16 Behavioral Observation and Coding

Katherine R. Thorson and Tessa West

How do people's emotions influence what they say when interacting with others? How are people's traits associated with their daily habits and health behaviors? How does a person's standing in a social hierarchy affect whether they help someone in need? At the heart of these questions – and many others – is the perennial quest to understand human behavior. Although the field of psychology largely focused on mental processes in the second half of the twentieth century (see Banaji, Chapter 1 in this volume), the drive to understand behavior has re-emerged in recent years, given the direct relevance of behaviors to everyday life and their clear implications for personal and societal well-being (Back et al., 2009; Baumeister et al., 2007; Furr, 2009; Hansen et al., 2022). In order for behavior to be studied, though, it must be measured well. In this chapter, we describe how to tackle the complex task of measuring behavior so that it can be scientifically examined. We introduce readers to opportunities, pitfalls, and best practices so that they can approach their questions about behavior with confidence and enthusiasm.

16.1 Chapter Overview

The goal of this chapter is to provide a practical overview of the most important steps of behavioral observation and coding, with a focus on how these processes are typically executed within social and personality psychology. We distinguish behavioral observation and coding from other methods of measuring behavior (such as automated coding done by computer programs). The chapter has several sections:

- *Definitions, strengths, and challenges.* We explain what we mean by behavioral observation and coding, and we outline strengths and challenges of this method.
- *Guiding principles.* We describe two guiding principles that apply throughout the process of observation and coding.
- *Aspects of observation and coding.* We highlight several aspects of observation and coding, many of which vary along a continuum, for researchers to consider (e.g., how much control do researchers want over the environment in which the behavior occurs? Are researchers interested in documenting whether a behavior is present or not or the strength with which it is displayed?).
- *Practical questions.* We discuss several practical questions regarding coding (e.g., how many coders are needed? How many items should each coder code?).
- *Analysis of behavioral data.* We describe the analysis of behavioral data – from establishing inter-rater agreement to running models with the coded behaviors as outcomes of interest.
- *Other topics and issues.* We discuss concerns related to automated processing of videos and text and topics related to the open-science movement (preregistration, transparency, and open data).

Of course, we cannot cover everything related to behavioral coding in one chapter. We aim to provide an overview and to focus on many of the "unspoken" or "unwritten" details that readers may be unlikely to find elsewhere. Throughout the chapter, we also point readers to more in-depth resources that are specifically dedicated to certain aspects of behavioral coding. We also include references to recently published studies that utilize behavioral coding, and we encourage readers to consult these papers to get a better sense of the range of behavioral coding that can be done and to see how behavioral coding is used to answer questions of interest to social and personality psychologists. Finally, there is nothing quite as useful as talking to other researchers who have successfully completed coding projects that are similar to your own, in terms of either the setting, the behaviors, the paradigms, or other features. We encourage researchers to seek help and get advice from more seasoned researchers whenever possible.

16.2 What Is Behavioral Observation?

We refer to "behavioral observation" as the process that occurs when a researcher sees or hears the actions of a person or multiple persons. (Of course, the behaviors of any organism can be observed, but in social and personality psychology, researchers are almost always focused on people.) We refer to "behavioral coding" as the process that occurs when other people systematically document those actions in quantitative form. We refer to this as "manual" coding, which stands in contrast to "automated" coding, which uses tools from computer science (e.g., NLP or natural-language processing) – rather than people – to quantify behaviors (Schmid Mast et al., 2015; see Ireland & Pennebaker, Chapter 14 in this volume; Schoedel & Mehl, Chapter 13 in this volume, for examples of this latter approach). Throughout the chapter, readers should assume that we are referring to manual coding that is done via visual or auditory observation, unless otherwise noted.

In addition, in this chapter, we focus on observation of direct behaviors and actions of a person, rather than the products or outcomes of a person's behaviors. For example, imagine studying how people interact with a new acquaintance. Two people talk to each other in the lab and complete a problem-solving task, and one person is given ten dollars to allocate between them. Direct observations of people include how much time they spend talking, how many questions they ask each other, and how friendly and anxious they appear (e.g., Bergsieker et al., 2010; Dumitru et al., 2022). Products or outcomes of people's behaviors include how many problems they solve correctly and how much money one person shares with another (Park et al., 2022; West et al., 2014). Because of the greater challenges and options inherent in direct observation of people, this chapter focuses on this aspect of behavioral observation. For people interested in measuring outcomes or products of behavior, we recommend consulting the relevant literature – for example, studies that use that particular measurement or paradigm – for best practices.

16.2.1 Strengths

Many theories and questions in social and personality psychology are directly concerned with people's behavior. How people behave with others, how behavior can be predicted by people's traits or life experiences, and how behaviors are tied to subjective experiences are perennial questions of interest to social and personality psychologists. Therefore, when researchers have questions about behavior, the primary strength of behavioral observation and coding is that it explicitly measures the variables which are of interest to researchers. Although prospective and retrospective self-reports about behavior can be informative, they are subject to numerous biases and, as such, do not show a one-to-one correspondence with real behavior (Gosling et al., 1998). Other methods, such as those used in neuroscience and psychophysiology, can also be informative, providing insight into whether people will behave in a particular way in the future or how they feel about particular behaviors or experiences, but they are, of course, not measures of observable behavior. Simply put: if a researcher's question is about behavior, then there really is no replacement for directly observing and measuring that behavior.

Another strength of behavioral observation is the relatively low level of interference required by researchers. Behavior can often be observed or recorded as it unfolds naturally "in real time," without researchers needing to interrupt the activities or tasks with which people are engaged. Thus the presence of the researcher can be minimized, reducing social desirability biases and experimenter effects. This can be particularly useful when researchers are studying social interactions and do not want to interrupt or intrude upon an ongoing dynamic between multiple people. Another strength is that behavior can be assessed with a high degree of temporal frequency; very often, researchers assess behaviors continuously throughout a study or, at least, at multiple timepoints. This method allows researchers to examine changes over time in how behavior unfolds and/or to obtain more stable estimates of typical behavior, by averaging across many observations. Lastly, because the initial recording of behaviors can often be done simply with a handheld video camera, behavioral data can be recorded in a wide range of settings, including those outside the lab.

16.2.2 Challenges

We see three primary challenges associated with behavioral observation and coding. First, behavioral coding is not immune to measurement difficulties. Although coding behavior may seem more "objective" than other measurements, like self-report, there are still many considerations researchers have to address to obtain reliable (e.g., would the same people judge this behavior in the same way?) and valid (e.g., is this a "real" or "true" measurement of this behavior?) measurements. For instance, imagine counting the number of times one person asks a question of another person (Thorson et al., 2019). At the outset, this seems simple. But what happens if the question gets interrupted? What happens if a question is framed as a statement, but with the intonation of asking a question ("That's right?"). Now take the more complicated case of a variable such as behavioral anxiety (West et al., 2017). What behaviors will you tell your coders to look for as indicators of behavioral anxiety? Will all coders be able to recognize these behaviors similarly? Are these behaviors "true" indicators of anxiety – will they correlate with subjective experiences of anxiety, for example? And might the behaviors that signal anxiety also signal other emotional experiences? In sum, just as with other methods, researchers using behavioral observation and coding have to address multiple challenges in order to obtain reliable and valid measurements.

Second, behavioral observation and coding are labor-intensive. Multiple research assistants must be hired, trained, and supervised throughout the coding process in order to ensure the reliability and generalizability of results. Therefore at least one person on the research team must have good project and people management skills, and the team must be able to fund or otherwise support (e.g., via course credit) several research assistants. With longer-term projects – for example, those which last for several years – significant changeover in the team of research assistants can add an extra layer of effort as new coders need to be continually added and trained.

Third, behavioral coding is time-intensive. Many trials are often needed to establish an effective coding scheme with high levels of *inter-rater reliability*, which is agreement between coders in their judgments of behaviors. Further, participants often need to be observed individually, and each person must be observed by multiple coders. The process of documenting codes can be slow, depending on the number of details required, and coders must take mental breaks from coding, meaning that they cannot spend a few weeks at the end of the semester "binge-coding" the data collected that semester.

Given these challenges, you may be hesitant to embark on a behavioral coding project. This is a reasonable concern, and you are not alone! However, in this chapter, we provide you with tools and guidelines that can streamline your process and help make this method easier, faster, and pain-free. We hope, at the end, that you'll feel that the benefits of this method are worth the time and effort involved.

16.3 Two Guiding Principles

Throughout this chapter, we emphasize two guiding principles for planning and implementing a research project with behavioral observation.

1 *Lead with your conceptual question*. It can be easy to get lost in the practical details of a large behavioral coding project and forget the theoretical questions you were interested in to begin with. Particularly for graduate students and early-career researchers who must be productive on a tight timeline, sometimes the feasibility of behavioral coding can seem overwhelming. As much as you can, try to start with a clearly outlined question or hypothesis that you want to test. Try not to build your research question around behaviors that are easy to record or collect – just because answering a question is feasible does not mean it is interesting or important. On the flip side, do not conduct a project just because it involves rigorous, thorough documentation of behavior. This approach will not necessarily yield answers to interesting or important questions. For instance, documenting a person's behaviors throughout an entire workday is probably not that interesting if the person is sitting at a computer writing a paper the whole time. As you think through the considerations outlined in this chapter, it is useful to keep returning to your conceptual question to guide your decision-making.

