negatively related to the duration of the session. In this example, it is less clear that rate is the best choice for the metric. Equating for duration of the session does not clarify the meaning of variability in number of communication acts because children who are observed for longer do not communicate more, in fact; they communicate less than children observed for shorter sessions.

Testing Whether the Data Fit the Assumption of Proportion Metrics

In single-subject and in group research designs, the assumption that there is a linear relation between the numerator and the denominator can be tested. Assuming there are sufficient data points (e.g., at least five), the test can be the Pearson product–moment coefficient for the association between denominator (e.g., duration) and numerator (e.g., number). The unit of analysis for group designs is the participant. The unit of analysis for single-subject designs is the session with each participant. One would conduct the test for each participant in a single-subject design and use the conclusion from the majority of the participants.

In application, it may be more useful to emphasize effect size, not statistical significance, when interpreting the results of such analyses.

For example, one might consider a correlation coefficient greater than .2 (i.e., between Cohen's benchmarks for "small" and "moderate" associations) as evidence the data fit the assumption. Similarly, a correlation coefficient of less than –.2 might be taken as evidence that the count represented in the numerator is negatively related to the denominator. A relation between .2 and –.2 might be considered evidence that no important relation between the numerator and denominator exists.

In single-subject designs in which there are insufficient data or when the investigator is uncomfortable with statistical analyses, the test could be logical. One could use reason to determine whether it is probable that higher denominator scores are likely to produce higher numerator scores because the former provide more opportunity for the latter. When conducting this test, it is useful to attempt to produce as good an argument as possible for both positive and negative relations. When the two arguments are equally convincing, we might conclude that different relations can exist for different sessions and that the assumption has not been met. When one argument is more persuasive for more participants, we might conclude that the relation is likely consistent with the more plausible argument. If greater denominator scores are most likely to produce or covary with lower numerator scores, then we might conclude that the data do not fit the assumption. If greater denominator scores are most likely to produce greater numerator scores, then we might conclude that the data are likely to fit the assumption.

Consequences of Using a Proportion When the Data Do Not Fit the Assumption

Technically, the stated assumption must be met for within-group and across-group data in a group comparison design and for within-phase and across-phase data in a single-subject design. The effect size of an experimental versus control group comparison on the proportion metric should be more accurate when the assumption is met than when it is incorrect. By generating a number of fictitious data sets that vary on the sign of the correlation between numerator and denominator (e.g., duration and number) within each group, we can compare the effect size when the assumption is met with the effect size estimates when the assumption is not met. Doing so allows us to demonstrate the consequences of using a proportion metric when the assumption is not met.

We used these principles to generate data sets in which the posttreatment number of the key behavior varies but is on average the same between experimental and control groups and the posttreatment observational session duration varies but has a shorter mean duration in the experimental group. We modeled situations in which the proportion variable could be and often was above 1.0 to provide an example that simulates rate metrics. We also modeled situations in which the proportion variable had to be between 0 and 1.0 inclusive to provide an example that simulates accuracy and style proportion metrics. We recognize that this is not a full-fledged simulation study because we have (a) used a small sample size within each experiment, (b) generated only one example of each condition, (c) investigated only two levels of within-group correlation (i.e., –.8 and .8), and (d) generated only three levels of across-group correlation (i.e., –.55, 0, and .55). However, this demonstration study of generated data is offered to illustrate that the relation between the numerator and denominator affects rates and other proportions equally and to demonstrate the consequence of violating the assumption of proportion metrics. We hope that this illustration will stimulate more comprehensive investigation into this important area of study.

Table 5.1 presents the results of this fictitious demonstration study. The bottom line of the study indicates that when the assumption is met within both groups and across groups the standard deviations for rate for both groups is relatively small, and the effect size for the between-group contrast of rate is relatively large. When the correlation between duration and number is negative within both groups and across groups, the standard deviation for rate is relatively large, and the effect size for the between-group contrast of rate is relatively small. When the correlation between duration and number is negative for one group and positive in the other group, the effect size is in the middle of the other two conditions. To determine whether this pattern holds for proportion metrics that can only be between 0 and 1.0 (e.g., accuracy and style), we added 5 to the numbers in the "duration" column to create a "denominator" column and divided the "number" column by the "denominator" column to create a new proportion variable. The pattern of results was substantively identical to that given for rate proportions.

The raw data for this set of experiments are in an Excel spreadsheet on the book's website (www.springerpub.com/yoder/supplements) entitled "Data used for proportion simulation study." The SPSS syntax for the simulation is provided in another file on the website, entitled "SPSS syntax for the proportion simulation." We encourage readers to import these data into SPSS (or some other statistical program) to confirm the results of the simulation (i.e., Table 5.1). Doing so is likely to

Table 5.1

RESULTS OF DEMONSTRATION SIMULATION STUDY ON PROPORTION METRICS WHEN MEAN NUMBER IS EQUAL BETWEEN GROUPS BUT DURATION IS SHORTER IN THE EXPERIMENTAL GROUP

	r FOR ASSOCIATION BETWEEN DURATION AND NUMBER			EXPERIMENTAL GROUP		CONTROL GROUP		
CONDITION	WITHIN EXPERIMENTAL GROUP	WITHIN CONTROL GROUP	ACROSS GROUPS	MEAN	SD	MEAN	SD	COHEN'S *d*
Proportion can be >1	0.8	0.8	0.55	1.1	0.1	0.86	0.07	2.8
Proportion can be >1	–0.8	–0.8	–0.55	1.1	0.29	0.87	0.2	1
Proportion can be >1	0.8	–0.8	0	1.1	0.1	0.87	0.2	1.45
Proportion can be >1	–0.8	0.8	0	1.2	0.29	0.86	0.07	1.61
Max proportion is 1.0	0.8	0.8	0.55	0.75	0.06	0.63	0.05	2.17
Max proportion is 1.0	–0.8	–0.8	–0.55	0.76	0.17	0.64	0.13	0.79
Max proportion is 1.0	0.8	–0.8	0	0.75	0.06	0.64	0.13	1.09
Max proportion is 1.0	–0.8	0.8	0	0.76	0.17	0.63	0.05	1.04

improve readers' understanding of what was done and the ramifications of the results.

Because we have shown that the consequences of violating the assumptions for rate are identical to those of doing so for other proportions, the next demonstration study was done only for rate. In the first demonstration study, there were no count differences between groups. This time we created data that would seem to favor one group if number was the metric used and would favor the other group if rate was the metric used. Specifically, the standardized difference between groups for number was –1.9, a very large effect size, favoring the control group. But when rate was the metric used, the standardized difference between groups favors the experimental group because the duration (i.e., the denominator for rate) was shorter in the experimental group. When the correlations of duration and number were positive (.8) in the control and experimental groups, the standardized mean between-group difference in rate was d = .42 (a small to moderate effect size). Because of the way we labeled the groups, the direction of the difference in rate favors the experimental group. Remember that if number is used as the metric, the difference favors the control group (by a great deal). Obviously, the computer does not understand group labels. We could just as easily have labeled the group with the higher mean number metric the experimental group and labeled the group with the higher mean rate metric (i.e., the shortest duration) the control group. Table 5.2 presents the results. The pattern of the effect sizes is similar to that provided in Table 5.1. Again, the effect size for between-group differences in rate is largest when the assumption for the proportion metric is met for both groups (i.e., the correlation between numerator and denominator is positive and strong). The effect size is smaller for rate when the assumption is violated in at least one group.

It is clear from these demonstration trials that at least three factors influence validity of proportion metrics: (a) which group's numerator (e.g., number) is higher, (b) which group's denominator (e.g., duration) is higher, and (c) whether the within-group correlation between numerator and denominator is positive and strong within both groups. The same is logically true for between-phase differences in single-subject designs. Future simulation studies are needed to specify these conditions in a more precise manner. However, it is clear that there are data patterns in which both the direction (i.e., which group or phase is superior) and magnitude (i.e., the amount of between-group difference) of group and phase differences will be quite different depending on whether rate or number is used and depending on whether the assumption for proportions is met.

Table 5.2

RESULTS OF DEMONSTRATION SIMULATION STUDY ON RATE METRICS WHEN NUMBER IS LARGER IN THE CONTROL GROUP (*d* = 1.9) AND DURATION IS SHORTER IN THE EXPERIMENTAL GROUP (*d* = 1.9)

r FOR ASSOCIATION BETWEEN DURATION AND NUMBER			EXPERIMENTAL GROUP		CONTROL GROUP		
WITHIN EXPERIMENTAL GROUP	WITHIN CONTROL GROUP	ACROSS GROUPS	MEAN	SD	MEAN	SD	COHEN'S *d*
0.8	0.8	0.91	1.19	0.11	1.15	0.08	0.42
–0.8	–0.8	.154	1.22	0.31	1.16	0.23	.22
0.8	–0.8	.53	1.19	0.11	1.16	0.23	.17
–0.8	0.8	.53	1.22	0.31	1.15	0.08	.31

On the basis of these initial results, it is our opinion that there are likely situations in which it makes more sense to use number than a proportion metric, even when the values in the denominator of the proportion vary across sessions or participants.

ALTERNATIVE METHODS TO CONTROL NUISANCE VARIABLES

Statistical Control

When using a group research design, an investigator may choose to control for nuisance influential variables such as duration of the session or opportunities to respond using a statistical method (e.g., analysis of covariance). Although it is beyond the scope of this book to discuss such methods in detail, they generally involve a process that is analogous to a two-step analysis: (a) derive the residuals of the association between the dependent variable and the covariate and (b) use these residuals as the dependent variable in the analysis to test the research question. In this context, residuals are equal to the predicted number of instances of the key behavior minus the observed number of instances of the key behavior. Just as meeting the assumption of the method of control for one group but not for another is problematic for proportion metrics, analysis of covariance requires that the relation between the nuisance variable and dependent variable is not statistically different between groups. If there is a significantly stronger relation in one group than in another, we have violated an assumption of the statistical procedure that controls for the nuisance variable (i.e., the assumption of homogeneity of covariance).

If the assumption of homogeneity of covariance is met, then using such a statistical method to control for nuisance variables is a reasonable alternative to using proportion metrics. It can be argued that because statistical methods of control quantify the degree to which there is a linear association between the nuisance variable and the dependent variable, instead of assuming a *perfect* relation, they are superior to proportion metrics even when the assumption for the proportion metric is generally met (i.e., a positive, albeit imperfect, association between duration and number in all groups and across groups).

Procedural Control

There are at least two reasons to consider controlling for nuisance variables using elements of the measurement system other than metric. First, both statistical control and proportion metrics assume a linear relationship between the key behavior and the nuisance variable. Such a simple relation is not always present. Controlling for the nuisance variable using elements of the measurement system will alleviate the need for either statistical control or proportion metrics. Second, the investigator may wish to sidestep the controversy regarding whether a proportion metric should be used. This controversy can be avoided if the denominator for the proportion being considered is constant across analysis units.

Duration and opportunity-to-respond differences can be controlled by structuring the measurement context. For example, the session administrator can provide the same number of opportunities to respond and keep sessions the same duration across sessions or participants. In chapters 1 and 2, we weighed the advantages and disadvantages of structuring the measurement context and concluded that there are situations in which doing so is very useful. Alternatively, the observer can be directed to code the same duration of the session for all sessions or participants. Practically, this means shortening the coded section of the observation session to the length of the shortest session in the study sample. In chapter 4, we considered and discarded this option, however. Controlling total behavior directed to a participant when attempting to measure individual style using various elements of the measurement system is generally not feasible.

TRANSFORMING METRICS OF OBSERVATIONAL VARIABLES IN GROUP STATISTICAL ANALYSES

Even though using a proportion or number (i.e., count) in a parametric analysis is often acceptable in terms of the assumptions underlying the statistical analysis method, doing so without transforming can result in loss of statistical power (i.e., an increased probability of Type II errors). The scores for a proportion are generally not normally distributed. Many parametric analyses are more powerful when the distributions of variables are approximately normal. This can be seen in Figure 5.2, which presents a distribution of scores from .002 to .998 incremented by .002 for each score.

In group statistical analyses, it is common to transform proportion metrics into a scale that better approximates the normal curve. This is particularly common when a large proportion of the scores are below .25 or above .75. The most common transformation for proportions is

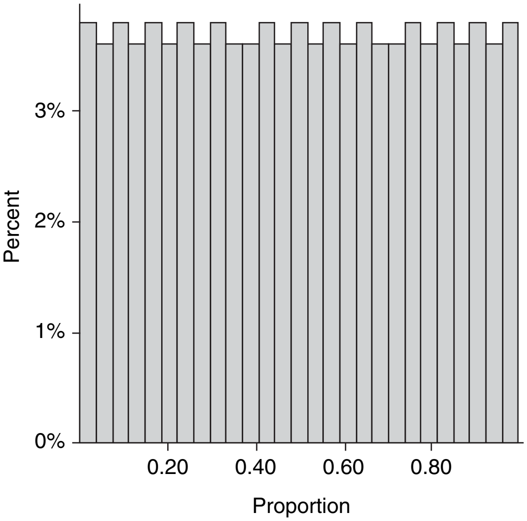


Figure 5.2 Illustration of a distribution of proportion scores from .002 to .998 incremented by .002.

the arcsine transformation (Cohen & Cohen, 1984). An Excel spreadsheet entitled "Arcsine transformation" is provided on the book's website (www.springerpub.com/yoder/supplements). The formula used to transform proportions is provided in the cells under the arcsine column label. Those less interested in the exact formula should note that the transformation involves more than simply taking the arcsine of the proportion. Figure 5.3 provides the distribution of the arcsine transformation of the proportion values illustrated in Figure 5.2. Comparing the two distributions makes it clear that the arcsine transformation provides a much better approximation of the normal curve than does the distribution of nontransformed proportion scores.

Similarly, number metrics are also frequently skewed. This is particularly true when we measure infrequent behaviors, as is commonly the case in observational research. In such cases, the distribution tends to be positively skewed (i.e., there are more cases on the left side of the distribution than on the right side of the distribution). One of the most commonly used transformations for variables with number metrics is the square root transformation (Cohen, Cohen, West, & Aiken, 2002). It should be noted that taking the square root of 0 and 1 does not change these values. Therefore, to treat 0 and 1 the same as other count values, we need to add 2 to each value before taking the square root (i.e., square root of [y + 2]). Doing so generally equalizes the variance, reduces skew, and linearizes relations with other variables (Cohen et al., 2002).

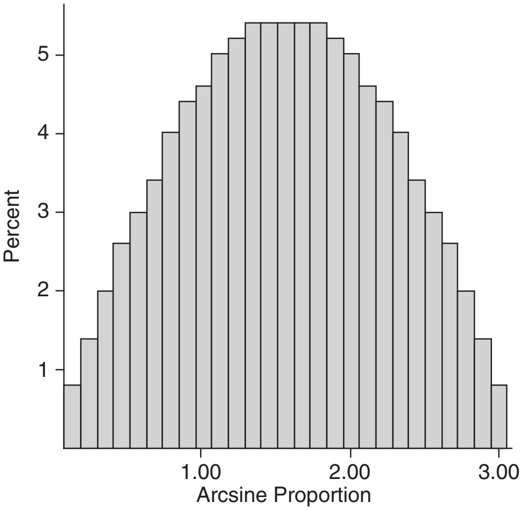


Figure 5.3 Illustration of a distribution of arcsine transformed proportion scores from .002 to .998 incremented by .002.

A discussion of transformations would not be complete without acknowledging that some investigators prefer not to transform scores because doing so reduces the clarity of communication and complicates interpretation of results. Next we discuss the scale of measurement of observational metrics, including transformed ones.

SCALES OF MEASUREMENT FOR OBSERVATIONAL VARIABLES

In our opinion, there is a long history of mislabeling observational variables as "nominal," probably because the act of coding what we are measuring is a nominal decision (Stevens, 1951). However, the metrics, and thus the object to which scales of measurement should be applied, for observational variables are most commonly a number metric (i.e., count), a time metric, a spatial metric, or a proportion metric. None of these is a nominal scale.

Using a modification of the classic questions used to identify scales of measurement, one can identify the measurement scale of observational variables (Bakeman, 2000). The questions are as follows:

I "Are intervals between values on the scale ordered?"
  - **A** If no, then the metric is nominal.
  - **B** If yes, then ask, "Are the intervals between values on the scale equivalent?"

1 If no, then the metric is ordinal.
2 If yes, then ask, "Does zero indicate that none of the quantifiable dimension occurred?"
  i If not, then the metric is interval (e.g., zero degrees does not indicate no temperature).
  ii If yes, then the metric is ratio.

If we use these questions to guide us, it becomes clear that variables with any of the metrics we have discussed are at least ordinal scales. Clearly, each of the metrics we have discussed has a zero score that indicates none of the quantifiable dimension. The key question is whether the intervals between values indicate the same amount of the dimension being measured. Regarding the scale of measurement for variables with nonproportion metrics (e.g., number), the scale depends on whether one is measuring a context-dependent behavior or a generalized characteristic.

When context-dependent behaviors are measured, it is easy to argue that the answer is yes: The difference between intervals on scales is equal regardless of where on the scale one looks. For example, the difference between 1 and 2 is the same as the difference between 9 and 10 (i.e., 1). This matches the idemnotic concept of measurement because the units of measurement have the same meaning regardless of where they occur on the scale. Therefore, when nonproportion metrics (e.g., number, duration) are used to measure context-dependent behaviors, the scale of measurement is a ratio scale.

When nonproportion metrics are applied to quantifying dimensions of behavior that are thought to be signs of a generalized characteristic, it is doubtful that the amount of generalized characteristic represented by the interval between values at the beginning of the scale equals that represented by the interval between values at the middle of the scale. With regard to whether a child tends to be aggressive, it is more meaningful to note that a child changes from 1 to 0 hits than it is to note that a child changes from 50 to 49 hits. Therefore, it is recommended that one consider nonproportion metrics for variables intended to represent generalized characteristics as ordinal scales.

When we are dealing with proportions, the scale is clearly ordinal. The difference between .01 and .05 is "more important" than the difference between .48 and .52, even though both intervals are separated by .04. The former is a 400% increase. The latter is an 8% increase.

Finally, all transformations of observational metrics (e.g., number or proportion) should be considered ordinal scales. That is, the intervals between transformed scales of number or proportion are, by definition, not equal, because the transformation normalizes distributions by changing the interval between values on the original scale (i.e., number or proportion) at the extreme of the original distribution of scores.

OBSERVATIONAL VARIABLES IN PARAMETRIC ANALYSES

It is useful to discuss the scale of measurement for observational variables because of the long history of assertions that group parametric statistical procedures should not be used with variables that have ordinal status (e.g., Stevens, 1951). This is not necessarily the case. Simulation studies have shown that most statistical procedures are robust to violations of measurement scale assumptions (Harris, 2001). The reason is that statistical procedures cannot tell whether the numbers in the data set are ordinal, interval, or ratio. If transformations of variables that originally had number or proportion metrics results in normalizing the distribution of residuals and one simply wants to speak to the statistical significance of associations, differences or change, as opposed to some more precise prediction of particular values, then parametric statistical results from ordinal-scaled data are interpretable.

RECOMMENDATIONS

We recommend using the metric that best fits the theory underlying the research question, assuming that the assumptions of the metric are met. It should now be clear that there are instances in which automatically converting number to proportion may not be necessary. By discussing a critical assumption underlying proportion metrics (i.e., a positive linear relation between the numerator and denominator), we wanted to make readers aware that there are instances in which using proportion metrics when this assumption is not met can result in Type I and Type II errors. We are not suggesting that proportion metrics should be avoided. Instead, we hope that this chapter has demonstrated the value of the conditional use of proportion metrics when the key assumption underlying proportions has been met. Regardless of whether the original metric is a proportion, transformations of the original metric may increase the statistical power of a group analysis for the observational variable of interest. Finally, we see no reason that observational variables cannot generally be used in parametric analyses provided general statements about the presence of associations, group differences, or changes are the goal of the research.

REFERENCES

Bakeman, R. (2000). Behavioral observation and coding. In H. T. Reis & L. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 138–159). New York: Cambridge University Press.

Cohen, J., & Cohen, P. (1984). *Applied multiple regression*. Mahwah, NJ: Erlbaum.

Cohen, J., Cohen, P., Aiken, L. S., & West, S. G. (2002). *Applied multiple regression: Correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Harris, R. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Erlbaum.

Haynes, S. N., & O'Brien, W. H. (1999). *Principles and practice of behavioral assessment*. New York: Kluwer.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

McDuffie, A., Yoder, P., & Krause, K. (2008). *Parent verbal responsiveness predicts language in preschoolers with autism*. Paper presented at the Symposium for Research in Child Language Disorders (June; Madison, Wisconsin).

Siller, M., & Sigman, M. (2002). The behaviors of parents of children with autism predict the subsequent development of their children's communication. *Journal of Autism and Developmental Disorders, 32*, 77–90.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 28–42). New York: Wiley.

This page intentionally left blank

6 Introduction to Sequential Analysis

OVERVIEW

Examining the sequence of behaviors within an observation session to infer whether the occurrence of one behavior is associated with the immediate or subsequent occurrence of another behavior is referred to as *sequential analysis*. Specifically, sequential analysis is concerned with the sequential or simultaneous occurrence of coded behaviors or of seconds or intervals in which a coded behavior has occurred within an observation session. The overall purpose of this chapter is to distinguish between sequential and nonsequential variables, to discuss the important and frequently misunderstood issue of exhaustive coding spaces (which is assumed by sequential analysis), to describe three major types of sequential analysis, to review contingency table construction, and to discuss issues concerning a common metrics used in sequential analysis (transitional probabilities).

We devote two chapters to sequential analyses because although it is an intuitively appealing approach, there are many common errors that can occur including (a) using event-lag sequential analysis when time-window analysis would be better, (b) using sequential analysis for sequences in which the target behavior can "only" occur after the proposed antecedent behavior, (c) using transitional probabilities when another index of sequential association would be better, and (d) trying to use sequential analysis for sequences that do not occur “enough” during the session to accurately estimate chance sequencing. This chapter will address the first two of these issues and chapter 7 will address the last two issues. Chapters 6 and 7 are written for the analysis of two-event sequences. Analyzing longer event sequences involve more complicated methods and are beyond the scope of the two chapters.

DEFINITIONS OF TERMS USED IN THIS CHAPTER

The following terms will be consistently employed here and in chapter 7 (see Yoder & Feurer, 2000, for more detailed definitions). To help us discuss the sequence of interest, we will call the hypothesized causal behavior, the hypothesized prompt, or hypothesized discriminative stimulus the *antecedent behavior* (sometimes referred to, in other sources, as the *given behavior*). The *target behavior* is the behavior hypothesized to be affected by the antecedent.

For example, when asking whether there is a sequential association between student–teacher instruction preceding student topic-continuing utterance, teacher instruction is the antecedent behavior and student topic-continuing utterance is the target behavior. A certain number of antecedent-preceding target behavior sequences will occur by chance. Therefore, to correctly interpret an index of sequential association, we need some quantification of the chance occurrence of the sequence of interest. We will call this value the *estimate of chance occurrences of the sequence.* Conceptually, a *sequential association* occurs when the antecedent precedes target behavior more or less often than would be estimated to occur by chance. An *index of sequential association* is a numeric expression for the sign (i.e., positive vs. negative) and magnitude of the sequential association.

SEQUENTIAL VERSUS NONSEQUENTIAL VARIABLES

The types of research questions that are appropriately addressed by sequential analysis are clarified by understanding the difference between *nonsequential* and *sequential* variables (Yoder, Short-Meyerson, & Tapp, 2004). To aid this discussion, suppose we hypothesize that teacher instructions have an immediate effect on student topic-continuing utterances. Before conducting an expensive experiment to determine whether increasing instructions results in greater student topic-continuing utterances, we might first want to test whether there is a close temporal association between teacher instruction and student topic-continuing utterances.

An approach that uses nonsequential variables to address this question might test whether there is a positive correlation between the total number of teacher instructions and the total number of student topic-continuing utterances. These two variables are nonsequential variables because each concerns only one behavior and neither variable expresses anything about the sequence of teacher instruction and student topic-continuing utterances within a particular session.

Testing the same hypothesis with sequential variables in a group design, one might test whether the extent to which a topic-continuing utterance occurred immediately after a teacher instruction was greater than one would expect by chance for the majority of teachers and students. The aspect of our theory that predicts that teacher instruction has an *immediate* effect on topic-continuing utterance is reflected in our choice of a sequential level variable.

The main point is that sequential variables reflect a very specific and close temporal association within an observation session. That is, the sequential variable quantifies the extent to which the target behavior (e.g., topic-continuing utterance) occurs within a specified number of coded behaviors or time units from the antecedent behavior (e.g., teacher instruction). The reader should note that sequential analysis requires specification of the time period (e.g., within 5 s) or the number of coded behaviors (e.g., the next coded behavior) from the antecedent behavior. Because sequential variables require the investigator to predict the number of behaviors after which, or the time window in which, the target behavior will occur after the antecedent behavior, very specific temporal relations can be tested. The degree of specificity implicit in sequential variables reduces the number of alternative explanations for the association of interest.

SEQUENTIAL ASSOCIATIONS ARE NOT SUFFICIENT EVIDENCE FOR CAUSAL INFERENCES

Statistically significant or large sequential associations do not necessarily mean that the antecedent behavior *caused* the target behavior to occur (Yoder et al., 2004). Similar to other indices of association, indices of sequential association can be high because of some previously occurring or simultaneously occurring behavior (i.e., a spurious association). For example, student topic-continuing utterances may stimulate the teacher to provide instructions in an attempt to inquire about the student's topic of interest and the student who has been engaging in topic-continuing utterances may continue doing so. Such a pattern would produce a large positive sequential association between teacher instruction and student topic-continuing utterance, but the direction of effect would be from the student to the teacher, not vice versa.

CODED UNITS AND EXHAUSTIVENESS

All sequential analyses require an exhaustive coding space. The concept of a "coded unit" is very helpful to understand when discussing an important observational measurement system concept known as "exhaustiveness." By a coded unit, we mean the entity that the observer identifies and (perhaps) classifies. When we are deriving number using event sampling, the coded unit is the behavior. When we are deriving duration using timed-event behavior sampling or we are using an interval sampling method, the coded unit is a time unit that is assigned to a behavior's presence or absence. In the case of duration from timed-event behavior sampling, the coded time unit is often seconds. In all interval coding methods, the coded time unit is the interval.

Exhaustive coding means that the record includes all "relevant" units that occurred in the observation session. Sessions are exhaustively coded when the observer uses a continuous timed-event behavior sampling method and duration is the derived metric, or when interval sampling in which all intervals are observed. In the former case, all seconds in the observation session are included in the sequential analyses and we ask whether observers agree on the coding of all seconds (see chapter 8). Similarly, in the latter case, all intervals in the observation session are included in the sequential analyses and we ask whether observers agree on the coding of all intervals. When this occurs, we say that we have an exhaustively coded observation session. Without meeting this assumption, the results of sequential analysis are uninterpretable.

In terms of sampling methods, almost all continuous or intermittent event sampling and almost all continuous timed-event sampling in which number is derived do *not* produce exhaustive coding of an observation session. Unfortunately, there is a common, but ill-advised, approach to try to create exhaustive sessions from continuous or intermittent event sampling. This is the use of an "other" (i.e., trashcan) category and code "all" unanalyzed behaviors in this category to make the coding exhaustive.

There are two reasons why, no matter how large the "other" category, using an "other" category does not really solve the problem. First, the observer will not be a perfect recorder of all occurrences of the behavior classes indicated in the coding manual. Second, the coding manual will not necessarily include all important behavior classes because present knowledge regarding potential members of any given behavior class of interest is limited. Determining which behaviors constitute an example of a "relevant" behavior is difficult to define because "relevancy" varies by topic area. For example, when we are examining the antecedents and consequences of aggressive behavior, the relevant behaviors are *all* potential antecedents and *all* potential consequences of *all* potentially classifiable aggressive behavior. If we do not include all potential antecedents, consequences, and examples of aggressive behavior, then estimates of agreement or sequential association among these classes will be incorrect. The math of estimates of chance (such as that used in sequential analysis and certain indices of interobserver agreement) requires that we quantify the total number of coded units (e.g., events) as a context (e.g., denominator) for interpreting the number of other behaviors. When time units (e.g., seconds or intervals) are the coded units, the total number of time units is the context against which the number of time units with the key behavior is interpreted. The same cannot usually be said when events are the coded unit.

This is not to say that continuous event coding is *never* useful as the basis for sequential analysis. There are situations in which the class of behaviors that are considered "relevant" is not in question and when observers are able to code the presence of all relevant behaviors with very high accuracy.

For example, transcriptions of conversations can provide a nearly exhaustive account of the spoken utterances in the conversation, particularly if the speakers are intelligible, the conversation is recorded, and the coding decision recording method allows stop-and-go transcription. In this case, the presence of an utterance can be identified with nearly perfect reliability. In another type of example, sequential analyses are even more likely to be accurate than if the number of events is preplanned and the plan is followed exactly. For example, if we are measuring children's approach behavior after a planned, unfamiliar event and planned adult message about the event (i.e., a social referencing task), the adult's message about unfamiliar events and the child's approach behavior are behaviors that can be coded in an exhaustive way using continuous event sampling, assuming there is 100% fidelity of administering the instructions.

