Published in final edited form as:

Res Nurs Health. 2009 August ; 32(4): 465–474. doi:10.1002/nur.20334.

Methods to Improve Reliability of Video Recorded Behavioral Data

Kim Kopenhaver Haidet, PhD, RNC[Assistant Professor],

School of Nursing, 307 Health & Human Development East, The Pennsylvania State University, University Park, PA 16802

Judith Tate, MSN, RN[Research Project Manager / Doctoral Candidate],

School of Nursing, University of Pittsburgh, Pittsburgh, PA

Dana Divirgilio-Thomas, MPH[Research Specialist / Data Manager],

School of Nursing, University of Pittsburgh

Ann Kolanowski, PhD, RN, FAAN[Elouise Ross Eberly Professor of Nursing], and

School of Nursing, The Pennsylvania State University, University Park, PA

Mary Beth Happ, PhD, RN[Associate Professor]

School of Nursing, University of Pittsburgh

Abstract

Behavioral observation is a fundamental component of nursing practice and a primary source of clinical research data. The use of video technology in behavioral research offers important advantages to nurse scientists in assessing complex behaviors and relationships between behaviors. The appeal of using this method should be balanced, however, by an informed approach to reliability issues. In this paper, we focus on factors that influence reliability, such as the use of sensitizing sessions to minimize participant reactivity and the importance of training protocols for video coders. In addition, we discuss data quality, the selection and use of observational tools, calculating reliability coefficients, and coding considerations for special populations based on our collective experiences across three different populations and settings.

Behavioral observation is a fundamental component of nursing practice and a primary source of clinical research data. The use of video technology to capture behavioral observations is becoming more prevalent in nursing research because of the advantage it provides in the ability to replay and review observational data, the control of observer fatigue or drift, the ability to achieve levels of observation and analysis not afforded by real-time observations, and the relative ease of using modern sophisticated recording equipment (Caldwell & Atwal, 2005; Heacock, Souder, & Chastain, 1996). The appeal of this method needs to be balanced, however, by an informed approach to reliability issues that include those universal to all observation methods and those unique to video recording.

In this paper, we describe techniques and strategies to ensure the reliability of behavioral data obtained via video recording. We emphasize non-technical factors influencing the reliability of such recordings. Our discussion is based on our collective experiences conducting the following studies involving three different subject populations and settings: *Improving Communication with Non-speaking ICU Patients; Symptom Management (PI: MBH)*, *Patient-Caregiver Communication*, and *ICU Outcomes (PI: MBH)*; *Non-pharmacological Interventions for Behavioral Symptoms in Nursing Home Residents with*

Dementia (PI: AK); and Patterns of Stress Reactivity in Low Birth Weight Preterm Infants (PI: KKH).

Selecting Video Recording As a Method of Data Collection

The decision to use video recording for data collection should be informed by the purpose and aims of the study as well as the unit of behavior being measured (Caldwell & Atwal, 2005). Video recordings are an excellent source of data that can be used to assess relationships between behaviors that occur in close temporal proximity to one another. For example, data from video recordings have demonstrated relationships among infant handling in the neonatal intensive care unit (NICU), facial expressions of the neonate, and simultaneously collected heart rate variability data (Haidet, Susman, West, & Marks, 2005). Video recording as a means of data capture has also been used in a study focused on testing the relationship between morning-care processes and pain expression in nursing home residents with dementia (Sloane et al., 2007). Video recording, however, may not be appropriate for all types of behavioral observations in clinical settings. The intrusiveness of the methodology may alter naturally occurring behaviors and, thus, limit its appropriateness in certain situations, such as interactions during emotionally-charged events or provision of direct care.

Video recording provides a high degree of reproducibility when measuring observations. Unlike real-time data collection, video recordings can be re-played any number of times. On the other hand, video recording provides only a window of what happens in real time and may lack important contextual data (Latvala, Vuokila-Oikkonen, & Janhonen, 2000). Even with multiple, strategically-placed cameras, it may be difficult to capture the gestalt of a targeted behavioral observation, such as team collaboration during an emergency in a busy clinical setting.