2 *Behavioral observation and coding are iterative processes*. You can and should adjust your conceptual questions, as well as your observation and coding processes, as you move through your research project. You should expect that you will make starts and stops within various stages of your project, as you refine and revise your procedures. For example, imagine that you are interested in bias toward racially minoritized people during conversations. You set out to code White people's use of overtly racist language during cross-race conversations. But then you observe a few conversations and find that use of such language is actually rare. In this case, perhaps you shift your coding to focus more on subtle expressions of bias, like appearing tense, uncomfortable, or avoiding eye contact (Dovidio et al., 1997). Of course, it is useful to do extensive piloting and pretesting before committing to specific data collection and coding procedures, but these are not the only times when revision is useful. For example, you may also find that you revise your overall coding plan after collecting your data and after your first "coding pass" of these data. For example, perhaps you initially intended on one coding pass to assess five variables, but after coding the whole data set, you decide to do a second coding pass as well – maybe to clarify the data obtained so far or to answer an extension of your original conceptual question. In sum, there are several key points in your observation and coding processes when it is useful to pause and consider whether any aspect needs revision or extension, and we will highlight these throughout the chapter.

16.4 Aspects to Consider When Observing and Coding Behavior

Below we highlight common aspects of observation and assessment of behavior, many of which vary along a continuum. Understanding these aspects, as well as the kinds of behavioral observation that exist, can help you clarify your research question and select an optimal method for examining it.

16.4.1 Live versus Recorded

Both observation and coding can occur “live” – with researchers watching and recording behaviors in real time, as they occur – or they can occur later, with researchers observing and making judgments from some documentation of those live actions, usually via videotape, audio recording, or potentially a text transcript. Recordings afford flexibility: they can be observed at any time and by anyone on the lab team, at the same time or separately. Recordings can be viewed repeatedly to optimize accuracy. If there are questions about how a behavior should be coded, multiple members of a team may view the same trial and discuss how to treat that behavior in upcoming, future trials. If codes require temporal precision, recordings can be slowed down (e.g., to capture particular facial expressions or to code behaviors that are occur quickly, such as fidgeting). Recordings can also be used for secondary data analysis, with new research questions and a new research team. Researchers do not even need to be present when recordings are collected: automatic recording devices, as well as devices that are activated by participants and automatically sent to researchers (e.g., webcam data from online studies) can be used. Lastly, advances over the past few decades have also made previous logistical concerns about recording largely obsolete: it is now possible to easily transport recording equipment, record participants unobtrusively, and securely store large amounts of data.

In our view, the main reason not to use recordings regards participant consent. (Modern video cameras are so small and unobtrusive that you need not worry about whether you can actually record behaviors in certain contexts; in almost all cases we can think of you will be able to find some recording solution.) Sometimes, you may find it possible to obtain consent for observation but not for recording. This is particularly true for vulnerable populations (e.g., children) or within sensitive contexts (e.g., doctor–patient interactions) and can also occur if recordings capture additional information that participants do not want documented (e.g., home recordings from webcams may include information about people’s living spaces that they would rather keep private). Lastly, you may choose not to record participants’ behavior because the recordings may require consent but the observations do not – for example, a field experiment where you observe people’s behaviors in public may not require obtaining consent from each participant (see Crandall, Giner-Sorolla, & Biernat, Chapter 2 in this volume). Guidelines vary across institutions, but, often, if you are observing behavior in a public space which is not expected to be private (e.g., holding the door open for others when going in and out of a coffee shop, and you are not capturing information that could be harmful to participants, you will not need to obtain consent from participants to be in the study. In these situations, it may make sense not to record participants’ behavior as that would require obtaining consent for the study, which you do not otherwise need.

16.4.2 Researcher Control over the Setting

Behavioral observation occurs in a variety of environments, ranging from those that are tightly controlled by the researcher to completely naturalistic contexts. Indeed, one of the benefits of studying behavior is that researchers can choose to exert more or less control over the setting, depending on their interest. This is in contrast to other methods, for example, which occur over such long periods of time (e.g., daily diary studies) or are so reliant on large, expensive equipment (e.g., fMRI research) that they can only be done in “real life” or lab environments respectively.

16.4.2.1 Naturalistic Observation

With naturalistic observation, researchers observe what occurs as naturally as possible, attempting to exert no influence over participants or the setting. This approach is used “in the field,” in the real environments where people live their lives. Thus the primary benefits are ecological validity and relevance (Maner, 2016; Paluck & Cialdini, 2014). Researchers need not argue or consider whether their results would apply to or have practical value for what occurs in the "real world" for "real people" (i.e., not just college students) as they are actually observing these processes.

Because this method requires that researchers observe people's behavior and wait for the behaviors of interest to occur, this approach can be quite time-consuming. For example, imagine you study workplace conversations at a company where employees work by themselves for 90 percent of the day. You would need to wait until the conversations naturally occurred on their own. Modern approaches can sometimes circumvent this problem with technology that automatically records people's behaviors – such as their language, location, and sleeping habits (Harari et al., 2016; Mehl, 2017; see Schoedel & Mehl, Chapter 13 in this volume). For example, to understand whether sleep contributes to relationship satisfaction in parents of young children, researchers measured sleeping behavior in parents by asking the parents to wear actigraphs on their wrists on a nightly basis (Härdelin et al., 2021). To understand how people's behaviors changed because of the COVID-19 pandemic, researchers used smartphone sensing apps to document changes in college students' behaviors – like their amount of physical activity and the number of physical locations they visited – in the weeks before and after the pandemic outbreak (Huckins et al., 2020). In both of these examples, behaviors are measured in their natural contexts – that is, the location in which those behaviors typically occur. One downside is that these methods can produce a substantial amount of data (e.g., many minutes of sleep per night per person) and so researchers may choose to code only random or specific segments of all recordings, or might aggregate information over time (e.g., examining total sleep time per night).

16.4.2.2 Quasi-naturalistic Observations

With quasi-naturalistic observation, researchers attempt to preserve ecological validity while placing some constraints that improve the quality and efficiency of data collection. When examining family interactions, for example, researchers might observe what goes on in the home for a two-hour period, while specifying that no one should use a device (television, phone, iPad, and so on) or leave the home during that time (Patterson, 1982). These instructions involve minimal researcher intervention, but they increase the amount of time when family members actually interact with each other, meaning that researchers are able to collect more data on family interactions than if they provided no instructions for family members at all. In addition, with quasi-naturalistic observation, researchers might not provide instructions to participants about how to behave but might instead constrain aspects of the setting. For instance, when studying preschoolers' math interest, researchers invited children to play with an educational math toy and then walked away, allowing children to engage with the toy or not. The researchers then documented the amount of time children played with the toy and coded their engagement, eagerness, and positive affect while doing so (Fisher et al., 2012). By constraining the situation – in other words, by providing a specific math toy to children – researchers were able to collect more data on behavioral math interest than if they had provided many different toys (some math-related, some not) or no toys at all.

16.4.2.3 Analog Observation

With analog observation, researchers use more control to construct situations of the kind in which the behaviors of interest occur. These are typically done in the lab and are generally designed so that researchers can efficiently elicit the behaviors in which they are interested. Within social and personality psychology, analog observations are often used when studying relationship dynamics. When studying couple conflict, researchers might ask couples to identify topics on which they disagree, and then ask them to discuss one of those issues for a specified amount of time (e.g., Gordon & Chen, 2016; Heyman et al., 2022; Kiecolt-Glaser et al., 1993; Williamson et al., 2013; Winczewski et al., 2016). Social support and friendship formation are also commonly studied through analog observation (Thorson et al., 2021; Zee & Bolger, 2022).

16.4.2.4 Experimental Manipulation

Researchers can exert more control by experimentally manipulating factors to examine how they influence the behaviors of interest. These designs are useful in that they allow researchers to hold all aspects of the environment constant and vary just the ones in which they are interested (see Smith, Chapter 7 of this volume). Experimental manipulations can occur in field or lab settings. In studies of interpersonal behavior, researchers may even control the interaction partner – meaning that researchers study social behavior by observing a participant interact with a research "confederate" – someone who is part of the research team but is pretending to be another participant. By doing this, the research team can hold visible characteristics (e.g., race or gender; Karremans & Verwijmeren, 2008; Mendes & Koslov, 2013) or behaviors (e.g., expression of positive versus neutral emotion or making supportive versus critical statements; Nils & Rimé, 2012; Yilmaz, 2016) of the interaction partner (the confederate) constant or they can systematically vary those characteristics or behaviors and observe their influence on participants. Some work has even used virtual-reality environments, where people interact with "virtual confederates," as a way to better understand social behavior (e.g., Rapuano et al., 2021).

A drawback of experiments within controlled laboratory environments is limited ecological validity. Researchers may create an environment that people are unlikely to encounter in real life (e.g., a room full of smoke), and so the degree to which a manipulated factor predicts a specific behavior in real life may be weak (because people rarely encounter smoky rooms). However, the goal here is to create situations that mirror real-life situations: for instance, the world might not be full of smoky rooms, but it is full of emergencies.

With confederate studies, typically only a handful of confederates are used, and so there is a danger that the interpersonal behaviors elicited by interacting with these few people will not extend to interacting with others more generally. In addition, confederates may not behave consistently across interaction partners. Their subtle nonverbal behaviors, for example, might vary depending on the behaviors of the participant, which is a threat to the internal validity of the study. To investigate this possibility, researchers can use analytic approaches to understand whether some confederates elicited certain behaviors more than others and adjust for this similarity in their analyses (see Kenny et al., 2001; and Thorson et al., 2020, who demonstrate how to treat an experimenter or confederate as a random effect in a multilevel model). Another common way of dealing with ecological-validity concerns in confederate studies is to demonstrate behavior within a confederate study and then show that these same effects also occur in interactions with real participants (e.g., Gaither et al., 2018; Sandstrom & Boothby, 2021).

