Observational Measurement of Behavior

Second Edition

Observational Measurement of Behavior

Second Edition

by

Paul J. Yoder, Ph.D.
Vanderbilt University
Nashville, Tennessee

Blair P. Lloyd, Ph.D., BCBA-D
Vanderbilt University
Nashville, Tennessee

and

Frank J. Symons, Ph.D.
University of Minnesota
Minneapolis, Minnesota

·P A U L·H·
BROOKES
PUBLISHING CO®

Baltimore • London • Sydney

Paul H. Brookes Publishing Co.
Post Office Box 10624
Baltimore, Maryland 21285-0624
USA

www.brookespublishing.com

Typeset by Progressive Publishing Services, York, Pennsylvania.
Manufactured in the United States of America by
Sheridan Books, Inc., Chelsea, Michigan.

All examples in this book are composites. Any similarity to actual individuals or circumstances is coincidental, and no implications should be inferred.

Library of Congress Cataloging-in-Publication Data

Names: Yoder, Paul Jordan, author. | Lloyd, Blair P., author. | Symons, Frank J., 1967– author.
Title: Observational measurement of behavior / by Paul J. Yoder, Ph.D., Vanderbilt University, Nashville, Tennessee, Blair P. Lloyd, Ph.D., BCBA-D, Vanderbilt University, Nashville, Tennessee, and Frank J. Symons, Ph.D., University of Minnesota, Minneapolis, Minnesota.
Description: Second Edition. | Baltimore, Maryland: Paul H. Brookes Publishing Co., [2018] | Includes bibliographical references and index.
Identifiers: LCCN 2017049681 (print) | LCCN 2017051969 (ebook) | ISBN 9781681252483 (epub) | ISBN 9781681252476 (pdf) | ISBN 9781681252469 (paper)
Subjects: LCSH: Behavioral assessment.
Classification: LCC BF176.5 (ebook) | LCC BF176.5 .Y63 2018 (print) | DDC 150.72/3—s dc23
LC record available at https://lccn.loc.gov/2017049681

British Library Cataloguing in Publication data are available from the British Library.

2022 2021 2020 2019 2018

10 9 8 7 6 5 4 3 2 1

Contents

Section II Advanced Topics

Section III Putting It All Together

About the Authors

Paul J. Yoder, Ph.D., Professor of Special Education, Department of Special Education, Box 220, Peabody College, Vanderbilt University, Nashville, Tennessee 37203

For more than 30 years, Dr. Yoder has used observational measurement to study communication and language development in children with disabilities and how parental interaction influences their immediate and sustained use of nonverbal and verbal communication acts. Throughout his career, Dr. Yoder has contributed to the empirical basis for decisions affecting the scientific utility of observational variables. He teaches graduate courses on observational measurement and research design at Vanderbilt University.

Blair P. Lloyd, Ph.D., BCBA-D, Assistant Professor of Special Education, Department of Special Education, Box 228, Peabody College, Vanderbilt University, Nashville, Tennessee 37203

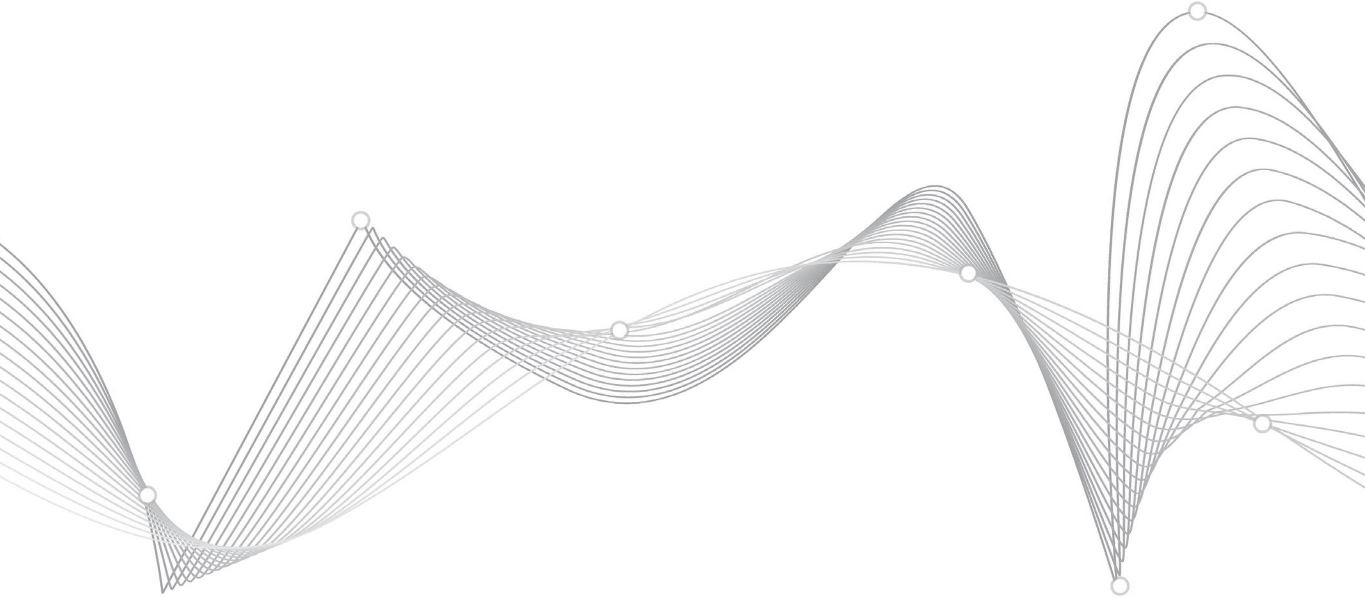
Dr. Lloyd's research focuses on individualized assessment and intervention for students with persistent challenging behavior. She is an active user of observational measurement and sequential analysis methods in her own research and has published multiple methodological papers on sequential analysis. She teaches graduate courses in experimental analysis of behavior and single-case research design.

