Published in final edited form as:

Assessment. 2016 August ; 23(4): 404–413. doi:10.1177/1073191116635807.

A Primer on Observational Measurement

Jeffrey M. Girard and
University of Pittsburgh

Jeffrey F. Cohn
University of Pittsburgh

Abstract

Observational measurement plays an integral role in a variety of scientific endeavors within biology, psychology, sociology, education, medicine, and marketing. The current article provides an interdisciplinary primer on observational measurement; in particular, it highlights recent advances in observational methodology and the challenges that accompany such growth. First, we detail the various types of instrument that can be used to standardize measurements across observers. Second, we argue for the importance of validity in observational measurement and provide several approaches to validation based on contemporary validity theory. Third, we outline the challenges currently faced by observational researchers pertaining to measurement drift, observer reactivity, reliability analysis, and time/expense. Fourth, we describe recent advances in computer-assisted measurement, fully-automated measurement, and statistical data analysis. Finally, we identify several key directions for future observational research to explore.

Keywords

observational measurement; computational behavioral science; affective computing; contemporary validity theory; inter-observer reliability

In day-to-day life, the success of any animal is heavily influenced by its ability to detect and interpret the behavior of other organisms. These crucial skills guide the search for life's basic necessities (e.g., food, water, and security), as well as a variety of social and intellectual pursuits (e.g., reproduction, group formation, and skill acquisition). These skills also play an integral role in the scientific method (Kosso, 2011) as the basis of *observational measurement*, which is a systematic approach to detecting and interpreting behavior. A wide variety of scientific endeavours rely on observational measurement including biology, psychology, sociology, education, medicine, and marketing.

Observational measurement nicely captures the unfolding of behavior over time, which is essential to understanding its functionality (i.e., antecedents and consequences) and the dynamic processes it contributes to (Bakeman & Quera, 2011). Observational methods also circumvent many of the response biases that survey methods are prone to such as self-presentation and social desirability (Stone et al., 2000). They can even be used with

Correspondence concerning this article should be addressed to Jeffrey Girard, 210 S. Bouquet St, Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260. j.girard@pitt.edu.

participants for whom surveys would be impractical such as nonhuman animals, very young children, or patients undergoing medical procedures.

The current paper provides an interdisciplinary primer on observational measurement. First, it details the various types of measurement instrument that can be applied to observational data. Second, it describes the various methods that can be used to validate inferences drawn from observational measurements. Finally, it outlines the challenges currently faced by observational researchers, several recent advances in observational methodology, and the directions that we hope future observational research will explore.

Measurement Instruments

Observational methods use trained individuals called *observers* to make quantitative judgments about behaviors of interest. These judgments are standardized across observers through the use of *measurement instruments*. Measurement instruments reflect researchers' theories about what aspects of behavior are important and focus observers' attention on specific types of behavioral information. Some instruments require observers to assign behaviors to one or more discrete categories, while others require observers to rate behaviors on one or more continuous dimensions. In the following subsections, we discuss several characteristics on which measurement instruments meaningfully vary.

Scale of Measurement

One of the fundamental characteristics of a measurement instrument is the statistical data type or "scale of measurement" (Stevens, 1946) that it yields (e.g., nominal, ordinal, interval, or ratio). The type of instrument chosen and its scale of measurement can have important consequences for data collection, validation, and statistical analysis. In general, instruments can be usefully characterized as either categorical or dimensional (Table 1).

Categorical instruments require observers to choose between a limited number of predefined options, which are often called *codes*. Codes are grouped into sets that are often (but not always) considered mutually-exclusive and exhaustive. Codes within a set may be treated as unordered or as existing on an ordered continuum. For example, a categorical instrument developed to assess teacher effectiveness might use a set of unordered codes to categorize a lesson's subject matter (e.g., math, science, or history) and a set of ordered codes to categorize its pacing (e.g., slow, medium, or fast).

Dimensional instruments, on the other hand, require observers to choose numerical values along continuous dimensions; these values are often called *ratings*. Each dimension has an upper and a lower bound; typically, any number between these bounds may be chosen, although numerical restrictions may be enforced within this range (e.g., only integers or multiples of 5 may be chosen). The choice of upper and lower bound values can influence how observers think about the dimension (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991). Including zero within the range communicates that the absence of behavior is possible, while excluding it communicates that it is not. Additionally, including negative numbers communicates that the dimension is bipolar, whereas excluding negative numbers communicates that the dimension is unipolar. For example, a dimensional instrument

developed to assess the effectiveness of video ads might have observers rate their brand loyalty before and after the ad on a dimension from −100 to 100 and their level of engagement during the ad on a dimension from 1 to 10.

The choice to use a categorical or dimensional instrument should be guided by theoretical consideration of the construct being measured. For instance, some researchers believe that emotion is best characterized using discrete categorical states, while others believe in measuring its underlying dimensions (Gunes & Schuller, 2013). It is also worth mentioning that the line between categorical and dimensional instruments can blur when a large set of ordered codes is used or when numerical restrictions reduce the number of possible ratings to several options. In such cases, the distinction becomes arbitrary.

Degree of Inference

Another fundamental characteristic of a measurement instrument is the degree of inference it requires. This characteristic can have important consequences for the validity of inferences drawn from its measurements. Instruments can be usefully categorized as either sign-based or message-based (Cohn & Ekman, 2005).

