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
Copyright © 2010 Springer Publishing Company, LLC

All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Springer Publishing Company, LLC, or authorization through payment of the appropriate fees to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978–750–8400, fax 978–646–8600, info@copyright.com or on the web at www.copyright.com.

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


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.
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

List of Figures xiii
List of Tables xv
Foreword xvii
Preface xix
Acknowledgments xxiii

1 Introduction and Measurement Contexts 1

Overview 1
Systematic Observation 2
Count Coding Systems 3
Importance of Falsifiable Research Questions or Hypotheses 4
Behavior as “Behavior” Versus Behavior as a Sign or Indicant of a
Construct 4
Two Interpretations of Operationalism 5
Distinction Between Context-Dependent Behavior and Generalized
Tendencies to Behave 7
Rationale for Identifying How We Are Conceptualizing Our Object
of Measurement 8
Influential Variables of a Measurement Context 9
Structuredness 9
Ecological Validity 10
Representativeness 10
Tension Between Structuredness and Ecological
Validity 11
Recommendations for Measuring Generalized Characteristics From
Observations 12
Potential Disadvantages of Systematic Observational Count
Measurement 13
Recommendations 14
References 15
2 Improving Measurement of Generalized Characteristics Through Direct Observation and Generalizability Theory 17

Overview 17
Two Concepts of Measurement 17
Generalizability Theory as a Measurement Theory for Vaganotic Measures 19
Example: Generalizability (G) Study With Multiple Sessions as a Single Facet 20
Consequences of a Low G Coefficient 23
Decision Studies 24
McWilliam and Ware as an Example of a Two-Faceted Decision Study 25
Practice Using a G Calculator on Data From a Two-Faceted G and D Study 26
Accuracy of D Study Projections 30
Implications of the Lessons of G and D Studies for Single-Subject Research 31
A Dilemma 32
Recommendations 33
References 33

3 Designing or Adapting Coding Manuals 35

Overview 35
Selecting, Adapting, or Creating a Coding Manual 36
Definition of a Coding Manual 36
Relation of the Coding Manual to the Research Questions and Predictions 36
Recommended Steps for Modifying or Designing Coding Manuals 37
Conceptually Defining the Context-Dependent Behavior or the Generalized Characteristic 37
Deciding the Level of Detail at Which the Behaviors Should Be Distinguished 38
Physically Based Definitions, Socially Based Definitions, or Both? 39
Defining the Lowest Level Categories 40
Sources of Conceptual and Operational Definitions 42
Defining Segmenting Rules 46
Defining When to Start and Stop Coding 47
The Potential Value of Flowcharts 48
Do Coding Manuals Need to Be Sufficiently Short to Be Included in Methods Sections? 49
Recommendations 49
References 51

4 Sampling Methods 53
Overview 53
The Elements of a Measurement System 53
Behavior Sampling 54
  Continuous Behavior Sampling 54
  Intermittent Behavior Sampling 55
Interval Sampling 56
  How Does Interval Sampling Estimate Number and Duration? 58
Participant Sampling 59
  Focal Sampling 59
  Multiple Pass Sampling 60
  Conspicuous Sampling 60
Reactivity 60
Live Coding Versus Recording the Observation for Later Coding 62
Recording Coding Decisions 64
Practice Recording Session 66
Recommendations 69
References 70

5 Common Metrics of Observational Variables 73
Overview 73
Definition of Metric 74
Quantifiable Dimensions of Behavior 74
Proportion Metrics 75
Proportion Metrics Change the Meaning of Observational Variables 75
  Scrutinizing Proportions 77
  An Implicit Assumption of Proportion Metrics 78
  Testing Whether the Data Fit the Assumption of Proportion Metrics 79
  Consequences of Using a Proportion When the Data Do Not Fit the Assumption 80
Alternative Methods to Control Nuisance Variables 85
  Statistical Control 85
  Procedural Control 85
Transforming Metrics of Observational Variables in Group Statistical Analyses 86
Scales of Measurement for Observational Variables 88
Contents

Sequential Association Indices as Dependent Variables in Group Designs 131
  Testing the Significance of a Mean Sequential Association 131
  Testing the Between-Group Difference in Mean Sequential Associations 132
  Testing the Within-Subject Difference in Sequential Associations 132
  Testing the Significance of the Summary-Level Association Between a Participant Characteristic and a Sequential Association Between Behaviors 133
Statistical Significance Testing of Sequential Associations in Single Cases 133
A Caveat Regarding the Use of Yule’s Q 136
Recommendations 137
References 138

8 Observer Training, Observer Drift Checks, and Discrepancy Discussions 141
Overview 141
Three Purposes of Point-by-Point Agreement on Coding Decisions 141
Two Definitions of Agreement 142
Agreement Matrices 145
Discrepancy Discussions 148
Criterion Coding Standards 149
Observer Training 151
Method of Selecting and Conducting Agreement Checks 153
Retraining When Observer Drift Is Identified 155
Recommendations 156
References 156

9 Interobserver Agreement and Reliability of Observational Variables 159
Overview 159
Additional Purposes of Point-by-Point Agreement 159
Added Principles When Agreement Checks Are Used to Estimate Interobserver “Reliability” of Observational Variable Scores 160
Exhaustive Coding Spaces Revisited 164
The Effect of Chance on Agreement 167
Common Indices of Point-by-Point Agreement
  Occurrence Percentage Agreement 168
  Nonoccurrence Percentage Agreement 168
  Total Percentage Agreement 169
  Kappa 169
Base Rate and Chance Agreement Revisited 171
Summary of Point-by-Point Agreement Indices 172
Intraclass Correlation Coefficient as an Index of Interobserver
  Reliability from the Vaganotic Concept of Measurement 174
Options for Running ICC With SPSS 175
Between-Participant Variance on the Variable of Interest
  Affects ICC 175
Using ICC as a Measure of Interobserver Reliability for
  Predictors and Dependent Variables in Group Designs 177
The Interpretation of SPSS Output for ICC 177
The Conceptual Relation Between Interobserver Agreement and
  ICC 178
Consequences of Low or Unknown Interobserver Reliability 178
Recommendations 180
References 181