Frank J. Symons, Ph.D., Professor, Department of Educational Psychology, College of Education and Human Development, 56 East River Road, Education Sciences Building, University of Minnesota, Minneapolis, Minnesota 55455

Dr. Symons is a Distinguished McKnight University Professor in Special Education and Educational Psychology at the University of Minnesota. His research agenda positions him in the crossroads of interdisciplinary inquiry in behavioral disorders and neurodevelopmental disabilities with several specific foci, including self-injury, pain, and Rett syndrome. Many of his approaches rely on direct observational methods.

SECTION I

FOUNDATIONAL TOPICS



Excerpted from Observational Measurement of Behavior, Second Edition
by Paul J. Yoder, Ph.D., Blair P. Lloyd, Ph.D., BCBA-D, & Frank J. Symons, Ph.D.
Brookes Publishing | www.brookespublishing.com | 1-800-638-3775

CHAPTER 1

Introduction to Systematic Observation and Measurement Contexts

The purpose of this chapter is to review a number of underlying issues involved in observational measurement of behavior. These issues, although not always explicitly articulated in a given research report, are critical to understanding the logic behind the different research approaches to quantifying behavior using systematic direct observation and the strategies used for doing so. In this chapter, we define the book's central topic: systematic observation using count coding. We then promote hypothesis-driven research as a general approach to maximize a study's scientific rigor and interpretability. Next, we discuss an important distinction between observed behavior as context dependent and observed behavior as a sign of a generalized person characteristic. These are two distinct types of objects of measurement. Because distinguishing between the two is difficult, we devote much of Chapter 1 to it. To illustrate why the distinction is important, we argue that each object of measurement has its own separate criteria for evaluating its scientific value. As part of this argument, we address the important concepts of ecological validity and representativeness. We wrap up the chapter with conclusions and recommendations regarding the issues discussed.

SYSTEMATIC OBSERVATION USING COUNT CODING

The *systematic observation* approach to measurement requires that before beginning data collection the following elements have been decided: the procedure (i.e., type of session) to observe, the definitions of key behaviors, and the type of number used to quantify the phenomenon of interest (Suen & Ary, 1989). An example of systematic observation is an observer recording the presence, quality, or amount of communication from a 15-minute parent–child interaction session. Other examples include observing engagement during a classroom activity or rating or counting key behaviors in a structured diagnostic evaluation, such as the Mullen Scales of Early Learning (Mullen, 1995). A final example includes transcribing utterances from a natural conversation and counting the occurrence and type of syntactic structures used therein. Systematic observation is contrasted with the type of observation used in qualitative research. The latter method requires fewer a priori decisions. Qualitative participant observational methods are covered in other sources (e.g., Taylor & Trujillo, 2001; Tracy, 2013) and will not be addressed in this book.

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.

Alternatives to Systematic Observation

Alternatives to systematic observation include *self report,* that is, asking the participants what they generally do, and *third-party report,* also known as *other* or *proxy report,* that is, asking people who have experience with the participant to make conclusions about the extent to which, or quality with which, the participant generally engages in particular behaviors. An example of a self report is a personality inventory, such as the Minnesota Multiphasic Personality Inventory, which asks participants to indicate the extent to which they generally engage in particular behaviors or experience particular events thought to be evidence of various personality disorders (Schiele, Baker, & Hathaway, 1943). An example of a third-party report is a parent inventory of words the child uses, for example, MacArthur-Bates Communicative Development Inventories (CDIs; Fenson et al., 2006). In both cases, the reporter is asked to draw from his or her memory of the target participant's behavior across many different contexts and periods. This book does not cover self-report or third-party report methods.

Self report: Measurement approach involving asking the participant what they do, feel, or think.

Third-party report: Measurement approach involving asking people who have experience with the participant to quantify some aspect of participant's general behavior.

Ways to Quantify Observations

Systematic observation can be used to quantify a phenomenon in three primary ways, the first of which is count coding, the focus of this book. *Count coding* involves indicating the occurrence of each instance or each instance's duration as it occurs during an observation. As such, count coding tends to quantify phenomena at a very detailed or microlevel. For example, a highly trained coder might count the number and duration of verbal responses to child vocal communication bouts as these responses occur in a 15-minute classroom activity. Results of count coding can produce various possible metrics (e.g., rates, proportions, indices of sequential association, latencies).

Transcribing observations requires a special note. Transcription is writing down what is said or occurs (or both). As such, it is a way to simplify what is observed to the elements considered critical for classifying the words, phrases, or utterances transcribed. The transcription is not count coding per se, but the transcription process identifies units that are often count coded. Therefore this process introduces error and thus needs to be subjected to the same rigorous standards as those used to monitor coding.

Within systematic observational measurement, two other alternatives used to quantify observations are rating scales and checklists. Relative to count coding, these methods tend to quantify the phenomenon at a more molar level. *Expert rating scales* often involve Likert-like scales on which an observer records global judgments about the quality or quantity of a particular class of behaviors after completing the entire observation. For example, after observing a parent and child interacting for 20 minutes, the observer rates the parent on parental responsivity by indicating where the parent fell on a 7-point scale. The design of the rating scale has assigned the behavioral anchors of *almost all of the time* and *almost never* to the two end points of the scale used to rate each item. The result is often a sum of Likert-like scores across a number of aspects of behaviors assumed to quantify a particular construct. (A *construct* is a psychological concept or process that is not directly observable, e.g., optimal parent–child interaction style.) *Observational checklists* involve having the observer indicate the presence or absence of key behaviors from a provided list. Checklists can be filled out during or after watching an observation session. For example, a trained observer might indicate which of 10 possible steps in an intervention protocol the interventionist uses. The result often indicates the percentage of desired steps completed.

Count coding: Indicating the occurrence of each instance or each instance's duration as it occurs during an observation.

Expert rating scale: A method of quantifying observations that often involves an expert observer using Likert-like scales to record global judgments about the quality or quantity of a particular class of behaviors after watching the entire observation session.

Construct: A psychological concept or process that is not directly observable.

Observational checklist: A way to quantify observations involving the indication of the presence or absence of key behaviors from a provided list of behaviors.

Rating scales and checklists are covered in detail in other sources (Cairns, 1979; Primavera, Allison, & Alfonso, 1997) and are not explored in this book. Figure 1.1 illustrates the relation of systematic observation using count coding among these other options for quantifying observations.