16.4.3 Observability of the Behavior

Behaviors and psychological constructs observed from behaviors (e.g., assertiveness, negativity, agreeableness, and so on) vary in their observability, which is important from both practical and conceptual perspectives (Brunswik, 1955; Carter et al., 2018; Funder, 1995). In general, more overt behaviors will be easier to code and to achieve reliable estimates for. However, research questions may not be about highly observable behaviors. To return to a prior example, if you are studying cross-race interactions, you may not be interested in overt expressions of bias, like racist language or slurs; instead, you may be interested in subtle expressions of bias, like appearing tense, uncomfortable, or avoiding eye contact (Dovidio et al., 1997). If you are unsure about the observability of a variable, (1) seek advice from other researchers in your area and (2) run pilot participants to see whether the behavior occurs frequently enough and whether you and your coding team can reliably code it. Beware, as well, that observability is often context-dependent. For example, you might find that the agreeableness of students is easier to judge when observing them during lunch with friends than during a lecture.

16.4.4 Topographical versus Dimensional

Behaviors can be judged by their presence or absence (with a *topographical* code) or along a particular dimension (with a *dimensional* code). The choice of code type should depend on the research question, which should, at least in part, depend on observations of your population of interest in the setting of interest. On more than one occasion we have thought a behavior should be coded dimensionally or topographically, and conversations with our research assistants and our own observations have convinced us otherwise.

In general, topographical codes are useful when there is a clear, meaningful distinction between the presence and absence of a behavior: for example, before a negotiation, did two people shake hands with each other or not (Schroeder et al., 2019)? Topographical codes are also useful for questions in which the intensity of a behavior matters less than whether it occurred at all. For example, you might not care *how long* two people shook hands for as long as you know whether they shook hands at all. In addition, topographical codes are useful when more fine-grained assessments of behavior are difficult. For instance, it might be challenging to evaluate hand-shaking along relevant qualities (e.g., warmth, authenticity, dominance), making a topographical code more appropriate than a dimensional one.

Dimensional codes are useful when behaviors vary in their level of intensity and when it is possible for coders to reliably and validly capture distinctions in intensity. For instance, during a conflict conversation between two romantic partners, all people will likely show some negative affect, and the variability in this negative affect can probably be captured on a scale of 1 to 7 (with clear meanings at each scale point). Dimensional codes are quite common in social and personality psychology and some popular items include negative and positive affect, friendliness, warmth, dominance, anxiety, body posture, agreeableness, and extraversion (e.g., Gordon & Chen, 2016; Hughes et al., 2021; Witkower et al., 2020). With dimensional codes, it is important that coders can identify examples at every level of the scale. If it is consistently difficult for a team to differentiate between particular levels, then reducing the number of points on the scale is likely a good idea.

With both code types, variability in measurements is important. Low variability can impair inter-rater reliability estimates (Hallgren, 2012). In addition, observing associations between variables with little variability is challenging. For topographical codes, if almost no one shows the behavior or almost everyone shows the behavior, then the code is not valuable. To improve variability, sometimes people assess the presence or absence of a behavior numerous times. For example, in a group decision-making context, you might evaluate whether or not a person participated by assessing their participation (yes or no) in each thirty-second interval of the discussion. More variability will exist in the sum of these codes than in one overall code. For dimensional codes, ideally, the average intensity in your sample would be the midpoint on your scale so that coders can capture sufficient deviation around that midpoint. If this is difficult, consider changing the anchoring points of your scale. A normal distribution around the midpoint, with all values of the scale used but at different frequencies, is also helpful.

Lastly, sometimes, a combination of topographical and dimensional codes is used to document multiple aspects of the situation or the same behavioral construct in multiple ways. For example, you could code whether a person spoke during a thirty-second interval (using a topographical code) and then code the positivity of what was spoken (e.g., not at all positive to extremely positive; using a dimensional code).

16.4.5 Macro versus Micro

Behavioral codes occur at different frequencies and in reference to different lengths of time. Macro coding systems (also known as molar or global) involve global or overall codes that are made over longer lengths of time. For example, when examining behavioral affiliation, coders might make one summary rating of how friendly a person appeared during a conversation (e.g., Moskowitz, 1988; Myaskovsky et al., 2005; Traupman et al., 2011). Micro coding systems (also known as molecular) involve more frequent coding, and the behaviors are often more specific and fine-grained than those examined in macro approaches. When examining friendliness, for example, a micro approach might ask coders to indicate every time a person smiles, laughs, or agrees with their conversation partner (e.g., Latu & Schmid Mast, 2016). Micro approaches might also ask coders to indicate whether or not any of these behaviors occurred within each five-second interval. Often, several micro codes are combined into one higher-level behavioral class for analysis. For example, researchers combined codes indicating the presence of fussing vocalizations, crying vocalizations, and gaze aversion away from a parent to create an overall code of "infant negative engagement" during parent–infant interactions (Feldman et al., 2011). Macro systems can be topographical or dimensional; micro systems are usually topographical.

Macro approaches are almost always faster – both in the training of coders and in the time to code. Sometimes, they can be easier for coders to grasp because they rely on overarching constructs – which we are used to judging in everyday life – and not on fine-grained behaviors – to which we are less used to paying close attention. That being said, sometimes macro approaches (especially if they are dimensional) involve a great deal of subjectivity, which can make it difficult to achieve agreement among coders. If every coder judges friendliness a little differently, for example, then the same behaviors in one person will lead to different summaries of friendliness across your coders. In this situation, you might decide to use a micro coding system, if your coding team can more easily agree on the micro behaviors that you think are components of friendliness. If your particular coding team has an easier time reaching inter-rater agreement, then your measurements may also have greater generalizability and predictive validity as well.

As with topographical and dimensional codes, elements of macro and micro systems can be used together. For example, you may choose to code some behaviors at the micro level, while also making more general ratings of a person's behavior throughout an entire task or activity. We used such an approach in a study investigating behavioral engagement within tutor–student dyads (Dumitru et al., 2022). Our coders documented each time students and tutors asked questions of one another (a micro rating), as well as providing overall ratings of the students' and tutors' levels of engagement (a macro rating). You may also choose to make macro ratings, but at higher frequencies; for example, perhaps your coders judge friendliness every thirty seconds of an interaction, rather than once at the end of it. If you believe that the behaviors you're assessing vary over time, such an approach helps capture that variability. Be careful, though: if a behavior does not actually vary over time (or is relatively rare), then this kind of coding can burn out your coders without providing additional value.

16.5 Basic Questions and Guidelines

We next describe several key questions that typically arise when people manually code behaviors using live observation or audiovisual recordings. None of these questions have one right answer, so we describe common considerations to help researchers make appropriate decisions for their particular study and research question. A common theme is the importance of thinking through these questions while designing and piloting a study. Small changes to a study's design or the way in which data are collected can (1) yield more reliable and valid information and/or (2) make coding easier and more streamlined, saving time and effort.

16.5.1 How Should Behavior Be Recorded?

Assuming you want or need to record behavior, how should you do it? In general, we suggest fully testing out a recording system and coding scheme before officially collecting data. If recordings do not adequately capture the behaviors you want to observe or if you cannot code the behaviors you want to code from your recordings, then your recordings are ultimately useless. Once, we collected an entire group interaction study without realizing that the audio and visual recordings were not aligned in our system. It was not until we started coding the data that we realized the trouble we were in. Take recordings of pilot participants and have your team code before you launch your study.

Other tips for recording:

1 Multiple camera angles can be helpful for capturing complex behaviors or multiple participants. It is easier to align multiple recordings while you are collecting the data (e.g., by hooking up all the cameras to the same recording station) than it is to align them with software post-collection.

2 Do not assume that keeping your cameras in the same position across all participants and sessions will work. For instance, we often adjust camera position and angle based on participant height or where people sit relative to each other.

3 Make sure that video recordings are of high enough quality to show the behaviors you want to code. Factors such as the zoom, the camera angle, the light in the room, and the resolution on your camera and recording software can affect the quality of video.

4 Many commercially available video cameras do not record audio well. Consider adding recording devices specifically for audio, if speech or other linguistic information is important.

5 If you are conducting a study with multiple tasks or segments, use visual cues to mark the beginning and end of tasks. For example, turn the lights off for a couple of seconds or flash a piece of colored paper in front of the camera. This makes it easier for coders to scan through videos quickly and find the start/stop times for different segments.

6 If participants are sitting, carefully consider your chairs. Stationary chairs with arm rests are better than ones in which participants can swivel or rock back and forth (unless you specifically care about these behaviors). Allowing participants to engage in these movements can affect your study in practical ways – for example, by obtaining recordings that capture the side of a person's face rather than the front – and theoretical ones – for example, by allowing some participants to sit closer to each other than others.