THREE MAJOR TYPES OF SEQUENTIAL ANALYSIS

The three types of sequential analysis are defined by how the immediate temporal relation between potential antecedent and the target behaviors is tested (Yoder et al., 2004). Immediacy is usually defined by a specified number of coded units between antecedent and target behaviors. The type of coded unit varies among the three types of sequential analysis. One may ask whether topic-continuing utterances follow teacher instruction by (a) a specified number of behaviors (event-lag sequential analysis), (b) a specified number of time units (e.g., seconds; i.e., time-lag sequential analysis), or (c) within a specific time window (e.g., within 5 s; time-window sequential analysis). In addition, finer distinctions can be made within each of the above three types by referring to (a) the direction of the analysis (i.e., forward, backward, concurrent) and (b) the number of coded units from the antecedent that the target is expected to occur (i.e., lags). One can designate both by assigning a sign to a lag number. For example, lag 1 means that the target is expected to occur exactly one behavior *after* the antecedent (i.e., a positive sign indicates a forward analysis). Lag –1 means that the target is expected to occur exactly one coded unit *prior to* the "antecedent." Lag 0 means that the target is expected to occur during the same coded unit as the "antecedent." In reality, only certain combinations of these types of sequential analyses tend to be used or useful. For example, we will posit in this chapter that backward sequential analysis is rarely, if ever, a better choice than forward sequential analysis. In our ongoing example of the sequential relation between "teacher instruction" and "topic-continuing utterances," the following sections review each of the three different sequential analysis types.

Event-Lag Sequential Analysis

Event-lag sequential analysis might be used to test whether the extent to which student topic-continuing utterances follow teacher instruction is more than expected by chance. The coded units in event-lag sequential analyses are behaviors or other events. The behavior sampling methods used to generate these coded units are either event or timed-event behavior sampling methods in which only the onset of the behavior is considered relevant. We have mentioned that very few event behavior sampling or timed-event behavior sampling methods using only frequency metrics can legitimately claim to include "all relevant behaviors."

For example, we conducted a sequential analysis of parent and child verbal conversations in which each spoken utterance was coded and analyzed to determine which type of parent utterance had the strongest sequential association with child conversational participation (Yoder, Davies, & Bishop, 1994). We implicitly claimed that the relevant behaviors were adult and child utterances in the observation session. However, because child utterances are not always intelligible, segmenting or separating the child utterances in the same conversational turn was not possible to accomplish with nearly 100% accuracy. Most event sequential analyses have an even more difficult problem. For example, it is not clear what the complete set of relevant behaviors might be when testing whether teacher instruction elicits student self-injury. Later, it will become apparent that our definition of the "relevant" behaviors to code is extremely important in estimating the chance sequential occurrence of the antecedent and target behaviors. Uncertainty about what constitutes the complete set of relevant behaviors is probably one reason why time-lag sequential analysis was created.

Time-Lag Sequential Analysis

A time-lag sequential analysis might test whether the onset of student topic-continuing utterance occurs *exactly* 1 s after the onset of teacher instruction more than is expected by chance. Timed-event sampling and interval coding behavior sampling methods are used to generate data for such analyses and the time unit (e.g., second or interval) is the coded unit. These analyses *do* typically meet the assumption of an exhaustive coding space because all relevant units are included in the analysis. Additionally, it is quite acceptable to code only the potential antecedent of interest and the target behavior of interest. For example, we might code an hour-long behavior sample for the time of onset of teacher instruction and the time of onset of topic-continuing utterance or we might code the intervals in which teacher instruction and/or topic-continuing utterance occur. Therefore, time-lag sequential analysis enables us to circumvent the problem of having to define all of the behaviors that should be considered "relevant to code." Later in the chapter, it will become apparent why all time units within the behavior sample are considered to determine whether topic-continuing utterance occurred after teacher instruction more than expected by chance. However, time-lag sequential analyses using timed-event data are extremely demanding because they require precise predictions regarding the exact number of time units (e.g., seconds) the investigator expects the target to occur after the antecedent. Therefore, it is more common to see interval data used for time-lag sequential analysis or to see concurrent analyses using timed-event data. Most of our research questions are not supported by sufficiently specific theory or knowledge to allow such precise predictions. This is probably why time-window sequential analysis was invented.

Time-Window Sequential Analysis

A time-window sequential analysis might test whether the onset of student topic-continuing utterance *occurs within a specific time window* (e.g., 5 s) from the onset of teacher instruction. When conducting a sequential analysis, timed-event behavior sampling is usually best analyzed with time-window analysis. The coded unit is the time unit (e.g., seconds). As in time-lag sequential analysis, the assumption of an exhaustive coding space is easily met because all time units are included in the analysis. That is, because a timed-event behavior sampling method is used, each time unit is coded for presence and type of a behavior of interest. This circumvents the problem of having to justify that all "relevant behaviors" are included in the analysis because the coded unit is a time unit, not behavior. However, the way the data are organized for analysis alters how chance is estimated in an important way that will become apparent when we discuss "contingency tables."

Note that the time-window lag sequential analysis allows less precision than the time or event-lag sequential analysis in the prediction of the exact number of coded units that the target (e.g., topic-continuing utterance) is expected to occur after the antecedent (e.g., teacher instruction). Therefore, time-window sequential analysis matches the complexity of human behavior and our limited state of knowledge of human interactions better than time- or event-lag sequential analysis (Yoder & Tapp, 2004). The "proper" duration of the time window is an empirical and theoretical matter. Because such information is generally unavailable, it is presently largely arbitrary. However, shorter windows (e.g., 5 s) are more common than larger windows (e.g., 1 min). The time-window sequential analysis was first introduced to the literature by Bakeman and Quera in 1995. In our opinion, time-window sequential analysis is likely to become the favored of the three methods as it becomes better known to investigators.

THE NEED TO "CONTROL FOR CHANCE"

Sequential frequency (i.e., the number of times the target follows the antecedent behavior) has been considered and discarded as a measure of sequential association because it does not control for chance occurrences of the sequence of interest (Bakeman & Gottman, 1997). An example of sequential frequency is the number of times the child's topic-continuing utterances occurred after the adult's questions. Suppose that two children, Joe and Lisa, continued the topic after the adult questions 5 times and 10 times, respectively, in a given period. Suppose further that the adult interacting with Joe used 10 questions and the adult interacting with Lisa used 20 questions. Even though Lisa's sequential frequency is twice that of Joe's, it is not clear what this means in terms of a sequential association between child talk and adult questions because Lisa's adult provided twice as many opportunities for the sequence of interest. This illustrates that the rate of occurrence of the antecedent, in this case adult questions, needs to be taken into account to know whether a sequential association between antecedent and target behaviors exists. Later in the chapter, we will show that dividing the sequential frequency by the base rate of the antecedent (i.e., the transitional probability) does not allow proper interpretation either because it is also influenced by chance.

It should be noted that the notion of "chance" or "probability" is not universally accepted as a scientifically useful concept because, as the argument goes, chance has multiple definitions (Johnston & Pennypacker, 1993). Despite this perspective, many have indicated that chance or probability is an essential concept for understanding sequential analysis (Bakeman & Gottman, 1997). As it applies to sequential analysis, one useful conceptual definition of a chance estimate of the sequential occurrence of behaviors is the mean sequential frequency out of a large number (e.g., 1,000) random sequencings of several (e.g., 4) types of behaviors (2 types are the target and the antecedent) that occur a specified number of times. At least one simulation study has shown that such a definition for a chance estimate of the sequential frequency is a remarkably accurate estimate of chance sequencing (Bakeman, Robinson, & Quera, 1996).

This commonly accepted estimate of chance occurrence of a sequence is computed from the base rates of *both* the antecedent and the target behaviors. For example, one index of sequential association estimates chance as the following: (simple probability of target) × (simple probability of antecedent) × (total number of behaviors) (Bakeman & Gottman, 1997). The reason simple probability is used for the target and antecedent behaviors in the formula is that the number of times a behavior occurs often has more meaning in the context of knowing how long the subject is observed or how many instances of other coded behaviors occurred in the session.

For example, if a topic-continuing utterance occurs 2 times out of 1,000 coded child behaviors, it means something quite different from 2 out of 5 coded child behaviors. Therefore, we quantify the extent to which the target behavior occurs in terms of probabilities, not frequency. A "simple probability" in event-lag sequential analysis is the number of times a behavior occurs divided by the total number of coded behaviors in the behavior sample. For example, if we have 100 utterances in a conversation and 20 of these are child topic-continuing utterances (i.e., 0.2) and 20 are adult questions (i.e., 0.2), the chance estimate of the adult question – child continuing utterance sequence is $(0.2) \times (0.2) \times (100) = 4$. Therefore, a sequential frequency of 5 is greater than one would expect by chance, but not by much.

Note the importance of the total number of behaviors in this formula. If this "total of all relevant behaviors" is not accurate, we will not compute chance correctly. A 2 × 2 contingency table is useful in organizing sequential data because such a table clarifies whether the estimate of chance occurrence of the sequence of interest really considers all of the instances of the antecedent and target behaviors. Before discussing contingency tables, however, we introduce and discuss two primary ways that data are represented prior to their tallying into contingency tables.

HOW SEQUENTIAL DATA ARE REPRESENTED PRIOR TO CONTINGENCY TABLE ORGANIZATION

We have mentioned that behaviors or time units are the coded units that are analyzed in sequential analysis. For two-event sequences, these may be represented in a single stream of behaviors (for behaviors that never or almost never co-occur) or in two (or more) streams of behaviors (for behaviors that co-occur regularly). During the coding phase of data collection, separate streams of coded units may represent the behavior or time units for different actors or they may represent different dimensions of the situation (e.g., classroom activity type vs. child behavior). Using the terms used in the ProcoderDV software (provided at www.springerpub.com/yoder/supplements), separate streams of behaviors can be represented as different "groups" in the ProcoderDV "code file" (see the exercise manual in chapter 4).

When considering timed-event sampled data, one must decide whether to analyze the duration of behaviors. It is possible to "sample" the onset and offset of behavior but still only "analyze" the onset. Imagine that "1" represents occurrence of the behavior in a time unit and that "0" represents nonoccurrence of the behavior in a time unit. One can ignore or not analyze the duration of behavior by converting the nononset time units to nonoccurrence values. The ProcoderDV software introduced in chapter 4 is one example of software (see Bakeman & Quera, 1995, for another) that will do this and its application to sequential analysis will be covered briefly in chapter 7. See Table 6.1 for an illustration of how the ProcoderDV software transforms timed-event sampled data that recorded both onset and offset of behaviors to data where only offset is represented for analysis.

CONTINGENCY TABLES

In the sequential analysis literature, a contingency table is frequently used to illustrate how raw sequential data are organized to compute the index of sequential association (Bakeman & Gottman, 1997). Although there are more complex contingency tables, we present the simplest case to illustrate the principles used to construct a proper contingency table for sequential analysis and to prepare the way to discuss two indices of sequential association: transitional probabilities and Yule's *Q*. Because misconstruction of contingency tables is one of the most common errors in published sequential analyses, three methods of proper construction of 2 × 2 tables will be covered in detail. The most important thing to remember about 2 × 2 contingency table construction is that all instances of the antecedent, all instances of the target, and all coded units are represented exactly once in the table. The method of tallying and labeling of

Table 6.1

ILLUSTRATION OF HOW SOFTWARE CAN CONVERT DURATION TO ONSET IN TWO-STREAM DATA

	ORIGINAL DATA		DATA REPRESENTATION AFTER THE SOFTWARE CONVERTS NONONSET SECONDS TO ABSENCE OF ONSET	
SECONDS	TEACHER BEHAVIOR	STUDENT BEHAVIOR	TEACHER BEHAVIOR	STUDENT BEHAVIOR
00:01	0	0	0	0
00:02	1	0	1	0
00:03	1[a]	0	0	0
00:04	1	1	0	1
00:04	0	1	0	0
00:05	0	1	0	0
00:06	0	0	0	0
00:07	1	1	1	1
00:08	1	1	0	0
00:09	1	0	0	0
00:10	0	0	0	0
00:11	0	0	0	0

[a]When “1” occurs immediately after “1,” it represents a continuation of the same instance of the key behavior whose onset is indicated by a “1” that is preceded by a “0.”

the columns and rows differ depending on whether the analysis that will follow is (a) a concurrent analysis of two streams of data, (b) an event-lag sequential of a single stream in which an instance of a behavior type can and does occur after another instance of the same behavior type (i.e., repeats), or (c) a time-window sequential analysis of two streams of data.

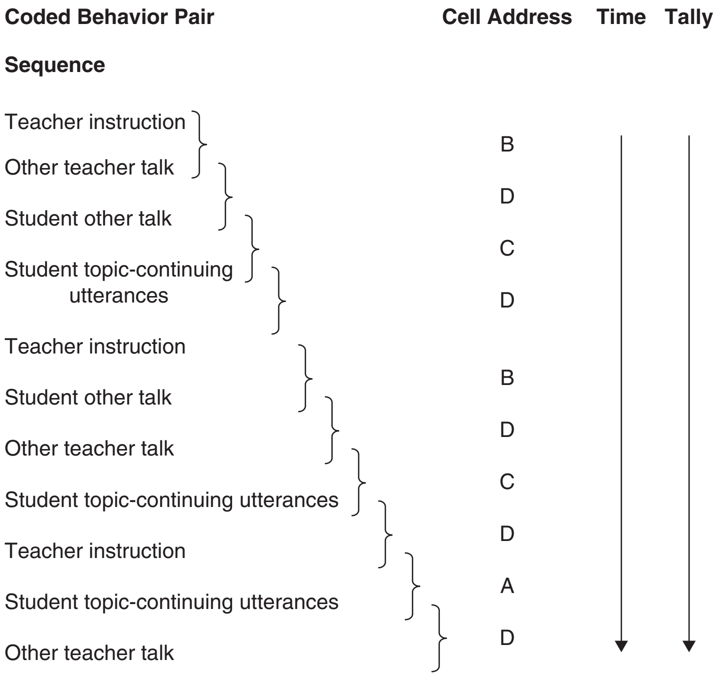
Proper 2 × 2 Contingency Table Construction of Two Streams of Data for Concurrent Analysis

The simplest way to conceptualize tallying pairs of coded units into a 2 ×2 table is to think about two streams of time units (e.g., intervals) that represent an exhaustive coding space of two behaviors: an antecedent (or not in one stream) and a target (or not in the other stream). The "pair" of coded units that is tallied is the presence or absence of two behaviors within the same time unit. For example, if there are hundred 10-s intervals in the observation session, then each interval is coded presence (1) or absence (0) for the antecedent in one ProcoderDV "group" and presence (1) or absence (0) for the target in another ProcoderDV "group." A convention in the sequential analysis literature is for columns to be used for coded units conceptualized as the target and rows to be used for coded units conceptualized as the antecedent. Another convention is for the first column to represent the target as present and for the first row to represent the antecedent as present. A third convention is for the cells to be labeled A, B, C, and D, moving from the upper left to upper right and then from the lower left to lower right cells, respectively. Therefore, the instances of the sequence of interest are tallied in the A cell. Its total (i.e., the number of tallies) is the observed sequential frequency.

Proper 2 × 2 Contingency Table Construction From One Stream of Data for Event-Lag Sequential Analysis

This method of 2 × 2 contingency table construction is complex and subject to error. Therefore, we cover it in detail. The proper 2 × 2 contingency table for this situation is indicated in Figure 6.1. Readers are strongly recommended to carefully read the general procedure by which pairs of behaviors are tallied into the cells of the 2 × 2 table.

The convention used in the sequential analysis literature is applied to this situation such that the rows are used to categorize each pair of behaviors according to whether or not the antecedent behavior is the first behavior in the pair and to use the columns to categorize the same pair of behaviors according to whether the target behavior is the second behavior in the pair. That is, except for the first and last, all coded behaviors are represented in *both* the rows and all the columns. The cell labels (A–D) for the four cells in the 2 × 2 table in Figure 6.1 should be noted. Particularly noteworthy is cell D (i.e., the cell in which nontarget behaviors follow nonantecedent behaviors). If the definition of "relevant" behavior is not accurate, then



These pairs of behaviors are tallied into a 2 x 2 table as follows:

Table A

		Behavior II		
		Student topic-continuing utterances	Not topic-continuing utterances (teacher talk or other student talk)	Total for rows
Behavior I	Teacher instruction	1 pair A	2 pairs B	3 pairs
	Not teacher instruction (other teacher talk or student behavior)	2 pairs C	5 pairs D	7 pairs
	Total for columns	3 pairs	7 pairs	Total of 10 observed pairs*

*These data are presented for illustrative purposes only.

Figure 6.1 Illustration of tallying behavior pairs into a 2 × 2 table for a forward event-lag sequential analysis.

the count in this cell will be "too low" and the estimate of chance occurrences of target behaviors after antecedents will be inaccurate. We call this the "D-cell problem." It is specific to event-lag sequential analysis.

In Figure 6.1, each behavior pair is labeled using the label for the cell into which it is tallied. It should be noted that behavior pairs are tallied into the 2 × 2 table in such a way that one behavior pair "overlaps" with the following behavior pair. That is, except for the first and last behaviors, each behavior is considered both a "first behavior" and a "second behavior" (i.e., the second behavior in one behavior pair is the first behavior in the next behavior pair). It has been demonstrated empirically that there are no disadvantages to using overlapping pairs of behavior (Bakeman & Dorval, 1989). A happy consequence of using overlapping behavior pairs is a doubling of number of tallies in the 2 × 2 table compared to what one would have if nonoverlapping pairs were tallied. As will be discussed in chapter 7, the number of tallies in the 2 × 2 table has important implications for the interpretability of sequential associations. We refer to the 2 × 2 table in Figure 6.1 as "Table A."

When the antecedent and target behaviors come from different people or participants, and the data are analyzed in an event-lag sequential analysis, a controversy exists concerning how to construct the 2 × 2 table. Some researchers have constructed 2 × 2 tables in which one person's behavior is tallied on the rows and the other person's behavior is tallied on the columns (Rocissano, Slade, & Lynch, 1987; Wampold & Kim, 1989). This method is illustrated in Table 6.2 and will be called "Table B."

For single-behavior streams in which behaviors cannot follow themselves (i.e., repeat) or other coded behaviors from the same actor, the

Table 6.2

ILLUSTRATION OF AN INAPPROPRIATE WAY TO ORGANIZE EVENT-LAG SEQUENTIAL DATA FROM A SINGLE DATA STREAM WHEN INSTANCES OF THE SAME BEHAVIOR TYPE CAN FOLLOW THEMSELVES

	STUDENT	
ADULT	TOPIC-CONTINUING UTTERANCE	OTHER UTTERANCE TYPE
Teacher instruction	A	B
Other utterance type	C	D

two types of 2 × 2 tables (represented by Tables A and B) yield exactly the same cell tallies. However, in many situations, behaviors *can* and *do* follow themselves (i.e., repeat). Therefore, to accurately reflect the *total* number of target and antecedent behaviors, not just those that occur in the hypothesized position (e.g., B follows A) in the behavior pair, the method of tallying behaviors into 2 × 2 tables must provide a cell for all instances of the target and antecedent behaviors, including those that occur after themselves and after other behaviors from the same actor. Failure to do so will result in inaccurate estimates of chance occurrence of the sequence of interest because the base rates of the antecedent and target behaviors will be inaccurate. For example, in Figure 6.1, the first instance of student topic-continuing utterance follows an instance of “other student utterances.” In a 2 × 2 table in which only teacher behaviors are counted in the first behavior position (as in Table B; Table 6.2), there is no place to tally this behavior pair. Next we investigate the consequence of using 2 × 2 tables like Table B when codes can follow themselves and other codes from the same actor.

Simulation Study to Compare Results From Two Ways to Construct Contingency Tables

We ran a simulation study to illustrate that the type of contingency table one constructs matters when instances of the same behavior type can follow themselves (Yoder et al., 2004). There were 1,000 streams of sequential data in which four behaviors could and did follow themselves, all generated from the same set of software commands where the true sequential link between behaviors was at chance level (i.e., the null condition). The data were tallied into 2 × 2 tables like Tables A and B (remember, Table A is based on the behavioral codes whereas Table B is based on the person). An index of sequential association (i.e., Yule’s Q, which will be covered in chapter 7) was computed for each table for each data stream. The range of the difference scores between Yule’s Q scores from each table from the *same behavior stream* was from –.9 to 1.0, illustrating that the indices of sequential association for Table A were sometimes very different from those for Table B (Yule’s Q can take on values ranging from –1 to 1).

To determine which table was the best, we focused on those behavior streams for which the difference in sequential association from the two different methods of tallying the data into 2 × 2 tables exceeded $|.20|$ (n = 505 behavior streams). Chance should produce difference scores in which the recommended table (i.e., A) produces the higher sequential association score about half the time. Additionally, because the population sequential association was zero, the mean sequential association closest to zero is the more accurate way to organize the data. The mean sequential association for Table B (mean Yule's Q = .04; SD = .45) was significantly higher than that for Table A (mean Yule's Q = .0006; SD = .24; t = −2.13; p = .03). The mean Yule's Q for Table A was closer to zero than was the mean Q for Table B. Thus, Table A is more accurate than Table B.

Contingency Tables for Time-Window Lag Sequential Analysis

One important question in time-window sequential analysis is whether the duration of the behavior is to be analyzed. Assuming "enough" coded units, empirical results indicate that it does not make a substantive difference whether duration is analyzed in sequential analysis if the durations of antecedents and targets are less than 5 s (Yoder & Tapp, 2004). When the durations of antecedents or targets are more often more than 5 s, the following logical decision rules are offered. If longer antecedents are thought to have more influence on the target than shorter antecedents, the duration of the antecedent should be analyzed. The duration of targets should be analyzed if longer targets are more important than shorter targets, and if the antecedent continues to influence the presence of the target behavior after the target's onset. Otherwise, analyzing the onset of behaviors better suits the theory behind the research question. If one cannot decide because both sides of the argument are justifiable, one can do an analysis with and without duration analyzed. Analyzing duration does provide more instances of coded units with the presence of key behaviors, which in turn increases the base rate for the key behaviors, which in turn increases the stability (i.e., replicability) of the sequential association.

The reader is referred to Table 6.1 to illustrate how two streams of timed-event data in which only onset of the behavior is considered for tallying coded units into a 2 × 2 table for time-window analysis. The label for the first row of the 2 × 2 table is the primary difference between the 2 × 2 table in Table 6.3 (onset-only tallying for time-window sequential analysis) and the table in Figure 6.1 (tallying for event-lag sequential analysis). By asking whether a student topic-continuing utterance occurs within 5 s of the onset of teacher instruction, we have subtly changed the research question to one that is a bit less specific than that which is posed

Table 6.3

ILLUSTRATION OF A CONTINGENCY TABLE FOR TIME-WINDOW SEQUENTIAL ANALYSIS

TIME OF ONSET	TEACHER BEHAVIOR	STUDENT BEHAVIOR	CELL ADDRESS
00:01	NA	NA	D
00:02	Instruction	NA	B
00:03	NA	NA	B
00:04	NA	Topic-continuing utterance	A
00:05	NA	NA	B
00:06	NA	NA	B
00:07	NA	NA	D
00:08	NA	Topic-continuing utterance	C
00:09	Instruction	NA	B
00:10	NA	NA	B
00:11	NA	NA	B
00:12	NA	NA	B

The Contingency Table

	SECOND II			
SECOND I	STUDENT TOPIC-CONTINUING UTTERANCE		OTHER	
At or within 5 s of onset of teacher instruction	1 s	A	B	8 s
Other	1 s	C	D	2 s

for event- or time-lag sequential analysis. No longer does student topic-continuing utterance have to occur a specified number of coded behaviors or an exact number of seconds after the onset of teacher instruction to be tallied in the A cell (i.e., the cell for the sequence of interest).

It should be noted that the base rate of the antecedent in time-window lag sequential analysis is the number of *time units within the specified time window*, not the number of antecedent behaviors. Therefore, the larger the time window, the larger the chance estimate of targets in the time window. The reduced precision allowed by time-window analysis is offset by the reduced sensitivity of the analysis. Simulation studies have demonstrated that this method of sequential analysis produces expected results (Yoder & Tapp, 2004).

TRANSITIONAL PROBABILITY

Informal observation indicates that the most frequently used index of sequential association is the transitional probability of the target following the antecedent behavior. This transitional probability is the proportion of instances of the antecedent behavior which are followed by an instance of the target behavior. Using the cell labels of the 2 × 2 table, the formula for this transitional probability is $A/(A + B)$. In the example in Figure 6.1, the transitional probability of student topic-continuing talk given teacher instruction is 1/3 or .33.

It is very important to note that a transitional probability is different from an accuracy or consistency proportion metric. The logic of sequential analysis requires that the base rate of the target and antecedent be free to vary from each other. The difference is that in a transitional probability, the target behavior is defined as one that *can* occur after behaviors other than the antecedent, even if this never occurs in the observation session. In contrast, in an accuracy or consistency proportion, the behavior represented in the numerator cannot, by definition, occur under any other conditions except that represented by the denominator. For example, a verbal response to child communication can only occur after child communication, by definition. Therefore, the proportion of child communication that is verbally responded to is a consistency proportion, not a transitional probability. Questions involving consistency or accuracy proportions are not subject to sequential analysis because chance occurrence of the sequence cannot be estimated using the math of sequential analysis.

Transitional probabilities are frequently used as an index of sequential association because they appear easy to interpret. Unfortunately, transitional probabilities are influenced by the target behavior base rate and thus make poor indices of sequential association in many situations.

To illustrate this, we conducted another simulation study. In this study, we generated 1,000 single data streams of 100 coded behaviors each. Each data stream contained four types of behaviors, each of which occurred as a random proportion of the 100 behaviors (i.e., the simple probability or "base rate" of each behavior varied). Instances of the four types of behavior were randomly sequenced, providing a mean or population sequential association of 0.0 (no association). We then computed the correlation between the transitional probability of behavior A after behavior B and the simple probability of each behavior. In sequential analysis, one wants an index of sequential association that is not influenced by the base rate of the behaviors it assesses. The results indicated that the association of the transitional probability with base rate of the antecedent was only –.15, but the analogous association between transitional probability and the base rate of the target was .48. The association between transitional probabilities and the base rate of the target was .7 when the population sequential association is generated to be strong (i.e., Yule's *Q* of .5). This means that almost half of the variance in transitional probabilities was influenced by the base rate of the target behavior.

Put conceptually, the higher the simple probability of the target behavior, the higher the transitional probability by chance processes alone. For example, in an observational session in which the child uses topic-continuing talk frequently, such talk will occur after teacher instruction very often by chance processes. This example makes it clear that an interpretable index of sequential association must be compared to an estimate of chance occurrences of the sequence.

In the next chapter, we will show that indices of sequential association can be computed for each participant in a group study and used as dependent scores in statistical analyses or are computed for each session and used as the dependent scores in graphs used to examine a potential treatment effect in a single-subject experimental design. Assuming that each participant or session has a different simple probability of the target behavior (a common occurrence), the meaning of the two identical transitional probabilities will vary. However, even in clinical practice, using the concept behind transitional probabilities as the basis for decision making is a real, but largely unrecognized problem. This will now be discussed in the context of backward sequential analyses.

Transitional Probabilities in Backward Sequential Analysis

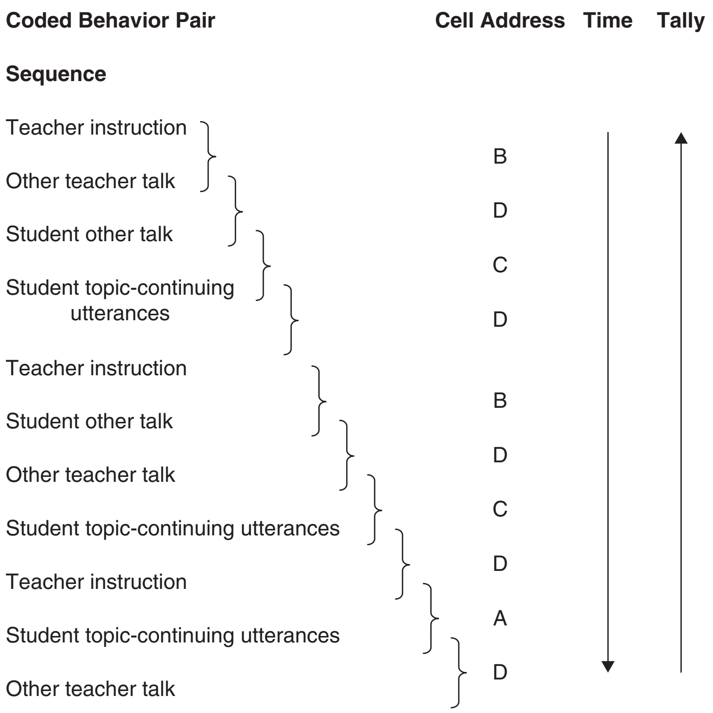
In this section, we will use an example first presented in another source to discuss why using transitional probabilities as the index of sequential association in backward sequential analyses tends to result in even more miscommunication than using transitional probabilities in forward sequential analyses designed to address the same research question (Yoder & Feurer, 2000).

By *backward sequential analysis*, we mean that the investigator tallies the number of times certain behaviors occur *before* the behavior of interest. For example, let us say that we use theory to guide our decision to code several teacher behaviors that may increase the probability of topic-continuing utterances (e.g., instructions, active ignoring, talking to other students, other talk to target student) because we ultimately want to increase the instances of the associated teacher instructional behavior as part of an intervention. Just as one application of functional analysis requires that we ask teachers what tends to occur before a key behavior, we might observe which of these teacher behaviors tends to *precede* topic-continuing talk most often.