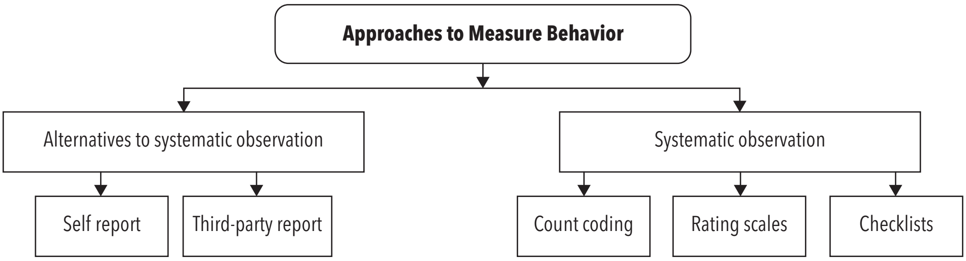


Figure 1.1. Illustration of how systematic observation using count coding (the focus of this text) is one of several approaches to measure behavior.

The Rationale for Systematic Observation Using Count Coding

There are three situations in which systematic observation might produce more scientifically useful scores than self report or third-party report. First, *systematic observations tend to be more accurate and therefore more valid than self report and third-party report when measuring the particular social and nonsocial contexts of behavior.* This advantage applies when the inferential goal is to relate the observed behavior, in part, to social and nonsocial contexts. For example, we may be interested in the behavioral antecedents or consequences of skillful student social initiations. Because exchanges in which the antecedent-behavior or behavior-consequence sequences often occur quickly, asking participants and others to note and report on such exchanges may not accurately capture the behavioral phenomenon of interest. In contrast, coding as it occurs can enable careful coding of the timing of contextual events relative to key behaviors.

Second, *systematic observations are often more valid than self report when the participant is preverbal or when cognitive impairments limit a person's ability to report on the behavioral phenomenon*. For example, nonverbal participants cannot use spoken language to self report on their interest in communicating for social reasons. In contrast, we can directly observe the frequency with which a participant uses behaviors that produce socially reinforcing consequences and are therefore inferred to have communicative function.

Third, *systematic observations are often more valid than self report and third-party reports of participant behavior when scores from those reports are affected by reporter characteristics*. For example, maternal reports of item-level vocabulary her children understand have been shown to reflect the mother's formal education level as well as characteristics of the participant (Yoder, Warren, & Biggar, 1997). The influence of reporter characteristics may explain, in part, why it is commonly found that different reporters often disagree in their responses concerning the same child (Smith, 2007). The training and highly specified coding system required for systematic observation using count coding can decrease the probability that scores reflect observer characteristics.

For the reasons described, systematic observation is potentially more useful than alternative methods in certain situations. In addition, count-coding measurement of systematic observations has four related advantages over the two other means of quantifying direct observations, rating scales and checklists. First, *count coding often provides a larger range of potential scores and more steps between values than* *do rating scales or checklists; these measurement properties, in turn, potentially provide a more sensitive measure of change or individual differences.* For example, the count of the number of communication acts from a 15-minute session might have a range of 0–100. In contrast a Likert-like rating of the amount of communication from the same session would likely have a smaller range of 0–7. A checklist record of whether communication occurred in the same session would have a still smaller range of 0–1.

Second, *compared with count coding, using Likert-like rating often demands that the investigator have more knowledge concerning the construct of interest.* Also, the concept being measured in rating scales is often more broad than those being measured by count coding. For example, suppose investigators wish to measure the construct "parent verbal responsivity." An instance of parent verbal responsivity, as measured by count coding, occurs when the parent vocalizes immediately after a target participant's vocalization (e.g., within 2 seconds) and in a way that is semantically related to it (e.g., puts into words the child's apparent referent). In contrast, a rater using a Likert-like method might rate his or her overall judgment of what the investigator defines as "sensitive, warm responsivity." Frequently, the rationale for using rating scales is that these scales attempt to measure concepts (or constructs) that are presumably more complex than those typically measured by count coding. However, the assumption that a rater is better able to quantify complex concepts than the count coder is based, at least in part, on the assumption that the rater has a deep understanding of the construct of interest. In contrast, the count coder might only have to apply a series of yes–no decisions, based on more specifically defined concepts than the rater uses. To put it another, more colloquial way, the difference between the approaches is "you'll know it when you see it" versus "count it and you'll know it."

Third, *compared with designers of Likert-like rating scales, designers of count-coding systems need not make as many arbitrary decisions regarding the amount of the variable needed to increment the variable score.* That is, for Likert values, the investigator must provide detailed descriptions or behavioral anchors. For example, how might the investigator decide the meaning of the behavioral anchor *most of the time* versus *almost always* when rating parental responsivity? Should the criterion dividing the two be 75% of *opportunities* or 75% of *time observed*? Or should the numerical criterion be 90% instead of 75%? Ideally, theory would guide these decisions, but usually this level of specificity is lacking.

Finally, *because count coding enables a greater level of specificity, it usually allows a more rigorous definition of interobserver agreement than is typically used in research relying on Likert-like rating.* Researchers using count coding can evaluate point-by-point agreement (i.e., agreement occurs if both observers see the same thing at the same time in the session). In contrast, researchers using Likert-like rating often consider observer ratings within 1 point as agreement. The latter is particularly problematic in light of the well-known tendency of observers to use a limited range on rating scales. For example, raters typically do not use the extreme negative value. If the rating scale involves 1–5, raters not using "1" will result in an actual range of 2–5. The result is that Likert-like rating, at an item level, produces a greater probability of appearing to achieve agreement through chance processes than does count coding.