The researcher needs to take all these factors into consideration when planning the study protocol. If the decision is in favor of using video recording for data collection, there are a number of ways to ensure the reliability of those data.

Methods To Improve Raw Data Quality

Participant reactivity is a persistent issue in any observational research, but the use of cameras means the presence of an additional “eye” (Gross, 1991). Reactivity is defined as a response between the researcher and participant during data collection that affects the natural course of behavior as a result of being observed (Paterson, 1994). The “Hawthorne effect” is a classic example of reactivity that can be decreased by exposing participants to longer periods of observation to acclimate them to the presence of the observer. This can be accomplished by conducting a sensitizing session or by analyzing data taped later in a predetermined “middle” segment of the session (e.g., after the first 3-5 minutes) (Elder, 1999). For example, Happ, Sereika, Garrett, and Tate (2008) conducted a sensitizing session simply by collecting one more video recording of nurse-patient interactions in the ICU than necessary for measurement. The researchers did not use the first of five video-recorded sessions for outcome measurement. These strategies enable the participant to move beyond an initial period of self-consciousness and the observer to capture a more representative sample of behavior (Rosenstein, 2002). Alternately, researchers might minimize participant reactivity by decreasing the presence of the additional “eye” through the use of a smaller or concealed camera (Gross).

Good planning in the observational protocol and training of research assistants are vital. Because of improvements in camera technology and the employment of trained video

technicians as research assistants, we have minimal missing or lost data due to poor camera angle, camera malfunction, or poor picture quality. Skilled video technicians can be found within a university, particularly if the institution has film majors. Local technical or art schools may provide suitably skilled personnel. If trained video technicians are not available, even data collectors with limited or no videography skills can be trained to meet study standards because video technology has eliminated the need for advanced skills to obtain high quality data. Training unskilled video data collectors can be accomplished with a checklist of both operational and video quality standards prior to entering the field. We recommend a series of guided practice sessions with the camera and testing to predetermined performance criteria of all research assistants who will be responsible for operating the camera.

The length of video-recorded observation needed will vary depending on the behavioral unit, research questions, and associated measures. A reasonable representation of the behavior of interest can be obtained by sampling either by time or by event (Casey, 2006). Selection of video segments for measurement should be guided by pragmatic and/or conceptually relevant criteria that can be consistently implemented across research participants and data samples. A predetermined time increment (e.g., beginning after the first 5 minutes of filming) or behavioral target (e.g., when dressing activities begin) can be used as the starting point for measurement. For example, surveillance video used to capture a naturally occurring event of interest will be longer than video recordings directed specifically to (and initiated with) a target event. A study examining gait among adults with Parkinson's disease video recorded subjects walking a predetermined distance. Length of video recordings varied by individual subject speed (Dijkstra, Zijlstra, Scherder, & Kamsma, 2008). Measures can be applied to a time sample or series of time samples (e.g., 1 minute of video recording sampled every 10 minutes). Alternatively, measurement may target a specific event within a longer recording, such as a parent's arrival in the child's hospital room (Dichter, 1999). There will likely be greater variability among surveillance video-recordings of hospitalized children's responses to parental visitation than in the preceding example of a planned test of gait.

Measures of human interaction and communication can be applied to relatively short, dense segments of videotaped observation. In a recent clinical trial, observations of nurse-patient interactions in the ICU were conducted systematically in the morning and afternoon based on the assumption that critical care nurses interact with patients on a regular basis (Happ et al., 2008). This sampling structure also accounted for the likelihood that behaviors may change based on the time of day. Event sampling is appropriate for those phenomena that are less frequently observed, such as bathing procedures in nursing home facilities. Researchers need to carefully define the temporal and spatial boundaries of the phenomena of interest. Once the behavioral unit is established, the next step is to select appropriate measurement instruments.