In a backward sequential analysis, one tabulates the sequence of behaviors into the 2 × 2 table moving backward in time. That is, the "first behavior" in the behavior pair (i.e., topic-continuing talk) actually occurs *after* the "second behavior" in the behavior pair. See Figure 6.2 for an illustration of this process using the same data that were presented in Figure 6.1. In accordance with "backward sequential analysis" principles, the tabulation of the first and second behaviors is reversed in Figure 6.2 when compared to Figure 6.1.

Assume that we decide to use transitional probabilities as the index of sequential association (a common practice). It should be noted that the transitional probability of continuing utterances (i.e., the first behavior in a backward analysis) *preceded by* teacher instruction is .50 ($A/[A + B]$ in Figure 6.2). Note that the transitional probability of continuing utterances *following* teacher instruction (i.e., the first behavior in a forward analysis) is different: .33 ($A/[A + B]$ in Figure 6.1). That is, the transitional probability for a forward sequential analysis is different from that for a backward analysis of the same behaviors.

The primary source of the miscommunication about backward sequential is the mismatch between the motivating theory for the study, the terms used in sequential analysis, and the backward sequential analysis process. The motivating theory of most studies employing sequential



These pairs of behaviors are tallied into a 2 x 2 table as follows:

Table B

		Behavior II		
		Teacher instruction	Any other teacher or student behavior	Total for rows
Behavior I	Student topic-continuing talk	1 pair A	1 pair B	2 pairs
	Any other student or teacher behavior	2 pairs C	4 pairs D	6 pairs
	Total for columns	3 pairs	5 pairs	Total of 8 observed pairs*

*These data are presented for illustrative purposes only.

Figure 6.2 Illustration of tallying behavior pairs into a 2 × 2 table for a backward event-lag sequential analysis.

analysis presumes a causal relation between the antecedent and target behaviors. Causality progresses forward in time. The term "target" or "second behavior" typically occurs *after* an "antecedent" or "first behavior." For example, assume we conduct a backward and forward analysis to identify the best candidates for antecedents of continuing utterances. The reason to conduct the study is to eventually increase continuing utterances by increasing the occurrence of the antecedent behaviors for continuing utterances. However, the targets or second behaviors in a backward analysis are the *possible antecedents*. The target or second behavior in a forward analysis to address the same question is *continuing utterances*.

Naïve readers may not realize that the target behaviors are typically different in the two or more sequences compared in this type of backward sequential analysis. If the target behaviors are different, the simple probabilities of the two target behaviors are almost always different. As always, differences in the transitional probabilities are not interpretable by themselves when we compare transitional probabilities for sequences with different simple probabilities of the target behavior.

The issue is not whether backward sequential analysis has a place in our armory of research tools. The issue is that many readers (and possibly researchers) may be less aware of the greater potential for misinterpreting transitional probabilities as indices of sequential association in the context of backward sequential analysis than in the context of forward sequential analysis.

Summary of Transitional Probabilities

Unfortunately, using transitional probabilities without reference to an estimate of chance occurrence is still frequently seen in the sequential analysis literature. We recommend not using transitional probabilities even for descriptive purposes, because (a) readers frequently make implicit comparisons between transitional probabilities of sequences from different sessions or groups or with different target behaviors and (b) Yule's *Q* provides a more interpretable descriptive index of sequential association in all of these situations. In the final analysis, transitional probabilities are *only* interpretable when they are compared with an estimate of chance occurrences of the sequence of interest. However, this practice is cumbersome to say the least. Alternatively, Yule's *Q* is an index of sequential association that reflects a comparison with an estimate of chance occurrence of the sequence. Yule's *Q* will be addressed in chapter 7.

RECOMMENDATIONS

When collecting data for sequential analysis, we recommend using timed-event behavior sampling and using the time units as the coded units so that the coding space will be exhaustive. For most research questions, we recommend using forward time-window analysis because the flexibility of the window fits the imprecision of our motivating theories better than event-lag or time-lag sequential analysis. Forward analysis almost always results in more easily communicated and interpretable results than does backward sequential analysis. To plan for sequential analyses, we recommend constructing the intended 2 × 2 table to make sure that the analysis is set up so that all instances of antecedent and target and all coded units are included in the table. When writing about sequential analyses, we recommend displaying the 2 × 2 table to aid readers in understanding what was done. Finally, we recommend that transitional probabilities be supplemented or avoided as the index of sequential association for most research questions. Chapter 7 will identify an alternative index of sequential association: Yule's *Q*.

REFERENCES

Bakeman, R., & Dorval, B. (1989). The distinction between sampling independence and empirical independence in sequential analysis. *Behavioral Assessment, 11*(1), 31–37.

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS & GSEQ*. New York: Cambridge University Press.

Bakeman, R., Robinson, B., & Quera, V. (1996). Testing sequential association: Estimating exact *p* values using sampled permutations. *Psychological Methods, 1*, 4–15.

Johnston, J. M., & Pennypacker, H. S. (1993). *Readings for strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Rocissano, L., Slade, A., & Lynch, V. (1987). Dyadic synchrony and toddler compliance. *Developmental Psychology, 23*(5), 698–704.

Wampold, B. E., & Kim, K. (1989). Sequential analysis applied to counseling process and outcome: A case study revisited. *Journal of Counseling Psychology, 36*(3), 357–364.

Yoder, P. J., Davies, B., & Bishop, K. (1994). Reciprocal sequential relations in conversations between parents and children with developmental delays. *Journal of Early Intervention, 18*(4), 362–379.

Yoder, P. J., & Feurer, I. D. (2000). Quantifying the magnitude of sequential association between events or behaviors. In T. Thompson & D. Felce (Eds.), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 317–333). Baltimore: Paul H. Brookes.

Yoder, P. J., Short-Meyerson, K., & Tapp, J. (2004). Measurement of behaviour with special emphasis on sequential analysis of behaviour. In E. Emerson, C. Hatton, T. Thompson, & T. Parmenter (Eds.), *International handbook of applied research in intellectual disability* (pp. 179–202). West Sussex, UK: Wiley.

Yoder, P. J., & Tapp, J. (2004). Empirical guidance for time-window sequential analysis of single cases. *Journal of Behavioral Education, 13*, 227–246.

This page intentionally left blank

7 Analyzing Research Questions Involving Sequential Associations

OVERVIEW

This chapter first leads the reader through the use of a software program designed to conduct sequential analysis. In doing so, it demonstrates further the steps involved in using sequential analysis to answer "sequential" research questions and confirms that a software program is necessary to conduct sequential analysis. Hand computation of the data would result in too much error and would be prohibitively time consuming.

We then discuss the relative superiority of Yule's *Q* over transitional probabilities as an index of sequential association. We also discuss the importance of having enough data to produce an interpretable Yule's *Q*. Then we indicate how research questions that involve a sequential association are tested using group and single-subject designs. Next, we indicate that when the research question involves an attempt to identify probable reinforcers and the putative reinforcer is a high probability event, another approach to quantifying sequential data, contingency space analysis, may be a useful tool. Finally, we summarize recommendations for conducting sequential analyses of observational data.

COMPUTER SOFTWARE TO AID SEQUENTIAL ANALYSIS

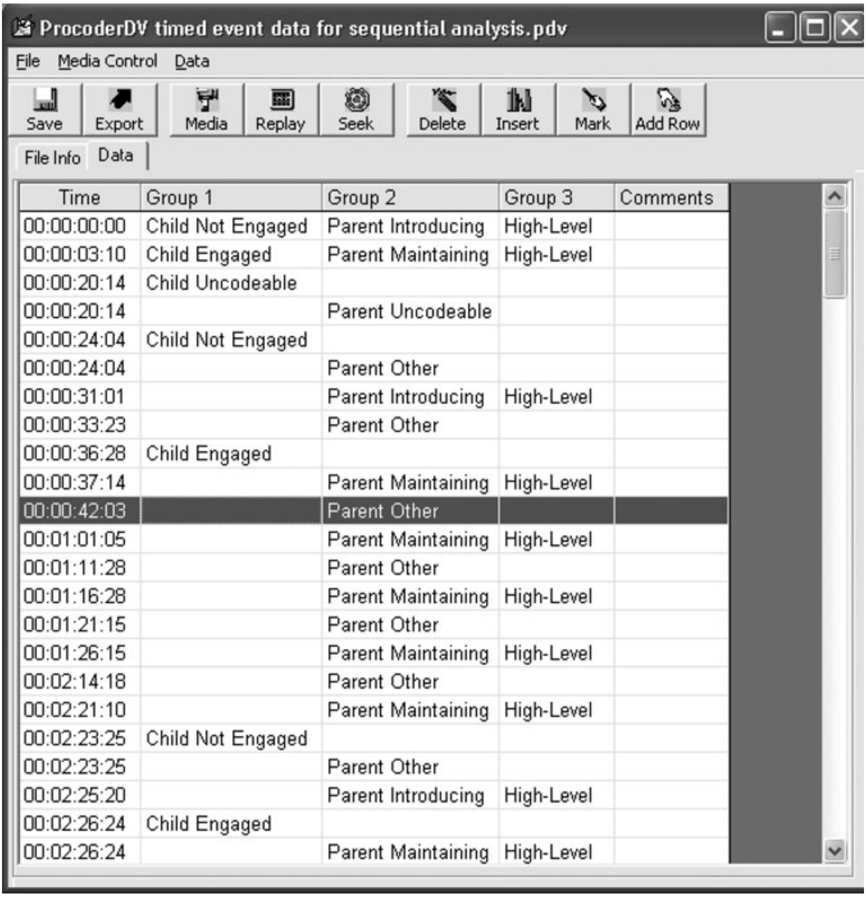
Computer programs designed to implement sequential analysis are necessary for a competent implementation of sequential analysis (Bakeman & Quera, 1995; Tapp, 1995). However, it is important to know what the software does to determine if it is doing what is needed. This chapter will bring the reader "behind the scenes" of what is seen on the computer screen to demonstrate what computer programs do (or some analogy of this) when conducting a sequential analysis. For the demonstration, we have selected timed-event data because it provides an exhaustive record of the observed session. We have selected time-window analysis as a type of sequential analysis to demonstrate because it is among the more promising analyses for the future, and we have selected a research question that lends itself to using onset of the antecedent and duration of the target as the objects of analysis to illustrate how such an analysis is performed using specialized observational software.

In this example, we use the software program titled Multiple Option Observation Software for Experimental Studies (MOOSES) (Tapp, 1995). Using this or other sequential analysis software (Bakeman & Quera, 1995), the observer does not have to judge whether a target occurs within a time window of an antecedent because the software program does that. All of the tallying for the 2 × 2 table is done "behind the scenes" so the user does not ordinarily get to see this. MOOSES has an option that allows the user to see how the tallying is completed and retained for analysis, making it an excellent software for demonstration purposes.

PRACTICE EXERCISE USING MOOSES SOFTWARE TO CONDUCT TIME-WINDOW ANALYSIS

We have copied on the book's website (www.springerpub.com/yoder/supplements) an example of timed-event data file that will be used for time-window analysis. This file can be opened in ProcoderDV and is called "ProcoderDV timed-event data for sequential analysis." The code file for this data is called "Code file for sequential analysis." A partial copy of this file is presented in Figure 7.1.

The code file for this observational data file was set up so that the child's state of engagement with an object ("e" for engaged and "u" for unengaged) was coded in "Group 1." Because we wanted to allow for



Time	Group 1	Group 2	Group 3	Comments
00:00:00:00	Child Not Engaged	Parent Introducing	High-Level	
00:00:03:10	Child Engaged	Parent Maintaining	High-Level	
00:00:20:14	Child Uncodeable			
00:00:20:14		Parent Uncodeable		
00:00:24:04	Child Not Engaged			
00:00:24:04		Parent Other		
00:00:31:01		Parent Introducing	High-Level	
00:00:33:23		Parent Other		
00:00:36:28	Child Engaged			
00:00:37:14		Parent Maintaining	High-Level	
00:00:42:03		Parent Other		
00:01:01:05		Parent Maintaining	High-Level	
00:01:11:28		Parent Other		
00:01:16:28		Parent Maintaining	High-Level	
00:01:21:15		Parent Other		
00:01:26:15		Parent Maintaining	High-Level	
00:02:14:18		Parent Other		
00:02:21:10		Parent Maintaining	High-Level	
00:02:23:25	Child Not Engaged			
00:02:23:25		Parent Other		
00:02:25:20		Parent Introducing	High-Level	
00:02:26:24	Child Engaged			
00:02:26:24		Parent Maintaining	High-Level	

Figure 7.1 Example of timed-event data from ProcoderDV that will be analyzed by time-window sequential analysis.

maximum precision in coding onset of the two engagement codes, we coded child engagement in a separate pass from other codes. A pass simply refers to coding exclusively within one group of codes for an entire session. In another pass, we coded whether the parent used a gesture, action, or verbalization to direct or maintain the child's attention on an object in "Group 2." In Group 2, we indicated whether the object to which the parent directed the child's attention was one the child was already attending to ("m" for maintains child's focus of attention), or was a new one. If the parent's focus of attention was to an object to which the child was not currently attending, we indicate whether the new object was introduced at a time the child was not attending to anything ("i" for introduce) or whether it was meant to redirect the child's attention to a different object ("r" for redirect). The codes in Group 3 were not analyzed in this run, but they represent another aspect of parental attentional directives. The research question was "Is there a positive and noteworthy sized sequential association of parental attentional directive that maintains the child's focus of attention (i.e., Parent Maintains) and the child's continued engagement with objects (i.e., Child Engages)?"

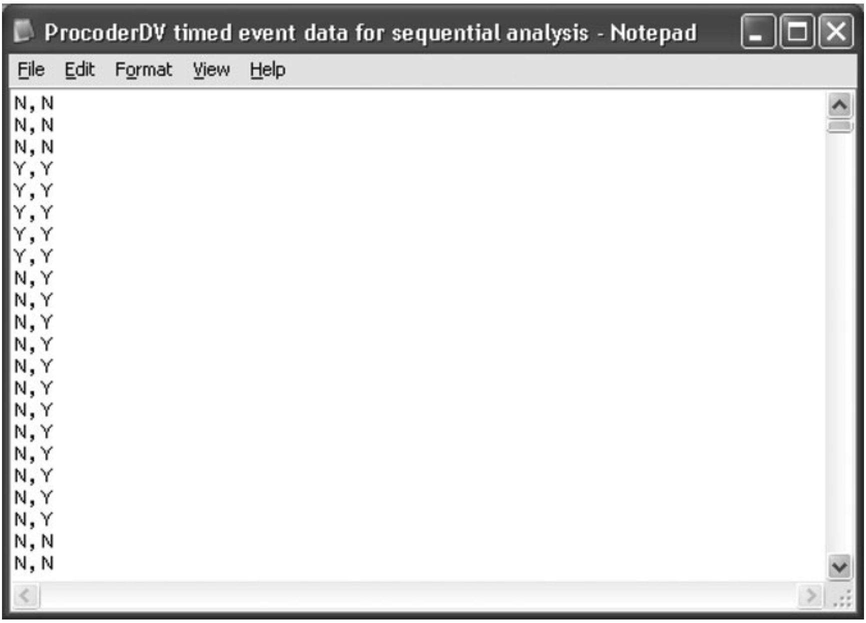
Because we have decided to analyze the duration of the target (engagement with objects), it is important that the codes in the group for the target (i.e., Group 1) are mutually exclusive and exhaustive. That is, the time of onset of one code *is* the time of offset for the preceding code and so forth. Because we have decided not to analyze the duration of the antecedent (coded in Group 2), we are only concerned with onset times of those codes and these codes are not exhaustive of all time units. In other words, some seconds are not coded for Group 2 but these uncoded seconds will still be analyzed in the sequential analysis. The time of occurrence of the onsets and offsets is the information that the software program uses to determine whether a target occurs within 5 s of the onset of the antecedent.

A manual directing the reader through a keystroke-by-keystroke execution of a demonstration version of MOOSES is available at www.springerpub.com/yoder/supplements. The manual is called "Instructions for using MOOSES to conduct a time-window sequential analysis." Readers are urged to execute the practice session to understand better the analytic consequences of different decisions that were reviewed in chapter 6. Before beginning the exercise, readers should set up the demonstration version of MOOSES software, which is also available at www.springerpub.com/yoder/supplements (the file called "Mooses.exe"), then open the other files on the website for the exercise in the folder for this chapter.

For those readers who choose not to engage in the exercise, it is useful to note that the software allows the investigator to communicate what type of sequential analysis is chosen. If the time-window method of sequential analysis is selected, then the investigator indicates (a) whether the time window begins at the onset (which reflects that the duration of the antecedent is not analyzed) or the offset of the antecedent event (which reflects that the duration of the antecedent is analyzed), (b) whether the target onset or duration is analyzed, and (c) the duration of the time window. It is important to note that this analysis is repeated for every sequence of interest, every condition within which a sequential association is derived, and for every participant in the study. As an option, one can have the file shown that demonstrates how the computer tallies the coded units for the 2 × 2 table.

In Figure 7.2, we present a partial record of how MOOSES keeps track of the timed-event data for time-window analysis of two streams of data. In this figure, the "Y" stands for "yes" and the "N" stands for "no." The first refers to the first stream of data and is a response to "Does the second of interest fall within the 5-s time window." The second refers to the second stream of data and asks, "Is the second of interest coded for the target?" MOOSES scores each second (or other time unit indicated) in this manner.

In Figure 7.1, note that the child is unengaged in the first 3 s, while the mother introduced a new object of attention to the child in the first second. Answering the two questions for maintains as an antecedent and engaged as a target, MOOSES records "N,N" for the first 3 s because none of the first 3 s was coded for the target or for the antecedent time window. However, within the fourth completed second, the child initiates an "engaged with object." The child continues to remain engaged

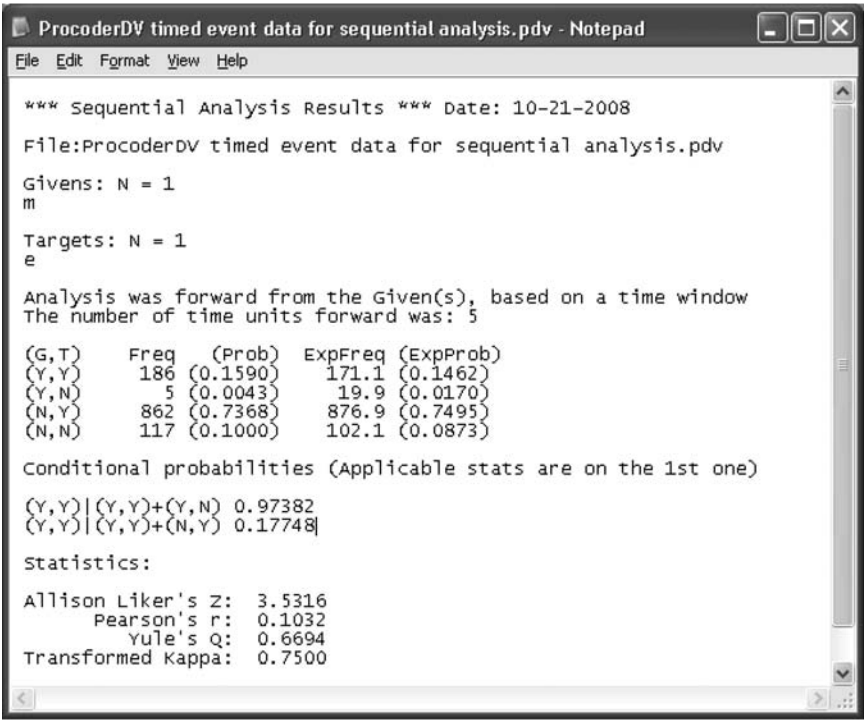

ProcoderDV timed event data for sequential analysis - Notepad
File
Edit
Format
View
Help

```
N,N
N,N
N,N
Y,Y
Y,Y
Y,Y
Y,Y
Y,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,Y
N,N
N,N
```

Figure 7.2 Example of how MOOSES tallies timed-event data for a 5-s time-window sequential analysis.

until the 21st second. Parent Maintains begins within the fourth completed second and the 5-s time window begins at the onset of the Parent Maintains. So, the answer to the key questions for next 5 s are "Y,Y." In Figure 7.2, the reader can verify this. This process continues until all 1,200 s in this 20-min session are tallied in this fashion.

In Figure 7.3, we have presented the output that one should get from running the exercise. In MOOSES, the "antecedent" is called the "given" event. The codes for the antecedent and target events are given near the top of the output. MOOSES provides the label for the 2 × 2 cells as (Y,Y) for the A cell, (Y,N) for the B cell, (N,Y) for the C cell, and (N,N) for the D cell. The observed cell count is listed under "Freq." For example, 186 s were (a) included in the 5-s window from the onset of the Parent Maintains *and* (b) coded with child engagement with object. This is the "observed sequential frequency" for Parent Maintains followed by Child Engages.


ProcoderDV timed event data for sequential analysis.pdv - Notepad
File
Edit
Format
View
Help

```
*** Sequential Analysis Results *** Date: 10-21-2008

File:ProcoderDV timed event data for sequential analysis.pdv

Givens: N = 1
m

Targets: N = 1
e

Analysis was forward from the Given(s), based on a time window
The number of time units forward was: 5

(G,T)     Freq   (Prob)  ExpFreq (ExpProb)
(Y,Y)      186 (0.1590)    171.1 (0.1462)
(Y,N)        5 (0.0043)     19.9 (0.0170)
(N,Y)      862 (0.7368)    876.9 (0.7495)
(N,N)      117 (0.1000)    102.1 (0.0873)

Conditional probabilities (Applicable stats are on the 1st one)

(Y,Y)|(Y,Y)+(Y,N) 0.97382
(Y,Y)|(Y,Y)+(N,Y) 0.17748

Statistics:

Allison Liker's Z:  3.5316
      Pearson's r:  0.1032
         Yule's Q:  0.6694
Transformed Kappa:  0.7500
```

Figure 7.3 Example output from time-window analysis of timed-event data from MOOSES.

As mentioned in chapter 6, the "observed sequential frequency" is generally not interpretable because it is clearly influenced by the base rate of the antecedent. In time-window analysis, the base rate of the antecedent is the number of seconds that occurs within the time window of the antecedent behavior. In our example, the base rate of the antecedent is the number of seconds that occurs within 5 s of the onset of Parent Maintains (i.e., $A + B = 191$). As indicated in chapter 6, the transitional probability controls for the antecedent base rate. On the MOOSES output, the transitional probability of interest is labeled under "conditional probabilities" and is the one for (Y,Y)/(Y,Y) + (Y,N). For this example, it is .97. This is a very high transitional probability. However, we indicated in chapter 6 that even high transitional probabilities can occur by chance if the target behavior occurs frequently enough, and child engagement occurs for 1,048 s (A + C) out of 1,200. This seems like a lot, but is it enough to discount the potential influence of Parent Maintains on Child Engages?

YULE'S *Q*

To answer the question posed above, Yule's *Q* is the currently recommended index of sequential association (Bakeman, McArthur, & Quera, 1996; Yoder, Short-Meyerson, & Tapp, 2004). It is equivalent to the odds ratio for the same 2 × 2 table (Reynolds, 1984). The primary difference is that, unlike the odds ratio, the Yule's *Q* has a potential range from –1.0 to 1.0. Using the cell addresses of the 2 × 2 table, the formula for Yule's *Q* is $(A \times D) - (B \times C)/(A \times D) + (B \times C)$. The reader can easily see that all four cells of the 2 × 2 table are used to compute Yule's *Q*. Using the data from the output of MOOSES that is presented in Figure 7.3, the *Q* for the example is $([186 \times 117] - [5 \times 862])/([186 \times 117] + [5 \times 862]) = 17{,}452/26{,}072 = 0.67$. As illustrated in Figure 7.3, the Yule's *Q* is the second to last number in the output of MOOSES.

There are some similarities in the interpretation of Yule's *Q* compared to that of Pearson's Product Moment correlation coefficient (i.e., *r*). A positive *Q* value means there is a positive sequential association (i.e., the observed sequential frequency is greater than expected by chance), and a negative *Q* value means there is a negative sequential association (i.e., the observed sequential frequency is less than expected by chance). Zero *Q* values mean that the observed sequential frequency is equal to that expected by chance (i.e., there is no association). Yule's *Q* values, like *r*, are effect size metrics for categorical data. Unlike *r* values, the benchmarks for large, moderate, and small Yule's *Q* values are 0.6, 0.43, and 0.2, respectively, and are derived from those given for odds ratio (Rosenthal, 1996). Therefore, in the example of sequential analysis, we found that the observed number of times Child Engagement fell within the 5-s window of the onset of Parent Maintains was much greater than expected by chance. In other words, there was a large, positive sequential association between Parent Maintains and Child Engagement with objects for this participant.

Importantly, the results of the simulation study that was described in chapter 6 also showed almost no relation of either antecedent or target base rate with Yule's *Q* (i.e., in both cases, $r < |.1|$). This is important because we want a sequential association index that is independent from the aspects of the session that are not intrinsic to the concept of a sequential association (e.g., base rates of the antecedent and target behavior and total number of coded units). Additionally, other simulation studies have shown that when the population sequential association is generated to be null, the mean Yule's *Q* is zero and is normally distributed (Bakeman, McArthur, et al., 1996). These simulation study results indicate that Yule's *Q* has attributes that make it a good fit for the types of dependent variables that perform best in parametric significance tests (e.g., *t*-tests, ANOVA, regression).

Finally, Yule's *Q* for a forward event-lag sequential analysis is identical to that for a backward event-lag sequential analysis (Yoder et al., 2004). Therefore, if an investigator uses a backward event-lag sequential analysis, it is strongly recommended that Yule's *Q*, not transitional probability, be used as the dependent variable (Yoder et al., 2004). In fact, it is generally better to use Yule's *Q* scores than it is to use transitional probabilities as the dependent variable for most types of research questions that are appropriate for sequential analysis. However, it should be noted that sufficient data is necessary for Yule's *Q* to be interpretable.

WHAT IS "ENOUGH DATA" AND HOW DO WE ATTAIN IT?

To determine whether the observed sequential frequency is substantially different from that which is expected by chance, we have to have "enough" data to estimate chance sequential frequency. When we do not, statisticians say that the contingency table is "sparse" (Reynolds, 1984). Some readers may recognize this concern as something general to categorical data analysis, not just sequential analysis. At this point, you may become confused regarding the measurement scale of observational data. In chapter 5, we indicated that observational variable metrics are on at least an ordinal scale. Indeed, when considering a group of Yule's *Q* values (e.g., 20 of them), the scale along which the *Q* values fall *is* at least ordinal. However, when considering the data in a single 2 × 2 table, the scale of measurement is nominal. The latter is the level of analysis that we are now addressing.

A sparse table is one in which there is an *expected value* equal to or fewer than 5 in any cell (Wickens, 1993). A 2 × 2 table with adequate data has an expected value of greater than 5 in all four cells. Recall from chapter 6 that an expected value is computed from three values: The two marginal frequencies that correspond to the cell in question and the total number of coded units. There are different ways to write the formula for expected value, but one way is as follows:

Simple *probability* of relevant row marginal × *frequency* of relevant column marginal.

Recall that simple probability is the frequency of the relevant row marginal/total number of coded units. In the MOOSES output in Figure 7.3, the expected frequencies of the cells are in the column marked "ExpFreq." In our example, all four cells had expected frequencies above 5. This means that the table had sufficient data to interpret values computed from it (e.g., Yule's *Q*). It is important to note that we are not using the observed (i.e., column labeled "Freq") cell values to determine whether we have sufficient data.

There is another way to detect a table with insufficient data. Whenever both of the cells in either of the diagonals have observed values of 0, Yule's *Q* cannot be computed due to the denominator of the computational formula being 0. One way that statisticians have addressed this problem is to add .5 to all frequency values in all four cells, not just the one with 0 (Reynolds, 1984). This is called Yate's correction.

Refer to Table 7.1 for six examples of 2 × 2 table cell values. Note what happens to Yule's *Q* when there are zeros in any cell (i.e., rows 1–3). The Yule's *Q* values are extreme (i.e., –1. or 1.0). Note what occurs when we add .5 to each cell in row 3 (i.e., see row 4). The Yule's *Q* is still extreme. Yate's correction did not really address the problem. The reason is that there is still insufficient data to compute an interpretable Yule's *Q*.

Table 7.1

EXAMPLES OF 2 × 2 CELL VALUES, EXPECTED VALUE OF A CELL, AND COMPUTED YULE'S *Q*

	CELLS				TOTAL CODED UNITS	BASE RATE OF		EXPECTED VALUE OF A CELL	YULE'S *Q*
CASES	A	B	C	D		ANTECEDENT	TARGET		
1	0	0	20	50	70	0	20	0	Not defined
2	0	10	10	50	70	10	10	1.4	–1
3	50	10	0	10	70	60	50	42.9	1
4	50.5	10.5	.5	10.5	72	61	51	43.2	.98
5	10	5	5	50	70	15	15	3.2	.90
6	20	10	10	30	70	30	30	12.9	.71

It is not indicated in the table, but the expected value for cell D is 3.2 (i.e., under the minimum acceptable expected value). Row 5 indicates a situation where the expected value of a cell (cell A) is under 5 even when there is not a 0 in any cell. None of the cases (1–5) contains sufficient data to interpret the Yule's *Q*.