16.5.2 How Much of My Observation Time or Recordings Should Be Coded?

People are often interested in how much of their observation time or recordings they should code. For example, if a researcher sits in a classroom and observes a teacher for one hour, should all of the teacher's behavior during that hour be coded? Or should the researcher code behavior only during specific or random segments of time? In more structured situations, with shorter tasks, or with macro or dimensional codes, it is common to code all observation time (e.g., a five-minute conflict conversation between two romantic partners). With longer tasks, researchers often select shorter intervals to code under the assumption that those segments are representative of longer periods of time. This "thin-slicing" approach can drastically reduce the amount of time and effort involved in a project. Thankfully, researchers have empirically examined the extent to which judgments made from slices of different lengths and for different behaviors are accurate, representative, and valid predictors of other outcomes (Murphy et al., 2015; Murphy et al., 2019; Murphy & Hall, 2021; Wang et al., 2021), and we recommend consulting this literature when deciding whether to use this approach. Some key considerations include the observability, consistency, and frequency of the behavior of interest. If you choose a thin-slice approach, make sure the behaviors aren't rare (e.g., interruptions occur on average three times during a thirty-minute interaction).

16.5.3 Should All Codes Be Done Simultaneously or Individually?

Almost always, researchers are coding multiple aspects of behavior. Thus, when behavior is recorded, researchers have the option to complete all codes at once or do them individually. In other words, they could do one "coding pass" or multiple coding passes. Coding passes represent separate instances of processing a recording for particular codes. Imagine that you video-record five team members as they work on several different tasks. On the first coding pass, you watch the recording and indicate the start and stop times for each task. On the second coding pass, you mark every time a group member asks a question. Within this coding pass, you also indicate which person was asking the question. On the third coding pass, you judge the overall emotional tenor (positive, negative, neutral) of the group every five seconds.

When you structure coding passes, the goal is to get as much information from coders as you can with as little effort as possible. You do not want coders' attention spread across too many codes in one coding pass because this can make it difficult to observe any of them reliably or validly. This is not to say that coding passes should only have one code, though. Sometimes, when you code one behavior, you can easily code additional information at the same time. In the prior example, in order to code whether a question was asked, the coder will also know who asked the question, and thus indicating who the question was asked by takes a minimal amount of effort and yields valuable information. For the sake of efficiency, we generally suggest coding variables that are related to one another in one pass, but you should explore the best process for your particular study.

16.5.4 How Should Coding Be Documented?

Where should coders record their judgments? You have two primary options. One, you could use any spreadsheet program. For example, you might have a column for participant ID number and individual columns for the different behavioral codes or coding passes. Spreadsheets are cost-effective, require minimal training to use, and have easily manipulable data (e.g., for exporting).

Two, you could use a software program specifically designed for behavioral coding (e.g., Datavyu or Noldus Observer). These programs are generally quite flexible, in terms of the kinds of codes that can be made and the frequencies at which they can be done. In addition, all information about a particular recording, along with that recording, is kept in one place, which minimizes burdens on coders by allowing information that is documented once (e.g., start and stop times of different parts of recordings) to be easily accessed and used many times. Software also often allows researchers to gain multiple pieces of information from one input. For example, when coding the presence of smiles, a researcher might click one button to indicate that a smile happened. The software then automatically records when the smile occurred, yielding information both about the frequency of smiles and their time course. Lastly, software allows one to combine information from different coding passes. For example, perhaps after coding smiling, you code where participants were looking. You could then easily combine these coding passes to understand how often smiles occurred when participants were looking in certain places – for example, at their spouse, their child, or the toy. Although specialized software has a learning curve – both in coding and in exporting the codes – in our experience, the benefits have been well worth the effort.

16.5.5 Who Should Coders Be?

One key question when manually coding behaviors is who should do the coding. Coders should not have extensive knowledge of the study hypotheses and research objectives. Thus coders are often junior members of the research team – for example, undergraduate research assistants or paid staff members. Coders should be conscientious, organized, and detail-oriented; they must also be willing and able to stay focused on what can be a tedious task, to execute their responsibilities as consistently as possible over the duration of a coding project, and to follow coding instructions as closely as possible, without applying their own interpretations of behavior (e.g., if the coding instructions specify that sarcasm should be viewed as an expression of disrespect, then it needs to be used as an indicator of disrespect, even if a particular coder personally disagrees). It is also critical to select coders who are willing to ask questions and report any issues that arise. Coding schemes are never perfect at the outset, and so coders who are willing to provide feedback or express confusion are important team members. Coders may also find deviations in study protocol that may not have been documented yet and can play a valuable role by making sure the lead researchers know about them.

Another question that often arises with regard to coders involves their social identities – for example, do their race, gender, age, nationality or social position matter? Should their identities match those of research participants? Should they match those of other coders? Decades of research on social perception have documented that people's own identities and social group memberships are associated with their perceptions of others' behaviors and emotions (Elfenbein & Luckman, 2016; Freeman et al., 2020; Hall et al., 2016). We recommend that researchers consult the literature regarding their theoretical question, the behavior of interest, and the context in which it is being examined to see which coder characteristics may have the most influence over the results of their study. There are some questions and behaviors for which coder identity is likely to matter much less than for others (e.g., topographical and micro observations are probably less affected than dimensional or macro observations). In practice, researchers often restrict their coder population to coders who are quite similar to each other to boost the reliability between coders. This is a reasonable choice; however, this may come at a cost to generalizability so it is worthwhile to consider how these concerns can best be balanced.

Note that in some cases you may have systematic bias due to coder identity that you can adjust for in your models. For example, if you find that female coders judge female participants more favorably relative to male participants than male coders do, you could adjust for coder identity in your analyses. However, you may also have times when you cannot adjust for the bias. Using the example above, if you only have female coders, then ratings of female participants are likely to be, on average, more favorable than those of male participants, but there is no clear way to adjust for this in your analyses. You could include participant identity in the analyses, but it will be unclear to you whether gender differences exist because of true differences in participants' behavior or because of biases from coders.

16.5.6 How Many Coders Should There Be?

In practice, the number of coders that researchers have varies widely so there is really no one standard. Two key questions to think about are inter-rater reliability (how strongly do your coders agree in the judgments they make) and generalizability (how strongly would your coders' judgments agree with those that other people would hypothetically make). Statistically, more coders working independently can produce greater reliability estimates. For instance, for average-ratings intraclass correlation coefficients (see section 16.6.1 below), rater-related variances are scaled (i.e., divided) by the number of coders. Because these rater-related variances are part of the total variance and are in the denominator of the statistic, all else being equal, having a greater number of coders will produce a larger reliability estimate overall. That being said, a large coding team may also be difficult to manage, and you may find it harder to train and establish shared understandings within a larger group of people. Also, in general, more coders improve generalizability, but you have to weigh this concern against the practical difficulties of a large coding team. We tend to worry most about generalizability when we have codes that involve more subjectivity (which isn't always intuitive; on more than one occasion we thought a behavior, such as smiling, was objective, and we were surprised to learn how subjective it is); in these instances, we try to have a larger number of coders.

16.5.7 How Much Overlap Should Coders Have with Each Other?

The question of how many coders to have is closely tied to the question of how much overlap coders should have with each other. We recommend thinking about overlap at the level of coding units or time intervals and not at the level of participants/dyads/families/groups and so on. This ensures that reliability estimates are based on how coders viewed participants in general and not on how much they agreed on only a subset of participants. In addition, this approach protects reliability estimates: if a few idiosyncratic (i.e., difficult-to-judge) participants end up in your reliability sample, reliability estimates can be severely impaired.

You must have some overlap among coders in order to assess inter-rater reliability; often people recommend 25 percent (Adolph et al., 2013). Sometimes coders completely overlap with each other (i.e., every coder codes everything), and then researchers average the responses across coders. Statistically, this approach boosts reliability estimates, and so this is one reason it is useful, though, of course, it involves more effort and time (Hallgren, 2012). You could also have the same set of coders code the same subset of units/time intervals (e.g., three coders each code the same 25 percent of data – which might be the first three minutes of every twelve-minute conversation). Then each of these coders might code a remaining separate 25 percent. One benefit of these two designs is that you can estimate and adjust for systematic bias between coders, which can also boost reliability estimates (Hallgren, 2012).

You could also have one lead coder who codes everything, and the remainder of your coders code different overlapping segments of observations (e.g., with fifteen-minute conversations, every person codes a different three-minute segment). This can be useful if you have one coder who works in your lab full-time (e.g., a lab manager) and has the time to dedicate to coding everything. Another approach is to have pairs or groupings of coders overlap in various ways. For example, maybe Diya and Kabir overlap with each other for minutes 1 through 3 of the first twenty conversations, as do Sage and Matías for minutes 4 through 6, but then for the next set of conversations, Diya and Matías overlap while Sage and Kabir overlap. There are many ways to mix and match coding overlap, and there is no one standard approach. Across all these approaches, the more your coders overlap and the less you rely on any one single coder, the better.

16.5.8 What Study Information Should Coders Have?

Researchers may wonder how much information coders should have about the study design and hypotheses. Generally, the less information coders have about the specific study hypotheses, the fewer chances there are for their judgments to inadvertently conform to or refute those hypotheses. There are two kinds of study information that are particularly worthy of consideration.

16.5.8.1 Experimental Manipulations

When observing behavior, sometimes an experimental manipulation can be heard and/or seen. The easiest thing to do is to capture the behavior of the participant, separate from any experimental manipulation, so that coders never have knowledge of the manipulation at all. For example, perhaps researchers structure the study so that the manipulation occurs prior to coders entering the room (in the case of live observation) or prior to the start of the videotape or offscreen from the recording.