Table 1.1. Attributes of systematic observation using count coding, compared to alternative measurement methods

Method	No. of sessions on which scores are based	Level of description of phenomenon of interest	Timing of recording judgment relative to observation	Typical amount of observer/reporter training	Level of memory demand on observer/reporter	Size of possible range of scores
Systematic observation						
Count coding	Fewer than reports	Micro	As it occurs	High	Low	Large
Rating	Fewer than reports	Macro	After session	High	Medium	Small
Checklist	Fewer than reports	Macro	Either	Low	Low	Small
Reports						
Self	More than observation	Either	Retrospective	None	High	Large
Other	More than observation	Either	Retrospective	None or low	High	Large

Despite the advantages of systematic observation using count coding, this method has some disadvantages. It must be said that count coding systems tend to require more time to implement than alternative methods, including self and third-party reports, rating scales, and checklists. Therefore, the precision gained by count coding comes with a cost in resources such as personnel time and training time. Furthermore, systematic observation is usually applied to a limited number of observations. In contrast, other and self reports are usually based on memory of many more observations. Table 1.1 summarizes the distinctions between systematic observation using count coding and the other measurement methods we have discussed, as well as the advantages of count coding relative to those methods.

THE IMPORTANCE OF FALSIFIABLE RESEARCH QUESTIONS OR HYPOTHESES

Systematic observation using count coding is particularly well-suited to testing very specific and highly falsifiable predictions. We call these predictions *falsifiable hypotheses*. The syntax used to formulate the hypothesis—that is, whether it is a statement or a question—is not important. What is important is that the statement specifies these elements: 1) the dependent and independent variables; 2) the investigator's expectations of an association, a difference, or a functional relation; and 3) the investigator's expectations regarding direction of the association (e.g., a positive one) or difference (e.g., the mean, trend, or variability of the experimental group [or phase] is greater than the contrast).

The more specific the hypothesis, the more guidance it will provide when designing the measurement system used to assess the independent and/or dependent variables. Creating such falsifiable hypotheses is important because findings that confirm very specific predictions are more likely to replicate than are findings that confirm vaguely stated predictions. This is not magic. When extant data and theory that support such specificity are sufficiently developed to generate confirmation, this suggests a field that is relatively mature. Falsifiable hypotheses are much easier to disconfirm than they are to confirm. There are many explanations for disconfirmations (e.g., poor design or measurement) and few explanations for confirmations (i.e., a scientifically useful motivating theory). This is a simplification of the positivist philosophy of science.

This book assumes that readers understand falsifiable hypotheses and are able to formulate them. If formulating a falsifiable hypothesis is not possible, research questions should be specified as theory and current knowledge allow. Less-specified research questions should be labeled as exploratory, and results of research examining such questions should be seen as hypothesis generating. The way we quantify the independent and dependent variables in these falsifiable hypotheses or research questions should be determined, in part, by the type of phenomenon we want to measure (i.e., object of measurement). The different types of objects of measurement are addressed in the next section.

Falsifiable research question: A prediction or question that specifies 1) the dependent and independent variables, 2) the investigator's expectations of an association or a difference, and 3) the investigator's expectations regarding direction of the association or difference prior to analyzing the data.

OBJECTS OF MEASUREMENT: THE CONTINUUM OF CONTEXT-DEPENDENT BEHAVIORS TO GENERALIZED PERSON CHARACTERISTICS

When investigators measure a person's behavior, the assumed or underlying phenomenon being measured (the object of measurement) may be transient and context dependent; it may be a *stable*, generalized characteristic of the person; or it may be something between the two. Prototypical context-dependent behavior changes are temporary, brief, and influenced by external circumstances; prototypical generalized person characteristics are stable, long-lasting, and influenced by internal variables (Chaplin, John, & Goldberg, 1988). The two extremes—context-dependent behavior and generalized characteristic—can be thought of as the two extreme ends of a continuum. Any observational variable exists somewhere along the continuum representing the extent to which the behavior is transient and context dependent. One of the most important decisions an investigator of a new study or reader of an extant study should make is where the observational variable as it is measured is located along this continuum.

In fact, most observational variables lie somewhere on a continuum between these prototypical extremes. However, understanding the extremes helps us place our object of measurement on this continuum. In this book, we attempt to show how understanding the variable of interest's location on the continuum should influence our decisions and interpretations. The following sections discuss in greater depth the terms *context-dependent behaviors* and *generalized person characteristics* as they apply to observational variables.

Stable: Rankings of participants' levels of a person characteristic are similar across ways or times of measuring the characteristic.

Context-Dependent Behaviors

Context-dependent behaviors are those that vary in number or duration due to eliciting or inhibiting attributes of the measurement context. The behavior is studied to learn about the environment's influence on the behavior. For example, suppose an investigator is interested in knowing whether visual reminders to attend to the teacher result in young children engaging in the teaching activity; these visual reminders might include items such as an illustration of children sitting on a carpet square and looking at the teacher in a small-group context. To study this question, the investigator measures children's instructional engagement with and without visual reminders present. The presence/absence of visual reminders could be manipulated in a variety of ways using different design approaches (single-case experimental design, within-group experimental designs).

Regardless of design type, participants experience both measurement conditions. It is important to note that the sequence of experiencing the conditions is counterbalanced or randomized across participants. Suppose that, regardless of sequence, between-condition difference in instructional engagement occurs;

that is, children are more engaged with the activity when the visual reminder is present, regardless of whether they experience this condition first or second. If this happens, it clearly signals the child's engagement is a context-dependent behavior. Within-child changes cannot explain such between-condition differences because order is counterbalanced, no sequence effects occur, and the time between conditions is brief. That is, the occurrence of the behavior is tied to or bound to the context. Without the particular contextual details, in this case a carpet square, the child is not likely to engage in the teaching activity. If these experiments are conceptualized as treatment studies, the studies would not test eventual generalization of instructional engagement to contexts in which visual reminders are absent, and this would not be of potential interest. Instead, the emphasis is on the aspect of the measurement context thought to influence occurrence or duration of the key behavior in the short run: visual reminders. The focus is on aspects of the environment that influence the context-dependent behaviors.

Measuring context-dependent behaviors requires a low *level of inference. Inference level* refers to the number of assumptions and level of evidence on which to base sound interpretations of the observational variable scores. This concept will be discussed more later in this chapter.

Context-dependent behaviors: Those that vary in number or duration because of eliciting or inhibiting attributes of the measurement context.

Inference level: The number of assumptions and level of evidence on which to base sound interpretations of the observational variable scores.