10 Validation of Observational Variables 183
Overview 183
The Changing Concept of Validation 184
Understanding Which Types of Validation Evidence Are Most
  Relevant for Different Research Designs, Objects of
  Measurement, and Research Purposes 185
Content Validation 186
  Definition of Content Validation 186
  Different Traditions Vary on the Levels of Importance Placed on
    Content Validation 187
  Weaknesses of Content Validation 188
Sensitivity to Change 188
  Definition of Sensitivity to Change 188
  Influences on Sensitivity to Change 189
  Weakness of Sensitivity to Change 190
Treatment Utility 190
  Definition of Treatment Utility 190
  Weaknesses of Treatment Utility 192
Criterion-Related Validation 193
  Definition of Criterion-Related Validation 193
  Primary Appeal of Criterion-Related Validation 193
  Weaknesses of Criterion-Related Validation 194
Construct Validation 194
  Definition of Construct Validation 194
  Discriminative Validation 195
  Nomological Validation 196
  Multitrait, Multimethod Validation 197
An Implicit “Weakness” of Science? 200
Recommendations 202
References 202

Glossary 205
Index 221
Figures

CHAPTER 1
Figure 1.1 Continuum of degree to which a measured entity is context and time dependent. 12

CHAPTER 2
Figure 2.1 Variance estimates for the example 2-faceted g study. 29

CHAPTER 3
Figure 3.1 Illustration of a flowchart for a coding manual on intentional nonverbal communication. 50

CHAPTER 5
Figure 5.1 Illustration of the relation between numerator and denominator in proportion metrics. 78
Figure 5.2 Illustration of a distribution of proportion scores from .002 to .998 incremented by .002. 87
Figure 5.3 Illustration of a distribution of arcsine transformed proportion scores from .002 to .998 incremented by .002. 88

CHAPTER 6
Figure 6.1 Illustration of tallying behavior pairs into a $2 \times 2$ table for a forward event-lag sequential analysis. 106
Figure 6.2 Illustration of tallying behavior pairs into a $2 \times 2$ table for a backward event-lag sequential analysis. 114
CHAPTER 7

Figure 7.1  Example of timed-event data from ProcoderDV that will be analyzed by time-window sequential analysis.  121
Figure 7.2  Example of how MOOSES tallies timed-event data for a 5-s time-window sequential analysis.  123
Figure 7.3  Example output from time-window analysis of timed-event data from MOOSES.  124

CHAPTER 9

Figure 9.1  Illustrating proportion reliability scores on the same graph that displays primary proportion scores.  163
Figure 9.2  Illustration of the relation between chance agreement and kappa as a function of the base rate of the key behavior.  170

CHAPTER 10

Figure 10.1  Illustration of flowchart of the scientific process in observational studies.  201
Tables

CHAPTER 2

Table 2.1 SPSS Spreadsheet for 1-Faceted Generalizability Study 21
Table 2.2 SPSS Syntax and Output of ANOVA 23
Table 2.3 Results of Practice 2-Faceted G Study 27
Table 2.4 Formulae for Variance Estimates From a Fully Crossed 2-Faceted ANOVA 29

CHAPTER 4

Table 4.1 Summary of Results of the Behavior Sampling Exercise 67

CHAPTER 5

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 82
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 ($d = 1.9$) 84

CHAPTER 6

Table 6.1 Illustration of How Software Can Convert Duration to Onset in Two-Stream Data 104
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 107
Table 6.3 Illustration of a Contingency Table for Time-Window Sequential Analysis 110

xv
CHAPTER 7

Table 7.1 Examples of $2 \times 2$ Cell Values, Expected Value of a Cell, and Computed Yule’s $Q$ 128

CHAPTER 8

Table 8.1 Illustration of Exact Agreement on Interval Data 143
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 144
Table 8.3 Agreement Matrix for Data in Table 8.1 146
Table 8.4 Agreement Matrix for Agreement Data in Table 8.2 147

CHAPTER 9

Table 9.1 Illustration That Agreement on Numerator and Denominator Does Not Ensure Agreement on Proportion 162
Table 9.2 Illustration of Collapsing an Exhaustive Agreement Matrix to a $2 \times 2$ Table for Category “2” 165
Table 9.3 Ill-Advised Agreement Matrix for a Nonexhaustive Coding Space 166
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 176

CHAPTER 10

Table 10.1 Illustration of Types of Validation by Research Design and Object of Measurement (and Variable Type) 185
Table 10.2 Illustration of Five Types of Validation Evidence by Purposes of Observational Research 186
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 198
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
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


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
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 non-social 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 opera-
tionalism 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 manifesta-
tions of a concept. Viewed from a natural science perspective, psycho-
logical 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 interpre-
tation 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., intel-
ligence is what intelligence tests measure). Second, there is a need to
make statements about people even though it is impossible to exhaus-
tively 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 com-
pletely 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 opera-
tionalism including the position that operationalism was never meant to
be interpreted in this manner (Feest, 2005).

In contrast, from a methodological operationalism perspective, opera-
tional 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 “hun-
ger” 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 psy-
chological 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 con-
cept 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
Observational Measurement of Behavior

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
Observational Measurement of Behavior

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
Observational Measurement of Behavior

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