However, it is simply not possible to prevent coders from seeing or hearing some study information. In these contexts, it may be worthwhile to consider a different study design. For example, imagine studying friendship formation. One well-known paradigm manipulates the degree to which participants disclose information about themselves by instructing them to ask and answer questions of different kinds with each other (Aron et al., 1997). In this paradigm, it would not be possible for coders to judge people's behavior during the conversation without hearing the content of their answers. After coding enough participants, coders would easily discover that some participants answered one set of questions and other participants answered a different set, and this knowledge might influence their judgments of people's behavior in the different situations.

In other situations, researchers could restrict coders' access to information about an experimental manipulation, but it might make coding difficult. For instance, imagine you want to code students' attention to their teachers when they are instructed to use a laptop versus pen and paper for taking notes (Barak et al., 2006). You film students while they are listening to a lecture, but because you do not want coders to know whether students have been instructed to use a laptop or pen and paper, you can't record visuals of their desk or their lower arms. These restrictions might make coding concentration difficult because coders have lost important information for their judgments of student concentration – what are students looking at? Are they doodling or scribbling notes? Are they on social media or taking notes in a word document? The bottom line is that if you have to restrict access to certain information in order to protect coders from knowing about your experimental manipulation, consider whether your coders still have enough information to reliably and accurately judge the behavior you want them to observe.

16.5.8.2 Social Behavior

Social and personality psychologists are often interested in understanding behavior during interactions with other people. In these situations, researchers should consider whether they want coders to see or hear the other interaction partners (study participants or confederates) when coding. In addition, should the coders see and judge all interaction partners simultaneously or individually? When outcomes are at the level of the dyad or the group, then coders need access to all interactants' behavior. For instance, if you were interested in group co-ordination while people solve problems together (e.g., "To what extent do group members pull together each other's ideas and suggestions?"), then you would need to simultaneously see and hear all group members (e.g., Dittmann et al., 2020).

When outcomes are at the level of the individual, researchers have to decide whether coders should have access to information about the other people in the interaction. This is critical because coders' judgments of someone may vary based on characteristics or behaviors of that person's interaction partners. This is a common issue in research on close relationships, where two members of a couple have a conversation together, and their behaviors are coded at the level of the individual afterward. In these situations, the potential for "spillover" – where information about one partner affects judgments of another – could be high. For example, would the same behavior by a man be considered withdrawal behavior (or as extreme withdrawal behavior) if the coder knew that his wife did not show a demand behavior first (Heavey et al., 1995)? As another example, imagine judging the politeness of a student when talking to someone else. While coding, the student says, "Hey, what's up?" Politeness judgments might vary depending on whom the student is speaking to. If the research question is whether people are more polite to higher- or similar-status others, the validity of the results could be affected if coders know whom the other person is. If the question is whether people are more or less polite depending on their self-reported moods, then knowledge about the other person may be less problematic.

To ensure that behaviors are judged consistently regardless of who or what a person's interaction partners do, one might choose to restrict coders' access to one person at a time. However, it can be hard to judge interactive behavior without knowing everyone's behaviors. We conducted a study, for instance, where two students solved math problems together, and we coded how often these students answered each other's questions (Thorson et al., 2019). We could not determine whether a person had answered another's question without knowing whether the other person had actually asked a question first. Visual cues can also be difficult to judge with access to only one partner. For example, studies of behavioral mimicry require that both people can be seen (in the same frame) in order to judge whether one person's behavior follows the other person's behavior (Poole & Henderson, 2022). We recommend that researchers consider, at the outset of a study, whether coding done with access to all participants is something they are comfortable with for their research design and the conclusions they will be able to make. Finally, once a decision has been made about the level of access that coders have to other interaction partners, it is important to keep this decision consistent across all participants and all coders to avoid unsystematic bias creeping into coders' judgments.

16.5.9 How Should Coders Be Trained?

The process of training coders is often inextricably tied to the process of establishing a coding scheme and a coding manual. After formulating our research question and observing our participants, we generally train coders and establish a coding scheme in a five-step process. These steps should serve as a rough guideline as there is no one right way to do this. Throughout this process, we rely on example observations of participants – for us, these are generally recordings of participants in the same or similar context. They could be from a similar, prior study, or they could be from the same study as we are coding, but of participants whose data will be excluded from analysis for reasons irrelevant to coding (e.g., participants we ran while piloting another part of the study procedures).

1 We develop an initial coding scheme, which describes all the behaviors we want documented and how we want them documented. It is useful here to keep one's conceptual question at the forefront of thinking: if coding proceeded based on these plans, would you be able to answer your theoretical question?

2 As lead researchers, we test this coding scheme ourselves. We ensure that we can observe the behaviors, and, with dimensional codes in particular, we make sure that we can achieve variability across the scale. We also develop a coding manual that includes any and all information about how to code. We include answers to likely questions, and we indicate specifics about how borderline or questionable cases should be coded. For example, when coding questions, we might indicate what coders should do if they hear a statement spoken like a question ("They live in New York?"). We strive for the manual to be useful not only for our current coding team, but also for other researchers who may want to replicate our process or learn what we did. We often include examples in the manual, either in text form or with pictures or video clips.

3 We explain our coding scheme and our coding manual to our coding team. We demonstrate how example observations should be coded.

4 Coders observe example participants and make judgments within small groups (without us). We discuss disagreements or questions – and review all ratings – as a full team, and adjust our coding scheme and/or manual. We repeat this process as needed until coders feel comfortable coding independently.

5 Coders independently code new example participants and then meet with each other to discuss questions and review discrepancies. We meet regularly with our coding team throughout this process and update the coding scheme and manual as needed. If we have enough example videos, we calculate inter-rater reliability statistics on coders' ratings and only proceed once coders have reached an acceptable level (see section 16.6.1 below).

16.5.10 How Should Coding Proceed Once Coders Are Trained?

Once coders have been trained and a coding scheme has been formalized, coders can proceed with coding "official" observations independently. Throughout this process, our full coding team meets weekly to minimize "coder drift" – a phenomenon that occurs when coders' rating processes change over time. For example, coders may differentiate behaviors less (potentially if they get bored or inattentive), resulting in less variability in their ratings over time. Coders can also "drift" to the edges of rating scales over time, seeing participants as more and more friendly, for example, or less and less likeable as time goes on. We discuss disagreements or questions, and we update our manual as needed – for example, adding instructions about new idiosyncrasies that were observed. Although we discuss disagreements in ratings, we do not update these ratings or "resolve" discrepancies in any way, as this would mean that the ratings were no longer independent of one another. In addition to reviewing all the raw data, someone from the lead research team regularly calculates inter-rater reliability to ensure that the coding teams are still achieving acceptable reliability. In addition to this general outline, there are some tips for leading a behavioral coding team that we've learned from other experts and our own experience along the way.

1 Engage coders intellectually with the research process. Coding can be tedious work, and coders may not feel they learn anything new or beneficial for their educational or career progress from week to week when completing a coding project. We actively involve our coders with the process of creating coding schemes, and, as much as we can, we explain the reasoning behind all steps of the coding process. We also have our behavioral coders involved with another non-coding project at the same time, and we provide other training and intellectual opportunities as much as possible, being careful not to reveal information about the hypotheses for the current coding project.

2 Cultivate an environment where coders feel comfortable asking questions or voicing

confusion. Coders can be afraid that asking questions signals incompetence, when quite the opposite is true. We explicitly tell our coders that if they are not asking questions, they probably are not paying close enough attention. Normalize rigorous, engaged check-in meetings – these should be active, working meetings and not situations where people "touch base" quickly and then leave.

3 Include sections in your coding spreadsheets for informal comments or questions. It is always easier to document these as they occur than to try to remember them later.

4 Randomize the order of observations that coders see. For example, coders should not code participants in the order in which they participated in the study. If coders observe multiple trials of something, the order in which they code these trials should also be randomized. If all coders do not code everything, randomize which participants or segments are part of the "reliability" or "overlap" set and randomly spread these out over time.

5 Prevent coders from seeing each other's coded data outside team meetings. Sometimes coders' intentions to code "correctly" are so strong that they are tempted to look at other coders' ratings as they are working. Coders' ratings are meant to be independent, so removing any temptations to see or otherwise use each others' ratings helps maintain independence.

6 Strive for continuous coding, without any weeks-long breaks in the process.

7 Prevent "binge-coding." We set limits on the amount of time that coders can continuously code, based on the study and the coding scheme. Coders can be tempted to complete all weekly coding in one long session, but this can easily compromise the quality of the data. We would much rather provide "extensions" for completing coding than have coders binge-code.

8 Provide quiet, uninterrupted spaces for coding. We block off special sections of our lab space specifically for video coding. This ensures participants' privacy, in line with ethical guidelines, and also facilitates coders' cognitive focus. Only in special, rare circumstances (e.g., when our labs and offices were closed due to the COVID-19 pandemic) do we allow coders to code remotely (i.e., outside the physical lab space). Before you do so, consult your institutional review board. Allowing people to code recordings at home, for example, could compromise the anonymity of study participants, if other people in the coding environment were to see the recordings.

16.6 Analysis of Behavioral Data

There are three major analytic issues relevant to behavioral data: inter-rater agreement and reliability, aggregated coded responses, and potential nonindependence of measurements.

16.6.1 Inter-rater Reliability

When two or more people code behavioral observations or recordings, researchers must assess the degree to which those coders agree in their ratings – often referred to as inter-rater reliability or inter-rater agreement. In general, reliability is the extent to which a measurement can be reproduced. Measurements can be unreliable for reasons other than disagreements between raters but the inter-rater aspect of reliability typically receives the most attention within behavioral observation.