In row six of Table 7.1, note that when the simple probabilities of the target and antecedent behaviors are high enough, we do not need nearly as many coded units as when the simple probability of the target and antecedent behaviors are relatively low. Recall that the simple probability of a cell is the count of the cell divided by the total number of coded units. One reason the *expected* frequency of a cell, instead of just row and column marginal frequencies corresponding to a cell, is used to decide whether there are enough data is because all three bits of information (i.e., the two marginals and the total number of coded units) need to be considered. We need this information for all four cells because Yule's *Q* is based on all four cells, not just the A cell.

In summary, the consequences of computing Yule's *Q* when there is insufficient data is to compute uninterpretable Yule's *Q* scores that will not replicate because they are artifacts of inaccurate estimates of chance level sequential frequencies. When we compare sequential associations or when sequential associations are needed from multiple participants, *all* analyzed Yule's *Q* scores must be based on 2 × 2 tables with sufficient data.

Proposed Solutions for Insufficient Data

Ideally, we would address insufficient data problems by having multiple sessions for each participant and concatenating the data from multiple sessions that were recorded temporally close to each other within each participant. By concatenate, we mean copy coded data from session one and paste it onto the end of the coded data for session two, and so on. We would analyze the resulting "session" as if it were one session. This could be done for each participant in a group design or for several sessions in a single-subject design. Such an action implicitly assumes that the sequential association within each session is nonsignificantly different from those in the other sessions. This assumption is called the *stationarity assumption.* It makes sense that such pooling would assume that the sequential association of interest is approximately constant across sessions to be pooled. Otherwise, negative sequential associations in some sessions may cancel out positive sequential associations in other sessions.

In reality, stationarity is difficult to test statistically because each session often has insufficient data to compute an interpretable Yule's *Q* scores. So, we usually have to use logic to determine whether it makes sense to assume that the data to be pooled are likely to produce similar sequential associations. Logically, stationarity is more likely to be met for within-participant pooling, when sessions occur temporally close to each other, and when sessions are observed under similar conditions. We recognize that this is an expensive solution.

Some older sequential analysis articles have pooled across participants (Symons & Moran, 1994). This is not generally accepted anymore for two primary reasons. First, concatenating data across participants into a single file logically violates the stationarity assumption. Logically, the sequential associations between events are more likely to vary among participants than are the associations from multiple sessions within a participant. Second, there is no entity in nature that is modeled by the "pseudosubject" that is created from pooling across different participants. In contrast, when we concatenate within a participant, it is analogous to coding a session that lasts a long time (e.g., seven daily sessions concatenated can be conceived of as a week-long session). In other words, there is an entity in reality that the within-participant concatenated session models. One incorrect interpretation of sequential association from a pseudosubject is that it is similar to the mean of a distribution of sequential associations. A pseudosubject is not the same as a mean of a distribution of sequential associations because the former weighs more heavily data from participants with more coded units than participants with fewer coded units. This is not true for means of a distribution.

Another option for dealing with insufficient data is to lump subordinate categories of either the target or the antecedent or both. Recall that when the base rates of the antecedent and target are high, the number of total coded units does not have to be as high as it does when the base rates of antecedents and targets are lower. For example, we had insufficient data to determine whether one type of adult question had a stronger sequential association with child replies than did another type of adult question. Therefore, we lumped the question types together to determine whether adult questions (as a unitary category) had a stronger sequential association with child replies than did adult comments (Yoder, Davies, & Bishop, 1994). This, of course, required changing the research question.

Another option for dealing with insufficient data is to drop participants from the group design whenever there is inadequate data for any condition or sequence of interest. We recognize that this usually means reducing the statistical power of the subsequent group analysis (i.e., increases the probability of Type II error). We consider this preferable to analyzing uninterpretable dependent variables, which can produce Type I errors.

SEQUENTIAL ASSOCIATION INDICES AS DEPENDENT VARIABLES IN GROUP DESIGNS

In the following types of research questions, Yule's *Q* or transitional probabilities are the dependent variable scores for each participant (and condition) and asymptotic tests (e.g., *t*-tests, ANOVA, regression) are used to test the research question. In this sense, the index of sequential association is an observational variable metric.

Significance testing on a distribution of sequential association scores is less controversial than is testing the significance of a sequential association or a difference in sequential associations within a single participant. The reason is that all significance tests derive a probability (*p*) value to help interpret whether the observed effect size for the average sequential association could have occurred by sampling error. These *p* values are derived assuming that the units of analysis do not influence each other (i.e., are independent; Hayes, 1996). In a group analysis, the unit of analysis is the participant. As long as each participant's Yule's *Q* score is derived on a different session from other participants' Yule's *Q* scores, the scores are likely to be independent (i.e., not influence other dyads' or participants' Yule's *Q* scores).

Testing the Significance of a Mean Sequential Association

Even if other research questions are addressed, usually an investigator using a group design and a sequential observational variable metric (e.g., a Yule's *Q*) will need to ask whether the sequential association that is furthest from zero has a confidence interval that includes zero (i.e., is significantly different from zero). For example, if we had 20 participants with data similar to that presented in Figure 7.1 and conducted a 5-s time-window analysis for each to quantify the extent to which Parent Maintains had a sequential association with Child Engages, we would have 20 Yule's *Q* scores. This distribution of *Q* scores has a mean and a standard deviation (SD). One way to test the significance of the mean *Q* score from this distribution is to use a one-sample *t*-test to derive the *p* value for whether the mean *Q* score could have been sampled from a population with a mean *Q* of 0. The effect size for this question is Cohen's *d* for a single mean (i.e., mean *Q* score/SD of the *Q* scores). A Cohen's *d* of 0.5 or above is considered a moderate effect size (Cohen, 1988). Most people consider a moderate effect size noteworthy. It is recognized that using external qualitative benchmarks for effect sizes is only a proxy until a field is sufficiently mature to produce its own qualitative benchmarks for effect sizes. There have not been a sufficient number of competently executed sequential analyses to produce such empirically derived benchmarks as of 2009.

Testing the Between-Group Difference in Mean Sequential Associations

One might want to know whether the sequential association between Parent Maintains and Child Engages is different between children with autism than in children with Down syndrome. To address this, one would compute Yule's *Q* scores for the sequential association in all participants (e.g., 20) in each group (e.g., $N = 40$). The test of significance for whether the confidence interval around each group's mean *Q* score overlaps would be an independent *t*-test. The Cohen's *d* for this contrast would be (mean *Q* for Autism—mean *Q* for Down syndrome)/(pooled SD for *Q*).

Testing the Within-Subject Difference in Sequential Associations

When testing a within-subjects contrast of two sequential associations using a group design, an index of sequential association would be derived for all participants in all conditions. In a within-subjects group design, the data from all conditions are collected for all participants. A within-subjects contrast tests the significance of mean difference scores. Specifically, the significance test for such questions is whether the confidence interval around the mean difference score contains zero. The effect size for this contrast is (mean of [sequential association for condition 1 – same sequential association for condition 2])/([SD of the above difference scores]/square root [1 – correlation between the two sequential associations]; Lipsey & Wilson, 2001).

The particular index of sequential association that is appropriate for this type of question varies according to whether the base rate of the target varies between conditions that are to be compared. When the sequential associations being compared have the same target, different antecedents, and are derived from the same observation session, one can use either the transitional probability or the Yule's *Q* as the within-condition dependent variable. For example, one might want to know whether the forward sequential association of Parent Maintains with Child Engages is greater than that for Parent Introduces with Child Engages. The target is Child Engages for both sequential associations. If the data for both sequential associations were from the same observation session, the base rate for Child Engages will be the same for both sequential associations. Therefore, the fact that the base rate of the target influences the transitional probability is nonproblematic for this situation.

When the sequential associations being compared have different targets *or* the target behavior has the same label but is derived from different observation sessions, then Yule's *Q* should be the index of sequential association. This is because the target base rate will be different for the two sequential associations being compared and Yule's *Q* is not influenced by the base rate of the target (or the antecedent). For example, an investigator may want to know whether the sequential association of Adult Maintains with Child Engages is greater when interacting with the mother than with the teacher.

Testing the Significance of the Summary-Level Association Between a Participant Characteristic and a Sequential Association Between Behaviors

For example, an investigator may want to know whether children's receptive language level has a summary-level correlation with the sequential association between adult question and child replies. The base rate of child replies will be different for each participant. Therefore, we must quantify the sequential association of interest with Yule's *Q* for each participant. We correlate these Yule's *Q* scores with the children's receptive language level as tested by a language test. The significance of this correlation coefficient can be tested in the usual way (i.e., a *t*-test). The effect size of this correlation is the *r* value. Cohen considers an *r* value of .25 as a moderate effect size (Cohen, 1988).

STATISTICAL SIGNIFICANCE TESTING OF SEQUENTIAL ASSOCIATIONS IN SINGLE CASES

Any observation session will only compute an *estimate* of the true sequential association that exists between behaviors. That is, an observation session is only a sample of behavior. We call such an estimate a "point estimate." Sometimes, we wish this sample of behavior to represent a generalized characteristic (e.g., a tendency for one person to affect another using a process that continues to operate outside of the observation session). Under such conditions, we would like to know whether our observed Yule's *Q* could have occurred due to behavioral sampling error (e.g., chance sampling of an observation session that happened to produce an atypically strong sequential association). This is what significance testing usually does for us. There are methods for testing the significance of sequential associations within a single participant (Bakeman, Robinson, & Quera, 1996) and there are published examples of doing so (e.g., Yoder et al., 1994). Despite our own past use of such methods, using them is controversial and we now recommend using an alternative approach.

Similar to the aforementioned assumption that units of analysis are independent, statistical tests that have been applied to testing significance of sequential associations within a single case assume that units of analysis are exchangeable (Good, 2000). The unit of analysis in single-subject tests of significance of sequential associations is the unit represented by the tally in our 2 × 2 tables (e.g., the pair of behaviors or the coded time units). When applied to testing the significance of sequential associations within a single participant or dyad, one frequently recommended method involves randomly shuffling the events and computing a sequential association from the postshuffled event sequence and repeating this many times (e.g., 1,000) to create an empirical probability distribution against which one compares the observed (i.e., real) sequential association (Bakeman, Robinson, et al., 1996). This method is called a permutation test or randomization test (Edgington, 1995; Good, 2000).

Shuffling the sequence of events assumes exchangeability (Good, 2000). For them to be exchangeable, each pair of behaviors or time units cannot influence following pairs of behaviors or time units. Unfortunately, it is unlikely that the assumption of exchangeability is met for most sequential analyses of behavior in observation sessions. The problem occurs when an instance of antecedent behavior influencing the occurrence of a target behavior within one pair of behaviors influences the probability that the target will be influenced by the antecedent later in the session. This is clearly the case in sequential analyses of behavior because the pairs of behavior come from the same session and are from the same participants.

To make matters worse, pairs of behavior are more likely to influence following pairs of behavior that are temporally close than they are to influence following pairs of behavior that are temporally far but still within the session. This is called "nonuniform" dependence. An important simulation study has shown that nonuniform dependency causes more severe Type I errors (i.e., produces apparent associations even when it is known that no association occurs) than does uniform dependency between units of analysis (Hayes, 1996). It is important to note that this occurs even when the type of statistical test used is one that generates its own empirical probability distribution (i.e., a permutation test; Hayes, 1996).

Just as randomization tests (i.e., permutation tests) in single-subject experimental significance testing require randomizing the onset of the phase change in a way that mirrors the random assignment used in the permutations, such tests, when applied to sequential analyses, require that the permutations do not alter the *structure* of the data in a way that does not exist in nature (i.e., that they are *exchangeable*). This is a different issue from that of the independence of residuals and serial dependency or autocorrelation. Additionally, the problem is *not* that one behavior may influence another behavior within the pair or that one person's behavior may influence another person's behavior within the analysis unit (Bakeman & Dorval, 1989).

However, we also recognize that others continue to promote the same statistical methods for other situations in which units of analysis are dependent and nonuniformly so (Bulte & Onghena, 2008). Perhaps, it is best to say that the use of significance tests in single cases is controversial and, if used, should be done so cautiously.

Therefore, we offer an alternative. We recommend using the "moderate" level benchmark for Yule's *Q* that was provided earlier to determine if the sequential association or difference between sequential associations is noteworthy. That is, if the Yule's *Q* or difference Yule's *Q* in question is greater than |.43|, then we suggest considering it worthy of further investigation (i.e., attempts a replication of the observed effect size or larger). We recognize that this is tantamount to interpreting a point estimate, but this is where the field is at present. As in other single-subject research, we are dependent on replication of the sequential association to help judge the generality of our observed finding within a participant.

For example, if the investigator wants to know whether the sequential association between Parent Maintains and Child Engages is noteworthy, he could compare the observed Yule's *Q* for this sequential association with the moderate effect size of |.43|. Similarly, if the investigator wants to know whether the sequential association between Parent Maintains and Child Engages is different enough to be considered noteworthy from the sequential association of Parent Introduces and Child Engages, he could compare the difference Yule's *Q* to |.43|. It is important to note that it is critical that each Yule's *Q* that is interpreted, even at the single case level, has sufficient data as indicated by the previous guidelines.

Another way to use Yule's *Q* as a dependent variable score in single-subject research is to graph it for each session (or group of sessions if needed to get enough data for each Yule's *Q*). In this way, one might test whether there is a change in level, trend, or variability of the sequential association between two behaviors as a function of some treatment.

A CAVEAT REGARDING THE USE OF YULE'S *Q*

There are conditions where using Yule's *Q* will probably not provide users with the information they want. For example, if the user wants to know whether a consequence (i.e., target or second) behavior is a probable reinforcer of a given (i.e., first) behavior, there are situations where one can have a high Yule's *Q* but a low "operant contingency." Behavioral learning theory indicates that one characteristic of a reinforcer is that it occurs "contingently" after the behavior of interest. Technically, this means that the consequence occurs if and only if the behavior of interest occurs. In reality, this type of perfect contingency rarely occurs. The term "*operant contingency*" describes a more realistic situation: one in which the probability of a consequence after a given behavior exceeds the probability of the consequence not after the behavior of interest (Hammond, 1980; Martens, DiGennaro, Reed, Szczech, & Rosethal, 2008).

A comparison of these two transitional probabilities has been called *contingency space analysis* (Martens et al., 2008). Contingency space analysis and Yule's *Q* will always provide the same algebraic sign. However, when the behavior considered to be the probable reinforcing consequence occurs almost all of the time, regardless of whether the preceding desired behavior of interest has occurred, Yule's *Q* will be much higher than the difference between the key transitional probabilities (i.e., the contingency space analysis index). It is the latter, the contingency space analysis, that more closely describes contingent relations in which the consequence is likely to function as positive reinforcement (i.e., increases the probability of the preceding behavior). In other words, following *Q* under these conditions would lead clinicians to select a consequence that would be unlikely to function as a reinforcer.

At present, a weakness of contingency space analysis is that we do not have benchmarks for how large the difference between transitional probabilities needs to be to consider it noteworthy. Additionally, we do not have guidelines for how much data is necessary to consider the contingency space analysis informative. Once these details have been provided, contingency space analysis is likely to be a very useful tool for those with an interest in using sequential analysis to discover probable reinforcers.

RECOMMENDATIONS

Software is needed to conduct sequential analyses. One such software, MOOSES, can be acquired at www.getmooses.com. MOOSES and other sequential analysis software compute and output Yule's *Q*, which is an index of sequential association. Yule's *Q* is a superior index of sequential association when compared to transitional probability. Yule's *Q* is computed (thus a separate 2 × 2 table is constructed) for each sequential association, condition, and participant in one's study. It is critical that one conduct sequential analyses with sufficient data to produce interpretable Yule's *Q* scores. If one must concatenate sessions to attain enough data to produce interpretable Yule's *Q* scores, then the investigator can concatenate sessions that are temporally close within each participant.

If using a group design to address research questions involving a sequential association, Yule's *Q* is usually the recommended dependent variable. One can then use the common tests of significance to test the research question. If using a single-subject design to address a research question involving a sequential association, we suggest relying on qualitative benchmarks to decide if a single Yule's *Q* is noteworthy and replication to inform us of the generality of the findings within a participant.

Finally, Yule's *Q* can be used as a dependent variable that is graphed as a function of experimental phases in a single-subject experimental design. In these ways, Yule's *Q* is a metric for observational variables designed to quantify sequential associations among two behaviors. When attempting to discover whether a consequence behavior is likely to act as a reinforcer and when the consequence behavior occurs very frequently, including in intervals or time units the "given" desired behavior does not occur, contingency space analysis offers a superior method to interpreting sequential data over Yule's *Q*.

REFERENCES

Bakeman, R., & Dorval, B. (1989). The distinction between sampling independence and empirical independence in sequential analysis. *Behavioral Assessment, 11*(1), 31–37.

Bakeman, R., McArthur, D., & Quera, V. (1996). Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared. *Behavior Research Methods, Instruments and Computers, 28*, 446–457.

Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS & GSEQ*. New York: Cambridge University Press.

Bakeman, R., Robinson, B., & Quera, V. (1996). Testing sequential association: Estimating exact p values using sampled permutations. *Psychological Methods, 1*, 4–15.

Bulte, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467–478.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Edgington, E. S. (1995). *Randomization tests* (3rd ed., revised and expanded). New York: Marcel Dekker.

Good, P. I. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd ed.). New York: Springer.

Hammond, L. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the Experimental Analysis of Behavior, 34*, 297–304.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H-sub-0: rho = 0. *Psychological Methods, 1*(2), 184–198.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Martens, B., DiGennaro, F., Reed, D., Szczech, F., & Rosethal, B. (2008). Contingency space analysis: An alternative method for identifying contingent relations from observational data. *Journal of Applied Behavior Analysis, 41*, 69–81.

Reynolds, H. T. (1984). *Analysis of nominal data* (2nd ed.). Beverly Hills, CA: Sage.

Rosenthal, J. A. (1996). Qualitative descriptors of strength association and effect size. *Journal of Social Science Research, 21*(4), 37–59.

Symons, D. S., & Moran, G. (1994). Responsiveness and dependency are different aspects of social contingencies: An example from mother and infant smiles. *Infant Behavior and Development, 17*, 209–214.

Tapp, J. (1995). *Multiple options for observation in experimental studies*. Nashville, TN: Vanderbilt Kennedy Center.

Wickens, T. D. (1993). Analysis of contingency tables with between-subjects variability. *Psychological Bulletin, 113*, 191–204.

Yoder, P. J., Davies, B., & Bishop, K. (1994). Reciprocal sequential relations in conversations between parents and children with developmental delays. *Journal of Early Intervention, 18*(4), 362–379.

Yoder, P. J., Short-Meyerson, K., & Tapp, J. (2004). Measurement of behavior with special emphasis on sequential analysis of behavior. In E. Emerson, C. Hatton, T. Thompson, & T. Parmenter (Eds.), *International handbook of applied research in intellectual disability* (pp. 179–202). West Sussex, UK: Wiley.

This page intentionally left blank

8 Observer Training, Observer Drift Checks, and Discrepancy Discussions

OVERVIEW

This chapter is about a process by which observers are trained and retrained to maximize the accuracy of their coding. It focuses on agreements and disagreements of *coding decisions*, not reliability of observational variables. We discuss the use of point-by-point agreement between an observer and a repeatedly and expertly coded example file (i.e., a criterion coding standard), as well as agreement between two observers. The most important activity that results from agreement checks is a discussion among observers or, ideally, between the content expert and the observers (i.e., a discrepancy discussion). The results of such a discussion may be (a) slight alterations in the coding manual or (b) (re)training of observers. We also discuss one method of training observers.

THREE PURPOSES OF POINT-BY-POINT AGREEMENT ON CODING DECISIONS

Point-by-point agreement is the extent to which two people categorize the same occurrence of a key behavior in the same category. It is in contrast to *summary level agreement*, which is the extent to which two people derive the same variable score (i.e., small estimate/large estimate proportion). The three purposes of point-by-point agreement checks that are covered in this chapter are (a) to provide a standard for observer training, (b) to indicate the degree to which a coding manual is adequate, and (c) to detect *observer drift* (i.e., the occurrence of an observer agreeing less often with a criterion coding standard than was the case immediately after he or she reached mastery level accuracy during initial training). These are purposes that serve idemnotic and vaganotic measurement concepts and group and single-subject research designs.

All observational measurement experts acknowledge that two observers can agree, even at a point-by-point level, and still be inaccurate. However, point-by-point agreement is a process by which we maximize the *probability* of accurate coding. This result is not magic. For point-by-point agreement to be high, observers need to be vigilant and the coding manual needs to be precise. In addition, the discussions regarding why disagreements occur help resolve and keep observers mindful of the difficult aspects of the coding task.

TWO DEFINITIONS OF AGREEMENT

To conduct a point-by-point agreement check, a definition of point-by-point agreement is necessary. There are two primary definitions of agreement: exact and time window. Exact agreement means that the time of and category for a coded unit is coded the same way by both observers. For example, both observers record that a smile occurs in interval 12. This is applicable when duration is the metric that will be derived from timed-event behavior sampling or when interval sampling is used (i.e., exhaustive coding spaces). Exact agreement is illustrated in Table 8.1.

When one is scoring point-by-point agreement for a *nonexhaustive coding space* (i.e., timed event sampling where number of the behavior is the metric of interest), we typically use time-window agreement. Time-window agreement is used when the onset of the behavior coded by one observer occurs within a prespecified time of the onset of the same behavior as indicated by the other observer (MacLean, Tapp, & Johnson, 1985). For example, if the prespecified time is set to 5 s and both coders code the behavior in that time frame, then time-window agreement has occurred.

There is no consensus on how long the time window should be. However, logic suggests that behaviors with brief interoccurrence

Table 8.1

ILLUSTRATION OF EXACT AGREEMENT ON INTERVAL DATA

INTERVAL	OBSERVER 1	OBSERVER 2	AGREEMENT SCORE (1 = AGREEMENT; 0 = DISAGREEMENT)	TYPE OF DISAGREEMENT (U = UNITIZING; C = CLASSIFYING)
1			1	
2	1		0	u
3	3	2	0	c
4			1	
5	2	2	1	
6			1	
7	1		0	u
8		1	0	u
9			1	
10	3		0	u

intervals require smaller time windows than behaviors with longer interoccurrence intervals. Time-window agreement is usually conducted by software. However, it is important to know what the software is doing to make sure the investigator agrees with the method used. Table 8.2 illustrates a useful method of conducting time-window agreement. The accuracy of computer-scored time-window agreement is greatly enhanced if humans are willing to aid the software program in deciding which coded behaviors "go together."

Ideally, the software program uses the time of occurrence and code for the act to "guess" which acts from Observer 1 "go with" the acts from Observer 2 and places its decision on a side-by-side screen display. The staff member uses the sequence of events, time of occurrence, and coding comments to support a judgment regarding whether to change or leave the software program's guess. A revision of MOOSES

Table 8.2

ILLUSTRATION OF 1-SECOND TIME-WINDOW AGREEMENT METHOD OF DEFINING POINT-BY-POINT AGREEMENT OF TIMED-EVENT DATA FOR WHICH BEHAVIOR IS THE CODED UNIT

TIME (S)	OBSERVER 1	OBSERVER 2	AGREEMENT SCORE (1 = AGREEMENT; 0 = DISAGREEMENT)	TYPE OF DISAGREEMENT (U = UNITIZING; C = CLASSIFYING)
1				
2	1		0	c
3		2		
4				
5		2	1	
6	2			
7				
8		1	0	u
9				
10	3		0	u
11				
12		3	0	c
13	1			
14				
15		1	0	u

(www.getmooses.com), which was introduced in the previous chapter, will enable this function. However, if the observer chooses not to use this feature or is using software that does not enable such human interface, then a number of observational software programs can score the agreements without human intervention regarding agreements.

AGREEMENT MATRICES

To use point-by-point agreement checks to their maximum benefit, it is extremely useful to construct an agreement matrix for each set of mutually exclusive codes (i.e., a "group" in ProcoderDV). An agreement matrix is a type of symmetrical matrix in which the rows and columns are the categories in the, coding group plus a row and column for "no coded behavior tallied" for each observer. Recall that a coding group is a set of categories for the same behavior dimension or actor. The rows represent one observer's coding, while the columns represent the other observer's coding. The tallies are the result of point-by-point agreement checks. Agreements are represented on the diagonal and disagreements are represented on the off-diagonal cells.

There are two types of disagreements. A *unitizing difference* occurs when one coder recorded an act while the other did not. A *classifying difference* occurs when one coder classified an act as Category A while the other classified the same act as Category B (Bakeman & Gottman, 1997). The purpose of agreement matrices is to identify (a) the times of disagreed upon acts, (b) disagreement types (i.e., unitizing vs. classifying), and (c) categories that tend to be confused in classification disagreement types. Doing so should increase the efficiency and efficacy of staff training by aiding discrepancy discussions.

Before discussing the details of agreement matrices, it is important to discuss the principles of proper agreement checks. First, it is important that all events coded by *either* observer are counted in the agreement matrix. Unfortunately, when examining point-by-point agreement for nonexhaustive coding spaces, some investigators have ignored disagreements caused by the "secondary" observer coding a behavior and the "primary" observer not coding it. It becomes clear that this is inappropriate when we create agreement matrices for exhaustive coding spaces because the number of events of the secondary coder does not equal to the sum of rows and it must. In point-by-point agreement checks on both exhaustive and nonexhaustive coding spaces, disagreements are counted regardless of who "caused" them. Second, it is extremely important to consider unitizing and classifying errors as equally important. A misguided approach that has occurred in some published articles is to examine agreement only on acts that both coders identify as relevant to count when counting disagreements. Either misguided practice inflates point-by-point agreement and undermines the discrepancy discussion process, the crux of the training and retraining process. For example, if we only count the events that both observers record in our record of agreement of classifying events (e.g., when we first create a transcript, have another person verify that the utterance is present and transcribed "correctly" and later classify the utterances), we inadvertently eliminate from our agreement estimate all events that one observer records but the other does not.

There are two types of agreement matrices: (a) one for exhaustive coding spaces (i.e., those in which agreement on nonoccurrence of any key behavior is defined) and (b) one for nonexhaustive coding spaces (i.e., those in which agreement on nonoccurrence of any key behavior is not defined). For exhaustive coding spaces, constructing the agreement matrix is very straightforward. The process is illustrated in Table 8.3. We need an indication of agreement on nonoccurrence of all key behaviors because this information is part of what we need to estimate nonchance agreement, as will be explained in greater detail in chapter 9. The numbers in parentheses in Table 8.3 indicate the numbers that label the intervals in Table 8.1 that are tallied into the cells of the agreement matrix in Table 8.3. The lower right cell is the total number of intervals coded and is the sum of the marginals.

Table 8.3

AGREEMENT MATRIX FOR DATA IN TABLE 8.1 (AN EXHAUSTIVE CODING SPACE) WITH INTERVAL NUMBER OF TALLIED BEHAVIOR IN PARENTHESES

		OBSERVER 2				
		1	2	3	NO BEHAVIOR CODED IN INTERVAL BY OBSERVER 2	MARGINAL FOR OBSERVER 1
OBSERVER 1	1				(2), (7)	2
	2		(5)			1
	3		(3)		(10)	2
	No behavior coded in interval by Observer	(8)			(1), (4), (6), (9)	5
MARGINAL FOR OBSERVER 2		1	2	0	7	10

The agreement matrix for nonexhaustive coding spaces is more complicated. Recall that the time-window definition of agreement is usually used in such cases. Unlike the agreement matrix for exhaustive coding spaces, this type of agreement matrix allows the sum of the marginals for Observer 1 to be different from the sum of the marginals for Observer 2. In the agreement matrix in Table 8.4, each of the set of values in parentheses is the range of seconds for the act(s) indicated in Table 8.2 that are tallied into a cell in the agreement matrix in Table 8.4. Note that the computer will score some of these events (e.g., 10–12 or 13–15) differently than a human might. An observer might decide to override the computer scoring and call such pairs of events agreements because he can clearly see that the problem is really just marking the onset of the same act at a slightly different time (a minor error for most research questions).

Table 8.4

AGREEMENT MATRIX FOR AGREEMENT DATA IN TABLE 8.2 (A NONEXHAUSTIVE CODING SPACE)

		OBSERVER 2				
		1	2	3	NO BEHAVIOR CODED BY OBSERVER 2	MARGINALS FOR OBSERVER 1
OBSERVER 1	1		(2–3)	(12–13)[a]		2
	2		(5–6)			1
	3				(10–11)	1
	No behavior coded by Observer 1	(8–9) (15)			Undefined	Undefined
MARGINALS FOR OBSERVER 2		2	2	1	Undefined	Potentially different sum between observers[b]

[a] Range of seconds of occurence of tallied behaviors shown within parentheses.
[b] Five behaviors coded for Observer 2; four behaviors coded for Observer 1.

DISCREPANCY DISCUSSIONS

A discrepancy discussion occurs when two observers discuss with each other or with a content expert the rationale for their coding particular coded units that are coded differently across observers. A note of caution is called for if a content expert is not a part of the discrepancy discussions. When parties disagree regarding how a discrepancy should be settled, the content expert must make the final decision.