Generalized Person Characteristics

We should measure the observational variable as a *person characteristic* when we test the following:

- Whether variance in a characteristic measured by systematic observation predicts future variance on an outcome or differs between intact groups (e.g., children with intellectual disability versus typically developing children)
- Whether effects of a treatment generalize from the treatment sessions to measurement contexts that differ from the treatment sessions on multiple dimensions simultaneously.

In the former case, we say that a group of individuals *has* a certain person characteristic. In the latter case, we are saying that the person has *changed* in the degree to which he or she exhibits evidence of the person characteristic. The phenomenon of interest is considered intrinsic to the participant rather than the measurement context; that is, the locus of influence is primarily the person, not the environment. One distinguishing feature of person characteristics, as opposed to context-dependent behaviors, is that measures of the former are estimates of what occurs outside a particular measurement context. Thus, we would expect to see evidence of the phenomenon in all valid measurement contexts.

Because we cannot practically collect all valid measures, we compromise by looking for measures with scores that are stable across ways or times of measuring the characteristic, with the term *stable* (as used in this book) meaning that rankings of participants' levels of a person characteristic are similar across ways or times of measuring it. For example, assume a person characteristic is measured in two observations in 10 people. If that measure is stable, then the scores for the first observation would be highly positively correlated with the scores in the second observation. Because this conception of stability inherently involves the relative rankings of participants across contexts, it is distinct from how single-subject researchers use this term (i.e., steady-state responding) (Sidman, 1960; Johnston & Pennypacker, 2009).

Some person characteristics are *constructs* (i.e., psychological concepts or processes). That is, 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. For example, we might observe children interacting with an examiner using a well-defined protocol and use this observation to infer the relative level of language or social ability among the children. There are two types of person characteristics that differ by the level of inference needed to interpret them accurately: 1) generalized behavioral tendencies and 2) skills.

Person characteristics: A person's stable, long-lasting characteristics that are presumed to be influenced primarily by internal variables.

Generalized Behavioral Tendencies *Generalized behavioral tendencies* are descriptors of what people usually do. As such, they are typically measured in the natural environment and are expected to be stable across valid measurement contexts. An example of a generalized behavioral tendency is loquaciousness. When we say that individuals are loquacious, we mean they exhibit high levels of talk relative to other individuals. Alternatively, when we say that a group of children is now more loquacious than in the past, we mean the children generally talk more than they used to. If the way we measure loquaciousness is, in fact, a generalized tendency to talk, we expect rankings of loquaciousness to be similar regardless of the valid measurement context we use to assess amount of talking. Because generalized tendencies to act in a certain way are intrinsically about what occurs in the natural environment, we acknowledge that the environment in which the behavior is measured is relevant. But the expectation is still that these objects of measurement represent within-person characteristics more than the contexts in which they are measured. The level of inference needed to interpret generalized behavioral tendencies is greater than needed for interpreting context-dependent behaviors but less than needed for interpreting skills.

Generalized behavioral tendency: Descriptor of what people usually do.

Table 1.2. Attributions of objects of measurement

Object of measurement	Locus of influence	Degree of control provided by setting of observation	Level of inference needed to interpret the variable
Context-dependent behavior	Environment	High	Low
Generalized behavioral tendency	Mostly person	Low	Moderate
Skill	Person	Either	High

Skills *Skills* are constructs that we call abilities or developmental achievements. Here, the term *skill* refers to a highly generalized ability that can be and is used in a wide variety of contexts, regardless of level of prompting from the environment. Examples of skills include language and reading. Even more than for generalized behavioral tendencies, variation in skill measures is thought to occur because of differences intrinsic to participants (e.g., IQ), not the environment in which skills are measured. Because variation in skills is thought to rely less on the environment in which they are assessed, and because skills represent constructs, the level of inference in accurately interpreting skill measures is high. It is higher than that of both context-bound behaviors and generalized behavioral tendencies. Table 1.2 indicates the different attributes of the various objects of measurement.

Skill: What a person does in a situation in which the effect of the context is made irrelevant by using a structured measurement context.

As shown in Table 1.2, context-dependent behavior measurement is usually conducted in studies in which the primary interest is environmental influence on the behavior. In contrast, person characteristics are usually measured in studies in which the primary interest is characteristics of people. However, in many studies, investigators want to interpret their observational variables as reflecting both environmental and within-person influences. This is where it becomes difficult to accurately place the object of measurement along the continuum of context dependency to generalized person characteristics. Some types of variables and studies provide good examples of where nuanced classification of the object of measurement is required.

When the observational variable is clearly dyadic, as in many parent–child variables, the variable is best placed in the middle of the continuum. Logically, for the predicted difference or association to replicate, contextual stability would have to occur. However, the nature of the variable is intrinsically about the parent (an aspect of the social environment) and the child (e.g., not all children will show the behavior when the parent interacts optimally).

Treatment studies also provide a good example of the complicating issues. In treatment studies, the treatment (an environmental influence) and change in participants' behavior are both important. However, two factors should determine the placement of the observational measure of the participants' behavior on the continuum.

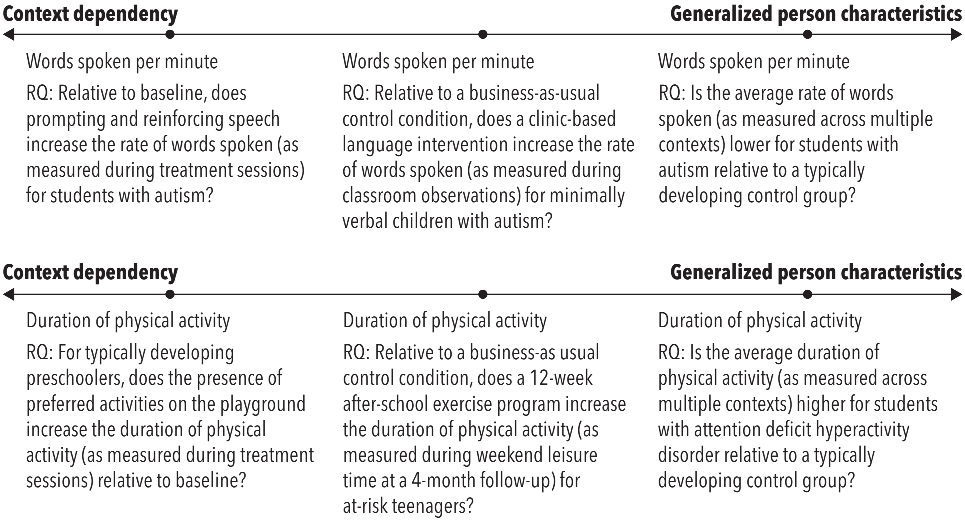
First, the degree to which behavior change reverses when the treatment is withdrawn should influence how we interpret the observational measure. If reversal is tested and observed, the object of measurement is clearly context dependent. But if reversal is not observed—either because it did not occur or because it was not tested—the object of measurement is probably best considered *potentially context dependent.* There is value in placing the object of measurement between the midpoint of the continuum and the end point marked context dependent.