Use of Observational Tools

Options for quantitative measurement of behaviors from observation span the full range of measurement levels, depending on the tool selected and the needs of the research. For example, researchers may choose to apply a nominal typology or categorization scheme such as "communication styles," or an ordinal Likert-type scale to rate degree of success (Happ et al., 2008). Behavioral measures can be as simple as frequency counts of the occurrence of a target behavior or a continuous measure of the duration of that behavior (Roter, Larson, Fischer, Arnold, & Tulskey, 2000; Spencer, Coiera, Logan, 2003).

The first decision in selection of observational measures is whether the level of measurement requires a microanalytic or macroanalytic approach. Microanalytic approaches can produce fine discrimination between units of analysis and a variety of information, but are very time consuming, resource-intensive, and produce large datasets. Light (1988) recommended microanalysis for single-subject case designs or small-sample studies. Macroanalysis, or analysis of an entire interactional process as a whole, using global measures may be more appropriate for studies evaluating the effectiveness of an intervention.

Observational measures employ predetermined coding or rating schemes to quantify behavior (Morrison, Phillips, & Chae 1990). Observational tools designed for use with video recording allow researchers to use manualized procedures and decision rules to assign codes to nonverbal behaviors, verbal expressions, or vocal utterances. For example, the Facial Action Coding System (FACS) is a microanalytic system for rating facial movements and emotional facial expressions (Ekman, Friesen, & Hager, 2002). FACS coders receive specialized instruction involving 100 hours of training and practice with final competency test and certification as a trained coder (Description of Facial Action Coding System [FACS], n.d.). Computer software available for conducting microanalysis of video-recorded behavioral interactions include Canopus ProCoder3™ (Thompson Grass Valley 2008, Inc., Burbank, CA), Noldus™ (Noldus Information Technologies, 2008, Wageningen, The Netherlands), Transana™ (Wisconsin Center for Education Research, University of Wisconsin-Madison), and Studiocode™ (Studiocode Business Group, Warriewood, Australia).

Observational checklist tools intended for use in direct observation can also be applied to video recordings, usually at the macroanalytic level. The QUALCARE Scale (Phillips, Morrison, & Chae, 1990), the Discomfort Scale in Dementia of the Alzheimer's Type (DC-DAT; Hurley, Volicer, Hanrahan, Houde, & Volicer, 1992), the Pain Assessment in Advanced Dementia (PAINAD) scale (Warden, Hurley, & Volicer, 2003), and the Premature Infant Pain Profile (PIPP) (Stevens, Johnston, Petryshen, & Taddio, 1996) are examples of such observational checklists. The QUALCARE Scale was developed for direct measurement of physical, psychological, and environmental dimensions of caregiving for dependent older adults (Phillips et al.). The DC-DAT; measures discomfort (Hurley et al), and the PAINAD measures pain by observation of nonverbal patients with advanced Alzheimer's disease (Warden et al.). The PIPP was designed to measure the physiological and behavioral indicators of pain in premature infants (Stevens et al.). Video exemplars were used to train coders in the use of the DC-DAT (Hurley et al.), QUALCARE Scale (Morrison et al., 1990), and the PIPP (Stevens, 1999).

Not all direct observation tools can be used with video data. This is particularly true of instruments that require the observer to capture contextual observations. For example, in the Passivity in Dementia Scale (PDS), one subscale is: interaction with the environment (Colling, 1999), but it is often impossible to see this interaction through the lens of a camera (i.e., resident responding to an off camera resident. One of us (A.K.) found that the PDS subscale had low reliability when used with video data in contrast to real-time observation. This experience highlights the importance of pilot testing instruments that will be used to code video data and to compute their reliability coefficients when used with video data so that appropriate selection of instruments can be made.