The focus of discrepancy discussions is on the reason for the disagreement. This process is considered critical for discovering where in the coding process coders need to be retrained, or eventually even fired or reassigned to a different task. Without such detailed information, we may not be able to prevent two well-known phenomena that result in reduced accuracy in coding after initial coding training ends: observer drift or consensual drift. *Observer drift* is used here to refer to a frequently documented occurrence of observers becoming less accurate (i.e., agreement with a coding standard) the longer it has been since their last training (Romanczyk, Kent, Diament, & O'Leary, 1973; Taplin & Reid, 1973). *Consensual drift* is when two observers agree with each other but neither agrees with a coding standard (Johnson & Bolstad, 1975). Although some have suggested prohibiting observers from discussing coding disagreements to prevent consensual drift (Repp, Nieminen, Olinger, & Brusca, 1988), it has never been shown that discrepancy discussions result in consensual drift. Indeed, informal experience leads us to conclude that such discussions may prevent both types of drift, particularly if the content expert is a part of many of these discussions.

The person running the discrepancy discussion (the discussion leader) shows the video clip of the behavior in question and attempts to elicit observers' rationale for their coding decision. Observational software such as ProcoderDV can use the time of occurrence or interval to efficiently replay key scenes. However, replaying parts of tapes to guide discrepancy discussions can be used without observational software. This is one reason why timed-event or interval behavior sampling is necessary for discrepancy discussions. Without replaying the scene that evoked the disagreement, discussions will be based on differing memories of what occurred, not on actual behavior and context. If the type of difference is a unitizing one, the discussion leader points out the accurate and inaccurate aspects of the rationale for coding the act and identifies the relevant parts of the coding manual that document the correct rules for coding the act. Many unitizing errors are due to lack of clarity in or insufficient attention to segmenting rules in the coding manual. If the type of difference is a classifying difference, the agreement matrix will indicate on which categories the trainees differ. The discussion leader talks about the relevant aspects of the coding manual and shows relevant videotaped examples from the session or from the criterion coding standard (see Criterion Coding Standards below) to help clarify the needed concepts. When observers fall below the agreement standard, they frequently need to recalibrate with a criterion coding standard.

CRITERION CODING STANDARDS

We rarely (some would say never) see or are assured of what is exactly true or correct. So, *estimating* accuracy is almost always the best we can do. Practically speaking, our "best estimate" of the true occurrence of events in a session is a repeatedly and expertly coded session. We call this a "criterion coding standard" (Sharpe & Koperwas, 2003). The purposes of criterion coding standards are to (a) initially train observers and (b) retrain observers when agreement is below interobserver agreement standards.

One process used to create criterion coding standards is presented here. It is an adaptation of a process described by Sharpe and Koperwas (2003). The first step is to create a videotape or media file of many scenes that illustrate at least four examples of all levels of all categories. Ideally, we need at least two criterion coding standards: (a) a demonstration coding criterion and (b) a training coding criterion. Therefore, as one is going through tapes to find training scenes, it is wise to identify at least eight examples for each category (four for each standard). A less expensive option is to have staff act out scripted scenes and edit these. This process will result in oversampling scenes of infrequently occurring categories and possibly undersampling scenes of frequently occurring categories. More than four examples are recommended for particularly difficult categories or aspects of the code (e.g., segmenting).

For example, if experience indicates that observers have difficulty segmenting communication acts when they occur rapidly and are interrupted by communication partners, then more than four examples of such acts should be provided. Within each criterion coding standard file, the scenes should be randomly sequenced. We understand this is an expensive process and may not be possible for all projects. However, it is useful to have an ideal from which to depart. Even if the first step is not used, one can use steps 2 through 4 on uneditted sessions selected for multiple instances of acts. The disadvantage of using real sessions, even of hand-selected sessions for training, is that they will frequently have insufficient numbers of examples of rare or difficult behaviors to classify to serve as optimal training stimuli.

Second, the first half of the criterion coding standard or selected session is repeatedly coded by experts. The experts should be very knowledgeable about the substantive area covered by the coding manual. They should code the session independently using timed-event behavior sampling. Timed-event behavior sampling is suggested even if other behavior sampling methods will be used by project staff to collect the data for the project because timed-event behavior sampling will provide optimal control over the media for discrepancy discussion and play back of key scenes for training purposes. Next, the observers identify acts that are discrepantly coded, including acts that only one observer coded and acts that both coded but coded differently. For each act on which there is disagreement, the experts hold a discrepancy discussion to (a) identify ambiguous terms in the coding manual and (b) decide on the most accurate way to code the act. If the problem is that certain terms in the coding manual are ambiguous, the coding manual is clarified through deeper levels of operational definitions and more examples or near nonexamples. After waiting for a sufficient period to prevent easy recall of specific coding decisions (e.g., 2 weeks), the first half of the first criterion coding standard file is recoded by the same two experts using the new coding manual. The discrepancy discussion process is repeated for any remaining acts on which there is disagreement. If the agreement standard (to be discussed in chapter 9) has not been met, this process is continued until the agreement standard is achieved, the agreement window is changed, the agreement standard is changed, or the measurement system is changed. The experts may decide to change the measurement system in one or more of the following ways: (a) collapse subordinate categories, (b) use a computer software program to aid coding, and (c) use a stop-and-go coding or multiple pass coding method. If so, then this new method becomes the coding method used by project staff.

Third, once criterion level agreement is reached, the experts code the second half of the first criterion coding standard using the newly revised measurement system. The above process is repeated until the agreement standard is achieved in the second half of the session. Eventually, the experts use discrepancy discussions to decide how discrepantly-coded acts should be coded for training the project staff. It is important that the discrepancy discussion process be implemented using logic and adherence to the coding manual, not force of personality, to make the final decisions.

Fourth, the same process is used to consensus code the second criterion coding standard session. The second one should be easier to code because of the alterations to the measurement system.

Fifth, the four segments of criterion coding standards are created: a short one for each third of the training criterion coding standard and one long one for the demonstration criterion coding standard. These will be used during the observers' training process. A new version of MOOSES will soon be available to create accuracy matrices (agreement between the trainee and the consensus-coded standard). If the observers will be using an interval coding system, then one of the project staff must transform the onset data for the consensus-coded standard data into an interval coded file.

OBSERVER TRAINING

Again, the following is an adaptation of a process outlined by Sharpe and Koperwas (2003). First, the lead investigator, presumably one of the content experts for the coding manual, holds discussions with the observer trainees about the coding categories and, if relevant, the generalized characteristics being measured. Part of this process is to show the demonstration criterion coding standard. If the project leader is using a program such as ProcoderDV, one can use the observation file that contains the consensus coding record, sort by the code that the leader is teaching about, and show the video scenes coded for that category that is on the demonstration coding standard. In this way, the sequence in which the first criterion coding standard is viewed does not have to be random but may be grouped by category to help staff trainees develop the key concepts. The sequence of delivering this presentation of the categories may be such that categories that are particularly likely to be confused can be shown back to back. The lead investigator uses questions and answers to attempt to draw the trainees into an active processing and discussion about the categories they are expected to learn.

The second step is for the project director or technician to demonstrate the recording method to the trainees. Then, the trainees are asked to engage in an exercise designed to get the trainees to use the coding recording method with support of the project director and other trainees so that questions can be addressed and manuals designed to guide the trainee through the use of the coding recording method. As part of this step, independent and accurate use of the coding recording method (e.g., use of the coding software) must be demonstrated. One hundred percent accurate use of the procedure is necessary. This level of accurate use is reasonable to ask from most people who will be expected to use the procedure while making coding decisions, often under rapid or fatiguing conditions.

Third, the observers independently code the first third of the training criterion coding standard file. For example, if the entire training criterion coding standard session is 30 min long, then the trainees code the segment for the first 10 min. Afterward, each observer independently uses the software to create an agreement matrix on the consensus coding of the training criterion as the comparison file to identify the time, number, and types of disagreements each observer has with the coding standard. The content expert then uses the discrepancy discussion process that is indicated above to address the types of errors the trainee displays.

Fourth, each trainee observer independently codes the second third of the training criterion coding standard file. The same process that was indicated above is repeated.

Fifth, the trainees independently code the last third of the training criterion coding standard file. The same process as was indicated above is repeated.

Sixth, for each accuracy matrix (i.e., agreement matrix for the trainee's coding with the expert consensus-coded file), compute agreement for *each* category using one of the methods indicated in chapter 9. If there is no improvement from the first third to the last third and all three training segments produce agreement with the consensus-coded file that is under criterion level agreement, the lead investigator needs to consider the nature of the disagreements. If the disagreements are probably due to the staff not learning the categories, then consider moving the trainee to a new duty or firing and rehiring staff. There are simply differences in the verbal facility and observational skills of adults who are hired as staff that cannot be easily changed within the resource limitations of most research projects. This process should result in some increased accuracy if the staff member is well suited to the job. If new staff members are hired or moved into observer positions, then the above process is repeated for them. If the disagreements are due to human or technical limitations (e.g., the observer cannot see participants' eyes, but need to in order to code accurately), then consider further changes in the measurement system to aid coding and repeat the above procedure until criterion level agreement is met.

Seventh, once improvement in point-by-point agreement is shown in the codings on the criterion coding standard file, the pair of trainees independently codes sessions that have not yet been coded by experts. During this stage of coding training, each trainee is coding independently entire sessions that may represent real data on which the research question will be addressed. After coding, the trainees produce an agreement matrix using the software indicated above that represents the degree to which the trainees agree with each other. The content expert meets with both (or more) trainees for a discrepancy discussion. With the expert, consensus coding decisions are made and the consensus-coded file is the basis for the primary observational variable score for each observational variable to be derived from that session. If interobserver agreement is above the criterion level for each observational variable, that session is counted toward the goal of achieving three consecutive sessions coded above criterion agreement level. If not, then this process is repeated until this goal is achieved. After establishing the required agreement level, we conduct regular agreement checks using the following principles.

METHOD OF SELECTING AND CONDUCTING AGREEMENT CHECKS

Agreement checks are generally conducted on a subset of the data because it is generally prohibitively expensive to conduct them on all sessions. Therefore, it is important that the process by which we select the sessions to be checked for agreement informs us of the level of interobserver agreement in the unchecked sessions. For this to occur, we need to be concerned about (a) agreement sample size (i.e., the number of sessions selected for agreement checks), (b) the method by which we select the agreement check sessions, (c) when during the study they are selected and checked for agreement, and (d) whether the primary coder knows which sessions are checked for agreement.

We know from sampling theory that smaller samples (i.e., number of sessions) tend to be less representative (i.e., produce estimates that are not close to the total sample mean) of the total study data than are larger samples. However, like all conventions, there will be disagreement regarding how large an agreement sample is needed for it to yield estimates that are representative of total data set. One reason for the controversy is that agreement proportions are only point estimates of what occurs in the population of data. That is, there is a confidence interval around these agreement estimates. A confidence interval is influenced by the size of the reliability sample (not the proportion with respect to the total data set) and the average variability of agreement estimates from the mean agreement estimate in the reliability sample (i.e., the SD of the agreement estimates). However, we do not know the SD of the agreement estimates until the study is over. And yet, we need to collect agreement checks as the study is being conducted to prevent observer and consensual drift. Therefore, conventions for reliability sample sizes are needed. We offer the following: For group designs, the reliability sample should be at least 20% of the total number of sessions; for single-subject designs, the reliability sample should be at least two sessions per design phase and at least 33% of the total data (Kazdin, 1982).

Sampling theory also tells us that randomly selected samples are more representative of the total data set than are nonrandomly selected samples. However, we want for our reliability sample to represent all design phases, groups, and conditions proportionally. Therefore, we generally, stratify (group according to the design element) prior to randomly selecting from each pool of sessions.

Sampling throughout the study and immediately conducting agreement checks and subsequent discrepancy discussions are necessary to prevent drift. If we wait to conduct such agreement checks at the end of the study, then we cannot retrain if drift is seen. If we do all of our agreement checks at the beginning of the study, we cannot test for, and retrain in the occurrence of, drift. We have found that doing a random agreement check for every fifth session coded is a useful way to determine whether retraining is necessary prior to doing more coding. If the agreement level is below criterion level, then we retrain and recode the sessions since the last time agreement was adequate (at most four sessions).

Finally, the primary coder should *not* know which sessions are to be checked for agreement. This is difficult to accomplish if live coding is conducted. Regardless, there is replicated evidence that primary coders are more accurate (Reid, 1970) and produce higher agreement scores (Romanczyk et al., 1973) when they believe they are being checked for accuracy or agreement than when they do not know whether they are being checked for accuracy or agreement.

It should be stated that several sessions that are double coded for the purpose of coder training and retraining are sometimes concatenated because there are so few instances of rare behaviors in any one session. This approach is useful for checking for observer drift because it is difficult to judge the seriousness of one or two disagreements on rare behaviors. However, this approach is not recommended when testing reliability of observational variables for reasons that will be discussed in chapter 9.

RETRAINING WHEN OBSERVER DRIFT IS IDENTIFIED

Retraining is necessary when agreement checks indicate that the interobserver agreement is below the criterion agreement level (see chapter 9 for computation of indices and setting criterion agreement standards). The retraining process is composed of (a) recalibrating with the demonstration (and perhaps training) criterion coding standard(s) followed by (b) consensus coding with another staff member who is still coding with acceptable levels of accuracy. Each of these is followed by discrepancy discussions with the content expert. After the observer codes the demonstration criterion coding standard, the content expert holds the discrepancy discussion. If the agreement level is below criterion level, then the observer codes the training criterion coding standard and another discrepancy discussion is held. Agreement with the coding standard is computed. If still below the agreement standard, then staffing decisions are revisited. If agreement is sufficient, the sessions that were coded by a single coder since the time that the agreement criterion level as last met (e.g., the other four sessions from the set of the most recently coded five sessions) are recoded by a second coder and the software indicated above is used to identify a list of times at which coding discrepancies occurred. The two, now calibrated observers use the discrepancy discussion process to produce consensus scores for the sessions that were recoded. Consensus scores replace the initial scores in the spreadsheet for the data that will be used to address the primary research question to "correct" the error that was detected during the agreement check. However, the scores *prior* to consensus coding are used to compute agreement and reliability for observational variable scores (see chapter 9 for details). Once retraining is complete, the observer needs to reestablish criterion-level agreement on three consecutive sessions before beginning the usual schedule of coding. These can be done on sessions that have not yet been coded.

RECOMMENDATIONS

If resources allow it, we recommend using edited files (i.e., a criterion coding standard) to initially train observers and retrain observers after drift occurs. If one cannot afford to create edited files as training stimuli, then we recommend using sessions selected for frequent use of rare behaviors as training stimuli. In either case, we recommend using a consensus coding by a pair of content experts as the best estimate of "accurate coding." These coding standards can be used to train and retrain observers. Retraining is necessary when a representative sampling of the sessions indicates subcriterion level agreement between pairs of observers. The sessions checked for agreement are most likely to represent the total sessions coded in the study when a sufficiently large number of sessions are selected in a representative fashion and the primary coder is blind to which sessions will be checked for agreement. Point-by-point agreement checks using agreement matrices and discrepancy discussions are the basis for checking for drift and retraining. It is critical that agreement matrices include (a) all sources of differences in the same matrix and (b) the disagreements caused by either observer are considered equally.

At the time this chapter was written, a new version of MOOSES was being developed to facilitate the agreement matrix and discrepancy discussion processes described in this chapter. When completed, it will be available at www.getmooses.com.

REFERENCES

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Johnson, S. M., & Bolstad, O. D. (1975). Reactivity to home observation: A comparison of audio recorded behavior with observers present or absent. *Journal of Applied Behavior Analysis, 8*, 181–185.

Kazdin, A. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

MacLean, W., Tapp, J., & Johnson, W. (1985). Alternate methods and software for calculating interobserver agreement for continuous observation data. *Journal of Psychopathology and Behavioral Assessment, 7*, 65–73.

Reid, J. B. (1970). Reliability assessment of observation data: A possible methodological problem. *Child Development, 41*, 1143–1150.

Repp, A., Nieminen, G., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. *Exceptional Children, 55*, 29–36.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. (1973). Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis, 6*, 175–184.

Sharpe, T., & Koperwas, J. (2003). *Behavior and sequential analysis*. Thousand Oaks, CA: Sage.

Taplin, P. S., & Reid, J. B. (1973). Effects of instructional set and experimental influences on observer reliability. *Child Development, 44*, 547–554.

This page intentionally left blank

9 Interobserver Agreement and Reliability of Observational Variables

OVERVIEW

In contrast to the purposes of training, retraining, and drift checking that were discussed in chapter 8, this chapter will focus on using the sessions selected for interobserver agreement checks to estimate the extent to which the different observers produce similar observational *variable scores* using the same measurement system. The level at which this similarity must occur and the relevance of whether participants vary on their observational variable scores to interpreting interobserver reliability estimates depends, in part, on the study's research design and the investigator's measurement perspective.

ADDITIONAL PURPOSES OF POINT-BY-POINT AGREEMENT

In addition to the purposes of point-by-point agreement checks provided in chapter 8, there is an additional purpose of interobserver agreement checks. In single-subject design studies and fidelity of treatment (FOT) scores assumed to be uniformly high, point-by-point interobserver agreement is often thought of as interobserver reliability. FOT is a coded description of the extent to which persons implementing the treatment do so as intended by the investigator. Most investigators want FOT measures to be uniformly high; therefore, variability among participants is not expected. Both these purposes of measurement are idemnotic in the sense that the investigator is interested in absolute, not relative, values of what occurs. Therefore, it is not appropriate that variance among units of analysis (i.e., sessions in single-subject experiments and participants in group-design FOT scores) affect these reliability estimates. Point-by-point agreement estimates do not reflect variance among units of analysis.

Point-by-point agreement in single-subject design and in FOT measures is arguably the most important evidence of "validity." Particularly when measuring context-dependent behavior, measuring occurrence accurately is one of the primary meanings of "validity" (Haynes & O'Brien, 1999; Johnston & Pennypacker, 1993). All observational measurement experts agree that agreement is not synonymous with accuracy (i.e., the extent to which an observer reports the true occurrence and nonoccurrence of behavior). However, observers with high agreement are thought to be more accurate than observers with low agreement.

One reason that point-by-point agreement is so important for single-subject and FOT variable measurement is the heightened possibility or certainty that the behavior of interest will be observed during treatment sessions. By definition, FOT measures always come from the treatment sessions. In many single-subject designs, the dependent variables are measured from treatment sessions (Kazdin, 1981). If presence of the treatment is apparent to observers, measuring the variable during the treatment sessions will inform the observer of the design phase or group to which the session belongs. This is a potential problem because it has been repeatedly shown that when observers believe the session is from the treatment phase or group, scores from these sessions are systematically inflated (Kent, O'Leary, Diament, & Dietz, 1974; O'Leary & Kent, 1977). Fortunately, it has also been repeatedly shown that proper training, which in part involves point-by-point agreement checks and standards, can reduce such expectancy effects on the accuracy and agreement of coding (Kent et al., 1974; Redfield & Paul, 1976).

ADDED PRINCIPLES WHEN AGREEMENT CHECKS ARE USED TO ESTIMATE INTEROBSERVER "RELIABILITY" OF OBSERVATIONAL VARIABLE SCORES

When sessions from which point-by-point agreement is used to estimate the reliability of observational variables, both concepts of measurement (idemnotic, vaganotic) and research design types (single -subject, group) require that the (a) variable metric and (b) unit of analysis on which reliability is estimated be the same as is used to test the research questions. For example, when a proportion metric is used to address the research question, it is the interobserver agreement on the proportion that is relevant. A common, but misguided, approach to addressing the reliability of proportions is only to indicate the point-by-point agreement of the numerator and denominator. Table 9.1 indicates that there are cases when agreement on either the numerator or denominator of a proportion can be *greater* than a common agreement criterion value (i.e., .8), while agreement on the *proportion* is *below* the same criterion value. This is true even when we use a "generous" index of agreement: the small/large ratio. A small/large agreement proportion is the result of dividing the smaller variable score estimate by the larger variable score estimate. The small/large proportion is not as stringent as point-by-point agreement because the behaviors that one observer counts to derive a variable score estimate do not have to be the same behaviors that the other observer counts to derive his or her variable score estimate.

A subtle but important point, concerning reliability of scores, occurs when investigators concatenate sessions (i.e., combine them) to derive a primary variable score (e.g., sessions are combined because of infrequent key behaviors). The reliability estimates should be derived from the same number of concatenated sessions as is used to derive the primary variable scores. Although this may seem obvious, it is important to discuss this point because, in chapter 8, we indicated that it is rather common for investigators to concatenate sessions within participant for point-by-point agreement checks on infrequent behaviors. While acceptable for training and retraining purposes, such an approach is inadvisable for interobserver agreement estimation of observational variables if primary data used to address research questions are derived on single sessions. In other words, agreement or reliability estimates on pooled sessions tell us nothing about agreement or reliability on the single-session level.

Another way that we sometimes see the issue of "pooled" or collapsed data played out is that agreement might be reported over all categories in the coding manual, even though we want to know whether a particular variable that is based on a single category has acceptable interobserver agreement. One may have good agreement on a noncollapsed agreement matrix without having good agreement on a particular category if instances of that category occur infrequently in the observation

Table 9.1

ILLUSTRATION THAT AGREEMENT ON NUMERATOR AND DENOMINATOR DOES NOT ENSURE AGREEMENT ON PROPORTION

NUMERATOR FOR OBSERVER 1	NUMERATOR FOR OBSERVER 2	DENOMINATOR FOR OBSERVER 1	DENOMINATOR FOR OBSERVER 2	PROPORTION FOR OBSERVER 1	PROPORTION FOR OBSERVER 2	SMALL/ LARGE AGREEMENT FOR NUMERATOR	SMALL/LARGE AGREEMENT FOR DENOMINATOR	SMALL/ LARGE AGREEMENT FOR PROPORTION METRIC
10	8	102	122	0.10	0.07	0.80	0.84	0.67
20	16	103	124	0.19	0.13	0.80	0.83	0.66
30	24	104	125	0.29	0.19	0.80	0.83	0.67
40	32	105	126	0.38	0.25	0.80	0.83	0.67
50	40	106	127	0.47	0.31	0.80	0.83	0.67
60	48	107	128	0.56	0.38	0.80	0.84	0.67
70	56	108	130	0.65	0.43	0.80	0.83	0.66
80	64	109	131	0.73	0.49	0.80	0.83	0.67
90	72	110	132	0.82	0.55	0.80	0.83	0.67
100	80	111	133	0.90	0.60	0.80	0.83	0.67

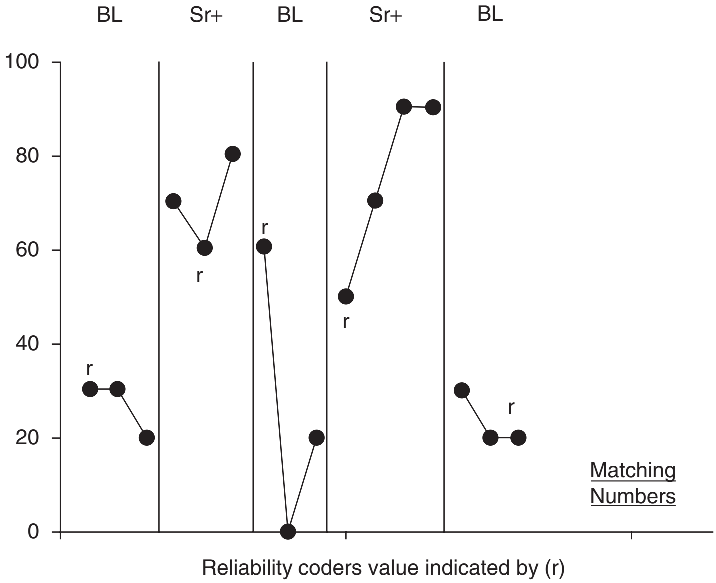


Figure 9.1 Illustrating proportion reliability scores on the same graph that displays primary proportion scores.

session. Therefore, when using interobserver agreement to inform the investigator or reader about interobserver reliability of an observational variable, we collapse our agreement matrices into 2 × 2 tables in which the upper left cell represents agreement on the category of interest (Kraemer, Periyakoil, & Noda, 2004).

In addition to reporting point-by-point agreement, we recommend that the secondary observer's data be graphed on the same graph as the primary observer's data in single-subject studies (Cooper, Heron, & Heward, 2007). The data graphed should be the metric and unit of analysis used to address the research question (see Figure 9.1). This practice allows the reader to judge the following: (a) whether the secondary data "tells the same story" as the primary data regarding presence or absence of a functional relation, (b) when, relative to design phases, the agreement check session occurred, (c) whether there is evidence that the secondary and primary observers differ systematically (e.g., whether secondary observer's data is consistently higher than the primary observer's data).

This simple practice, although uncommon, would allow more transparency than does reporting mean and range percentage agreement and could reduce a number of misguided practices (e.g., reporting agreement on different metrics or units of analysis than primary data, separate reporting on unitizing versus classifying agreement percentages, asymmetrical counting of secondary observer's disagreements over primary observer's disagreement). Using this method of displaying reliability data makes it clear that large changes between phases can be detected even with lower interobserver agreement, while smaller changes between phases need higher interobserver agreement to show the same data pattern (Cooper et al., 2007).

EXHAUSTIVE CODING SPACES REVISITED

Table 9.2 illustrates how an agreement matrix for an *exhaustive* coding space can be collapsed into a 2 × 2 table for a particular category. In this table, we reprint the agreement matrix from chapter 8 for exhaustive coding spaces and collapse it to highlight occurrence and nonoccurrence agreement for Category "2." Collapsing agreement matrices to 2 × 2 tables cannot be done properly for nonexhaustive agreement matrices because the sums of the marginals for the two observers do not have to be equal for properly constructed, noncollapsed, nonexhaustive agreement matrices (see chapter 8 for how to construct these). For 2 × 2 agreement tables, however, the sums of the marginals *do* have to be equal. The 2 × 2 table makes it clear how many units are considered agreements and how many are considered disagreements (Table 9.2) about the presence and absence of Category "2." The cell (a–d) and marginal (f and g) labels are provided for use later in this chapter.

Table 9.3 illustrates why it is not appropriate to collapse nonexhaustive agreement matrices into a 2 × 2 table. Two of the "marginals" for this nonexhaustive agreement matrix indicate something quite different from the other marginals: the total number of times the observer did *not* record a behavior. While constructing nonexhaustive agreement matrices are very useful for training observers and checking observer drift (chapter 8), adding across these different types of marginals produces a sum that does not mean what most sums mean (i.e., it violates an assumption behind summing scores). Therefore, we do not advise doing so.

Recall that one way to determine whether we have an exhaustive coding space is to note that such coding allows agreement on nonoccurrence of *all* key behaviors. We determine this from the full agreement matrix, not the collapsed 2 × 2 table. The reason is that the D cell in a collapsed 2 × 2 table indicates agreement that the "x category" is absent, not whether there is agreement on nonoccurrence of all key behaviors.

Table 9.2

ILLUSTRATION OF COLLAPSING AN EXHAUSTIVE AGREEMENT MATRIX TO A 2 × 2 TABLE FOR CATEGORY “2”

		OBSERVER 2				
		1	2	3	NO BEHAVIOR CODED IN INTERVAL BY OBSERVER 2	MARGINAL FOR OBSERVER 1
	1				2	2
	2		1			1
	3		1		1	2
OBSERVER 1	No behavior coded in interval by Observer 1	1			4	5
MARGINAL FOR OBSERVER 2		1	2	0	7	10

		OBSERVER 2		MARGINALS FOR OBSERVER 1
		2 PRESENT	2 ABSENT	
OBSERVER 1	2 present	1 a	0 b	1 g1
	2 absent	c 1	d 8	9 g2
MARGINALS FOR OBSERVER 2		2 f1	8 f2	10 N

a = instances observers agree “2” was present; b = instances Observer 1 counted a behavior as “2” but Observer 2 did not; c = instances Observer 2 counted a behavior as “2” but Observer 1 did not; d = instances observers agree “2” was not present. Letters are provided to enhance communicating about the cell values.

When using a time-window definition of agreement and timed-event behavior sampling in which frequency is the metric to be derived (a nonexhaustive agreement matrix), agreement on nonoccurrence of *all* key behaviors is not defined. When agreement on nonoccurrence of *all* key

Table 9.3

ILL-ADVISED AGREEMENT MATRIX FOR A NONEXHAUSTIVE CODING SPACE

		OBSERVER 2				
		1	2	3	NO BEHAVIOR CODED BY OBSERVER 2	MARGINALS FOR OBSERVER 1 CODED BEHAVIOR AND INSTANCES OF UNITIZING DISAGREEMENTS
OBSERVER 1	1		(2–3)	(12–13)		2
	2		(5–6)			1
	3				(10–11)	1
	No behavior coded by Observer 1	(8–9) (15)			Assigned 0	2[a]
MARGINALS FOR OBSERVER 2'S CODED BEHAVIOR AND INSTANCES UNITIZING DISAGREEMENTS		2	2	1	1[a]	6

[a]These "marginals" are for behaviors the observer did *not* record, while the marginals for the other rows or columns are for behaviors the observer *did* record. This mixing of the logic of what the marginals measure violates the assumption behind estimating chance agreement.

behavior is not computable, nonchance agreement cannot be estimated accurately. We will return to this point later in this chapter.