Why is inter-rater reliability important? Many behavioral codes have some degree of subjectivity, and it is important to ensure that measurements are not simply based on one coder's unique, idiosyncratic judgment of behavior. Ideally, a coding procedure can be used by many different people to produce similar ratings of the same set of behavioral observations or recordings. Assessing inter-rater reliability is one way to establish this. Inter-rater reliability is also important for statistical power because unreliable measurements make it harder to detect true relationships among variables. In addition, inter-rater reliability is important for validity given that unreliable measurements are likely to be poor representations of the theoretical construct one is trying to measure.

There are many theoretical approaches and analytic techniques regarding inter-rater reliability and, of course, reliability more broadly (see Revelle & Garner, Chapter 20 in this volume, and Shrout & Mogami, Chapter 21 in this volume). Below, we briefly describe two of the most common analytic techniques for assessing inter-rater reliability: Cohen's kappa and the intraclass correlation coefficient (ICC). We explain these here to provide a general sense of what they are, with the expectation that readers will consult more detailed resources to better understand the variants of these statistics, when they are appropriate, and how to calculate them (e.g., Hallgren, 2012; Heyman et al., 2014).

16.6.1.1 Cohen's Kappa and Related Variants

Cohen's kappa (Cohen, 1960) and similar variants (see de Mast, 2007; Hallgren, 2012; Xu & Lorber, 2014) are appropriate if the final, analyzed behavioral outcome is measured on a nominal scale (e.g., did the participant ask a question (yes or no)?, what kind of affect did the participant display (positive, negative, neutral)?). They range from –1 (complete disagreement) to 1 (complete agreement), and values between 0 and 0.2 are usually considered slight agreement, between 0.2 and 0.4 fair, between 0.4 and 0.6 moderate, between 0.6 and 0.8 substantial, and between 0.8 and 1 almost perfect (Landis & Koch, 1977). Note, though, that these cutoffs should be considered alongside the behavior: a value of 0.6 might indicate "substantial" agreement, but readers would certainly question coding practices if a value of 0.6 were obtained for overt behaviors such as whether or not a person spoke at all. In general, these measures assess agreement between coders, while adjusting for agreement expected due to chance. Several considerations can help determine which kappa variant to use, including the number of coders, the type of overlap among coders, the marginal distributions of ratings, and whether certain disagreements should be more strongly weighted than others.

16.6.1.2 Intraclass Correlation Coefficients

The ICC is another popular inter-rater reliability statistic and is used for variables measured on ordinal, interval, or ratio scales (Shrout & Fleiss, 1979). The ICC is rooted in generalizability theory, which focuses on understanding the sources contributing to variation in measurements: some variation is due to "true" variability between people in the construct of interest and some is due to measurement error (Cronbach et al., 1972; Shavelson & Webb, 2006). Applying this approach to inter-rater reliability, variability in coders' ratings is decomposed into variation from participant to participant, variation from coder to coder, and variation due to the interaction of participant and coder. These variances are then used to compute ICCs, which represent the proportion of variance in ratings that can be explained by differences between participants – or, put another way, the proportion of variance in ratings that is independent of coders. Note that with clustered or nonindependent data (e.g., with repeated measures or dyadic data; see below for a definition), accounting for variation due to the clustered nature of the data is also necessary when calculating ICCs (Ten Hove et al., 2021). An ICC of 1 corresponds to complete agreement and 0 to only random agreement; ICCs below 0.4 are considered poor, between 0.4 and 0.6 fair, between 0.6 and 0.75 good, and between 0.75 and 1 excellent (Cicchetti, 1994).

Several considerations should be addressed for appropriate calculation and interpretation of ICCs (Hallgren, 2012; McGraw & Wong, 1996; Shrout & Fleiss, 1979). These are useful to understand ahead of time because they can shape decisions about coding procedures.

1 Are coders considered random or fixed effects? Generally, random – meaning that coders have been randomly selected from a larger population of interest – is the appropriate answer, unless you have a rare situation where the coders in your study are the only people whose ratings you care about.
2 Did a different or the same set of coders code each participant? If a different set of coders is randomly selected for each participant, a one-way model is necessary. If the same set of coders rate the same participants (whether all of your sample or just a subset), a two-way model, which accounts for systematic deviation due to coder, should be used. Two-way models generally provide higher ICCs.
3 Is it important for coders to provide the exact same values or are ratings that are similar in rank order acceptable? If you want coders to provide the same values (which is generally the case), absolute agreement is necessary. If you just want ratings that are similar in rank order, consistency is acceptable. Consistency ICCs are generally higher than absolute-agreement ICCs, but it is rarely the case that a strong argument for a consistency ICC can be made.
4 What is the unit of analysis? If each participant was coded by multiple coders and you intend on analyzing the average of their ratings, the ICC should be based on average measures. If some participants were only coded by one person, the ICC must be based on single measures. Average-measure ICCs are higher than single-measure ICCs.

16.6.1.3 Additional Tips

Regardless of the statistic you use, there are a few tips worth keeping in mind. One, before conducting reliability analyses, remove mistakes. For example, a coded value of 55 for a variable on a 1 to 5 scale should be recoded or marked as missing. Everyone makes coding mistakes, and you do not want these mistakes to artificially bias reliability estimates. Two, calculate inter-rater reliability with the final values that you are going to use in your analyses (Hallgren, 2012). For example, perhaps you code whether an event happened on a nominal scale (yes or no), but then you analyze a sum of these codes per participant. In this case, your final analysis measure is on a ratio scale and so you should choose a statistic suitable for a ratio scale and not a nominal scale. Three, report as much information about your inter-rater reliability analysis as possible, including the statistic that was calculated and the variant of that statistic (e.g., see Hallgren, 2012; ten Hove et al., 2021), even if only in a supplement.

16.6.2 Aggregation

You often have the option of aggregating coded responses. You could average responses across coders and use these averages as your final analysis values. If all coders coded all participants, then averaging across coders can allow you to incorporate the number of coders into your reliability calculation, which can boost reliability estimates. For example, with average-ratings intraclass correlation coefficients, rater-related variances are scaled by the number of coders, ultimately decreasing the total variance, which is the denominator of the statistic. If all coders did not code all participants, you can decide whether to aggregate when you can (i.e., when you have ratings from multiple coders) or simply use the responses from one coder only as the final analysis values and use the additional coders' responses to calculate reliability only. Averaging when you can is generally a good idea unless you have reason to trust one coder more than another or to prefer final analysis values from one person who judged all participants, rather than coders who saw only a portion of participants.

You can also aggregate responses across codes. For example, imagine you coded every time a baby fusses, cries, or averts his/her eyes away from a parent (as was done in the study described by Feldman et al., 2011 cited above). Rather than treating each of these codes as an individual outcome, you might sum frequencies across these codes to create a more general measure of infant negative engagement. This may lead to more reliable estimates with more predictive validity, as well as a more parsimonious analysis. Just as with self-report items that are averaged to create scales, if you take this approach, you want to make sure that the behavioral codes hang well together by assessing internal consistency – for example, by using a variance decomposition approach, as mentioned above for inter-rater reliability (see Shrout & Lane, 2012, for a discussion of assessing internal consistency among multiple self-report items).

Another approach is aggregating responses across time. Again, this can lead to more reliable estimates with more predictive validity, as well as a more parsimonious analysis. This approach can also be useful if you assess behavior at different times or with unequally spaced intervals. For example, if you code every time participants smile, this will unfold temporally in different ways for different people. One way of dealing with this is summing the number of times people smiled over the entire conversation so that the outcome aligns across participants. Researchers may also aggregate over time when unfamiliar with analyzing repeated measures data. We encourage you, however, not to shy away from these models because there are interesting conceptual questions that you can answer when you analyze behavioral repeated measures. Learning techniques for analyzing repeated-measures data or finding a collaborator with relevant expertise allows you to take advantage of all that your data provide (e.g., see Gordon & Thorson, Chapter 22 in this volume, on repeated measures).

16.6.3 Nonindependence

Nonindependence exists when outcomes cannot be considered fully independent of one another and can occur in behavioral coding projects in several ways. First, as noted above, you may have repeated measures over time within a particular unit – usually, person is considered the unit. Measures from or about one person are likely to be more highly correlated than measures that were collected from separate people. Measurements that are ordered in time are also likely to have specific temporal dependencies (e.g., behaviors assessed during minutes 1 and 2 of a conversation are likely to be more similar than behaviors assessed during minutes 1 and 5; Bolger & Shrout, 2007). Second, your measures may come from separate people who are part of a larger unit, like a dyad, family, or team. Again, measures from people who are linked in some way are likely to be more highly correlated than those from people who are not linked. Traditional statistical approaches, like ANOVA and linear regression, assume independent observations, and so you often cannot use these approaches when analyzing coded behavioral data. Two chapters in this handbook provide in-depth explanations for analyzing nonindependent data (for repeated measures, see Gordon & Thorson, Chapter 22 in this volume; for dyadic and group data, see Kenny, Ackerman, & Kashy, Chapter 23 in this volume).