When measuring phenomena for which there are no established observational tools, researchers may choose to create adaptations of checklist tools that are typically applied by clinicians or self-rated by patients. For example, items from the revised Condensed-Memorial Symptom Assessment Scale (C-MSAS; Chang, Hwang, Feuerman, Kasimis, &

Thaler, 2000) are currently being applied to video observations of nurse-patient communication in the ICU to measure symptom assessment and management (K24-NR010244). Nelson et al. (2001) revised the C-MSAS for use with nonspeaking critically ill patients. In this revision, patients endorse the presence or absence of each symptom on a checklist of 14 physical and emotional symptoms (e.g., shortness of breath, pain, thirst, worry) followed by rating the level of distress from each physical symptom or frequency of each emotional symptom experienced. In applying this symptom checklist to video observation of nurse-patient communication, only symptom presence as identified in the observed communication interaction is counted. Judgments about symptom distress or intensity are not made from the video recording. Although the previously established symptom constructs are foundational to the application of the C-MSAS checklist to video observation, new psychometric testing should be conducted for this application. In conjunction with employing appropriate observational tools for measurement, it is imperative to establish an effective training regimen for video coders.

Training, Re-training, and Retaining Video Coders

In order to achieve acceptable levels of reliability between coders, a standardized training procedure should be implemented to provide instruction on the use of instruments and the video coding process (Castorr et al., 1990). We developed detailed training manuals that provide specific instructions for using each instrument. The training includes operational definitions of terms and specific interpretation of the behaviors to be rated. Training reduces variability in scoring by minimizing personal interpretation (Curyto, Van Haitsma, & Vriesman, 2008; Washington & Moss, 1988) and the effects of prior experience with the study population (Kaasa, Wessel, Darrah, & Bruera, 2000).

Some projects, do not require any specific credentials or experience for video coders. In fact, naïve coders are often the best research assistants because they may not have preconceived notions about which behaviors are “appropriate” in specific situations, and would, therefore, be more likely to note these behaviors. In other cases, clinical experience would be an asset, if not a necessity, when coding highly technical procedures (Treloar et al., 2008). Regardless of background, research assistants need detailed study information to achieve reliability performance in the field (Cambron & Evans, 2003). The content and length of training programs will vary according to the complexity of the protocol and the reliability of the instruments used to code data. In general, instruments with lower reliability coefficients will require that video coders have more training in their use before acceptable levels of reliability are achieved.

We begin our training by discussing the difference between an objective behavioral observation and a behavioral inference. We illustrate this critical difference by describing situations in which the context could influence coders' rating of behaviors: for example, inferring that a positive affective behavior is being expressed by a subject who is sitting next to individuals who are smiling and laughing, without noting the actual facial expression of the subject. Once this concept is grasped, coders commit to memory the behaviors and the definitions of behaviors that are included on each instrument they use. Additionally, we instruct our coders to access the data collection manuals we develop that describe in detail the behavioral units we are measuring, and we include short definitions of these behaviors on our data collection forms as memory prompts.

We then practice using each instrument with actual video data from the project. This is initially done in a group setting so we can discuss differences across coders. When coders feel confident in identifying behaviors, we ask them to rate video data individually and compare their ratings to a pre-established gold standard. Practice continues until coders

achieve 80% agreement with the gold standard. The cut point for competency reliability ($\geq .80$) is commonly used as acceptable inter-rater reliability in research involving observational coding from videos (de los Rios Castillo & Sosa, 2002; Morse, Beres, Spiers, Mayan, & Olson, 2003; Topf, 1986), however, some researchers use $>70\%$ inter-observer agreement as acceptable for new instruments (Le May & Redfern, 1987).

Across all studies, we train video coders to avoid fatigue when coding tapes and to take frequent breaks to counteract drift. Coders are also asked to avoid coding tapes on days when they are emotionally upset or ill, as their own emotional state can influence the behavioral observations they make. To avoid intra-rating bias during video coding, we assign different coders to rate different behaviors in the same subject. Calibration reliability checks are completed on 10% of recordings throughout the project on each video coder. If reliability falls below 80%, retraining takes place. In one of our projects, routine booster sessions are conducted in which all raters are brought together when there is a lapse of 1 month or more in video coding because of slow enrollment. The purpose of these sessions is briefly to review behavioral definitions and their expression in order to prevent low reliability due to loss of rating skill.