Sign-based instruments attempt to describe the features of behavior and require relatively low degrees of inference. Such instruments have also been termed “atomic” because they focus on small units of behavior, such as utterances and gestures. One quintessential example of a sign-based instrument is the Facial Action Coding System (FACS; Ekman, Friesen, & Hager, 2002), which provides codes to describe facial behavior in terms of muscle contractions. An observer trained in FACS might see an interviewer smile at an interviewee and measure this behavior as contraction of the zygomatic major muscle.

Message-based instruments, on the other hand, attempt to interpret the meaning of behavior and require relatively high degrees of inference. Such instruments rely on observers to be “cultural informants” who can see the distinctions delineated by their categories or dimensions (Bakeman & Quera, 2011). One quintessential example of a message-based instrument is the Specific Affect Coding System (SPAFF; Coan & Gottman, 2007), which provides codes to interpret facial behavior in terms of its affective and interpersonal meaning. To return to the previous example, an observer trained in SPAFF might view this same interviewer behavior and, based on context, measure it as affection or interest.

It is worth mentioning that sign-based instruments still require a degree of inference and that some message-based instruments require more inference than others. Thus, although the distinction between signs and messages is a useful one, this property of measurement instruments may be better characterized as a continuum from low to high inference.

Temporal Representation

Measurement instruments require observers to represent behaviors in time. This representation can be accomplished in several ways and with varying degrees of granularity. These characteristics can have important consequences for the data’s temporal precision and for the types of statistical analysis that are possible. Instruments can be usefully categorized as either event-based or interval-based (Bakeman & Quera, 2011).

Event-based instruments require observers to identify behavioral *events* and assign measurements to them. Events are discrete occurrences of a behavior that have detectable starting and stopping (i.e., transition) points. Because of the need to specify transition points in addition to measurements, event-based methods can be time-consuming and challenging for observers. However, the benefit to this approach is that event-based methods allow researchers to answer questions about the number, duration, and sequencing of events with high temporal precision. For example, an event-based instrument designed to measure parental involvement in children's homework might require observers to identify discrete occurrences of children asking questions and parents providing answers. Research questions could be then be explored regarding the number and duration of these behaviors, and additional measurements could be obtained about the quality of the identified events (e.g., the interpersonal warmth or level of technical detail in a parent's answer).

Interval-based instruments, on the other hand, prompt observers to provide measurements at predetermined time intervals. Measurements may be made either about the moment of each prompt (i.e., what is happening right now?) or about the period between prompts (i.e., what has happened since the last prompt?). Prompts can be configured to occur at intervals of any length. Shorter intervals provide higher temporal precision, but are typically more expensive. Longer intervals are less burdensome for observers, but might collapse distinct behaviors into a single observation or miss a fluctuation in behavior entirely. In the previous example, an interval-based instrument might prompt observers once per minute to indicate if the child has asked a question about the homework and if the parent has provided an answer. The same ratings (e.g., of warmth or level of detail) could also be collected at these moments.

Types of Observer

Earlier, we defined an observer as an individual trained in the use of a measurement instrument. Training is an iterative process and the length of training is largely determined by the complexity of both the instrument and the behavior of interest. Research on observational skills training has found that providing immediate feedback to trainees about their measurements improves their accuracy and reduces the development of response biases (Boice, 1983). It may also lead to deeper understanding to have trainees role-play behaviors that would receive different codes or ratings (Scheffler, 1977).

Measurements can be collected from several different types of observers. Observers may be researchers or staff members with extensive training (i.e., expert observers), or they may be study participants with minimal training (i.e., participant-observers). Furthermore, observers may be unrelated to the individuals whose behavior they are observing, or they may be observing recordings of their own or a loved-one's behavior. For example, Levenson and Ruef (1992) collected ratings of marital interactions from unrelated participant-observers, while Gottman and Levenson (1985) collected similar ratings from the married couples themselves. Finally, a large number of participant-observers may be recruited through crowdsourcing platforms such as Amazon's Mechanical Turk (Mason & Suri, 2012).

The choice of observer type has important implications for the objectivity, reliability, feasibility, and usefulness of the resulting measurements. Expert observers are typically more likely to use measurement instruments as intended. However, expert observers are

more difficult and costly to acquire. There are also reasons to prefer participant-observers in some cases, such as when researchers want to capitalize on the privileged knowledge that participant-observers have of their own or of loved ones' behavior. For instance, evidence suggests that patients' perspectives on the therapeutic alliance are more predictive of treatment outcome than observers' perspectives (Horvath & Bedi, 2002).

Validity in Observational Measurement

Journal reviewers, policy makers, and instrument users require evidence that the inferences drawn from observational measurements (e.g., that a student has mastered a concept, that a patient is improving, or that a participant feels happy) are valid. It has even been argued that validity is "the most fundamental consideration in developing tests and evaluating tests" (AERA, APA, & NCME, 2014, p. 11). Validity is a multifaceted construct that has evolved in many ways since its inception (see Geisinger, 1992). Contemporary validity theory emphasizes a single, unified construct that captures "the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (Messick, 1989, p. 13, emphasis in original). Thus, validity is a property of inferences made from measurements and not a property of the measurements themselves or of the instruments that yielded them. Validity is also importantly a dimensional and changeable property; an inference lies somewhere between the extremes of 'wholly valid' and 'wholly invalid,' and its specific position may shift over time as more evidence in favor of (or against) validity is gathered and as theoretical understanding of the focal construct evolves (Cizek, 2015).