16.7 Other Topics and Issues

16.7.1 Automated Coding

Manual observation and coding of behavior is the traditional, dominant approach for assessing behavior within social and personality psychology. However, advances in technology and computer science increasingly allow for automated coding of behavior – either live or via recordings (Schmid Mast et al., 2015). Many of these methods train computer models to recognize behaviors of interest based on a subset of information – for example, recordings of a sample of participants. These models then attempt to recognize these behaviors within new recordings from the same or similar context (e.g., Bartlett et al., 2005; Carcone et al., 2019; Chakravarthula et al., 2021). With these approaches, researchers can examine more behavioral information and with more efficiency than might be feasible with human coders – for example, from more participants or for longer periods of time. For example, one recent automated coding tool involves computer vision and machine learning (CVML) to code facial affect intensity (Haines et al., 2019). To test this tool, trained human coders first rated the degree of positive and negative emotional intensity in the faces of participants who viewed emotionally evocative images. Next, CVML was trained to identify facial actions that strongly corresponded with human ratings of affect intensity, using a subset of participants' data. Researchers then tested the model on a different subset of participants' data, showing that this tool could predict affect intensity (as rated by humans) with reasonable accuracy.

One drawback of these approaches is that they generally require substantial expertise in computer science; thus psychologists hoping to implement these methods often need additional training or collaborators to implement them. In addition, it can be challenging to automatically code macro behaviors or constructs in ways that are congruent with human perceptions (Black et al., 2013). For example, people may generally agree about whether a person shows more friendliness in one conversation than another, but it may be hard to train a computer on the specific behavioral cues that generate human perceptions of friendliness. Despite these drawbacks, automated coding holds a great deal of potential for expanding our knowledge about human behavior and is likely to increase in prevalence in the coming years.

16.7.2 Text Analysis

Automated coding can also be used on transcriptions of participant speech, and common options within social and personality psychology include (1) dictionary- and rule-based approaches and (2) statistical and machine learning techniques (Boyd & Schwarz, 2021; Brady et al., 2021; Tausczik & Pennebaker, 2010). Thus far, within social and personality psychology, these approaches have most frequently been applied to data that are originally text data (e.g., written communication and social media data) and not to text transcriptions of spoken conversations, though this is becoming more common (see Ireland & Pennebaker, Chapter 14 in this volume). Outside psychology, fields such as spoken-language processing in computer science apply automated methods specifically to spoken language (Haghani et al., 2018). For researchers interested in automated text processing from audio or video recordings of participants, we recommend consulting Ireland & Pennebaker, Chapter 14 in this volume.

16.7.3 Preregistration

When done well, behavioral observation and coding are iterative processes, which makes it difficult to preregister an entire study method and analytic approach up front. Therefore, if you want to preregister, we recommend outlining the general stages of your observation and coding processes and describing how you will make decisions during each of these stages, rather than indicating exactly what all the decisions will be. For example, after piloting your recordings and your coding, you could reasonably preregister some of the following: your process for assessing reliability, your process for resolving discrepancies or discussing disagreements, and your process for deciding whether to aggregate codes. Also, keep in mind that preregistration is not all-or-nothing. If behavioral recordings have already been collected or even if they have already been coded, you can still preregister analyses. This can help you gain some benefits of preregistration without losing the ability to examine potentially rich and informative recordings which have already been collected.

16.7.4 Transparency

In our view, behavioral observation and coding seem to lag behind other methodological approaches with regard to recent advances in transparency (at least in social and personality psychology). Behavioral coding often receives just a few sentences in methods sections, generally focusing on the items coded, the number of coders, and the reliability statistic. Rarely do you see details about the myriad other decisions that produce the final analysis data, including how coders were trained, whether or how coder drift was handled, and the specific kind of intraclass correlation or other statistic used to indicate reliability. This is a threat to our fields' trust in behavioral observation and coding – other researchers may find it hard to reproduce similar findings based on so little information, and readers may become skeptical with so few details. Therefore we encourage transparency as much as possible, even if only in a supplement or on a publicly accessible website.

16.7.5 Open Data

Generally, when people think of "open" or "publicly available" data, they think of the quantitative data that are used in published analyses. However, researchers may also be interested in accessing video or audio recordings – maybe to understand how quantitative codes align with specific recorded behaviors, for instance. To facilitate this, you could make all recordings available or you could make a subset of recordings available – perhaps examples representing the endpoints and average of a dimensional code. Or, perhaps, researchers are interested in coding your recordings to answer a different question. If you want to facilitate this, you could consider posting your recordings on an open repository specifically for behavioral data, such as Databrary (Adolph, 2016). Of course, whether you share your recordings or data with anyone relies on participant consent, and you must work with your local institutional review board to figure out best practices. We recommend asking participants about all potential uses of their data (e.g., sharing with collaborators, sharing with other researchers, posting on certain websites, sending to transcription services, and so on) either immediately before or after you obtain recordings. It is much easier to obtain this information close to data collection than it is years later when you decide you want to share data with new collaborators, for instance.

16.8 Conclusion

Understanding human behavior has always been one of the core goals of psychology, but assessing behavior is not always the easiest or most straightforward process. In this chapter, we outlined the major steps of behavioral observation and coding, while providing critical tips and tricks along the way. We hope this introduction will inspire social and personality psychologists to embark on new projects regarding human behavior, and we look forward to new discoveries in the future.

References

Adolph, K. (2016). Video as data: from transient behavior to tangible recording. *APS Observer*, 29(3), 23–25.

Adolph, K., Gilmore, R. O., Staff and Databrary Admin (2013). Databrary sponsored workshops and events. Databrary, http://doi.org/10.17910/B7159Q (accessed March 3, 2023).

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., and Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 23(4), 363–377.

Back, M. D., Schmukle, S. C., and Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, 97(3), 533–5418.

Barak, M., Lipson, A., and Lerman, S. (2006). Wireless laptops as means for promoting active learning in large lecture halls. *Journal of Research on Technology in Education*, 38(3), 245–263.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 568–573.

Baumeister, R. F., Vohs, K. D., and Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2(4), 396–403.

Bergsieker, H. B., Shelton, J. N., and Richeson, J. A. (2010). To be liked versus respected: Divergent goals in interracial interactions. *Journal of Personality and Social Psychology*, 99(2), 248–264.

Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C.-C., Lammert, A. C., Christensen, A., Georgiou, P. G., and Narayanan, S. S. (2013). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, 55(1), 1–21.

Bolger, N., and Shrout, P. E. (2007). Accounting for statistical dependency in longitudinal data on dyads. In T. D. Little, J. A. Bovaird, and N. A. Card (eds.) *Modeling Contextual Effects in Longitudinal Studies*. Lawrence Erlbaum Associates Publishers.

Boyd, R. L., and Schwarz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1), https://journals.sagepub.com/doi/full/10.1177/0261927X20967028.

Brady, W. J., McLoughlin, K., Doan, T. N., and Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7, DOI:10.31234/osf.io/gf7t5.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.

Capps, K. P., Updegraff, J. A., Foust, J. L., O'Brien, A. G., and Taber, J. M. (2022). Field experiment of signs promoting hand hygiene during the COVID-19 pandemic. *Health Psychology*, 41, 826–832.

Carcone, A. I., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K., and Kotov, A. (2019). Developing machine learning models for behavioral coding. *Journal of Pediatric Psychology*, 44(3), 289–299.

Carter, N. T., Carter, D. R., and DeChurch, L. A. (2018). Implications of observability for the theory and measurement of emergent team phenomena. *Journal of Management*, 44(4), 1398–1425.

Chakravarthula, S. N., Baucom, B. R. W., Narayanan, S., and Georgiou, P. (2021). An analysis of observation length requirements for machine understanding of human behaviors from spoken language. *Computer Speech & Language*, 66, 101162, https://doi.org/10.1016/j.csl.2020.101162.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons Inc.

de Mast, J. (2007). Agreement and kappa-type indices. *American Statistician*, 61(2), 148–153.

Dittmann, A. G., Stephens, N. M., and Townsend, S. S. M. (2020). Achievement is not class-neutral: Working together benefits people from working-class contexts. *Journal of Personality and Social Psychology*, 119, 517–539.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., and Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, 33(5), 510–540.

Dumitru, O. D., Thorson, K. R., and West, T. V. (2022). Investigating gender differences among tutors and students during STEM peer tutoring: Women are as behaviorally engaged as men but experience more negative affect. *Contemporary Educational Psychology*, 70, 102088, https://doi.org/10.1016/j.cedpsych.2022.102088.

Elfenbein, H. A., and Luckman, E. A. (2016). Interpersonal accuracy in relation to culture and ethnicity. In J. A. Hall, M. Schmid Mast, and T. V. West (eds.) *The Social Psychology of Perceiving Others Accurately*, 1st ed. Cambridge University Press.

Feldman, R., Gordon, I., and Zagoory-Sharon, O. (2011). Maternal and paternal plasma, salivary, and urinary oxytocin and parent–infant synchrony: Considering stress and affiliation components of human bonding. *Developmental Science*, 14(4), 752–761.

Fisher, P. H., Dobbs-Oates, J., Doctoroff, G. L., and Arnold, D. H. (2012). Early math interest and the development of math skills. *Journal of Educational Psychology*, 104(3), 673–681.

Freeman, J. B., Stolier, R. M., and Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. *Advances in Experimental Social Psychology*, 61, 237–287.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 10(4), 652–670.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality*, 23(5), 369–401.

Gaither, S. E., Babbitt, L. G., and Sommers, S. R. (2018). Resolving racial ambiguity in social interactions. *Journal of Experimental Social Psychology*, 76, 259–269.

Gordon, A. M., and Chen, S. (2016). Do you get where I'm coming from? Perceived understanding buffers against the negative impact of conflict on relationship satisfaction. *Journal of Personality and Social Psychology*, 110(2), 239–260.

Gosling, S. D., John, O. P., Craik, K. H., and Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology*, 74, 1337–1349.