Because the training and re-training of video coders is so intense in terms of time and labor, we minimize video coder attrition and subsequent fluctuations in reliability by compensating video coders for their training after 6 months of service to the project. We have found this to be a very effective and acceptable method for retention of our well-trained video coders. In addition to the establishment of an effective training protocol and acceptable levels of reliability between video coders, it is important to estimate reliability coefficients using standard methods.

Computation of Reliability Coefficients

In order to establish reliability of the instrument, it is important to measure internal consistency, and consistency of measures across time when rated by different individuals. Use of Cronbach's alpha allows the researcher to estimate the average correlation of each scale item with the other scale items used (Kerlinger & Lee, 2000). The reliability of observation systems is most often defined as agreement among two or more independent raters of behavior, also known as inter-rater reliability (IRR). Kappa-type statistics, coefficient kappa (also known as Cohen's kappa) and intraclass correlation (ICC) are commonly used to quantify IRR (Hulley, Cummings, Browner, Grady, & Newman, 2007).

Kappa is an estimation of the degree of consensus between raters. The calculated value will be in the range of -1 to +1 (Landis & Koch, 1977). If the raters are in perfect agreement, kappa equals 1, in contrast to perfect disagreement between raters where kappa equals -1. A kappa of zero represents the same likelihood that the agreement between raters is equal to chance. An acceptable coefficient kappa (good agreement) would be a value above .75, moderate agreement is between .4 – .75, and low agreement is below .4 (Bellieni et al., 2007; Laschinger, 1992). Kappa is calculated using the following equation:

$$\frac{\text{Observed agreement (\%)} - \text{Expected agreement (\%)}}{100\% - \text{Expected agreement (\%)}}$$

Coefficient kappa allows for quantification of IRR as well as rater consistency, and provides a more precise determination of the reliability estimate than a simple correlation coefficient (Hulley et al., 2007).

Reliable Video Coding in Special Populations

In addition to general issues for improving the reliability of video-recorded behavioral data, there are specific issues that relate to special populations.

Critically Ill Patients

There are special conditions to consider when coding data from critically ill and non-speaking participants. In the case of video-recorded communication interactions with intensive care unit (ICU) patients, several factors contribute to the complexity of the task.

An initial challenge of analyzing video recordings in critical care is to distinguish which of the participants is speaking due to camera angles that may obscure a visual record of the speaker, co-occurring conversations, and the effect of environmental noise. These environmental noises can be a radio, a television set, family members sitting in the room, clinicians outside of the room, equipment noises or monitor alarms. Yet, it is very important to listen for these environmental noises, even if this means listening to the audio several times and in smaller segments, because patients sometimes react to them. For example, a patient in the Study of Patient-Nurse Effectiveness with Assisted Communication Strategies (SPEACS) study reacted to a news current event on the television by mouthing words, but his conversational message was misinterpreted by the nurse who was concerned with the patient's immediate physical comfort needs. Data coders reviewed the video many times to accurately lip read the patient's message using the context of the background noise of the television news program to aid interpretation.

Critically ill patients tend to have oral endotracheal tubes or tracheostomy tubes preventing speech. The tube and fixation tape masks the participant's mouth movements, making it difficult to read the patient's lips. Beards can also contribute to the distortion of oral movements. Furthermore, when participants no longer have their own teeth or exhibit oral tremors, the enunciation of words may be affected, adding an additional layer of difficulty to the analysis.

In addition to mouthing words as a means of communicating, critically ill patients tend to gesture. The patient's ability to use upper motor movements is not always precise; much of the time, movements are gross. Patients' arms and hands can be swollen creating difficulty in deciphering the gestures. The coder must discern whether the patient is truly intending to use voluntary gestural communication or whether the movement is involuntary. The wrists of ICU patients may be restrained, which prevents, limits, or distorts the gesture for communication purposes. Furthermore, the position of the patient (in the bed or a bedside chair), blankets on the patient, and bed rails can also hinder or obstruct the patient's gestures.