Although extensive discussions of validity and the process of validation are beyond the scope of the current paper, this section will provide an overview of the primary threats to validity and the sources of validity evidence that are most common in observational measurement. Readers interested in learning more about these topics are directed to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Threats to Validity

Two major threats to validity come from *construct under-representation* and *construct-irrelevant variance* (Messick, 1989). When faced with either threat, an inference runs the risk of misrepresenting certain individuals and, as a result, inspiring actions that may lead to unfortunate individual, societal, and scientific consequences.

Construct under-representation occurs when an assessment is too narrow and fails to capture important aspects of the construct being measured. For example, the validity of conclusions drawn from an assessment of depression severity might be called into question if the assessment captured some aspects of depression (e.g., cognitive symptoms) but omitted other important aspects (e.g., social and appetitive symptoms).

Construct-irrelevant variance, on the other hand, occurs when an assessment is too broad and contains excess variance that is associated with other, distinct constructs or with extraneous characteristics of the measurement situation. For example, the validity of conclusions drawn from an assessment of depression severity might be called into question if the assessment

inadvertently captured aspects of anxiety and psychosis or if its results varied significantly based on the ordering of the questions, the setting in which the assessment was administered, or the clinician who administered it.

Evidence of Validity

Evidence for validating inferences based on measurements can come from several different sources; three particularly important sources are test content, response processes, and hypothesized relationships among variables (Cizek, 2015). The responsibility for collecting and presenting such evidence is shared by both the test developer and the test user (AERA, APA, & NCME, 2014). The role of the test user is especially important when the test is applied in settings or for uses different than those intended by the developer.

Evidence based on test content derives from analysis of the relationship between the content of a test and the construct it is meant to measure. In observational measurement, test content typically refers to the naming, description, and criteria for an instrument's behavioral categories and dimensions, as well as the details of its implementation (e.g., temporal resolution and observer type). This form of evidence often involves logical and empirical analyses of the relationship between test content and theoretical construct, as well as expert judgments of test adequacy. The threats of construct under-representation and construct-irrelevance increase to the extent that content and construct fail to align.

Evidence based on response processes derives from analysis of the cognitive processes engaged in by test takers. In observational measurement, the response processes of the observers are of central importance. The validity of test content is inconsequential if observers do not use the appropriate criteria to assign their measurements or are unduly influenced by construct-irrelevant factors. This form of evidence often involves questioning test takers about their general response strategies and examining their responses to particular items using "think aloud" protocols (Van Someren, Barnard, & Sandberg, 1994). Measurements of test taker behavior (e.g., response times and eye tracking on individual items) may also reveal information about their response processes.

Finally, evidence based on hypothesized relationships among variables derives from analysis of the internal structure of test variables and their relationships to external variables of interest. In observational measurement, the internal structure of test variables refers primarily to the relationships between behavioral categories and dimensions (as well as to the reliability of measurements, which will be discussed separately). Validity is supported to the extent that these internal relationships align with the hypotheses of accepted theories. External variables of interest often include measures of outcomes that the test is expected to predict, as well as the results of other tests hypothesized to measure similar or distinct constructs. Validity is supported to the extent that behavioral categories and dimensions predict what they are expected to, are related to accepted measures of similar constructs, and are unrelated to accepted measures of distinct constructs.

Inter-Observer Reliability

The most commonly-provided evidence of validity in observational studies comes from studies of inter-observer reliability. Although validity and reliability are considered distinct

constructs, contemporary validity theory recognizes that reliability has important implications for validity and can be considered a source of validity evidence based on hypothesized relationships among variables (Cizek, 2015). By quantifying the extent to which multiple observers assign similar measurements to the same items, these studies reveal whether or not observers are a problematic source of construct-irrelevant variance.

Estimating inter-observer reliability with a categorical measurement instrument involves assessing the extent to which observers assign items to the same (or similar) categories. Many approaches to estimating inter-observer reliability that 'adjust' for random guessing by observers have been proposed and widely-used. Two recent articles illuminate the advantages and disadvantages of existing approaches (Feng, 2013; Zhao, Liu, & Deng, 2012). Both suggest that, although no ideal approach yet exists, Bennett, Alpert, and Goldstein's (1954) *S* index[1] appears to be the least-biased option currently available, especially when the number of categories in each set of categories is relatively small.

Estimating inter-observer reliability with a dimensional measurement instrument involves calculating the degree of association between the measurements of each observer. The intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) is typically used for this purpose. The ICC relies on partitioning the measurement variability into several components (e.g., variability due to items, observers, and measurement error). There are many different ICC formulations, but the general idea shared among them is that measurements can be considered reliable when the variability associated with error constitutes a relatively small proportion of the total variability. For a discussion of the different ICC formulations, readers are referred to McGraw and Wong (1996).