Haghani, P., Narayanan, A., Bacchiani, M., Chuang, G., Gaur, N., Moreno, P., Prabhavalkar, R., Qu, Z., and Waters, A. (2018). From audio to semantics: Approaches to end-to-end spoken language understanding. *2018 IEEE Spoken Language Technology Workshop (SLT)*, 720–726, https://doi.org/10.1109/SLT.2018.8639043.

Haines, N., Southward, M. W., Cheavens, J. S., Beauchaine, T., and Ahn, W.-Y. (2019). Using computer-vision and machine learning to automate facial coding of positive and negative affect intensity. *PLOS ONE*, 14(2), e0211735, https://doi.org/10.1371/journal.pone.0211735.

Hall, J. A., Gunnery, S. D., and Horgan, T. G. (2016). Gender differences in interpersonal accuracy. In J. A. Hall, M. Schmid Mast, and T. V. West (eds.) *The Social Psychology of Perceiving Others Accurately*, 1st ed. Cambridge University Press.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34.

Hansen, P. G., Larsen, E. G., and Gundersen, C. D. (2022). Reporting on one's behavior: A survey experiment on the nonvalidity of self-reported COVID-19 hygiene-relevant routine behaviors. *Behavioural Public Policy*, 6(1), 34–51.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., and Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 11(6), 838–854.

Härdelin, G., Holding, B. C., Reess, T., Geranmayeh, A., Axelsson, J., and Sundelin, T. (2021). Do mothers have worse sleep than fathers? Sleep imbalance, parental stress, and relationship satisfaction in working parents. *Nature and Science of Sleep*, 13, 1955–1966.

Heavey, C. L., Christensen, A., and Malamuth, N. M. (1995). The longitudinal impact of demand and withdrawal during marital conflict. *Journal of Consulting and Clinical Psychology*, 63(5), 797–801.

Heyman, R. E., Lorber, M. F., Eddy, J. M., and West, T. V. (2014). Behavioral observation and coding. In H. T. Reis and C. M. Judd (eds.) *Handbook of Research Methods in Social and Personality Psychology*, 2nd ed. Cambridge University Press.

Huckins, J. F., daSilva, A. W., Wang, W., Hedlund, E., Rogers, C., Nepal, S. K., Wu, J., Obuchi, M., Murphy, E. I., Meyer, M. L., Wagner, D. D., Holtzheimer, P. E., and Campbell, A. T. (2020). Mental health and behavior of college students during the early phases of the COVID-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *Journal of Medical Internet Research*, 22(6), e20185, https://doi.org/10.2196/20185.

Hughes, B. T., Flournoy, J. C., and Srivastava, S. (2021). Is perceived similarity more than assumed similarity? An interpersonal path to seeing similarity between self and others. *Journal of Personality and Social Psychology*, 121, 184–200.

Karremans, J. C., and Verwijmeren, T. (2008). Mimicking attractive opposite-sex others: The role of romantic relationship status. *Personality and Social Psychology Bulletin*, 34(7), 939–950.

Kenny, D. A., Mohr, C. D., and Levesque, M. J. (2001). A social relations variance partitioning of dyadic behavior. *Psychological Bulletin*, 127, 128–141.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Latu, I. M., and Schmid Mast, M. (2016). Male interviewers' nonverbal dominance predicts lower evaluations of female applicants in simulated job interviews. *Journal of Personnel Psychology*, 15(3), 116, https://doi.org/10.1027/1866-5888/a000159

McGraw, K. O., and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46.

Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, 66, 100–106.

Mehl, M. R. (2017). The electronically activated recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science*, 26(2), 184–190.

Mendes, W. B., and Koslov, K. (2013). Brittle smiles: Positive biases toward stigmatized and outgroup targets. *Journal of Experimental Psychology: General*, 142, 923–933.

Moskowitz, D. S. (1988). Cross-situational generality in the laboratory: Dominance and friendliness. *Journal of Personality and Social Psychology*, 54(5), 829–839.

Murphy, N. A., and Hall, J. A. (2021). Capturing behavior in small doses: A review of comparative research in evaluating thin slices for behavioral measurement. *Frontiers in Psychology*, 12, www.frontiersin.org/articles/10.3389/fpsyg.2021.667326.

Murphy, N. A., Hall, J. A., Ruben, M. A., Frauendorfer, D., Schmid Mast, M., Johnson, K. E., and Nguyen, L. (2019). Predictive validity of thin-slice nonverbal behavior from social interactions. *Personality and Social Psychology Bulletin*, 45(7), https://journals.sagepub.com/doi/10.1177/0146167218802834.

Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., Roter, D. L., and Nguyen, L. (2015). Reliability and validity of nonverbal thin slices in social interactions. *Personality and Social Psychology Bulletin*, 41(2), 199–213.

Myaskovsky, L., Unikel, E., and Dew, M. A. (2005). Effects of gender diversity on performance and interpersonal behavior in small work groups. *Sex Roles*, 52(9), 645–657.

Nils, F., and Rimé, B. (2012). Beyond the myth of venting: Social sharing modes determine the benefits of emotional disclosure. *European Journal of Social Psychology*, 42(6), 672–681.

Paluck, E. L., and Cialdini, R. B. (2014). Field research methods. In H. T. Reis and C. M. Judd (eds.) *Handbook of Research Methods in Social and Personality Psychology*, 2nd ed. Cambridge University Press.

Park, J., Woolley, J., and Mendes, W. B. (2022). The effects of intranasal oxytocin on black participants' responses to outgroup acceptance and rejection. *Frontiers in Psychology*, 13, 916305, https://doi.org/10.3389/fpsyg.2022.916305.

Patterson, G. R. (1982). *Coercive Family Process*. Castalia.

Poole, K. L., and Henderson, H. A. (2022). Shyness, self-focused attention, and behavioral mimicry during social interaction. *Journal of Research in Personality*, 98, 104225, https://doi.org/10.1016/j.jrp.2022.104225.

Rapuano, M., Sbordone, F. L., Borrelli, L. O., Ruggiero, G., and Iachini, T. (2021). The effect of facial expressions on interpersonal space: A gender study in immersive virtual reality. In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, and E. Pasero (eds.), *Progresses in Artificial Intelligence and Neural Systems*. Springer.

Sandstrom, G. M., and Boothby, E. J. (2021). Why do people avoid talking to strangers? A mini meta-analysis of predicted fears and actual experiences talking to a stranger. *Self and Identity*, 20(1), 47–71.

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., and Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science*, 24(2), 154–160.

Schroeder, J., Risen, J. L., Gino, F., and Norton, M. I. (2019). Handshaking promotes deal-making by signaling cooperative intent. *Journal of Personality and Social Psychology*, 116(5), 743–768.

Shavelson, R. J., and Webb, N. M. (2006). Generalizability theory. In J. L. Green, G. Camilli, and P. B. Elmore (eds.), *Handbook of Complementary Methods in Education Research*, 3rd ed. Routledge.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

Shrout, P. E., and Lane, S. P. (2012). Psychometrics. In M. R. Mehl and T. S. Conner (eds.) *Handbook of Research Methods for Studying Daily Life*. Guilford Press.

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.

ten Hove, D., Jorgensen, T. D., and van der Ark, L. A. (2021). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, 27(4), 650–666.

Thorson, K. R., Forbes, C. E., Magerman, A. B., and West, T. V. (2019). Under threat but engaged: Stereotype threat leads women to engage with female but not male partners in math. *Contemporary Educational Psychology*, 58, 243–259.

Thorson, K. R., Ketay, S., Roy, A. R. K., and Welker, K. M. (2021). Self-disclosure is associated with adrenocortical attunement between new acquaintances. *Psychoneuroendocrinology*, 132, 105323, https://doi.org/10.1016/j.psyneuen.2021.105323.

Thorson, K. R., Mendes, W. B., and West, T. V. (2020). Controlling the uncontrolled: Are there incidental experimenter effects on physiologic responding?*Psychophysiology*, 57(3), e13500, https://doi.org/10.1111/psyp.13500.

Traupman, E. K., Smith, T. W., Florsheim, P., Berg, C. A., and Uchino, B. N. (2011). Appraisals of spouse affiliation and control during marital conflict: Common and specific cognitive correlates among facets of negative affectivity. *Cognitive Therapy and Research*, 35(3), 187–198.

Wang, M., Chen, K., and Hall, J. (2021). Predictive validity of thin slices of verbal and nonverbal behaviors: Comparison of slice lengths and rating methodologies. *Journal of Nonverbal Behavior*, 45, 1–14.

West, T. V., Koslov, K., Page-Gould, E., Major, B., and Mendes, W. B. (2017). Contagious anxiety: Anxious European Americans can transmit their physiological reactivity to African Americans. *Psychological Science*, 28(12), 1796–1806.

West, T. V., Pearson, A. R., and Stern, C. (2014). Anxiety perseverance in intergroup interaction: When incidental explanations backfire. *Journal of Personality and Social Psychology*, 107(5), 825–843.

Witkower, Z., Tracy, J. L., Cheng, J. T., and Henrich, J. (2020). Two signals of social rank: Prestige and dominance are associated with distinct nonverbal displays. *Journal of Personality and Social Psychology*, 118(1), 89–120.

Xu, S., and Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology*, 82, 1219–1227.

Yilmaz, G. (2016). What you do and how you speak matter: Behavioral and linguistic determinants of performance in virtual teams. *Journal of Language and Social Psychology*, 35(1), 76–97.

Zee, K. S., and Bolger, N. (2022). Physiological coregulation during social support discussions. *Emotion*, 23(3), 825–843.