Despite these barriers to the interpretation of patient communication in the ICU, videography offers a unique and powerful tool for recording communicative behaviors among nonspeaking critically ill patients. The ability to pause, replay, and study the interaction permits a level of understanding unavailable during real-time observation. We recommend the use of close-up camera shots and wide angle lenses, hand-held cameras to follow the action/actors, and directional microphones to optimize the recording quality. Data collection protocols that reduce extraneous background noise before the start of recording sessions should be considered: for example, muting the television, closing the door, and posting a "video recording in progress" sign on the door to discourage interruptions. Finally, because critically ill patients can have dramatic fluctuations in illness severity, cognition, and attention, we recommend that researchers always consider measures of illness severity or acuity and delirium for use in describing the sample, eligibility screening, and/or as covariates.

Nursing Home Residents with Dementia

A second program of research focused on interventions for improving the behavioral symptoms of nursing home residents with dementia. Important outcomes in this work are displays of both positive and negative affect. A growing literature documents that persons with dementia experience a full range of emotions and not just the more negative ones of anger, anxiety or sadness (Cotrell & Hooker, 2005; Hubbard, Downs, & Tester, 2003; Kolanowski, Hoffman, & Hofer, 2007). Video coders, however, often have trouble distinguishing normal age changes or changes brought on by chronic conditions, medications, or loss of dentition from facial, voice, and postural indicators of negative affect. It is essential that video coders be trained to differentiate, for example, the stooped posture resulting from degenerative arthritis from the stooped shoulders indicating depression; the wrinkled, loose skin around the eyes and horizontal furrows in the forehead due to gravitational effects of aging from the sagging facial muscles and horseshoe furrows in the forehead due to sadness; or, the loud, coarse voice quality due to hearing loss from the increased voice volume indicative of anger (Ekman & Rosenberg, 1997). Although it seems intuitive that video recording coupled with blinded video coders would eliminate bias related to familiarity with the subject, the gain in objectivity may be at the expense of inaccuracy in rating. A baseline period is critical to help video coders differentiate the behavioral signs of affect from the age and disease changes typical in older adults with dementia.

Because fluctuations in affect and other behavioral outcomes are related to cognitive status, it is important to describe the sample along these characteristics. We have also found that variability in displayed negative affect (but not positive affect) increases with increasing severity of disease (Kolanowski et al., 2007). This necessitates more intense measurement models (more frequent collection of data) accurately to capture the day-to-day variability in affect in this population.

Preterm Infants in Neonatal Intensive Care

A third program of research focused on the integration of biological and behavioral responses of preterm infants to environmental stressors associated with nursing care in the NICU. The goal is to investigate patterns of early stress response and stress recovery to determine the extent to which individual patterns of stress response are predictive of negative short-term health consequences. A growing literature supports the use of behavioral indicators for pain/stress measurement in preverbal infants and young children (Ballantyne, Stevens, McAllister, Dionne, & Jack, 1999; Gibbins et al., 2008; Grunau & Oberlander, 1998; Hunt et al., 2004; Hunt et al., 2007; Peters et al., 2003).

The neonatal facial coding system (NFCS) is a behavioral rating scale with demonstrated clinical validity as a measure of pain/stress in infants (Grunau & Oberlander, 1998). The measure consists of 9 distinct facial dimensions defined by upper and lower facial muscle activity. The activities include the following: brow bulge, eye squeeze, nasolabial furrow, open lips, vertical mouth stretch, horizontal mouth stretch, lip purse, taut tongue and chin quiver (Grunau & Oberlander). In the preterm infant, the lower facial muscles develop later than the upper facial muscles; there is a normal progression of development from head to toe. Thus, the preterm infant with limited lower facial muscle development most commonly displays only the upper facial activities (brow bulge, eye squeeze, and nasolabial furrow) during stress or pain evoking events (Peters et al., 2003). Infant behavioral states (quiet sleep, active sleep, transition, quiet awake, active awake) as well as current level of illness are important covariates that may contribute to a diminution or robustness of behavioral responses. We recommend that researchers document behavioral state and acuity of illness at the time of observation.