Current Challenges

Observational researchers inevitably confront several challenges. One is drift or the fact that, over time and experience, observers' measurements may vary in systematic or stochastic ways. Second is reactivity or the fact that observers' response processes may change in the face of overt evaluation. Third is the fact that estimating inter-observer reliability is not always straight-forward. Finally, there is the high cost of collecting observational measurements.

Drift may occur for a single observer due to fatigue, forgetting, a loss of motivation, or the accumulation of bad habits (Boice, 1983; Campbell & Stanley, 1966). It can also occur for a group of observers who, after working and training together, become more reliable with each other but less reliable with observers outside that group (O'Leary & Kent, 1973). Group drift is especially concerning when the same measurement instrument is used by multiple research groups, as it can prevent meaningful comparison between studies. Researchers can detect drift by periodically comparing observers' measurements to external measurements that are accepted as 'correct.' However, this solution is rarely used due to the difficulty of collecting (and the relative absence of) accepted measurements. Drift may also be

[1]Since 1954, the *S* index has been reinvented many times and given many different names. For the most detailed instructions on calculating it, see Gwet's (2014) handbook, where it is denoted both κ_{BP} and κ_{Q}.

exacerbated in group-specific (i.e., unshared) databases, which would not be detected by this approach. Instead, researchers sometimes detect drift from an initial baseline by having observers assign measurements to the same items over the course of time. Using this data, *intra*-observer reliability analyses can be performed to reveal changes in observers' responses over time. However, without accepted measurements to compare to, it can be difficult to determine whether such changes are due to drift or whether the newer measurements are actually more accurate than the older ones.

Reactivity can be thought of as a specific case of the Hawthorne effect (Adair, 1984; Boice, 1983). It occurs when observers modify their response processes when they know they are being evaluated. Reactivity is a serious challenge when collecting evidence of validity based on response processes and inter-observer reliability. If observers use different response processes when they are being overtly and covertly evaluated, then evidence of validity based on their overtly measured response processes may not pertain to any measurements collected in the absence of overt evaluation. Similarly, if observers are more reliable during overt than covert evaluation, as previous research has found (Reid, 1970), then evidence of validity based on overt reliability analyses may not pertain to any measurements collected in the absence of overt evaluation. If reactivity is suspected, then researchers can devise covert or indirect means of assessing response processes and inter-observer reliability. As a general rule, it seems worthwhile to keep observers blind to the items that will be included in reliability and response process analyses when possible. Furthermore, we recommend the use of frequent 'observer meetings' where measurements are randomly selected from each observer to be viewed and discussed by the group. These meetings serve a didactic function and encourage observers to remain consistent in their response processes given that any of their measurements may be evaluated. Researchers must take care not to encourage group drift during such meetings however.

Inter-observer reliability is an important source of validity evidence in observational measurement. However, researchers currently face several challenges pertaining to its use. First, a wide range of reliability indexes are used by different researchers and by different fields. This heterogeneity creates confusion and, in studies where a reliability index is the dependent variable, makes comparison between studies difficult or impossible. Second, reliability indexes are commonly applied and reported incorrectly (Feng, 2013). These mistakes are understandable given the number of different options available, but are very problematic from a validity perspective. Third, although numerous criteria have been proposed (see Gwet, 2014, pp. 166–168), there is no agreed-upon criterion for what constitutes an 'acceptable' reliability score. For example, Fleiss (1981) suggested that chance-adjusted reliability scores above 0.40 are "intermediate to good" and scores above 0.75 are "excellent." However, there is little consensus on such criteria and many researchers (e.g., Bakeman & Quera, 2011) have challenged the notion that any criteria could be universally applicable. This issue is complicated and deserves more attention. Finally, as mentioned previously, existing approaches to estimating reliability with categorical instruments are problematic. New categorical reliability indexes and widely adopted standards for their use, reporting, and evaluation are sorely needed.

Perhaps the most limiting challenge faced by observational researchers to date has been the sheer cost of training observers and collecting measurements. As an exemplar, training in FACS takes several months and coding a single minute of video with FACS can require an hour or more of observer time (Cohn & Ekman, 2005). While these estimates are likely greater than what is required for many measurement instruments, the temporal and financial burden of observational measurement is a serious obstacle. Crucially, this burden often imposes limits on the number of participants that can be included in an observational study, reducing its statistical power and generalizability. While no easy solution to this problem exists, advances in computer-assisted and fully-automated measurement help to mitigate its impact. The next section describes these, and other, recent advances.

Recent Advances

Computer-Assisted Measurement

The tools for recording observational measurements have greatly advanced in recent years. Early observers recorded measurements using paper-and-pencil forms that were typically organized as grids with time intervals represented as rows and behavioral categories or dimensions represented as columns (Bakeman & Quera, 2011). This intuitive and parsimonious grid-based format has been preserved in many of the more recent tools.

However, researchers have increasingly adopted computerized alternatives to paper-and-pencil forms. The benefits of computer-assisted measurement generally include increased ease of use, efficiency, and temporal accuracy. Nowadays, observers typically assign measurements to audiovisual records of participant behavior. This approach restricts observers to the information captured on the record (e.g., behavior occurring off-camera would not be visible), but offers substantial benefits in exchange. Recordings enable multiple observers to view identical information, even if they are separated in time and space. Using computer-assisted measurement tools, observers can easily control playback of the record: playing it at various speeds, rewinding it, and pausing it when necessary. Well-designed tools also synchronize playback and measurement recording automatically, increasing temporal accuracy and reducing clerical errors.