THE EFFECT OF CHANCE ON AGREEMENT

How frequently or, conversely, how infrequently a target behavior occurs can affect how much of the agreement estimate is caused by chance processes, and not by accurate coding. For high-frequency behaviors, even random coding will produce inflated occurrence agreement estimates relative to that which occurs for the same behavior occurring less frequently (House, House, & Campbell, 1981). Conversely, when coding infrequently occurring behaviors, even random coding will produce inflated *non*occurrence agreement estimates relative to that which occurs for the same behavior occurring more frequently. We call the portion of the agreement that could occur even when observers code randomly "chance agreement." Chance agreement can occur for occurrence and nonoccurrence of instances of target behavior.

As was discussed in chapter 6, chance is only accurately estimable when we have an exhaustive coding space. One direct implication, then, is that chance agreement is only estimated accurately when our definition of point-by-point agreement is exact agreement derived from exhaustive coding spaces (see chapter 8 for elaboration). Chance, however, *still* influences agreement when agreement is estimated on nonexhaustive coding spaces and when the definition of point-by-point agreement is the time-window agreement definition. This is an important point that is easily overlooked or misunderstood.

In exhaustive coding spaces, chance agreement for the A (i.e., occurrence) or the D (i.e., nonoccurrence) cells in a 2 × 2 agreement table is the product of each observer's estimate of the base rate of the key behavior × the total number of coded units. When applied to agreement matrices, it is useful to compute base rate as a probability or proportion. This probability is the number of times a behavior type occurs/total number of coded units. The chance agreement estimates for the A (occurrence) and D (nonoccurrence) cells in Table 9.2 are 0.2 (0.1 × 0.2 × 10) and 7.2 (0.9 × 0.8 × 10), respectively.

Because point-by-point agreement is meant to be an estimate of the accuracy of coding, we need an index of interobserver agreement that is not influenced by chance (i.e., that is not attributable to random coding). We call this "nonchance agreement." It is sometimes recommended that researchers focus on nonoccurrence agreement (i.e., cell D) when they expect the base rate of the key behavior to be "high" and to focus on occurrence agreement (i.e., cell A) when they expect the base rate of the key behavior to be "low" (Ayres & Gast, 2009), but this suggestion does not solve the problem that chance agreement affects our indices of point-by-point agreement. Additionally, a problem arises when the base rate of the behavior is low for some periods or sessions in the study and high for other periods or sessions in the study. As discussed in the next section, different types of indices of point-by-point agreement have been derived to address this problem.

COMMON INDICES OF POINT-BY-POINT AGREEMENT

The following is not an exhaustive list of the many ways to compute point-by-point agreement. Several others are covered in other sources, but they do not solve the problems discussed in this chapter (Ary, Covalt, & Suen, 1990; Primavera, Allison, & Alfonso, 1997). We have chosen to focus on four methods because they are common and make a substantive point about the nature of chance agreement and indices of point-by-point agreement. These four methods are occurrence percentage agreement, nonoccurrence percentage agreement, total percentage agreement, and kappa.

Occurrence Percentage Agreement

Using the cell address for the 2 × 2 table in Table 9.2, the formula for occurrence agreement is $(a/[a + b + c]) \times 100$. This is the percentage of agreements divided by agreements plus disagreements on the occurrence of a particular category. A percentage metric is used to enhance interpretation. This is the most common type of point-by-point agreement statistic. One reason for its popularity is that it can be computed from nonexhaustive agreement matrices. Its biggest disadvantage is that it does not control for chance agreement. However, if the occurrence percentage agreement is lower than a criterion level agreement the investigator selects, it can still indicate that retraining is necessary. After all, chance agreement cannot be the reason why occurrence agreement is low!

Nonoccurrence Percentage Agreement

Nonoccurrence percentage agreement can only be correctly computed on exhaustive coding spaces. This occurs because agreement on the absence of a particular category necessarily includes the coding units for which both observers code no key behavior. This is a key point that is sometimes overlooked. Using the cell addresses for the 2 × 2 table in Table 9.2, the formula for nonoccurrence percentage agreement is $(d/[d + c + b]) \times 100$. Again, chance agreement is not controlled in the nonoccurrence percentage agreement method, but, again, low observed nonoccurrence percentage agreement is still indicative that additional training is necessary.

Total Percentage Agreement

To compute total percentage agreement for a particular category using the cell addresses for the 2 × 2 table in Table 9.2, the formula is $([a + d]/[a + b + c + d]) \times 100$. Another way to convey this formula is $a + d/N$, where N represents $a + b + c + d$. Because nonoccurrence of Category "2" is included in the formula for total percentage agreement and because nonoccurrence agreement can only be computed correctly on exhaustive coding spaces, total percentage agreement can only be computed correctly from exhaustive coding spaces. Although this approach may seem to address the problem of the base rate of the behavior changing from session to session during the study, it still does not control for chance agreement.

Kappa

Kappa was created to control for chance agreement (Cohen, 1960). Conceptually, kappa is the proportion of potentially available nonchance agreement (i.e., 1—total chance agreement) that is attained (i.e., observed total agreement—chance total agreement). As mentioned above, the formula for observed total agreement is $(a + d)/N$. Again, note that observed total agreement includes agreement on occurrence and nonoccurrence. The formula for estimated chance total agreement, $(f1 \times g1 + f2 \times g2)/N$ 2 is also based on chance occurrence and chance nonoccurrence agreement. Figure 9.2 illustrates the observed kappa for the example data in the Excel spreadsheet on the website associated with this text that has an observed total agreement of 0.8 (i.e., when transformed from 80%). By comparing the total chance agreement and kappa lines, it is apparent that observed kappa is lower when total chance agreement is higher. This is what we would expect for an index that "controls for chance agreement."

The formula for kappa includes f2, g2, and N. All three of these values are based in part on the D cell, the agreement on nonoccurrence of

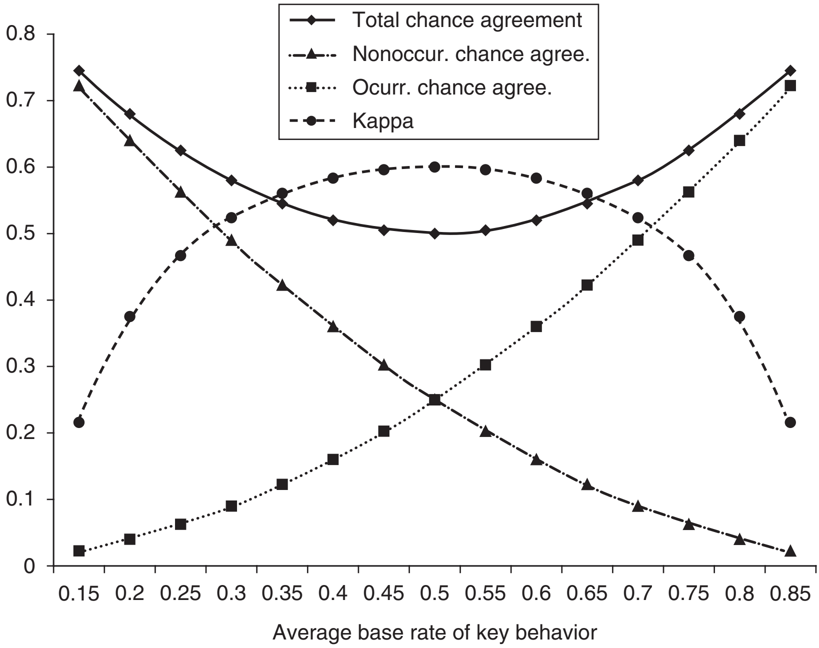


Figure 9.2 Illustration of the relation between chance agreement and kappa as a function of the base rate of the key behavior.

the key behavior. Again, the D cell value in the collapsed 2 × 2 table must come from an exhaustive coding space. For example, if nonexhaustive agreement matrices were collapsed to 2 × 2 tables, the "D cell" would be based on agreement on other behaviors and disagreements on other behaviors, but not on coding units in which there is agreement of nonoccurrence of *all* coded behavior. Therefore, kappa assumes an exhaustive coding space. This assumption can be met when exact agreement is used and the coding space is exhaustive (e.g., behavior sampling is time event, and duration is the metric of interest or when interval behavior sampling is used).

Unfortunately, kappa also assumes that one behavior does not influence the occurrence of another behavior (i.e., independence of analysis units; Cohen, 1960). This is an unrealistic assumption in observational agreement data because all the behaviors come from the same session and the tallies are from the same two observers. Despite these issues, kappa continues to be used frequently in the social sciences. One might justify this by pointing to the fact that there are other times that social scientists use mathematical functions that are based on unrealistic assumptions (e.g., some applications of inferential statistics in education or psychology). We do so because we need the information that the mathematical functions are meant to provide. However, there is another important issue that further reduces the utility of kappa.

Using an exhaustive 2 × 2 agreement matrix as the context, it has been shown many times by experts in many fields that the probable range of kappa varies greatly when the occurrence base rate of a target behavior is much less than or much greater than the target behavior's nonoccurrence base rate (Bakeman, McArthur, Quera, & Robinson, 1997; Bruckner & Yoder, 2006). Recall that accuracy refers to an observer's agreement with the true occurrence and nonoccurrence of the key behavior (i.e., something we cannot know with certainty in reality). Therefore, observer accuracy is a higher standard of what we want from our observers than is interobserver agreement. Even when accuracy of both observers is modeled to be .9 (a "good" accuracy level; Bakeman et al., 1997), the kappa is 0.39 when the occurrence base rate is 0.1 or 0.9, whereas, the kappa is 0.64 for the same accuracy level when the occurrence base rate is 0.5 (Bruckner & Yoder, 2006). To prevent readers incorrectly using Figure 9.2, it is important to note that Figure 9.2 is based on 0.80 interobserver's agreement, not 0.80 accuracy. Readers who wish to read more about this issue and to see figures of the obtained kappa values as a function of the base rate of the key behavior are referred to Bruckner and Yoder (2006).

Thus, the meaning of kappa (i.e., whether observed kappa is "good" or "poor") varies as a function of the base rate of the key behavior. This complicated state of affairs is similar to the problem that occurred with total percentage agreement. An appropriate criterion level of total percentage agreement needs to change as the base rate of the key behavior changes.

Base Rate and Chance Agreement Revisited

We turn to a discussion of how base rate of the key behavior affects chance agreement and an index that explicitly controls for chance agreement, kappa. Figure 9.2 illustrates the relation between occurrence, nonoccurrence, and total chance agreement as a function of the base rate of a key behavior when the observed total percentage agreement is 80%. Note that the relation of the chance *occurrence* agreement by base rate of the behavior is a mirror image of the relation of the chance *nonoccurrence* agreement by the base rate of the same behavior. The lines that represent these relations cross at the base rate of 0.5. When occurrence and nonoccurrence chance agreement are added (as they are in kappa), total chance agreement is at its lowest when the base rate is the point at which these lines cross. Readers who would like to understand (and check!) the math behind this figure are invited to do so by examining the formulae in the Excel spreadsheet on the website associated with this text entitled "chance agreement and kappa by base rate."

The implications of Figure 9.2 are disturbing if taken seriously. Note that many dependent variables of interest will at least sometime during the study or for some participants have a low (e.g., communication in children with communicative impairments) or high (e.g., challenging behavior in children with behavior disorders) base rate. In such cases, percentage agreement indices are *greatly* influenced by chance. For example, Figure 9.2 indicates that at the base rate of .15, there is 75% chance agreement. This means that 94% of the observed total percentage agreement of 80% could have been achieved purely by chance processes! Another lesson of this figure is that the observed or reported percentage agreement estimate is the upper limit of the potential nonchance agreement. That is, if the reported percentage agreement estimate is "low," then we can know that the nonchance agreement (what we really care about) is even lower.

Summary of Point-by-Point Agreement Indices

The fact that achieving a criterion level of point-by-point agreement requires more observer vigilance and coding manual adherence than does achieving the same level of small/large agreement ratios suggests that point-by-point agreement indices are likely to be more useful for training and retraining purposes than are small/large agreement ratios. At first glance, it may seem that readers, reviewers, or investigators can be confident that they know the nonchance point-by-point agreement of an observational variable when several conditions exist. First, as discussed in chapter 8, an adequate reliability sampling method must be used. This includes (a) a sufficient sample size, (b) a random or systematic sampling of the total data set, (c) independent agreement checks, and (d) the primary observer to be blind to when agreement checks occur. Second, the variable and unit of analysis used to estimate the index of point-by-point agreement should match that used to generate the variable scores which are then used to address the research question. Third, the method of estimating point-by-point agreement is more informative regarding accuracy if it controls for chance agreement. At first glance, it appeared that the latter criterion could be met by using kappa. However, we end up being stuck with an index of agreement (kappa) that is still difficult to fully interpret because its meaning is influenced by the base rate of the key behavior and because it is based on an assumption that is almost never true in observational studies (i.e., independence). One option is for investigators to use the estimated base rate of the key behavior to select the criterion level of kappa that corresponds with a set level of estimated accuracy (Bakeman et al., 1997). For those wishing to do so and who have the requisite set of conditions allowing reasonable kappa computation, the reader is referred to Bruckner and Yoder (2006) for guidance. It should be noted, however, there may be opposition philosophically to "correction for chance" because, from some perspectives (a) the concept of "chance" has little scientific utility due to multiple definitions (Baer, 1977; Johnston & Pennypacker, 1993) or (b) observational data rarely meet all requirements for precise estimation of chance. For those who prefer not to correct for chance agreement, percentage agreement will be the index of choice. It is true that when percentage agreement is high, it may be due to chance agreement. However, when percentage agreement is low, it cannot be due to chance agreement. In the final analysis, percentage agreement can be used to indicate when data are not sufficiently accurate, but it provides less certainty in judging when data are sufficiently accurate.

At present, many direct observation studies in which single-subject designs are used tend to rely exclusively on point-by-point agreement indices to judge the "trustworthiness" of data. This is problematic because the required level of point-by-point agreement that is necessary to trust a demonstrated functional relation (a) varies as a function of the size of effect (i.e., the magnitude of the change in the dependent variable among design phrases) and (b) whether observers are blind to design phase. The latter point will be addressed later in this chapter.

Fortunately, the graphing of secondary and primary data on the same graph has a number of advantages over only reporting mean and range of point-by-point agreement indices for a particular variable. At present, graphing of secondary and primary data is not common practice. This may (or may not) change in the future.

Regardless, there is, and probably will continue to be, a need to use an agreement index to guide decision making regarding observer (re)training. There is an additional need to set a criterion level of agreement below which retraining or continued training will occur. Because we cannot know the magnitude of the changes in the dependent variable between phases or effect sizes until all data are collected, knowing that the tolerable level of agreement varies by the effect size is not helpful in setting a criterion level of agreement. For this and other reasons given above, it is extremely difficult to set a criterion level of agreement that is informed and mindful. This difficulty does not change our need for a criterion level of agreement. This continued need for a criterion level of agreement is probably why at least one professional consensus group has suggested 0.8 for total percentage agreement and 0.6 for kappa as criterion levels (Horner et al., 2005), despite multiple warnings against their use in absolute terms. Using 0.8 agreement or 0.6 kappa is no better, or worse, than any other arguably high level of agreement for making (re)training decisions.

An argument can even be made for using a small/large agreement ratio as a method of deciding *when* retraining needs to occur. Point-by-point agreement is then represented in the form of agreement matrices and discrepancy discussions are used for the actual retraining. As we will discuss in detail later in this chapter, these recommendations are particularly defensible when the observers are kept blind to design elements such as when the treatment phase begins or which participants belong to treatment versus control groups.

At present, it appears there is no satisfactory solution to the dilemma of (a) needing a criterion level of agreement to make retraining decisions on one hand, but (b) not having a universally reasonable criterion level of agreement available on the other hand. This state of affairs should tell us (a) to graph primary and secondary observer's data on the same graph in single-subject studies to determine if they follow the same data path, (b) to avoid rigidity in our use of any criterion level of agreement, and (c) to seriously consider that the primary value of point-by-point agreement checks are more likely to be found in the discrepancy discussion process than in the "interobserver agreement" index. Ironically, in our opinion, the majority of attention has been given to the latter.

INTRACLASS CORRELATION COEFFICIENT AS AN INDEX OF INTEROBSERVER RELIABILITY FROM THE VAGANOTIC CONCEPT OF MEASUREMENT

In the vaganotic concept of measurement, interobserver reliability means that the ranking of participants on a particular variable is very similar regardless of the observers used to code the observation sessions

(Shavelson & Webb, 1991). Many observational measurement experts recognize the value of intraclass correlation coefficients (ICC) as an index of interobserver reliability for dependent variables and predictors in group designs (Bakeman & Gottman, 1997; Mitchell, 1979; Primavera, Allison, & Alfonso, 1997; Suen & Ary, 1989).

The ICC is identical to the *g* coefficient in a single-facet generalizability study with only two observers (see chapter 2 of this book; Shavelson & Webb, 1991). Like the *g* coefficient, the conceptual meaning of the ICC when measuring interobserver reliability is the proportion of the variability in the reliability sample that is due to between-participant variance in true score estimates of the behavior of interest (Shavelson & Webb, 1991). As a group-design statistic, ICC cannot be run on a single agreement check session. It takes at least five agreement check sessions that are coded by at least two observers to derive a reasonable ICC statistic. As will be seen, the larger the number of sessions on which reliability is estimated, the more confident we can be that the reliability estimate represents what occurs in the total data set.

Options for Running ICC with SPSS

SPSS and other statistical software programs that use ICC to estimate interobserver reliability have various options for running the software. The command statements provided in Table 9.4 indicate how to run ICC using SPSS to estimate interobserver reliability. We have selected (a) a mixed model in which the participant factor is treated as random and the observer factor is treated as fixed, (b) the ICC as the reliability coefficient, and (c) "absolute" as the agreement measure. These are in accordance with the recommendations of McGraw and Wong (1996) and Nichols (1998). Briefly, these guidelines indicate that it is appropriate to treat observers as a fixed factor because observers are not randomly selected from a population and that doing so means one should restrict one's generalization about the reliability of scores to those particular observers (Nichols, 1998). The absolute agreement measure is selected to detect whether one observer consistently scores more than the other across participants.

Between-Participant Variance on the Variable of Interest Affects ICC

An important attribute of ICC is that it reflects the variance among participants, unlike agreement indices. The latter are based on single sessions; therefore, they cannot reflect the among-participant variability

Table 9.4

SPSS SYNTAX AND RESULTS FOR THE PRACTICE EXERCISE THAT ILLUSTRATES THE EFFECT OF VARIABILITY ON ICC EVEN WHEN MEAN DIFFERENCE AND SMALL/LARGE AGREEMENT IS RELATIVELY STABLE

SPSS syntax for exercise:

```
RELIABILITY
/VARIABLES=DVwsmallvariance1 DVwsmallvariance2
/SCALE('small variance DV') ALL/MODEL=ALPHA
/ICC=MODEL(MIXED) TYPE(ABSOLUTE) CIN=95 TESTVAL=0.
RELIABILITY
/VARIABLES=DVwlargevariance1 DVwlargevariance2
/SCALE('large variance DV') ALL/MODEL=ALPHA
/ICC=MODEL(MIXED) TYPE(ABSOLUTE) CIN=95 TESTVAL=0.
```

SUMMARY OF RESULTS

	OBSERVER 1	OBSERVER 2	MEAN SMALL/LARGE AGREEMENT	ICC
	MEAN (SD)	MEAN (SD)		
DV WITH SMALL VARIANCE	4.3 (0.95)	3.5 (1.3)	0.69	0.37
DV WITH LARGER VARIANCE	4.4 (2.4)	3.5 (2.2)	0.70	0.84

on the variable of interest. On the website that is associated with this textbook, there is an Excel spreadsheet entitled "Data for exercise demonstrating the variance effect on ICC in Excel." The reader is invited to import these data into SPSS or other statistical software and run an ICC on the two pairs of dependent variables. The scores for each member within the pair are estimates from a different observer. The SPSS commands for ICC and a summary of the results for this example are provided in Table 9.4. The primary difference between the two variables is that one has about twice as much variance between participants as the other. Note that the two variables are very similar on (a) the means within observer, (b) the mean difference between observers, and (c) the mean small/large agreement proportion. Despite these similarities, the ICC for the variable with small variance is less than half that of the variable with larger variance.

Using ICC as a Measure of Interobserver Reliability for Predictors and Dependent Variables in Group Designs

The inclusion of variance between participants in the formula for ICC is important because it is much more difficult to show a relation or mean difference on a variable when the true score variance among people on the variable is very limited. In other words, the smaller the variance among participants in their true scores, the higher the interobserver agreement must be to detect the true differences among participants. Conversely, even when interobserver agreement between observers is low, we can still detect differences among participants as long as the variance among participants in their true scores is large enough.

This is just another way to say that we can detect the signal even when there is a lot of noise if the signal is clear enough. The signal in group designs is the association or difference indicated in our research question. The clarity of that signal is the effect size. The effect size of the expected association or difference is greatly influenced by the variance in the variable of interest. This is easiest to explain for associations. Part of the definition of an association is the degree to which we can predict the ranking of participants on a variable given knowledge of their ranking on another variable. To understand the concept that the variance of the variable should influence the reliability estimate, imagine a variable with a normal distribution and another variable with a distribution of scores with very little variance (i.e., a sharp peak with many scores clustered closely around the mean). All things being equal, we have lower confidence in the measured ranking of the participants on the variable with the sharply peaked distribution than on the variable with the normal distribution because there is much less difference between participants in the former variable.

The Interpretation of SPSS Output for ICC

In the SPSS output for the ICC, the "single measure" ICC is the interobserver reliability estimate for a single observer for the relevant variable. In SPSS output, the "average measure" ICC is that for the average of the observers' estimates for the variable. It is only appropriate to use the latter when the investigator has all sessions coded by more than one observer and uses the average score across observers as the variable score to answer the research question. The 95% confidence interval for the ICC point estimate means that the actual ICC in the total data set is somewhere between the lower and upper bound of the given interval.

One can tighten the confidence interval around the estimated ICC by increasing the size of the reliability sample. The probability value (i.e., under "sig" in SPSS output) of the ICC is associated with a significance test whether the confidence interval around the ICC includes zero. This probability value is only minimally relevant. The more important information is the absolute magnitude of the point estimate for ICC. The minimally acceptable ICC is relative to the area of study and the effect size for the expected association or group difference. As a benchmark, some consider an ICC of 0.7 as "very good" (Mitchell, 1979).

The SPSS output for the exercise (see Table 9.4) indicates the single measure ICC to be 0.84, which means that 84% of the variance in the variable in the reliability sample is due to between-participant variance. On the expanded SPSS output (not shown), the 95% confidence interval for this ICC point estimate is 0.48–0.96. This means that the ICC in the total data set is somewhere between 0.48 (some would interpret this as "poor") to 0.96 (outstanding).

THE CONCEPTUAL RELATION BETWEEN INTEROBSERVER AGREEMENT AND ICC

The "noise" in variable scores is measurement error. In group designs, measurement error due to observers is the extent to which observers disagree on the mean and ranking of participants on the variable of interest. Disagreement on the "proper" mean and ranking of participants on a variable is influenced by agreement between observers within a participant. More agreement will result in better ICCs, all things being equal. The group design concept of reliability requires only summary level agreement (e.g., small/large proportion agreement). However, we improve our summary level agreement by maximizing our point-by-point agreement and by reducing our observer drift. Additionally, if observers are not blind to group membership status, then the discrepancy discussions that point-by-point agreement checks stimulate are particularly important because they may reduce the probability of Type I error by maximizing accuracy.

CONSEQUENCES OF LOW OR UNKNOWN INTEROBSERVER RELIABILITY

The two types of scientific error are Type I and Type II. Type I error is detecting a difference or relation, when one does not really exist. Type

II error is failing to detect a difference or relation that is present. When interobserver agreement or reliability of an observational variable is unknown because one or more of the principles of conducting sound reliability checks has not been followed, it is most conservative to assume interobserver reliability or agreement on the observational variable to be "low" (meaning below the investigator's or reader's standards of minimally desirable). To evaluate the consequences of "low" interobserver agreement or reliability, we must know whether the observers are blind to when the design phase changes (e.g., from baseline to treatment phase) or to membership of participants to groups (e.g., control vs. experimental groups) or to participants' scores on other variables listed in the research question.

"Blindness" to these design elements is important information because not being blind is the primary source of "correlated measurement error" in observational variables. In observational research, this occurs when the observer systematically overestimates the true score in the predicted superior group or phase while systematically underestimating the true score for those in the predicted inferior group or phase. For example, assume the observer knows that Participants A, B, and C are from the experimental group, Participants D, E, and F are from the control group, and the observer systematically overestimates the true score for those in the experimental group and systematically underestimate the true score for those in the control group. This would create a group difference even if there were not one in reality. An analogous process can occur in single-subject designs or group correlational designs. No one is accusing the observers in such a situation of being dishonest. Such bias can and does occur even when observers are competent and well intentioned (Reid, 1970). The consequence of correlation measurement error is an elevated probability of Type I error.

One way to test for whether the primary observer's data has correlated measurement error is to examine whether one observer's data is consistently higher than the other observer's data in the group or design phases. This can be seen in by the graphing of secondary and primary data as suggested earlier. A similar process can be used for group data. Unless both secondary and primary observers are biased in the same way (an unlikely event), such a pattern is consistent with the finding that correlated measurement error occurs. If this type of pattern does not occur, then the "reliability data" is not consistent with the hypothesis of correlated measurement error and can thus probably not explain any apparent functional relations in single-subject experiments or significant differences or relations in group designs.

If observers are blind to these design elements or if examination of consistently higher scores by one observer over another indicates no systematic bias, then the extent to which observers overestimate and underestimate true score is likely to be randomly distributed across design groups or phases (Thompson & Vacha-Haase, 2000). We call this type of measurement error "uncorrelated measurement error." The consequence of "high uncorrelated measurement error" is an elevated probability of Type II error (Thompson & Vacha-Haase, 2000).

When coders are not blind to treatment phase or group, high point-by-point agreement does not *ensure* that measurement error is uncorrelated with design phase but it does increase the probability that it is. Additionally, as indicated above, it is quite difficult to show with confidence that *non-chance* point-by-point agreement is high. If point-by-point agreement is low, however this is defined, and observers are unblind to design phase or group, then the probability that measurement error is correlated with design phase is higher than if point-by-point agreement is high.

High ICCs are also insufficient to ensure that correlated measurement error has not occurred when observers are not blind to design features. This is because ICCs do not tell us about point-by-point agreement, much less accuracy. It is the more detailed level of agreement and subsequent discrepancy discussions that are relevant for reducing the probability of correlated measurement error when observers are not blind. Therefore, group designs should (a) include point-by-point agreement checks with their accompanying discrepancy discussions (chapter 8) to maximize the accuracy of observers, (b) use ICC to estimate reliability of summary level observational variables in a way that reflects variability among participants, and, whenever possible, (c) keep observers blind to design features. If observers cannot be blind, then secondary and primary observers' scores should be reported or tested for systematic bias to enable readers to judge the probability of correlated measurement error.

RECOMMENDATIONS

Ironically, most of the emphasis related to reliability in observational measurement has been on point-by-point agreement indices. This is unfortunate because such indices are largely uninterpretable with regard to the real topic of interest: accuracy. In contrast, almost no discussion has been provided about a very useful process: discrepancy discussions (i.e., the topic of chapter 8). Discrepancy discussions have been largely ignored in large part due to the lack of computer software programs enabling them. The software features indicated in chapter 8 will be available in new observational software (e.g., the 2010 version of MOOSES). However, even with software-enabled discrepancy discussions, investigators need a criterion agreement level to guide (re)training decisions. Unfortunately, there are no universally accepted criterion agreement levels. Investigators must decide on an agreement index and criterion level value based on their design and specific area of study.

In both designs (group and single-subject experiments), observers should be kept blind to design features when possible, and discrepancy discussions from regular point-by-point agreement checks should be used to maximize the accuracy of coding decisions. When using the idemnotic measurement perspective (e.g., single-subject experiments and FOT measures), we recommend graphing primary and secondary data on the same graph to allow readers to judge whether both observers' data show the same general pattern (e.g., support an inference of a functional relation). When using the vaganotic measurement perspective (e.g., when measuring dependent and predictor variables in group design studies), we recommend reporting ICCs on each observational variable indicated in the research questions. If observers are blind to design elements, low point-by-point agreement and low ICCs should lead us to expect heightened probability of Type II errors. If observers are *not* blind to design elements, we should test whether the relation of primary to secondary observers' data shows evidence of correlated measurement error. If so, then we should be cautious about accepting the causal or even correlational nature of the study findings. Additionally, low point-by-point agreement and unblinded observers increase the probability that measurement error could be correlated with design phase or group and thus should lead us to be cautious about accepting causal or relational claims.

REFERENCES

Ary, D., Covalt, W., & Suen, H. K. (1990). Graphic comparisons of interobserver agreement indices. *Journal of Psychopathology and Behavioral Assessment, 12*, 151–156.

Ayres, K., & Gast, D. L. (2009). Dependent measures and measurement procedures. In D. Gast (Ed.), *Single-subject research methodology in behavioral sciences* (pp. 129–165). New York: Routledge.

Baer, D. M. (1977). Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis, 10*, 117–119.

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*(4), 357–370.

Bruckner, C. T., & Yoder, P. J. (2006). Interpreting kappa in observational research: Baserate matters. *American Journal of Mental Retardation, 111*, 433–441.

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Columbus, OH: Pearson.

Haynes, S. N., & O'Brien, W. H. (1999). *Principles and practice of behavioral assessment*. New York: Kluwer.