The Premature Infant Pain Profile (PIPP) is a comprehensive pain/stress scale that includes both behavioral and physiological parameters and has established validity and reliability for use in preterm infants (Bellieni et al., 2007; Stevens, 1999). The PIPP instrument consists of seven indicators, each scored on a 4-point scale. Indicators include the following: gestational age, behavioral state, heart rate and oxygen saturation responses, and upper facial behavioral responses (brow bulge, eye squeeze, nasolabial furrow; Stevens et al., 1996). To determine responses to stimuli, behaviors are recorded during baseline, stimulus, and recovery times. Reliability for facial coding can be established by determining the percentage of agreement per occurrence for each facial activity within 2-minute epochs (Gibbins et al., 2008). Because gestational age is the important contextual indicator for rating distress in preterm infants, it is critical that video coders using PIPP are instructed on the key developmental differences between infants of varied gestational ages (Gibbins et al.). Attention should be directed at establishing intra- and inter-rater agreement for evaluating infants within each of the preterm gestational age categories under observational study (< 27 weeks, 27-29 weeks, 30-33 weeks and 34-36 weeks gestation).

Summary

The use of video technology in behavioral research offers important advantages to nurse scientists in assessing complex behaviors and relationships between behaviors. The reliability of data obtained using this method needs to be weighed against the reliability of using real-time observations. Video recordings of behavioral responses are preferable as they afford the rater with the ability to freeze frame and replay data for review and coding of behavior, thus, enhancing reliability. An additional advantage of video-recorded behavioral observations over real-time observations is the ability of the coder to take breaks as needed and minimize problems associated with observer fatigue and drift. Video-recorded behavioral data also facilitates independent rating of responses by a coder who can be blinded to the scoring of other coders, reducing intra-rater bias. In addition, video recording provides an alternate method to real-time observation when complex multi-level systems comprise the behavioral units, as with complicated physiological and behavioral scoring systems where managing all the data at once in real time would be difficult. Further, this method allows for the implementation of training of naïve observers without clinical expertise and, thus, facilitates rating to be done without preconceived judgment based solely on clinical knowledge.

The limitations associated with the use of video recording in behavioral studies include intrusiveness of the equipment, potential for higher participant reactivity, and potential loss of the larger environmental context outside the view of the lens. Researchers can minimize the difficulties associated with these limitations by using expert technicians to set up and manage the equipment and to employ additional strategies to capture the broader environmental context.

In summary, when the decision has been made to use video recording as a means to collect behavioral data, researchers need to give careful consideration to factors that enhance raw data quality, the selection of appropriate observational tools to use with video data, the development of a strong training program for video coders that includes special consideration of the population studied, and the methods used for calculation of reliability coefficients. An informed approach to video recording as a method of data collection will enhance the quality of data obtained and the findings that ultimately inform nursing practice.

Acknowledgments

Kim Kopenhaver Haidet was supported by a Johnson & Johnson Health Behaviors and Quality of Life: 2006-2007 grant, Autonomic and Behavioral Stress Responses in Low Birth Weight Preterm Infants. Ann Kolanowski was supported by a grant from the National Institute of Nursing Research (R01 NR008910), A Prescription For Enhancing Resident Quality of Life. Mary Beth Happ was supported by NINR (K24 NR010244) Symptom Management, Patient-Caregiver Communication, and Outcomes in ICU; NICHD (R01 HD042988) Improving Communication with NonSpeaking ICU Patients.