One relatively recent advance that bears mentioning is the development of continuous measurement systems. Inspired by Gottman and Levenson's (1985) affect rating dial, these systems collect dimensional measurements continuously (i.e., with very short time intervals) as observers view audiovisual records in real-time. Observers typically adjust the numerical value of their measurements by manipulating a physical input device such as a dial, lever, or joystick. The primary benefits of such tools are their efficiency (e.g., one minute of video only requires one minute to measure) and that the distributions of such measurements tend to be far less skewed than those from most categorical coding systems. Such tools have been used to great effect in the study of affective and interpersonal processes (e.g., Baker, Haltigan, Brewster, Jaccard, & Messinger, 2010; Cowie, McKeown, & Douglas-Cowie, 2012; Lizdek, Sadler, Woody, Ethier, & Malet, 2012).

General purpose computer-assisted measurement systems are now widely-available in both commercial and freeware models. Popular commercial tools include Noldus' The Observer

and Mangold's INTERACT, while popular freeware tools include ELAN and ANVIL. More specialized software is also available that excels at certain tasks. For instance, the freely-available ChronoViz was designed to enable easy synchronization and visualization of multiple data types including audio, video, digital pen, and geolocation data. For continuous measurement, one of the current authors developed two open-source software packages: CARMA for measuring a single dimension and DARMA for measuring two dimensions simultaneously (Girard, 2014). These latter systems also include powerful options for analyzing inter-observer reliability both quantitatively and qualitatively.

Fully-Automated Measurement

Recent advances in computer science have yielded algorithms capable of performing certain observational measurement tasks without human intervention. While the majority of this work has focused on automatic detection of facial expressions, a growing body of literature is exploring fully-automated measurement of other behavioral constructs, such as emotional and cognitive states, physical pain, and depression (Cohn & De la Torre, 2014). Given the considerable cost of collecting observational measurements, fully-automated measurement could represent an enormous increase in research efficiency. And with their promise of nearly-infinite scalability and real-time analysis speeds, such tools also have the potential to open up entirely new avenues of research and application.

Although numerous approaches exist for fully-automated measurement, most researchers have converged on the same basic structure of analysis. In this structure, an algorithm is trained using two types of information. First, trusted human observers provide categorical or dimensional measurements on a subset of items. Second, quantitative measurements of these items, called *features*, are extracted using computer vision and signal processing techniques. The algorithm then attempts to learn a high-dimensional mapping between the features and the trusted measurements. For example, an algorithm might learn that items categorized as 'smiles' tend to have certain combinations of features, while items categorized as 'non-smiles' tend to have different combinations. This learned mapping can then be used (i.e., extrapolated from) to generate predicted measurements for novel items.

While the majority of work on fully-automated measurement has focused on visual sign-based instruments like FACS, subsets of work have focused on training algorithms to make message-based predictions (e.g., Gunes & Schuller, 2013) and to integrate features from multiple behavioral modalities (e.g., face, posture, and speech; Dibeklioglu, Hammal, Yang, & Cohn, 2015; Pantic & Rothkrantz, 2003). As message-based measurements are highly inferential and sensitive to context, it may be hard to imagine how an algorithm could perform this type of task (cf., Bakeman & Quera, 2011, pp. 20–21). However, as the field of computational behavioral science matures, its ability to measure such contextual information increases. Armed with such rich information, we believe that researchers will continue to close the gap between fully-automated and human observers over time.

One of the exciting benefits of fully-automated measurement is that algorithms are immune to drift and reactivity. They do not become fatigued, distracted, or self-conscious; once trained, they do not change their minds. However, this blessing can become a curse when its

implications are not fully appreciated. Because current algorithms typically do not continue to learn over time, they are entirely dependent on their initial training.

Girard, Cohn, Jeni, Sayette, and De la Torre (2015) demonstrated that, when provided reliable and representative training data, current algorithms are able to perform well on even difficult observational tasks such as fully-automated FACS coding of an unstructured social interaction. However, they also found that algorithms, like human observers, have a range of dependability beyond which their accuracy degrades. Specifically, they found that the accuracy of their algorithm was significantly degraded by extreme head pose (i.e., by participants turning away from the camera). It is thus imperative for researchers who develop or purchase fully-automated measurement tools to gather evidence of validity using *their own* particular data sets.

Statistical Analysis

Analyzing the data from an observational study is relatively straight-forward when measurements are assigned to a small number of time intervals per participant. Things become more complex, however, when behavioral events or a large number of time intervals are used. One common approach is to pool repeated measurements from each participant into *summary scores* such as the mean of each behavioral dimension or the proportion of items assigned to each behavioral category. Groups of participants (or conditions) can then be compared using mean summary scores. For example, Girard et al. (2014) compared depressed participants before and after treatment on the amount of time they contracted different facial muscles during a clinical interview. While this type of approach is simple and intuitive, a number of more sophisticated methods for statistical analysis have been developed and applied to observational data in recent years. Several notable methods are multilevel modeling, sequential analysis, and dynamic systems analysis.