The second factor is the degree to which behavior change as a function of treatment is shown to be highly generalized. This should influence how we interpret the object of measurement. Within a treatment study, in the context of an internally valid research design, an observational dependent variable can be considered in the middle of the continuum if behavior change is shown not only in the treatment session but also in a measurement context that differs from the treatment session on all primary dimensions that might restrict the generalized use of the behavior. This is known as *far transfer*. For example, measurement contexts for a behavior may differ in location, activity, materials, interaction style, or person with whom the participant interacts. The behavior is therefore considered *malleable* (i.e., influenced by the environment). The behavior also appears to represent characteristics of a person in the sense that the behavior change is stable across treatment and the far transfer generalized measurement context. The degree to which the characteristic is placed near the generalized person characteristic end of the continuum should be influenced by how much intervention was needed to produce the far transfer.

Far transfer: Behavior change that is shown to occur in a measurement context that differs from the treatment session on all primary dimensions that might restrict the generalized use of the behavior.

Malleable: Used to describe a generalized person characteristic that is influenced by the environment.

The same behavior or set of behaviors can be measured as a context-dependent behavior in one study and a person characteristic in another study. An example is the amount of talking a child does. Talking may be measured as a context-dependent behavior when an intervention study shows that prompting and reinforcing a child for talking helps the child do so only during the treatment sessions. In this instance, we identified talking as a potentially context-dependent behavior because generalization was not tested or shown. Now, suppose a test of far transfer showed that the behavior change, more talking, generalized to measurement conditions that differed from the treatment session on all major dimensions of generalization. In that instance, we would conclude that the amount of talking represented a characteristic in the center of the continuum. Similarly, suppose the amount of talking predicted reading or was different between intact groups, such as children with cognitive impairment versus those who are typically developing. In that instance, we would position the amount of talking near the generalized



Context dependency **Generalized person characteristics**

Words spoken per minute

RQ: Relative to baseline, does prompting and reinforcing speech increase the rate of words spoken (as measured during treatment sessions) for students with autism?

Words spoken per minute

RQ: Relative to a business-as-usual control condition, does a clinic-based language intervention increase the rate of words spoken (as measured during classroom observations) for minimally verbal children with autism?

Words spoken per minute

RQ: Is the average rate of words spoken (as measured across multiple contexts) lower for students with autism relative to a typically developing control group?

Context dependency **Generalized person characteristics**

Duration of physical activity

RQ: For typically developing preschoolers, does the presence of preferred activities on the playground increase the duration of physical activity (as measured during treatment sessions) relative to baseline?

Duration of physical activity

RQ: Relative to a business-as usual control condition, does a 12-week after-school exercise program increase the duration of physical activity (as measured during weekend leisure time at a 4-month follow-up) for at-risk teenagers?

Duration of physical activity

RQ: Is the average duration of physical activity (as measured across multiple contexts) higher for students with attention deficit hyperactivity disorder relative to a typically developing control group?

Figure 1.2. Examples of how the same behavior can potentially be a context dependent and a generalized characteristic, depending on how it is studied.

person characteristic end of the continuum. Figure 1.2 provides a visual representation of how the same behaviors can be placed at different points along the continuum, depending how the behavior is studied and what the research question and research design indicate it is supposed to represent.

Once the investigator has determined, or at least estimated, the location of an observational variable he or she wishes to measure on the context-dependent-to-generalized person characteristic continuum, he or she can evaluate the relative value of alternative ways to measure the phenomenon of interest. That is, the criteria by which one judges alternative ways to measure the phenomenon of interest should be informed by the phenomenon's placement on the continuum.

JUDGING THE RELATIVE SCIENTIFIC VALUE OF DIFFERENT MEASURES

When we say that we want the best measure of something, we are referring to the concept of scientific utility. Scientific utility has two components: reliability and validity. Although the topics of reliability and validity will be covered in more detail in later chapters, it is necessary to introduce them here to illustrate why it is so important to identify our object of measurement.

Reliability

Reliability is the degree to which a measure is consistent with another measure of the same thing. The most relevant types of reliability to observational measurement are 1) interobserver agreement and 2) stability of scores (in the group-design sense of the term). The first of these is widely understood and is discussed in detail in Chapter 8. Here we introduce the concept of stability because it is underreported for observational variables, despite its importance.

There are two types of stability that are relevant to observational measurement: *contextual stability* and *temporal stability*. A contextually stable measure ranks participants' scores of the person characteristic similarly across valid measurement contexts. For example, consider what is meant by a contextually stable measure of loquaciousness. A long interaction session is judged to produce this contextually stable measure of loquaciousness when the degree of similarity is high (e.g., .80) in ranked scores of 10 participants' number of verbal utterances across structured versus unstructured interactions. That is, loquaciousness remains stable even when the context varies in its degree of structure. When referring to contextual stability, we expect stability across contexts that realistically evoke the key behaviors and not just any possible context. We would not expect a count of aggressive acts from the playground to be stable with a count of aggressive acts in the movie theater. Context variables present in a movie theater may inhibit aggression, whereas those on the playground may evoke aggression. A *temporally stable* measure ranks participants' scores from the same measurement context similarly across two or more testings. In this context, the length of interval between testings is expected to be short. For example, a procedure with a well-defined protocol is judged to produce a more temporally stable measure of vocabulary diversity if the degree of similarity is high (e.g., .8) in ranked scores for 10 participants' number of different words used on Monday versus Tuesday.