Horner, R. H., Carr, E. G., Halle, J., McGee, G. G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Behavioral Assessment, 3*, 37–57.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kazdin, A. (1981). Drawing valid inferences from case studies. *Journal of Consulting and Clinical Psychology, 49,* 183–192.

Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology, 42*, 774–780.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2004). Kappa coefficients in medical research. *Psychometrika, 44*, 461–472.

Mitchell, S. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86*(2), 376–390.

Nichols, D. P. (1998). Choosing an intraclass correlation coefficient. *Technical support paper from the SPSS tech support web site.* Retrieved 10/2008 from http:// www.spss.com/tech/stat/articles/whichicc.txt.

O'Leary, K. D., & Kent, R. N. (1977). Sources of bias in observational recording. In B. C. Etzel, J. M. LeBlanc, & D. M. Baer (Eds.), *New developments in behavioral research, theory, method, and application* (pp. 231–236). Hillsdale, NJ: Erlbaum.

Primavera, L., Allison, D. B., & Alfonso, V. C. (1997). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–90). Mahwah, NJ: Erlbaum.

Redfield, J., & Paul, G. L. (1976). Bias in behavioral observation as a function of observer familiarity with subjects and typicality of behavior. *Journal of Consulting and Clinical Psychology, 44*, 156.

Reid, J. B. (1970). Reliability assessment of observation data: A possible methodological problem. *Child Development, 41*, 1143–1150.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Erlbaum .

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174–195.

10 Validation of Observational Variables

OVERVIEW

Generally speaking, "validity" means scientifically useful. From one perspective, it has been said that when it comes to behavioral observation, one only needs to be concerned with accuracy (Johnston & Pennypacker, 1993). Alternatively, most researchers who use observational variables acknowledge that because of disagreements regarding what constitutes a "legitimate" instance of the behavior class of interest, validation is also important even when we are measuring context-dependent behavior from an idemnotic perspective (Haynes & O'Brien, 1999; Primavera, Allison, & Alfonso, 1997). In this, our last chapter, we (a) suggest that the relevant validation evidence varies by research design, object of measurement, and purpose of the research, (b) indicate the primary purposes of observational variables that organize the common types of validation evidence, and (c) summarize the five types of validation evidence that are especially relevant for observational variables.

THE CHANGING CONCEPT OF VALIDATION

Since 1954, the year of the first published standards of psychological and educational measurement, the concept of validity has been formally revised four times (Goodwin & Leech, 2003). Each revision has made it increasingly clear that the concept of "validity" does not refer to a measure or test but to a particular variable, use, and population. An observational variable score is the end product of all of the decisions and procedures made to produce the score (i.e., measurement context, number of sessions averaged, coding manual, behavior sampling method, coding decision recording method, observational session recording method, and metric). The principle that the object of validation is specific to a variable, use, and population means that providing support for one purpose or one population does not necessarily provide support for using the same observational variable for other purposes or with other populations. Finally, validation support for one aspect of the measurement system (e.g., a variable derived from one behavior sampling method or quantified with one metric) does not provide evidence of the scientific value for the variable using a different set of decisions about the measurement system (e.g., a variable derived from a different behavior sampling method or quantified by a different metric). In summary, validation evidence is conditional and highly contextualized (Haynes & O'Brien, 1999).

There are different types of validation processes. These differing processes used to be called "types of validity," but that terminology has been revised to reflect the modern concept of validation as a (a) purpose-specific, (b) ongoing, and (c) cumulative process. In this chapter, we cover five types of validation evidence that are relevant to the most common uses of observational variables. These five types of validation evidence are (a) content, (b) sensitivity to change, (c) treatment utility, (d) criterion related, and (e) construct. Although some taxonomists have combined sensitivity to change and treatment utility under one category (clinical utility; Haynes & O'Brien, 1999), we discuss them separately because sensitivity to change has relevance for all objects of measurement, whereas treatment utility has relevance only for generalized characteristics. Before describing each type of validation method, we present an organizational framework that is designed to help readers recognize that some aspects of validation are emphasized more than others for different research designs, objects of measurement, and purposes.

UNDERSTANDING WHICH TYPES OF VALIDATION EVIDENCE ARE MOST RELEVANT FOR DIFFERENT RESEARCH DESIGNS, OBJECTS OF MEASUREMENT, AND RESEARCH PURPOSES

Table 10.1 illustrates a crossing of two factors: type of research design (single subject vs. group) and object of measurement (context-dependent behavior vs. generalized characteristic). The cells of the resulting 2 × 2 matrix indicate the types of validation evidence that are most relevant to the object of measurement when studied through a particular type of research design. Table 10.2 illustrates the types of validation evidence that are most relevant to four of the purposes of observational research. We

Table 10.1

ILLUSTRATION OF TYPES OF VALIDATION BY RESEARCH DESIGN AND OBJECT OF MEASUREMENT (AND VARIABLE TYPE)

		OBJECT OF MEASUREMENT	
		CONTEXT-DEPENDENT BEHAVIORS (VARIABLE TYPE)	GENERALIZED CHARACTERISTIC (VARIABLE TYPE)
Research design	Single subject	Content (all) Sensitivity to change (all)	Content (dependent) Sensitivity to change (dependent) Treatment utility (participant characteristic or treatment context variable) Criterion related (dependent) Construct (dependent)
	Group	Content (FOT)	Content (predictor and dependent) Sensitivity to change (dependent) Treatment utility (participant characteristic or treatment context variable) Criterion related (predictor and dependent) Construct (predictor and dependent)

FOT: fidelity of treatment.

Table 10.2

ILLUSTRATION OF FIVE TYPES OF VALIDATION EVIDENCE BY PURPOSES OF OBSERVATIONAL RESEARCH

	TYPES OF VALIDATION EVIDENCE				
	CONTENT	SENSITIVITY TO CHANGE	TREATMENT UTILITY	CRITERION RELATED	CONSTRUCT
PURPOSES					
Describing	x				
Understanding Variability	x			x	x
Demonstrating that Treatment Affects Change in Dependent Variable (Treatment Effect)	x	x		x	x
Predicting Differential Treatment Response	x	x	x	x	x

begin our discussion with one type of validation evidence that is important to all objects of measurement and all research designs: content validation.

CONTENT VALIDATION

Definition of Content Validation

As applied to a coding manual, *content validation* is the expert rating of the relevance and representativeness of the examples and instances identified by the definitions in the coding manual to the stated object of measurement. The coding manual is part of the measurement system. Content validation can also include expert judgment regarding the adequacy of the measurement context, the number of sessions averaged, behavior sampling method, the coding decision recording method, the observation session recording method, and the metric with regard to measuring what the investigator says she wants to measure (Messick, 1989; Primavera et al., 1997).

Different Traditions Vary on the Levels of Importance Placed on Content Validation

Interestingly, content validation is the only type of validation that does not involve empirical examination of the participants' variable scores. Instead, it focuses exclusively on the measurement system (Geisinger, 1992). Because of this, measurement experts in the vaganotic and thus group design traditions have questioned whether content validation is really a validation process at all (Cronbach, 1988; Guion, 1977). Even when combined with accuracy information, these critics do not consider content validation sufficient to support a measurement system for any use to which vaganotic measurement approaches are applied (Messick, 1989).

In contrast, some single-subject research measurement experts consider content validation the only necessary companion of accuracy to support the scientific value of observational measurement for some purposes. Quoting one of the texts on behavioral measurement for single-subject research,

> Content validity…for a particular assessment purpose is one of the most important psychometric evaluative dimensions in behavioral assessment. (Haynes and O'Brien, 1999, p. 201)

To understand this perspective, it is important to recall that much single-subject research is focused on describing what occurs in certain contexts (stimulus–response/response–stimulus relations) or after certain behaviors (response–response relations). There may be no need to generalize past the measurement context. Therefore, accurately coding and classifying a behavior by the correct name is paramount. Using an example from Primavera et al. (1997) to illustrate why content validation is so important, we might ask, "Are all movements of the head toward a solid surface 'head banging'? Or is cranial contact necessary?" (p. 50). One might add, "Is cranial contact that is sufficiently hard to produce an audible sound necessary for the instance to be 'head banging'?"

Weaknesses of Content Validation

One of the main weaknesses of content validation is that the definition of who is considered an expert is somewhat subjective and varies by content area (Johnston & Pennypacker, 1993). For example, if we are studying "challenging behaviors" because we want to reduce these to address the concern of a consumer of clinical services, we might consider the experts to be parents or teachers. If we are studying "hyperactivity" because we want to understand what behaviors covary because of a presumed common genetic or neurological cause, we might consider the experts to be particular professionals who are highly knowledgeable about a substantive area. Even within a particular group of experts, there is likely to be some disagreement. Such disagreement is often addressed through consensus or majority vote. The process by which a consensus or vote is called often involves discussion of differences. Experts differ in their ability to communicate effectively, however. Therefore, the outcome of such discussions may reflect dominant experts' communication skills more than, or in addition to, content expertise and knowledge. Finally, expert knowledge changes over time. What seems "true" today is not necessarily considered "true" tomorrow.

SENSITIVITY TO CHANGE

Definition of Sensitivity to Change

Sensitivity to change is the degree to which a measure changes in a therapeutic direction after participation in treatment (Vermeersch, Lambert, & Burlingame, 2000). As this definition implies, sensitivity to change is not just about measuring change in any direction or measuring change when no formal treatment occurs. The concept is explicitly linked to a formal treatment, implying that a portion of the change is thought to be due to a treatment. Therefore, the aspects of the research design (e.g., differences between control and experimental group or differences between design phases or conditions) that are controls over nontreatment influences on change are an essential aspect of testing sensitivity to change. One of the reasons one might consider observational variables worth the effort is that observational variables are often more sensitive to change than more global measures (Haynes & O'Brien, 1999). Sensitivity to change has been called a type of construct validity (Vermeersch et al., 2000), but we do not classify it as such because the general concept is relevant for context-dependent behaviors (which are not constructs) as well as generalized characteristics.

Influences on Sensitivity to Change

As mentioned in chapters 1 and 2 of this book, measures of generalized characteristics will be more difficult to change after a brief treatment than will be context-dependent behaviors. This is due, in part, to the fact that measures of generalized characteristics are relatively stable over contexts, by definition. Because the internal validity of single-subject experiments that use comparison between a baseline and a treatment phase (AB variants) benefits from a rapid shift in the dependent variable immediately after the onset of the treatment phase (Kazdin, 1981), group designs, which do not require immediate changes in the dependent variable to show treatment effects, are often used to address treatment effects on generalized characteristics. This covariation of research design with object of measurement influences how measurement systems are designed, which in turn influences the sensitivity to change for many observational variables.

When different types of behaviors are lumped into a single category, they may be less likely to show sensitivity to change than when the different types of behavior are measured separately. The lower sensitivity to change in lumped categories occurs, in part, because participants in the control group may also change on some of the behavior types lumped into the same category with the types of behavior the treatment affects. However, categories that lump behavior types are used when changes due to a treatment are measured in a group because the behaviors that change in some members of the group are often different from the behaviors that change in other members of the group. That is, the category may change as a whole by incrementing any of a number of behaviors that are affected in different individuals. Thus, there is a balance between making the category in the coding manual sensitive to change and yet inclusive of the different ways that members of a group are likely to show change on different manifestations of the category.

When considering sensitivity to change, it is also important to select measurement systems that avoid floor (i.e., most sessions' or participants' scores are near the empirical minimum) or ceiling (i.e., most sessions' or participants' scores are near the empirical maximum of the variable) effects. Floor or ceiling effects can mask changes that would otherwise have been detectable. When we are measuring generalized characteristics as the dependent variable for treatment efficacy studies, the measurement context must remain constant in all design phases or groups. Because of this requirement, we have a need to show variance in the observational dependent variable's scores in all design groups or phases at all periods in a study.

We can inadvertently create a floor effect if (a) the measurement context is not evocative of the key behaviors, (b) the observation sessions are too short or there are not a sufficient number of sessions concatenated within participants, (c) the coding manual is too restrictive about the allowable instances of behavior into the behavioral category, (d) the participant sampling method does not allow the observer to note the behaviors of all key participants, (e) the behavior sampling method does not detect legitimate instances of the key behavior, and (f) the coding decision or observation session recording methods leaves observers doubtful about what they have observed. The latter can lead observers not to score legitimate instances of the key behavior due to the conservative ethic in science. We can inadvertently create a ceiling effect by using a measurement activity that is "too easy" for many participants or in many sessions.

Weakness of Sensitivity to Change

The more specific the definition, the more sensitive the observational variable will be to contextual differences and thus environmental manipulation. This is fine for context-dependent behavior. When wanting to measure generalized characteristics that change in a therapeutic direction while being stable over relevant contexts, there is a tension between being too specific to be socially important and being specific enough to achieve adequate interobserver agreement.

TREATMENT UTILITY

Definition of Treatment Utility

Another clinically related validation process is *treatment utility*. Treatment utility has been defined as the degree to which assessment is shown to contribute to beneficial treatment outcome (Nelson-Gray, 2003). We can break down the methods of examining treatment utility into two categories defined by the research questions they address.

The first question is "Does a treatment produce better outcomes when it is based on a particular assessment outcome (i.e., the variable of interest) than when it is based on a different assessment outcome?" One example of such a study comes from the functional assessment literature (Carr & Durand, 1985). In this study, functional assessment information (the variable of interest) was used to identify two small groups of children: (a) those whose challenging behavior was identified as supported by "escaping a difficult task" (i.e., negatively reinforced challenging behavior) and (b) those whose challenging behavior was identified as supported by "seeking positive attention" (i.e., positively reinforced challenging behavior).

The two treatments involved teaching the children to say either of the following phrases: (a) "Help me" or (b) "Am I doing good work?" depending on the design phase and identified function of the challenging behavior for the child. The former phrase was hypothesized to match the escape function, and the latter phrase was hypothesized to match the attention-seeking function. Design phases in which the hypothesized function of challenging behavior matched the taught replacement phrase resulted in the most reduction in challenging behavior. This is an example of seeking to examine whether the functional assessment information had treatment utility validation evidence by testing whether the function–treatment match condition resulted in better outcomes than the function–treatment mismatch conditions in both types of children. This single-subject design approach to testing treatment utility can be effective when the participant characteristic-treatment match produces almost perfectly predictable results, as was the case in the Carr and Durand study.

A similar research question posed by treatment utility validation is "Does pretreatment level on a particular participant characteristic predict which of two treatments is most effective in facilitating posttreatment scores on the dependent variable?" The participant characteristic variable is the variable whose treatment utility is being tested. This type of question is most frequently tested in a group experimental design using random assignment to groups with a pretreatment measure of the participant characteristic and is most appropriate when the participant characteristic-treatment match relation is less than perfect. These participant characteristics can be measured using observational methods.

An example of such a study asked whether pretreatment level of joint attention predict which of two communication treatments are most effective in facilitating post-treatment joint attention (Yoder & Stone, 2006). The results of this randomized experiment found that children with autism learned to use joint attention more frequently through a particular treatment better than through an alternative treatment if they already used about seven joint attention acts during the pooled pretreatment communication samples designed to elicit joint attention. These results provided treatment utility support for the pretreatment measure of joint attention because the results corresponded to the predicted pattern of results.

Another type of research question that is sometimes posed by the treatment utility validation process is "Does therapy produce better results when the therapist is supplied with feedback and client progress data than when the therapist is not provided such feedback and client progress data?" The feedback and client progress data are the variables whose treatment utility is being tested. This is a subtle variant on the above example in the sense that an explicit match versus mismatch between treatment types and participant characteristics is not being tested. Therapist feedback and client progress data could be, and often is, measured through observational means. An example of a study examining this type of question is a group design examining the amount of change in depressive symptoms following therapy sessions designed to treat depression (Lambert, Hansen, & Finch, 2001). Two randomly assigned groups varied in terms of whether the therapists received client progress data and therapist feedback or not. In clients who showed initially no change, the group whose therapist received feedback and client progress data made more therapeutic change than the group whose therapist received no feedback or client progress data.

Weaknesses of Treatment Utility

This type of validation information is only relevant for a small proportion of the observational research conducted at present. Treatment utility validation is highlighted here because we wish to stimulate more such applications of observational research in the future. At present, many treatments used to address observational variables do not have a sufficiently rich or specific theory to make prior predictions regarding which participant characteristics should predict differential treatment efficacy for particular treatments. Another issue is that it is likely that participant characteristics relevant for predicting differential treatment efficacy are specific to the treatments being compared (Yoder & Compton, 2004). This makes using extant studies less useful for posing predictions for future studies using different treatments.

CRITERION-RELATED VALIDATION

Definition of Criterion-Related Validation

Perhaps the most commonly used method of validation is criterion-related validation. Such a process involves examining the magnitude of the association between the variable of interest with "gold standard" measures of the *same* generalized characteristic of interest. Ideally, a gold standard has much evidence that it is a scientifically useful measure of the generalized characteristic of interest for particular purposes and populations (Cronbach & Meehl, 1955). In practice, measures used as gold standards are often ones that have been on the scene for a long time or are used by many professionals to measure the generalized characteristic of interest (Haynes & O'Brien, 1999). One might be interested in a new measure because it is shorter or less expensive than the gold standard. However, is seldom the case for observational variables. Criterion-related validation can be implemented between two variables that are measured at the same measurement period (i.e., concurrent) or at different measurement periods (i.e., predictive). A more likely application of criterion-related validation is when the investigator asks whether an indirect and earlier measure of a process is related to a later more direct and more widely accepted measure of the same process. For example, one might ask whether the presence of a *single* declarative intentional communication act in a newly designed screening instrument predicts later *number* of declarative intentional communication acts in a gold standard communication assessment given several months later.

Primary Appeal of Criterion-Related Validation

Criterion-related validation is by far the most frequently used method of validation for observational variables. This may be because it does not require direct appeal to a particular theory. An atheoretical approach to validation may be attractive to some because any given theory is likely to be unattractive to some professionals (Geisinger, 1992). Another possible reason that criterion-related validation is a popular approach to validation of observational variables may be that some researchers question the accuracy of other- or self-report measures and wish to convince readers that observational measures provide some of the same information as the more widely used self- or other-report measures. However, it appears that both measures, the proposed gold standard and the new measure, are under scrutiny in such cases. Technically, these are not really criterion-related tests of validation, even if the investigator thinks of them as such.

Weaknesses of Criterion-Related Validation

The overreliance on criterion-related validation has been much maligned (Geisinger, 1992). An association between two questionable measures is not the logic of criterion-related validation. However, this misapplication of the criterion-related validation logic is quite common. The major issue is that the presence of a true gold standard is rare. Most measures have small or equivocal records of validation. In addition, many people who have attempted to weave arguments that a particular measure is a gold standard have fallen prey to the logical error that one type of validation evidence supports the validity of the measurement system for a different purpose (Geisinger, 1992). More importantly, at some point in history, *every* generalized construct lacked a gold standard measure, including general intelligence. This obvious point begs the question, "How do we examine the validity of *any* measure when no gold standard exists?"

CONSTRUCT VALIDATION

Definition of Construct Validation

Construct validation methods use correlational, group, or experimental studies to test hypotheses regarding whether the measure of the construct of interest has theoretically predictable associations or group differences (Cronbach & Meehl, 1955). Confirmation of predictions provides support for the particular use of the score tested in the study. When predicted findings are tested, but are not found, the study weakens the evidence regarding the value of the proposed use of the test score.

In response to the growing concern that many psychologists of the day were settling for circular logic when asked to define the words they were using (e.g., "intelligence is what intelligence tests measure"), Cronbach and Meehl (1955) elaborated on the principles set forth by the American Psychological Association's (APA) first set of officially endorsed recommendations for judging the soundness of psychological measurement (American Psychological Association, 1954). Among other points, Cronbach and Meehl pointed out that a new measure can be better than what is considered a criterion measure. As an example, the authors pointed out that prior to the invention of the mercury-based thermometer, human-based judgment was the criterion by which other measures of temperature were judged. After inventing the mercury-based thermometer, it was easily shown that the new "test" was a better measure of temperature (i.e., the construct of interest) than the "gold standard." The majority of the members of the APA embraced Cronbach and Meehl's ideas about construct validity (Waller, Yonce, Grove, Faust, & Lenzenweger, 2006). Construct validation is applicable to generalized characteristics and vaganotic concepts of measurement but less relevant for context-dependent behaviors and idemnotic concepts of measurement.

The collection of predicted links between a measure of interest to (a) measures of other constructs and (b) group memberships is called the *nomological net* (Cronbach & Meehl, 1955). The general idea behind this important concept is that if the scores of the observational variable correlate with variables they are predicted to correlate with and discriminate membership in groups they are predicted to discriminate, then this is an evidence that we have measured what we have intended to measure. In addition to the same subtypes applied to criterion-related validation (concurrent vs. predictive), construct validation has been subdivided into types according to whether group differences (discriminative) or associations (nomological) are expected.

Discriminative Validation

Discriminative validation should not be confused with *discriminant validity* (see the discussion of multitrait, multimethod validation [MTMM] theory later in this chapter for a definition of the latter). An example may help to understand the former. One study attempted to identify the accuracy with which children could be identified as having Fetal Alcohol Syndrome Disorder (FASD) from a particular language feature (nominal reference errors in oral story telling) in preschoolers (Thome & Coggins, 2008). There were two groups with known diagnostic status: FASD (n = 16) and typically developing (n = 16) matched on several important variables. As hypothesized, the language feature discriminated membership to diagnostic groups with 88% accuracy.

In some contexts, one problem with such research is that diagnosis is a dichotomous variable, whereas, the construct of interest may be continuous. Additionally, the method by which many observational variables will be used to predict group membership is an exploratory one (i.e., response optimization curves) that maximizes accuracy through an iterative process. Such a process maximizes the probability of identifying cut-off scores that are sample specific (i.e., do not replicate in other samples of the population). Finally, there is only a single outcome (i.e., group membership) that is being predicted. These issues make discriminative tests of construct validation less falsifiable than other approaches. For purposes other than diagnostic ones, discriminative tests of construct validity are often less convincing than other approaches.

Nomological Validation

Nomological validation involves testing whether the variable of interest is associated with multiple, continuously measured constructs that are theoretically related to the construct of interest. As an example of nomological validation, we asked whether a new measure of "breadth of interests" in children with autism was predictive of the three latter-measured abilities that theory suggests breadth of interest should predict in nonverbal children (Bruckner & Yoder, 2007).

In this situation, there was no gold standard measure of breadth of interest or its opposite—"restricted interest"—for our population. All of the proposed measures of restricted interest in the extant literature were developmentally inappropriate for our population (i.e., nonverbal children with autism). We reasoned that the "interests" of nonverbal children are demonstrated through children's play. Without going into detail about the particular play variable, suffice it to say that we reasoned the number of objects on which children used a particular type of play was a potential measure of breadth of interest. Because the opposite of broad interests, restricted interests, has been claimed to result in ignoring social input about objects, we reasoned that broad interests in objects should be related positively to other child skills that require attention to an object and to a person: (a) responding to other's attentional directives about objects, (b) imitating other's actions on objects, and (c) directing gaze or gestures to both objects and the message recipient during communication. The results supported the predictions.

Some may be dissatisfied with nomological validation evidence because of the tendency in nature for positive variables to covary and for negative variables to covary. Adding hypotheses about variables that should *not* be associated with the variable of interest improves the falsifiability of the set of predictions and may increase the persuasiveness of the validation evidence.

Multitrait, Multimethod Validation

One attempt to improve the falsifiability of construct validation is the MTMM approach. The MTMM approach involves testing whether the measure of interest correlates more strongly in a positive direction with (a) other measures of the same construct that use a different method of assessment than (b) measures of another construct that use the same method of assessment (Campbell & Fiske, 1959). The "multitrait" part of this approach involves using the same general methods of assessment to measure two or more different generalized characteristics.

For example, we might wish to measure children's expressive vocabulary and children's pragmatic language skills (e.g., extent to which children use language in socially appropriate ways). The "multimethod" part of this approach involves using two different methods of assessing the same construct. For example, we might use parent report and direct observation to measure these two constructs. The use of two measures of the same trait in the MTMM method is somewhat similar to the logic used in criterion-related validation. The difference between the two validation approaches is that the MTMM method does not require claiming one method to be a gold standard. In this sense, the MTMM approach fits reality better than the criterion-related validation approach. Additionally, the MTMM approach acknowledges that one can get high correlations between two measures of different constructs because they share a method of assessment. This issue is not addressed in the traditional nomological net approach to construct validation and is the "multitrait" part of the approach. To complete the MTMM matrix, we cross all methods and all constructs being tested. Table 10.3 provides a fictitious example of the type of results one would expect if direct observation measures of expressive vocabulary and pragmatics showed MTMM validation. This pattern of findings would also support the MTMM validation of parent-report measure of vocabulary and pragmatics.

Because the MTMM matrix is a square and (mostly) symmetrical correlation matrix, we only have to examine one (e.g., the bottom) triangle

Table 10.3

ILLUSTRATION OF FICTITIOUS EXAMPLE OF EXPECTED ASSOCIATIONS IN A MULTITRAIT, MULTIMETHOD (MTMM) MATRIX FOR TWO METHODS OF ASSESSING EXPRESSIVE VOCABULARY AND PRAGMATICS

		OBSERVATION		PARENT REPORT	
		VOCABULARY	PRAGMATICS	VOCABULARY	PRAGMATICS
Observation	Vocabulary	.89 (Reliability)			
	Pragmatics	.38 (Method-related error)	.70 (Reliability)		
Parent report	Vocabulary	.57 (Validity)	.08 (Different-trait, different-method)	.65 (Reliability)	
	Pragmatics	.10 (Different-trait, different-method)	.51 (Validity)	.43 (Method-related error)	.68 (Reliability)

of the square matrix. The exception to the symmetry of the MTMM correlation matrix is that the same-trait, same-method diagonal contains the reliability coefficients instead of perfect correlations (i.e., 1.0). For example, it might contain the intraclass correlation coefficients (ICCs) for the variable from different measurement contexts or different reporters. Alternatively, it could contain the ICCs for the variable from different time periods. These should be the highest in the matrix because generalized characteristics are expected to be stable over contexts and reporters. Such reliability coefficients are seen as evidence of the extent to which a measure is related to itself.

The same-trait, different-method diagonal contains the correlation between two different measures of the same construct. Because our example has two "traits," it has two validity coefficients: one for vocabulary and one for pragmatics. If our measures have "strong MTMM validation evidence," these coefficients will be the next highest of the four types of correlation coefficients (after the reliability coefficients). We also expect them to be positive and statistically significant (i.e., the confidence interval does not include zero). These are called "convergent validity coefficients."

By having predictions of both high and low associations, the MTMM provides more convincing evidence of construct validation than the nomological network approach because there are more predictions to verify (i.e., it is more falsifiable than the nomological network approach). The other two types of coefficients in the MTMM matrix represent examples of discriminant validity (i.e., variables that are predicted to have lower correlations with the variables of interest). The different-trait, same-method triangle contains the measure of "methods-related variance" (i.e., the extent to which using the same way to assess different constructs has an effect on scores). From a construct validation perspective, methods-related variance is considered measurement error.

For our example, if the variables we consider measures of vocabulary are both construct valid measures of vocabulary, their correlation with each other should be greater than the two within-method correlations between pragmatics and vocabulary. Similarly, if our measures of vocabulary have strong MTMM validation evidence, we expect the convergent validity coefficients to be higher than the two different-trait, different-method coefficients. In Table 10.3, the measures of vocabulary have strong MTMM validation evidence because the validity coefficient exceeds the two methods-related variance (.57 > .38 and .43) and the validity coefficient exceeds the two different-trait, different-methods coefficients (.57 > .08 and .10). A similar pattern supports the validity of the pragmatic measures. Note that one measure of each construct does not have more support for validity than the other because neither measure is considered a "gold standard" measure of the construct.

AN IMPLICIT "WEAKNESS" OF SCIENCE?

Figure 10.1 is an illustration of the iterative observational research process. We have adapted this from a figure presented for the general psychological measurement process (Whitley, 1996). The scientific enterprise requires an iterative process in which all of the decisions discussed in this book are applied to create variable scores to test highly falsifiable a priori hypotheses. Ultimately, theories and constructs are neither "true" nor "false." Either they are scientifically useful or they are not (Waller et al., 2006). When we confirm falsifiable a priori hypotheses, we conclude that the theory generating the hypotheses is scientifically useful.

When a priori predictions are not confirmed (i.e., the null hypotheses are not rejected), it is usually the case that we have multiple explanations for the results. These explanations can be divided into two main classes: (a) the theory generating the hypotheses is in need of modification and (b) the measurement system did not quantify the context-dependent behavior or generalized characteristic in a reliable or scientifically useful manner. Immediately after the study, we are often left with guesses as to which explanation is more likely.

Often, it is not until years of tweaking measurement systems, minor elaboration of the theory, and continued failures to confirm theoretically motivated predictions that theories are discarded. Usually, theories are not discarded until a more scientifically useful theory (i.e., one that requires fewer assumptions about unmeasured variables that also accounts for more or the same amount of empirical data) is put forth. This is what occurred when Einstein's theory of relativity replaced Aristotle's theory that a fictitious substance called "ether" was necessary to transmit the stars' light to earth.