Multilevel modeling (e.g., Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002) enables researchers to account for and ask research questions about the hierarchical structure of "nested data." A hierarchy consists of lower-level observations nested within one or more higher levels. Examples include students nested within classrooms and workers nested within departments; these classrooms and departments may, in turn, be nested within schools and within corporations. Hierarchies are very common in observational data and this structure must be taken into account if analyses are to be accurate; of particular importance here is the nesting of repeated measurements within individual participants. For example, Girard et al. (2015) used multilevel models nesting video frames within participants to examine the influence of frame-level (i.e., illumination and head pose) and participant-level (i.e., gender and ethnicity) variables on the accuracy of a fully-automated facial expression analysis system.

Sequential analysis (Bakeman & Quera, 2011) enables researchers to ask questions about the temporal ordering and contingent relationships between behavioral categories. Importantly, these behavioral sequences may occur within or between individuals. As an early example, Bakeman and Brownlee (1980) found that children tended to transition from parallel activity to group play at rates greater than would be expected by chance, suggesting that children "size up" potential playmates before committing to group play. More recently, Knobloch-

Fedders et al. (2014) examined behavioral sequences between romantic partners, finding that relationship quality was negatively associated with sequences of demanding behavior from one partner being met with either withdrawing or submissive behavior from the other partner. These important relationships between behaviors would have been missed by non-sequential analyses.

Finally, dynamic systems analyses (Salvatore & Tschacher, 2012) enable researchers to ask questions about the relationship of a behavioral process to time; these include analyses of periodicity (i.e., repeating cycles), nonlinear change over time, deterministic chaos (i.e., quasiperiodic cycles), and self-organization. Of particular interest in observational measurement are the “attractor states” that a dynamic behavioral system tends to return to when perturbed and the “phase transitions” that it goes through when reorganizing. Examples of dynamic systems analyses involving observation of one and two individuals are provided for illustration. Hayes and Yasinski (2015) found that more variability in patients’ thoughts, behaviors, emotions, and somatic functioning in the later phase of cognitive therapy for personality disorders predicted more symptom reduction at termination, suggesting that destabilization of old patterns may be necessary for new, healthier patterns to develop. Ramseyer and Tschacher (2011) used dynamic systems analyses to model the nonverbal interaction of patients and psychotherapists as a self-organizing system characterized by the emergence of body movement synchrony; overall, they found that more synchrony predicted higher relationship quality and symptom reduction.

Future Directions

We are entering an exciting new era of behavioral science in which computer-assisted and fully-automated measurement tools are beginning to yield unprecedented increases in the efficiency and scalability of observational measurement. We would like to highlight several research directions that are particularly important to pursue in this new era.

First, observational researchers can improve the rigor and comprehensiveness of their validation methods by reaching beyond inter-observer reliability. Although inter-observer reliability is an important source of validity evidence, it is by no means sufficient. Evidence from content, response processes, and hypothesized relationships among variables is sorely needed. In particular, observational researchers can begin by demonstrating that their observational measurements of a given construct are related to accepted measures of similar constructs and are unrelated to accepted measures of distinct constructs. Ongoing validation is especially important for fully-automated measurement tools, which may have a restricted range of dependability based on their training data.

Second, observational researchers can help to standardize the use of popular measurement instruments and detect group drift by increasing the sharing and comparing of observational data (i.e., audiovisual records and measurements) across research groups. The facial expression analysis community has advanced several relevant practices that are worth replicating in other areas of observational research. One is the ‘certification test’ (e.g., the FACS Final Test; Ekman & Friesen, 1978), which provides a standardized set of observational data for trainees to demonstrate their proficiency on. Another is the

'community challenge' (e.g., the FERA Challenge; Valstar et al., 2015), which provides a standardized set of observational data and validation methods for researchers to use in comparing the performance of their fully-automated measurement tools.

Third, observational researchers can continue to improve fully-automated measurement tools by increasing their range of dependability and leveraging contextual and multimodal information. Of particular importance is for researchers to demonstrate that a given algorithm can maintain its accuracy when applied to data sets very different from the one(s) it was trained on. Researchers can also use these tools to push the envelope of what's possible in observational research. For instance, algorithms may be capable of quantifying subtle changes in behavior that human observers struggle with, such as the amplitude and velocity of motion. In the facial expression analysis literature, such properties have already demonstrated utility in differentiating different types of smiles (Ambadar, Cohn, & Reed, 2009; Schmidt, Ambadar, Cohn, & Reed, 2006) and different mental health states (Juckel et al., 2008; Mergl, Mavrogiorgou, Hegerl, & Juckel, 2005). Hybrid tools that automate some aspects of observational measurement and leave the rest to humans are another promising avenue of research (e.g., De La Torre, Simon, Ambadar, & Cohn, 2011).

Finally, more research is needed to explore how specific behavioral signs (e.g., muscle movements and gestures) relate to intended and interpreted behavioral messages (e.g., affective and cognitive states). Of particular importance is establishing the specificity and generality of any identified relationships (Cacioppo & Tassinary, 1990). Research on how signs are interpreted by observers can be an excellent starting place, but studies of when (and why) signs are produced by participants are also necessary. Only through diligent measurement and careful examination of well-designed observational data can we decode the complex world of meaning contained in behavior.