Although we have used the term "high" in our examples, there is no threshold level of stability one must achieve for variables to be acceptable. It is the relative stability of measures that enables us to select among alternatives. The measure with the greater stability tends to be more scientifically useful, all other things being equal.

Reliability: The degree to which a measure is consistent with another measure of the same thing.

Contextual stability: The degree to which a measure ranks participants' scores of the person characteristic similarly across valid measurement contexts.

Temporal stability: The degree to which a measure ranks a group of participants' scores from the same measurement context similarly across two or more testings.

Validity

Validity is the degree to which a measure represents what we believe it represents. To put it a slightly different way, a measure's validity exists in regard to the types of evidence that support warranted inferences from the measure in relation to a given purpose or construct. Three types of validity and corresponding types of validity evidence to support an inference are briefly discussed here: *content validity, sensitivity to change,* and *construct validity.* These apply to observational measurement as follows:

- *Content validity* (also commonly referred to as *content validation*) is the extent to which experts agree that the definitions used to code the observation session conform to known information and beliefs about what the variable label means. (For example, if we say we are measuring "aggression," experts should agree that the behaviors considered evidence of aggression in the coding manual are examples of aggression.)

- *Sensitivity to change* is the extent to which a measure changes with intervention.
- *Construct validity* (also commonly referred to as *construct validation*) is the degree to which a measure produces a pattern of correlations or group differences that are predicted by theory.

We judge the relative scientific utility of observational variables by different types of reliability and validity criteria depending on where our variable is located on the continuum of context-dependent behavior-to-generalized person characteristic. For context-dependent variables, relative scientific utility is based on interobserver agreement, content validity, and sensitivity to change. For skills, relative scientific utility is based on temporal stability and construct validity. Because measuring context-dependent behavior does not require scores to be stable across context or time, there is more flexibility about where and in how many sessions to obtain measures. Because measuring skills requires an inference about a specific construct, there is a greater need to measure in contexts that control for contextual variables that might vary across participants and contexts. Thus, skills are often measured in a more controlled setting than is possible within the home or community, using procedures that control contextual variables that influence scores. For this reason, one needs to average across relatively few procedures to yield temporally stable scores. (Measuring generalized behavioral tendencies presents special challenges that will be addressed in the next section on ecological validity.)

Validity: The degree to which evidence and theory support the interpretations of observational variable scores as measuring a particular construct or concept in a particular population.

Content validation: As applied to a coding manual, its most frequent object of validation, this 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.

Sensitivity to change: As a validation concept, this is the degree to which a measure changes in a therapeutic direction after participation in treatment.

Construct validation: A cumulative process by which empirical studies test whether particular measurement systems yield variables that perform as expected by theory and logic.

Ecological Validity

Generalized behavioral tendencies present a special case that highlights the importance of two concepts: *ecological validity* and *representativeness* (defined in the next section). *Ecological validity* has been used to refer to the extent to which measurement contexts resemble or take place in naturally occurring (unmanipulated) and frequently experienced contexts (Brooks & Baumeister, 1977). We use the term *naturalistic* to refer to contexts that are familiar to the participant and *contrived* to refer to contexts that are unfamiliar to the participant and are often set up by the researcher. There is a legitimate societal need to know the extent to which participants use key behaviors in uncontrolled conditions that the individual frequently experiences (Brooks & Baumeister, 1977). Generalized behavioral tendencies are measured in ecologically valid contexts. *Ecologically valid* is a descriptor of a procedure and the variables that it generates; however, it is not synonymous with representativeness.

Ecological validity: The extent to which measurement contexts resemble or take place in naturally occurring (unmanipulated) and frequently experienced contexts.

Naturalistic: Used to describe contexts that are familiar to the participant.

Contrived: Used to describe procedural contexts that are unfamiliar to the participant and are often set up by the researcher.

Representativeness

The lay definition of the word *representative* differs from that used in measurement theory. The lay definition is "typical" or "usual" (*Shorter Oxford English dictionary*, 2002). However, a single ecologically valid measurement context rarely produces scores on an observational variable that are similar to those produced by other ecologically valid measurement contexts. This lack of reliability for observational variable scores from multiple ecologically valid measurement contexts is problematic in the scientific realm. The complex relation between the scientific concept of representativeness and ecological validity will be discussed in detail in Chapter 3.

When applied to generalized behavioral tendencies, classical measurement theory defines the term *representativeness* to mean the degree of similarity of the observational variable scores to that derived from averaging all valid measures of the generalized behavioral tendency (Cronbach, 1972). We cannot examine any phenomenon in *all* valid contexts. Thus, classical measurement theory asserts that the within-person average across as many ecologically valid measurement contexts as possible is the best estimate of "what a person usually does" (Crocker & Algina, 1986; Cronbach, 1972).

When applied to group design logic, a measure is more representative than another if it is more contextually stable. When applied to single-case design logic, a measure is more representative if it is more similar to the within-person, across-multiple-procedure mean of the generalized behavioral tendency. An example of the single-case design concept of representativeness is as follows: The within-person mean of on-task behavior was computed from ten 15-minute observations of small-group activities made across 5 days and was found to be 15% of the total observed time. An observation in the first 15-minute small-group lesson (i.e., 20% of the observation) was judged to be more representative than the tenth 15-minute small-group lesson (i.e., 5% of the observation) because the former is closer to the estimate based on all available observation (i.e., 20% is closer to 15% than is 5%).

Many particular naturalistic contexts vary greatly among participants and over time, and such variation could cause scores to be ranked differently across naturalistic observations. For this reason, single naturalistic contexts are unlikely to produce observational variable scores that are representative in the scientific sense of the word. Thus, there is a tension between the need for measures of generalized behavioral tendencies to be both ecologically valid and representative.