Some will find this state of affairs unsatisfying because it seems somewhat circular, slow, and nondefinitive regarding an externally defined "truth." Scientists believe that centuries of nonscientific alternatives to

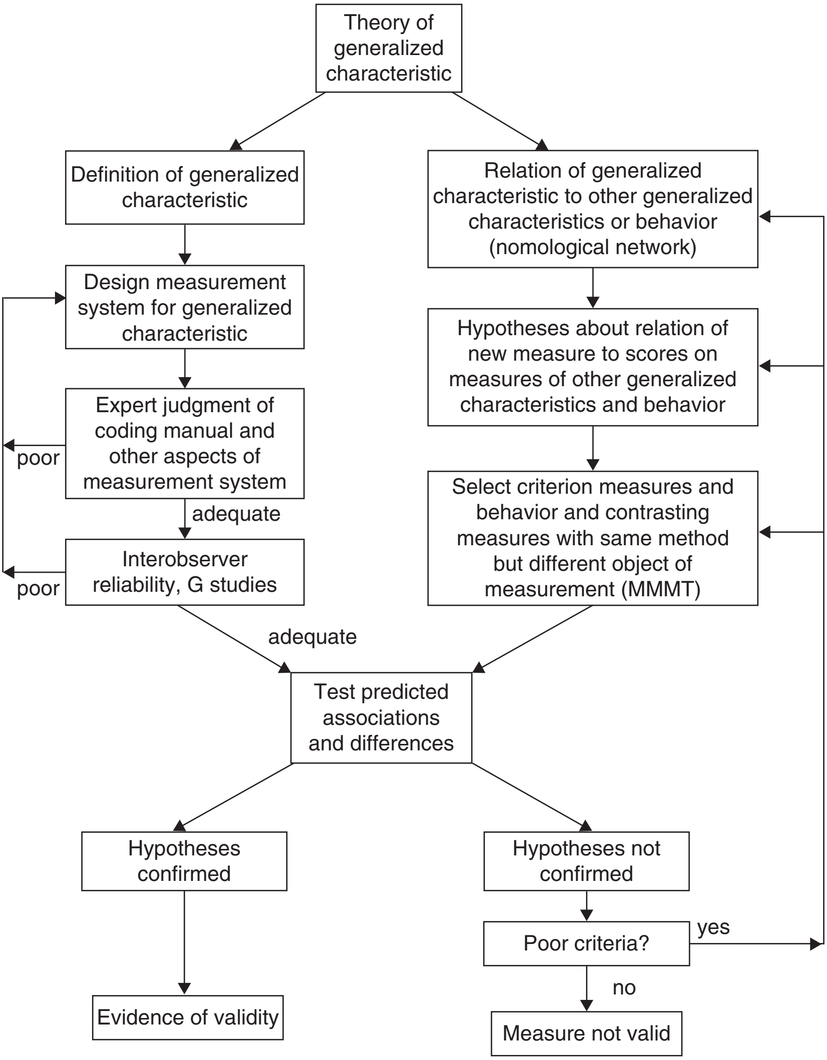

Theory of generalized characteristic
Definition of generalized characteristic
Design measurement system for generalized characteristic
Expert judgment of coding manual and other aspects of measurement system
poor
adequate
Interobserver reliability, G studies
poor
adequate
Relation of generalized characteristic to other generalized characteristics or behavior (nomological network)
Hypotheses about relation of new measure to scores on measures of other generalized characteristics and behavior
Select criterion measures and behavior and contrasting measures with same method but different object of measurement (MMMT)
Test predicted associations and differences
Hypotheses confirmed
Hypotheses not confirmed
Poor criteria?
yes
no
Evidence of validity
Measure not valid

Figure 10.1 Illustration of flowchart of the scientific process in observational studies.

knowledge development have proven even more dissatisfying from the perspective of generating a replicable knowledge base. We believe that the scientific method, even if not completely satisfying to some, is the best we have to offer.

RECOMMENDATIONS

Readers should seek and investigators should provide psychometric evidence regarding the degree to which observational variable scores are scientifically useful for a particular purpose and population. Part of this psychometric evidence is validation evidence. Validation evidence is important for all variables, including observational variables. The amount and types of validation evidence that are relevant vary by the research design, the object of measurement, and the research purpose of the study. This is particularly true for the object of measurement. When the object of measurement is abstract (i.e., a construct) and is thought to represent a generalized characteristic, the critical reader should demand greater validation evidence than when the object of measurement is a context-dependent behavior.

Sufficient validation evidence will not be provided in a single study. The validation process is an ongoing one that occurs across many studies. It is the job of the investigator who aims to assess an object of measurement as a generalized characteristic to contribute to this evidence. Occasionally, a professional or professional group will summarize and critically analyze the validation information for a particular observational variable for a particular purpose and population, but this is rare. Most often, it will fall to the critical reader to accrue and assess the cumulative validation evidence. Ultimately, measurement systems are reflections of motivating theories for the studies. When theoretically motivated predictions are not confirmed, we are left with two primary explanations: the theory needs modification; the measurement system needs modification. Each observational study is part of the noble enterprise to improve both.

REFERENCES

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, 67–78.

Bruckner, C., & Yoder, P. (2007). Restricted object use in young children with autism: Definition and construct validity. *Autism, 11*, 161–171.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carr, E. G., & Durand, C. M. (1985). The social-communicative basis of severe behavior problems in children. In S. Reiss & R. R. Bootzin (Eds.), *Theoretical issues in behavior therapy* (pp. 219–254). New York: Academic Press.

Cronbach, L. (1988). Five perspectives on the validity argument. In R. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Geisinger, K. F. (1992). The metamorphosis to test validation. *Educational Psychologist, 27*, 197–222.

Goodwin, L., & Leech, N. (2003). The meaning of validity in the new standard for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181–191.

Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.

Haynes, S. N., & O'Brien, W. H. (1999). *Principles and practice of behavioral assessment*. New York: Kluwer.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ:Erlbaum.

Kazdin, A. (1981). Drawing valid inferences from case studies. *Journal of Consulting & Clinical Psychology, 49,* 183–192.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting & Clinical Psychology, 69*, 159–172.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: ACE and MacMillan.

Nelson-Gray, R. O. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15*, 521–531.

Primavera, L., Allison, D. B., & Alfonso, V. C. (1997). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–90). Mahwah, NJ: Erlbaum.

Thome, J., & Coggins, T. E. (2008). A diagnostically promising technique for tallying nominal reference errors in narratives of school-aged children with Fetal Alcohol Syndrome Disorders (FASD). *International Journal of Language & Communication Disorders, 43*, 570–594.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*, 242–261.

Waller, N., Yonce, L., Grove, W., Faust, D., & Lenzenweger, M. (2006). *A Paul Meehl reader: Essays on the practice of scientific psychology*. Mahwah, NJ: Erlbaum.

Whitley, B. E. (1996). *Principles of research in behavioral science*. Mountain View, CA: Mayfield.

Yoder, P. J., & Compton, D. (2004). Identifying predictors of treatment response. *Mental Retardation and Developmental Disabilities Research Reviews, 10*, 162–168.

Yoder P. J., & Stone, W. (2006) Randomized comparison of two communication interventions for preschoolers with autism spectrum disorders. *Journal of Consulting and Clinical Psychology*, 74, 426–435.

This page intentionally left blank

Glossary

Accuracy: When applied to coding, agreement between an observer with an expert consensus-coded file (criterion-coding standard).

Accuracy matrix: Agreement matrix for the trainee's coding with the experts (criterion-coding standard).

Accuracy proportion: The number of correct responses/number of opportunities for correct responding.

Agreement matrix: A type of symmetrical matrix in which the rows and columns are the categories in the coding manual for the same dimension plus a row and column for "no coded behavior tallied." The rows represent one observer's coding, while the columns represent the other observer's coding. The tallies are the result of point-by-point agreement checks. Agreements are represented on the diagonal and disagreements are represented on the off-diagonal cells.

Antecedent: A term used in sequential analysis to refer to the hypothesized causal behavior, the hypothesized prompt, or hypothesized discriminative stimulus. Sometimes referred to in other sources as the "given" behavior.

Backward sequential analysis: A type of sequential analysis in which the investigator tallies the number of times certain behaviors occur *before* the behavior of interest. One tabulates the sequence of behaviors into the 2 × 2 table moving backward in time. That is, the "first behavior" in the behavior pair actually occurs *after* the "second behavior" in the behavior pair.

Behavior sampling methods: A set of methods by which observers decide how to code an observation session. There are several types of behavior sampling methods including continuous and intermittent behavior sampling and interval sampling with two or more subtypes under each of these.

Ceiling effects: A situation in which most sessions' or participants' score are near the empirical maximum of the variable.

Chance agreement: The type of agreement that could occur when both or one of the observers records every instance of key behaviors at random. Chance agreement is only estimable when we have an exhaustive coding space. Total chance agreement is the sum of the chance agreement for the occurrence of a behavior plus chance agreement for the nonoccurrence of a behavior.

Chance occurrences of the sequence: A term used in sequential analysis to mean an estimate of the random occurrence of the sequence of interest. It is computed as follows: (simple probability of target) × (simple probability of antecedent) × (total number of coded units).

Classifying error: Disagreement among observers on which type of relevant behavior occurred.

Coded unit: The entity that the observer identifies and (perhaps) classifies. When we are deriving number (i.e., count) using event sampling, the coded unit is the behavior. When we are deriving duration using timed-event behavior sampling or when we use an interval sampling method, the coded unit is a time unit that is assigned to a behavior's presence or absence. In the case of duration from timed-event sampling, the coded time unit is often seconds. In all interval coding methods, the coded time unit is the interval.

Coding manual: The set of rules, definitions, examples, and near examples that guide the observers in counting and/or indicating the duration of the behaviors of interest.

Concatenating sessions: To copy coded data from session one and paste it onto the end of the coded data for session two, and so on, to create a single session for analysis.

Conceptual definitions: Meanings for coding terms that provide the theoretical framework observers need to judge whether marginal examples fit the rationale for inclusion in the category.

Consensual drift: When two observers agree with each other but neither agrees with a criterion coding standard.

Consensus coding: A process by which observers discuss each disagreed-upon act and decide through discussion and application of the manual how the act "should be coded."

Consistency proportion: The number of responses to an eliciting stimulus/number of opportunities for the response.

Conspicuous participant sampling: The observer watches the entire group and notes which individual is engaged in any predefined, conspicuous, and rarely occurring behaviors.

Construct validation: A cumulative process by which empirical studies test whether particular measurement systems yield variables that perform as expected by theory and logic. Expected performance is tested via testing a prior prediction regarding associations, group differences, and changes overtime.

Content validation: As applied to a coding manual, content validation is the expert rating of the relevance and representativeness of the examples and instances identified by the definitions in the coding manual to the stated object of measurement. Content validation can also include expert judgment regarding the adequacy of the measurement context, the number of sessions averaged, behavior sampling method, the coding decision recording method, the observation session recording method, and the metric with regard to measuring what the investigator says she wants to measure.

Context-dependent behaviors: Behaviors that are considered important for their own sake and behaviors that are thought to be reflections of states. These are *not* expected to inform us of what occurs outside of the measurement context.

Contingency space analysis: A method of sequential analysis in which (a) the transitional probability of a behavior that is hypothesized to act as a reinforcer after a target behavior is compared with (b) the transitional probability of the hypothesized reinforcer occurring after behavior other than the desired behavior. To the extent this difference is positive, there is a heighted probability that the consequence will function as a reinforcer for the target behavior. To the extent this difference is negative, there is a heighted probability that the consequence will function as a reinforcer for not using the target behavior (differential reinforcement for other behavior). This logic has also been called "operant contingency."

Contingency table: In sequential analysis, a method of organizing the sequence of pairs of coded units when all relevant events are included in the coded data.

Continuous event sampling: This behavior sampling method requires counting or tallying the number of instances of each key behavior that occurs during the observation session.

Continuous timed-event sampling: This behavior sampling method requires observing the entire observation session; it also requires indicating the time of occurrence of the onset (and sometimes, offset) of each instance of a key behavior.

Correlated measurement error: In observational research, this occurs when the observer systematically overestimates the true score in one group or phase, while systematically underestimating the true score for those in the contrasting group or phase.

Count coding systems: Methods of quantifying variables that are designed to lead the observer to count the number of instances and/or duration of instances of the key behaviors.

Criterion coding standard: A repeatedly and expertly coded session. Our best estimate of the true occurrence of events in a session.

Criterion-related validation: A process that involves examining the magnitude of the association between the variable of interest with a "gold standard" measure of the *same* generalized characteristic of interest. Ideally, a gold standard has much evidence that it is a scientifically useful measure of the generalized characteristic of interest for particular purposes and populations.

D cell problem: In event-lag sequential analysis and 2 × 2 interobserver agreement matrices, this is the issue of what is considered a "relevant" behavior to code. If this definition is not accurate and complete, then the count in this cell (e.g., nonantecedent and nontarget cell in sequential analysis or point-by-point interobserver agreement on nonoccurrence in agreement matrices) will be "too low" and the estimate of chance will be inaccurate.

Decision studies: A group design, statistical method that allows us to posit different scenarios (e.g., number of sessions and/or number of observers) to estimate how many sessions and/or observers we need to achieve a criterion level of group design reliability (i.e., g or intraclass correlation) coefficient.

Discrepancy discussions: These occur when two observers discuss with each other or with a content expert the rationale for their coding particular coded units that are coded differently across observers.

Discrete events: Events or behaviors with clearly perceptible beginnings and endings.

Discriminant validation: Associations within a multitrait, multimethod (MTMM) approach matrix that are predicted to have lower correlations (e.g., measures of different traits) than associations predicted to have high association (e.g., measures of same trait).

Discriminative validation: A type of construct validation in which members of known groups are shown to be different on the variable of interest.

Duration: The time from onset of a behavior to offset of the same instance of the behavior.

Ecological validity: The extent to which measurement contexts resemble or take place in naturally occurring (unmanipulated) and frequently experienced contexts.

Event-lag sequential analysis: A type of sequential analysis in which one measures the extent to which a target behavior occurs a specified number of behaviors from an antecedent.

Exchangeability assumption in single case significance testing: Each pair of behaviors or time unit cannot influence the following pairs of behaviors or time units.

Exhaustive coding: The record includes all "relevant" coded units that occurred in the observation session. These occur when agreement on nonoccurrence of any key behavior is defined in the interobserver point-by-point agreement matrix. It is most common when time or interval is the coded unit.

Expected value of a cell: In the context of sequential analysis and kappa, this is the value expected by chance. As a count value, it is computed as (simple probability of relevant row marginal) × (count of relevant column marginal). In terms of probability, it is computed as (simple probability of relevant row marginal) × (simple probability of relevant column marginal).

Experts: People with both explicit (those who teach others to do the skill) and implicit (those who practice the art or skill being studied) knowledge of the context-dependent behavior or generalized characteristic of interest.

Facets: Factors in a generalizability study that represent measurement error.

Falsifiable research question: A prediction or question that specifies (a) the dependent and independent variables, (b) the investigator's expectations of an association or a difference, and (c) the investigator's expectations regarding direction of the association or difference (e.g., positive association, experimental group [or phase] as superior [i.e., greater]) prior to analyzing the data.

Fidelity of treatment: A measurement of the extent to which persons implementing the treatment do so as intended by the investigator.

Floor effects: A situation where most sessions' or participants' scores are near zero or the empirical minimum on the scale.

Focal participant sampling: Coding one participant for a predetermined period, then coding a different participant in the group for the same period of time, and so on, until all selected participants have been coded.

Frequency: When used in the context of sequential analysis, it means "number of instances."

*Generalizability (*g*) coefficient*: Conceptually, a g coefficient is the proportion of between-participant variance in observed scores that is true score. It is computed as the person variance on the variable of interest/ total variance in the reliability sample on the same variable. It is an intraclass correlation coefficient.

Generalizability (G) studies: A group design, statistical approach to quantifying reliability of predictors and outcomes. It focuses on quantifying sources of measurement error and reliability of between-participant variance.

Generalized characteristics: These are the psychological constructs (e.g., skills or attributes) that are assumed to be stable (in the group design sense of the word) over relevant contexts and time. Behaviors we observe are mere signs of levels on such constructs.

Gold standard: A measure that has much evidence that it is a scientifically useful measure of the generalized characteristic of interest for particular purposes and populations.

Homogeneity of covariance assumption: An assumption of statistical control methods in which the association between the controlled variable (i.e., covariate) and the outcome is assumed to be nonsignificantly different among groups in the research design.

Idemnotic: A concept of measurement that requires that the phenomenon of interest is (a) measured along a continuum, (b) has an absolute and often preexisting possible minimum, and (c) uses units or steps whose existence is established independently of variability in the phenomenon being measured. This concept of measurement is particularly common when measuring dependent variables in single-subject designs and fidelity of treatment measures in group and single-subject designs.

Influential variables: Variables that affect the occurrence of the key behaviors in an observation session.

Interobserver reliability: When referring to reliability of predictors and dependent variables in a group design predictors or dependent variables, this means that the ranking of participants on a particular variable is very similar regardless of the observers used to code the observation sessions. It is often quantified by the intraclass correlation coefficient. In a single-subject design measuring dependent variables or in single-subject and group designs measuring fidelity of treatment, this means point-by-point agreement on the variable of interest at the level of analysis used to address the research question.

Interoccurrence or interresponse time: The time from the offset of an event to the onset of a second occurrence of an event from the same class of behavior (e.g., the time between communication acts).

Interval scale of measurement: An ordinal scale of measurement in which the intervals between values on the scale indicate the same amount of the dimension being measured at the extremes of the scale as at the middle of the scale.

Intraclass correlation coefficient (ICC): In the context of group design interobserver reliability, it is also a g coefficient in a single facet generalizability study with only two observers. Conceptually, it is the proportion of total variance in a reliability sample due to between-person variance in the true score.

Kappa: An index of point-by-point agreement that is designed to control for chance agreement. Conceptually, it is the proportion of potentially available nonchance agreement (i.e., 1 – total chance agreement) that is attained (i.e., observed total agreement – chance total agreement).

Lags: A way of specifying within sequential analysis (a) the direction of the analysis (i.e., forward, backward, concurrent) and (b) the number of coded units from the antecedent that the target is expected.

Latency: The time from the offset of a behavior or event to the onset of a second, different behavior (e.g., time from the start shot of a race to onset of a sprinter's run).

Live (in situ) coding: The observer codes the behavior while it is occurring.

Measurement error: Conceptually, this is the portion of the observed score variance that is *not* due to true score variance (e.g., differences among observers or measurement contexts). At an individual level, measurement error within a participant is the average deviation of observed scores around the estimated true score for that participant. At the group level, measurement error is reflected in different rankings of the participants on the dependent variable depending on the measurement context or observer that generated the observed score.

Measurement system: This is comprised of (a) a measurement context, (b) a coding manual, (c) a behavior sampling method, (d) a participant sampling method, (e) a session recording method, and (f) a coding decision recording method.

Methodological operationalism: An interpretation of operationalism that asserts that operational definitions of concepts are partial and temporary specifications used to study the real concept of importance.

Metric: The unit of measurement or type of number that indicates the level of a quantifiable dimension about a property of behavior or generalized characteristic.

Momentary interval coding: A method of interval sampling in which the observer marks a behavior as present if and only if the behavior occurs at the boundary of the interval (e.g., the end of the interval).

Multiple pass participant sampling: When the observer selects one participant and codes the entire session for only that one participant, then repeats the procedure for another participant by watching the recorded observation session again. This is repeated until all relevant participants are coded.

Multitrait, multimethod (MTMM) approach: A type of construct validation that attempts to improve the falsifiability of the construct validation method by testing whether the measure of interest correlates more strongly in a positive direction with (a) other measures of the same construct that use different methods of assessment (convergent validation) than (b) measures of another construct that use the same method of assessment (discriminant validation).

Near nonexamples of a category: In a coding manual, these are behaviors that are provided to help the observer define the boundaries of the coding concept. These are typically superficially similar in form or topography to true examples, but differ from true examples in an important way.

Nominal scale: A scale of measurement in which values are not ordered or ranked.

Nomological net: The collection of predicted links between a measure of interest to (a) measures of other constructs and (b) group memberships.

Nomological validation: This involves testing the extent to which the variable of interest is consistently associated in the predicted direction (i.e., positive vs. negative) with multiple, continuously measured constructs that are theoretically related to the construct of interest.

Nonexhaustive coding spaces: Coded records for which point-by-point interobserver agreement on nonoccurrence of any key behavior is *not* defined. Often occurs in event behavior sampled data.

Noninfluential variables: Variables that do not affect the occurrence of the key behaviors in an observation session.

Nonoccurrence chance agreement: Conceptually, this is the estimated agreement between observers' coding of the same session on the nonoccurrence of a key behavior that could occur through a random process. Its formula is the probability base rate of nonoccurrence of the key behavior as estimated by Observer 1 × the probability base rate of nonoccurrence of the key behavior as estimated by Observer 2. Using the cell addresses of an exhaustive 2 × 2 agreement matrix, the formula is $([b + d]/N) \times ([c + d]/N)$. Nonoccurrence chance agreement can only be computed on an exhaustive agreement matrix.

Nonoccurrence percentage agreement: Using the cell addresses for an exhaustive 2 × 2 agreement matrix, the formula for nonoccurrence percentage agreement is $(d/[d + c + b]) \times 100$. Nonoccurrence percentage agreement can only be correctly computed on exhaustive coding spaces.

Nonsequential metric: A number scale used to quantify the object of measurement that does not reflect the sequence of behaviors within an observation session.

Number: In the context of nonsequential metrics, this is the count of instances of a key behavior within an observation session. Technically, it is the number of onset–offset cycles of instances of the same category of behavior.

Observational variable score: The end product of all of the decisions and procedures made to produce the score (i.e., measurement context, number of sessions averaged, coding manual, behavior sampling method, coding decision recording method, observational session recording method, and metric selection). It is labeled by stating the object of measurement and metric (e.g., rate of hand raising).

Observer drift: The occurrence of an observer agreeing less often with a criterion coding standard than was the case immediately after he reached mastery level accuracy during initial training.

Occurrence chance agreement. Conceptually, this is the estimated agreement between observers' coding of the same session on the occurrence of a key behavior that could occur through a random process. In probability form, this is the base rate probability of the key behavior as estimated by Observer 1 × the base rate probability of the key behavior as estimated by Observer 2. Using the cell labels in an exhaustive 2 × 2 agreement table, the formula is $([a + b]/N) \times ([a + c]/N)$.

Occurrence percentage agreement: This is the percentage of agreements divided by agreements plus disagreements on the occurrence of a particular category. Using the cell addresses for the 2 × 2 agreement matrix, the formula for occurrence agreement is $(a/[a + b + c]) \times 100$. This index of point-by-point agreement can be computed on a nonexhaustive agreement matrix because it does not consider chance occurrence agreement.

Offset: The end of a behavior.

Onset: The beginning of a behavior.

Operant contingency: See contingency space analysis.

Operational definitions: In the context of coding manuals, these are the meanings for coding terms that use only words with observable referents to define concepts.

Ordinal scale of measurement: A scale in which values are ordered, but the intervals between values do not represent equal amounts of the object of measurement.

Other report: Behavioral ratings or reports completed by asking others who know what the participant does (and often how often or consistently).

Partial interval coding: A type of interval sampling in which the observer marks one and only one occurrence of a key behavior when the behavior occurs *anytime* during the interval.

Participant sampling: A method used to decide which participant to code when there is more than one participant to be coded from a single observation session. There are several types of participant sampling, including focal, multiple pass, and conspicuous.

Permutation test: A type of statistical significance test that generates its own empirical probability distribution.

Physically based categories: Definitions of behaviors that rely only on detection of the presence or absence of stated behaviors. This type of category is very narrowly defined and is often composed of an *exhaustive* list of the behaviors that constitute the whole of the category. There is a greater emphasis on operational definitions than on conceptual definitions in physically based categories.

Point-by-point agreement: The extent to which two people see the same occurrence of the same example of the same category.

Proportion metrics: These are metrics that result from dividing one number by another. Three commonly used proportion metrics are rate style, accuracy, and consistency. For example, accuracy or consistency is the number of correct responses/number of opportunities for correct responses.

Prototypical examples of a category: These are examples of a category that share all critical attributes of the category and are quite common. They are provided to help the observer relate the operational definitions at a level of analysis that is more commonly used by educated consumers of the research.

Rate: The number of acts/duration of codeable portion of the observation session.

Ratio scale of measurement: An interval scale of measurement with a meaningful zero value.

Reactivity: Participants acting differently when watched than when not.

Reliability: In classical measurement theory: True score variance/observed score variance.

Representativeness: Stable (in the group design sense of the word) across contexts that evoke behaviors of interest.

Segmenting rules: A set of rules used to define the onset and offset of events.

Self-report: Asking the participant what they do, feel, or think.

Semantic operationalism: An interpretation of operationalism that asserts that the meaning of a concept can be *exhaustively* defined by stating particular observable manifestations of a concept.

Sensitivity to change: As a validation concept, this is the degree to which a measure changes in a therapeutic direction after participation in treatment.

Sequential analysis: A method of quantifying the sequential or simultaneous occurrence of coded behaviors or of seconds or intervals in which a coded behavior has occurred within an observation session.

Sequential association: Such associations occur when the antecedent and target behaviors occur more or less often than would be estimated to occur by chance.

Sequential frequency: In the context of sequential analysis, the number of times the target follows the antecedent behavior.

Sequential variable: A variable that quantifies the extent to which the target behavior occurs within a specified number of coded behaviors or time units from the antecedent behavior.

Significance testing of a sequential association within a single case: The testing of whether the observed Yule's *Q* could have occurred due to behavioral sampling error (i.e., chance sampling of an observation session that happened to produce an atypically strong sequential association).

Simple probability: In event-lag sequential analysis, it is the number of times a behavior occurs divided by the total number of coded behaviors in the behavior sample. In time-lag sequential analysis, simple probability is the number of time units (e.g., second) a behavior is coded divided by the total number of time units coded in the total behavior session.

Socially based categories: This categories of behavior differ from physically based categories in that the former tends to have categories with more exemplars or behavioral forms and requires observers to make a judgment regarding whether the behavior in question has a particular function or meets a series of conceptual criteria. There is a greater emphasis on conceptual definitions in socially based categories than in physically based categories.

Sparse table: A contingency table in which there is an expected value equal to or fewer than five in any cell.

Stationarity: An assumption made when sessions are concatenated for sequential analysis. It is an assumption that the sequential association within each session is nonsignificantly different from those in other sessions.

Structuredness: An adjective applied to measurement context that refers to the degree to which we keep influential variables constant across sessions or participants.

Summary level agreement: The extent to which two people derive the same variable score (i.e., small estimate/large estimate proportion).

Systematic observation: A method of quantifying variables in which a coding manual, context of measurement, sampling methods, and metric are decided prior to collecting data.

Target: A term used in sequential analysis to refer to the behavior that is hypothesized to be affected by the antecedent.

Time-lag sequential analysis: This type of sequential analysis quantifies the extent to which the onset of a target behavior occurs *exactly* at a prespecified number of time units from the antecedent at a rate that is different from chance.

Time-window sequential analysis: This type of sequential analysis quantifies the extent to which a target behavior *occurs within a specific time window* (e.g., 5 s) from an antecedent at a rate that is different from chance.

Total percentage agreement: The total percentage of occurrence plus nonoccurrence agreement. Using a 2 × 2 agreement matrix, the formula for total percentage agreement for a particular category is $([a + d]/[a + b + c + d]) \times 100$. Total percentage agreement can only be computed from exhaustive coding spaces.

Transitional probability: In sequential analysis, it is the proportion of instances of the antecedent behavior which is followed by an instance of the target behavior. Using the cell labels of the 2 × 2 contingency table that follows the conventions for sequential analysis, the formula for this transitional probability is $A/(A + B)$. Transitional probabilities are different from accuracy or consistency proportions in that in a transitional probability the target behavior can, by definition, occur after other behaviors than that represented by the denominator.

Treatment utility: A method of validation that means the degree to which assessment is shown to contribute to beneficial treatment outcome.

True score: Theoretically, this is the mean of the observed scores from all valid measurement contexts for the generalized characteristic of interest. In generalizability theory, we *estimate* a participant's true score by averaging all *available* observed scores for that participant.

Type I error: Detecting a difference or relation when one does not really exist.

Type II error: Failing to detect a difference or relation that is present.

Uncorrelated measurement error: In observational research, the extent to which observers over- and underestimate true score is randomly distributed across design groups or phases.

Unitizing: A part of the coding process that indicates when and how many coded units are present during the observation session.

Unitizing error: Disagreement among observers regarding whether a relevant behavior occurred.

Vaganotic: A concept of measurement that is the dominant implicit or explicit concept in studies by investigators who are most interested in individual differences and developmental changes. It is the measurement approach used to conceptualize the assessment of the predictors and dependent variables in most group designs. The meaning of high and low is relative to a group. The group can either be the sample of participants in the study or another reference group (e.g., a standardization sample in a norm-referenced test).

Whole interval coding: A type of interval sampling in which the key behavior must occur during the entire interval for it to be coded as "present."

Yate's correction: One way that statisticians have addressed the sparse contingency table problem: adding .5 to all frequency values in all contingency table cells.

Yule's Q: The most widely appropriate index of sequential association. Using the cell addresses of the 2 × 2 contingency table, the formula for Yule's Q is $(A \times D—B \times C)/(A \times D + B \times C)$. It is equivalent to the odds ratio for the same 2 × 2 table, with the primary difference being that Yule's Q has a potential range from –1.0 to 1.0.

This page intentionally left blank

Index