Acknowledgments

Preparation of this article was supported in part by the National Institute of Mental Health of the National Institutes of Health under Award Number MH096951 and Army Research Laboratory Collaborative Technology Alliance Program under Cooperative Agreement W911NF-10-2-0016. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health or the Army Research Laboratory.

References

Adair JG. The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology. 1984; 69(2):334–345.

AERA, APA & NCME. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.

Ambadar Z, Cohn JF, Reed LI. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of Nonverbal Behavior. 2009; 33(1): 17–34. [PubMed: 19554208]

Bakeman R, Brownlee JR. The strategic use of parallel play: A sequential analysis. Child Development. 1980; 51(3):873–878.

Bakeman, R., Quera, V. Squential analysis and observational methods for the behavioral sciences. New York, NY: Cambridge University Press; 2011.

Baker JK, Haltigan JD, Brewster R, Jaccard J, Messinger DS. Non-expert ratings of infant and parent emotion: Concordance with expert coding and relevance to early autism risk. International Journal of Behavioral Development. 2010; 34(1):88–95. [PubMed: 20436947]

Bennett EM, Alpert R, Goldstein AC. Communication through limited response questioning. The Public Opinion Quarterly. 1954; 18(3):303–308.

Boice R. Observational skills. Psychological Bulletin. 1983; 93(1):3–29. [PubMed: 6828601]

Cacioppo JT, Tassinary LG. Inferring psychological significance from physiological signals. The American Psychologist. 1990; 45(1):16–28. [PubMed: 2297166]

Campbell, DT., Stanley, JC. Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally; 1966.

Cizek GJ. Validating test score meaning and defending test score use: Different aims, different methods. Assessment in Education: Principles, Policy & Practice. 2015:1–14.

Coan, JA., Gottman, JM. The specific affect coding system (SPAFF). In: Coan, JA., Allen, JJB., editors. Handbook of emotion elicitation and assessment. Oxford University Press; USA: 2007. p. 267-285.

Cohn, JF., De la Torre, F. Automated face analysis for affective computing. In: Calvo, RA.D'Mello, SK.Gratch, J., Kappas, A., editors. Handbook of affective computing. New York, NY: Oxford; 2014.

Cohn, JF., Ekman, P. Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In: Harrigan, JA.Rosenthal, R., Scherer, KR., editors. The new handbook of nonverbal behavior research. New York, NY: Oxford University Press; 2005. p. 9-64.

Cowie R, McKeown G, Douglas-Cowie E. Tracing emotion: an overview. International Journal of Synthetic Emotions. 2012; 3(1):1–17.

De La Torre F, Simon T, Ambadar Z, Cohn JF. Fast-FACS: A computer-assisted system to increase speed and reliability of manual FACS coding. Lecture Notes in Computer Science. 2011; 6974(PART 1):57–66.LNCS

Dibeklioglu H, Hammal Z, Yang Y, Cohn JF. Multimodal detection of depression in clinical interviews. Proceedings of the acm international conference on multimodal interaction. 2015

Ekman, P., Friesen, WV. Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press; 1978.

Ekman, P., Friesen, WV., Hager, J. Facial action coding system: A technique for the measurement of facial movement. Salt Lake City, UT: Research Nexus; 2002.

Feng GC. Factors affecting intercoder reliability: A Monte Carlo experiment. Quality & Quantity. 2013; 47(5):2959–2982.

Fleiss, JL. Statistical methods for rates and proportions. 2nd. New York, NY: John Wiley & Sons; 1981.

Geisinger KF. The metamorphosis of test validation. Educational Psychologist. 1992; 27(2):197–222.

Girard JM. CARMA: Software for continuous affect rating and media annotation. Journal of Open Research Software. 2014; 2(1):e5.

Girard JM, Cohn JF, Jeni LA, Sayette MA, De la Torre F. Spontaneous facial expression in unscripted social interactions can be measured automatically. Behavior Research Methods. 2015; 47(4):1136–1147. [PubMed: 25488104]

Girard JM, Cohn JF, Mahoor MH, Mavadati SM, Hammal Z, Rosenwald DP. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. Image and Vision Computing. 2014; 32(10):641–647. [PubMed: 25378765]

Gottman JM, Levenson RW. A valid procedure for obtaining self-report of affect in marital interaction. Journal of Consulting and Clinical Psychology. 1985; 53(2):151–160. [PubMed: 3998244]

Gunes H, Schuller BW. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing. 2013; 31(2):120–136.

Gwet, KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. 4th. Gaithersburg, MD: Advanced Analytics; 2014.

Hayes AM, Yasinski C. Pattern destabilization and emotional processing in cognitive therapy for personality disorders. Frontiers in Psychology. 2015

Horvath, AO., Bedi, RP. The alliance. In: Norcross, JC., editor. Psychotherapy relationships that work: Therapists contributions and responsiveness to patients. New York, NY: Oxford University Press; 2002. p. 37-69.

Juckel G, Mergl R, Prassl A, Mavrogiorgou P, Witthaus H, Moller HJ, Hegerl U. Kinematic analysis of facial behaviour in patients with schizophrenia under emotional stimulation by films with “Mr. Bean”. European Archives of Psychiatry and Clinical Neuroscience. 2008; 258(3):186–191. [PubMed: 18071625]

Knobloch-Fedders LM, Critchfield KL, Boisson T, Woods N, Bitman R, Durbin CE. Depression, relationship quality, and couples’ demand/withdraw and demand/submit sequential interactions. Journal of Counseling Psychology. 2014 Apr; 61(2):264–279. [PubMed: 24749515]

Kosso, P. A summary of the scientific method. Springer Science & Business Media; 2011.