Good observational measurement studies address this tension by averaging scores within participants and across multiple ecologically valid measures that differ in how much they control for influential contextual variables. The theory behind this practice is that some of these procedures will underestimate and others will overestimate the most representative score. Averaging scores across underestimating and overestimating procedures is thought to cancel out the direction of measurement error, thereby producing a mean that is closer to the most representative score than any one procedure would produce (Cronbach, 1972). This point will be addressed further in Chapter 3. The number of contexts that one needs to average across is judged by the number needed to generate a contextually stable measure. In Chapter 11, we address the method used to determine the number of contexts needed to yield this criterion level of contextual stability.

Representativeness: For single-case researchers, the concept of representativeness has been operationalized as proximity of the score in question to the score from a very long observation that occurs across many measurement contexts. In a group research design context, the representativeness is operationalized as contextual stability.

Table 1.3 summarizes the criteria by which we judge the relative scientific value of the three objects of measurement and recommendations related to setting and number of sessions across which to average.

Table 1.3. Summary of criteria for evaluating relative scientific value for each object of measurement

Object of measurement	Type of reliability	Type of validity	Type of measurement setting	Number of sessions needed to average across to yield scientifically useful attributes
Context-dependent behavior	Point-by-point interobserver agreement	Content and sensitivity to change	Naturalistic or contrived	One
Generalized behavioral tendency	Contextual stability	Sometimes construct	Naturalistic	Many
Skill	Temporal stability	Construct	Naturalistic or contrived	One to few

CONCLUSIONS AND RECOMMENDATIONS

Despite our preference for observational measurement for many purposes, it is not everyone's preference when measuring generalized person characteristics. Now that we have covered the reason why single observations are often inadequate to reliably measure person characteristics—that is, the single observation may produce a score that is a poor estimate of the mean score—the rationale for using third-party reports (e.g., parent reports) when measuring generalized behavioral tendencies is strengthened. Specifically, third-party reports about the participant's behavioral tendencies potentially draw on a wide range of experiences with the participant. If reporters are able to synthesize across their experience with the participant while keeping their biases from influencing their report, then third-party reports can potentially produce valid estimates of person characteristics. Because sampling many observational sessions and averaging scores to produce a single estimate is expensive, and thus rare, many investigators prefer third-party reports over systematic observation when measuring person characteristics.

On the other hand, if parents or third-party reporters are not able to keep their biases from influencing their report of the participant's behavior, then using the average of many observation sessions may produce a more valid estimate of the person characteristic than will third-party reports or self reports. In addition, systematic observation will almost always be a more valid way to report on context-dependent behaviors than is a third-party report of the participant's behavior.

If one is not aware of the distinction between context-dependent behaviors and person characteristics, one might mistakenly overgeneralize and believe that systematic observation is *always* more valid than third-party report. Ultimately, the relative validity of third-party reports versus systematic observational measures of person characteristics is an empirical question. Furthermore, these empirical comparisons of relative validity will need to occur for each combination of population and person characteristic. This is arguably impractical. Therefore, for the foreseeable future, investigator preferences will surely affect the selection of systematic observation versus third-party report when measuring person characteristics. Others have written about the advantages and disadvantages of systematic observation versus third-party report methods for measuring person characteristics (Jacobson, 1985). One approach to this ongoing debate is to measure a person characteristic using multiple methods (e.g., both third-party report and observational measurement) and look for converging evidence across methods or aggregate them if they are correlated (Cook & Campbell, 1979).

A context-dependent variable may be measured in any context that evokes the key behavior. However, cultural values will often influence investigators to use naturally occurring contexts. If the design is negatively influenced by highly variable scores (e.g., single-case experimental designs relying on comparison of adjacent conditions), it is best to select a naturally occurring measurement context that tends to remain relatively constant across time and people on contextual variables that influence scores.

When measuring a generalized behavioral tendency, often the investigator is best served by averaging across many different types of naturalistic measurement contexts that evoke the key behavior. Doing so will produce a more contextually stable estimate of each participant's generalized behavioral tendency than would a single observation.

When measuring a skill, we recommend using one or more procedures that control for contextual variables that influence scores. Doing so often produces more temporally stable scores than do procedures that vary on influential contextual variables. Given there is no need to measure these objects of measurement in a way that is representative, it is not necessary to measure skills in the natural environment. Naturalistic measurement contexts often prevent control of influential contextual variables. Temporally stable scores are more likely to be construct valid than are unstable scores. However, generalization of results to other conditions is restricted due to the use of one type of procedure (e.g., structured). If it is desirable to measure skills in a representative way, then we recommend averaging across multiple naturalistic measurement contexts that tend to remain relatively constant across time and people on contextual variables that influence scores.

In this chapter, we defined what we mean by systematic direct observation using count coding. We also discussed the distinction between measuring a context-dependent behavior and measuring two types of person characteristics. The distinction is very important for proper framing and interpretation of a study and for many measurement decisions that affect the measures' psychometric properties.

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*, 407–416.

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

Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of states and traits: Dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology*, *54*(4), 541–557. doi:10.1037/0022-3514.54.4.541

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

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

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

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). *MacArthur-Bates Communicative Development Inventories (CDIs): User's guide and technical manual*. Baltimore, MD: Paul H. Brookes Publishing Co.

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

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.

Mullen, E. M. (1995). *Mullen scales of early learning* (pp. 58–64). Circle Pines, MN: AGS.

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: Lawrence Erlbaum Associates.

Schiele, B. C., Baker, A. B., & Hathaway, S. R. (1943). The Minnesota Multiphasic Personality Inventory. *Lancet, 63*, 292–297.

Shorter Oxford English dictionary (5th ed., Vol. 2). (2002). Oxford, United Kingdom: Author.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York, NY: Basic Books.

Smith, S. (2007). Making sense of multiple informants of child and adolescent psychopathology: A guide for clinicians. *Journal of Psychoeducational Assessment, 25*, 139–149. doi:10.1177/0734282906296233

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

Taylor, B. C., & Trujillo, N. (2001). Qualitative research methods. In F. M. Jablin & L. L. Putnam (Eds.), *The new handbook of organizational communication: Advances in theory, research, and methods* (pp. 161–194). Thousand Oaks, CA: Sage Publications.

Tracy, S. J. (2013). *Qualitative research methods*. London, United Kingdom: Wiley-Blackwell.

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. doi:10.1044/1058-0360.0601.59