Kreft, IGG., de Leeuw, J. Introducing multilevel modeling. Thousand Oaks, CA: Sage Publications; 1998.

Levenson RW, Ruef AM. Empathy: a physiological substrate. Journal of Personality and Social Psychology. 1992; 63(2):234–246. [PubMed: 1403614]

Lizdek I, Sadler P, Woody E, Ethier N, Malet G. Capturing the stream of behavior: A computer-joystick method for coding interpersonal behavior continuously over time. Social Science Computer Review. 2012; 30(4):513–521.

Mason W, Suri S. Conducting behavioral research on AmazonâĂŹs Mechanical Turk. Behavior Research Methods. 2012; 44(1):1–23. [PubMed: 21717266]

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996; 1(1):30–46.

Mergl R, Mavrogiorgou P, Hegerl U, Juckel G. Kinematical analysis of emotionally induced facial expressions: a novel tool to investigate hypomimia in patients suffering from depression. Journal of Neurology, Neurosurgery, and Psychiatry. 2005; 76(1):138–140.

Messick, S. Validity. In: Linn, RL., editor. Educational measurement. 3rd. New York, NY: Macmillan; 1989. p. 13-103.

O’Leary, KD., Kent, RN. Behavior modification for social action: Research tactics and problems. In: Hamerlynck, LA.Handy, LC., Nash, EJ., editors. Behavior change: Methodology, concepts and practice. Champaign, IL: Research Press; 1973. p. 69-96.

Pantic M, Rothkrantz LJM. Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE. 2003; 91(9):1370–1390.

Ramseyer F, Tschacher W. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. Journal of Consulting and Clinical Psychology. 2011 Jun; 79(3):284–95. [PubMed: 21639608]

Raudenbush, SW., Bryk, AS. Hierarchical linear models: Applications and data analysis methods. 2nd. Thousand Oaks, CA: Sage; 2002.

Reid JB. Reliability assessment of observation data: A possible methodological problem. Child Development. 1970; 41(4):1143–1150.

Salvatore S, Tschacher W. Time dependency of psychotherapeutic exchanges: The contribution of the Theory of Dynamic Systems in analyzing process. Frontiers in Psychology. 2012; 3(253):1–14. [PubMed: 22279440]

Scheffler LW. Being is believing: Playing the psychiatric patient. Journal of Psychiatric Education. 1977; 1(1):63–67.

Schmidt KL, Ambadar Z, Cohn JF, Reed LI. Movement Differences between Deliberate and Spontaneous Facial Expressions: Zygomaticus Major Action in Smiling. Journal of Nonverbal Behavior. 2006; 30(1):37–52. [PubMed: 19367343]

Schwarz N, Knauper B, Hippler HJ, Noelle-Neumann E, Clark L. Rating scales: Numeric values may change the meaning of scale labels. Public Opinion Quarterly. 1991; 55:570–582.

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979; 86(2):420–428. [PubMed: 18839484]

Stevens SS. On the theory of scales of measurement. Science. 1946; 103:677–680.

Stone, AA., Turkkan, JS., Bachrach, CA., Jobe, JB., Kurtzman, HS., Cain, VS. The science of self-report: Implications for research and practice. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

Valstar MF, Almaev T, Girard JM, Mckeown G, Mehu M, Yin L, Cohn JF. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. Proceedings of the ieee international conference on automatic face & gesture recognition. 2015

Van Someren, MW., Barnard, YF., Sandberg, JAC. The think aloud method: A practical guide to modelling cognitive processes. London: Academic Press; 1994.

Zhao, X., Liu, JS., Deng, K. Assumptions behind intercoder reliability indices. In: Salmon, CT., editor. Communication yearbook. Routledge; 2012. p. 418-480.

Table 1

Definitions for different approaches to observational measurement

Categorical Measurements	vs.	Dimensional Measurements
Observers choose between a limited number of predefined options (e.g., basic emotions, attachment styles)		Observers choose numerical values along a continuous dimension (e.g., pleasure, arousal, dominance)
Sign-Based Measurements	vs.	Message-Based Measurements
Observers describe the features of behavior in terms of small units (e.g., movements, utterances, positions)		Observers interpret the meaning of behavior using cultural knowledge (e.g., emotions, motives, mental states)
Event-Based Measurements	vs.	Interval-Based Measurements
Observers identify discrete behavioral events and then make measurements (i.e., they find start and stop points)		Observers make measurements at or about predetermined time intervals (e.g., they measure twice per minute)
Expert Observers	vs.	Naïve Observers
Observers who have received extensive training and have met quality-criteria (e.g., have passed a certification test)		Observers who have received little or no training and may be study participants (e.g., self-ratings, crowdsourced-ratings)
Computer-Assisted Measurement	vs.	Fully-Automated Measurement
Humans provide measurements of behavior using computer software (e.g., continuous rating software)		Algorithms provide measurements of behavior after some initial training (e.g., head or eye tracking software)