References

- Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the Premature Infant Pain Profile in the clinical setting. *Clinical Journal of Pain*. 1999; 15(4):297–303. [PubMed: 10617258]
- Bellieni CV, Cordelli DM, Caliani C, Palazzi C, Franci N, Perrone S, et al. Inter-observer reliability of two pain scales for newborns. *Early Human Development*. 2007; 83(8):549–552. [PubMed: 17161923]
- Caldwell K, Atwal A. Non-participant observation: Using video tapes to collect data in nursing research. *Nurse Researcher*. 2005; 13(2):42–54. [PubMed: 16416979]
- Cambron JA, Evans R. Research assistants' perspective of clinical trials: Results of a focus group. *Journal of Manipulative & Physiological Therapeutics*. 2003; 26(5):287–292. [PubMed: 12819624]
- Casey D. Choosing an appropriate method of data collection. *Nurse Researcher*. 2006; 13(3):75–92. [PubMed: 16594371]
- Castorr AH, Thompson KO, Ryan JW, Phillips CY, Prescott PA, Soeken KL. The process of rater training for observational instruments: Implications for interrater reliability. *Research in Nursing & Health*. 1990; 13(5):311–318. [PubMed: 2236654]
- Chang VT, Hwang SS, Feuerman M, Kasimis BS, Thaler HT. The memorial symptom assessment scale short form (MSAS-SF). *Cancer*. 2000; 89(5):1162–1171. [PubMed: 10964347]
- Colling KB. Passive behaviors in dementia: Clinical application of the need-driven dementia-compromised behavior model. *Journal of Gerontological Nursing*. 1999; 25(9):27–32. [PubMed: 10776141]
- Cotrell V, Hooker K. Possible selves of individuals with Alzheimer's disease. *Psychology and Aging*. 2005; 20(2):285–294. [PubMed: 16029092]
- Curyto KJK, Van Haisma K, Vriesman DK. Direct observation of behavior: A review of current measures for use with older adults with dementia. *Research in Gerontological Nursing*. 2008; 1(1): 1–26.
- de los Rios Castillo J, Sosa JJS. Well-being and medical recovery in the critical care unit: The role of the nurse-patient interaction. *Salud Mental*. 2002; 25(2):21–31.
- Description of Facial Action Coding System (FACS). n.d. Retrieved March 19 2009, from <http://face-and-emotion.com/dataface/facs/description.jsp>
- Dichter, CH. Dissertation Abstracts International. Vol. 59. 1999. Toddler's responses to parental presence in the pediatric intensive care unit. (Doctoral dissertation, University of Pennsylvania, 1999); p. 5784
- Dijkstra B, Zijlstra W, Scherder E, Kamsma Y. Detection of walking periods and number of steps in older adults and patients with Parkinson's disease: Accuracy of a pedometer and an accelerometry-based method. *Age and Aging*. 2008; 37(4):436–441.
- Ekman, P.; Friesen, WV.; Hager, JC. The Facial Action Coding System. 2nd. Salt Lake City, UT: Research Nexus eBook; 2002.
- Ekman, P.; Rosenberg, E. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). New York: Oxford University Press; 1997.
- Elder JH. Videotaped behavioral observations: Enhancing validity and reliability. *Applied Nursing Research*. 1999; 12(4):206–209. [PubMed: 10589109]
- Gibbins S, Stevens B, Beyene J, Chan PC, Bagg M, Asztalos E. Pain behaviours in extremely low gestational age infants. *Early Human Development*. 2008; 84(7):451–458. [PubMed: 18243593]
- Gross D. Issues related to validity of videotaped observational data. *Western Journal of Nursing Research*. 1991; 12(5):658–663. [PubMed: 1949769]

- Grunau RE, Oberlander T. Bedside application of the Neonatal Facial Coding System in pain assessment of premature neonates. *Pain*. 1998; 76(3):277–286. [PubMed: 9718246]
- Haidet KK, Susman EJ, West SG, Marks KH. Biobehavioral responses to handling in preterm infants. *Journal of Pediatric Nursing*. 2005; 20(2):128.
- Happ MB, Sereika S, Garrett K, Tate JA. Use of quasi-experimental sequential cohort design in the Study of Patient-Nurse Effectiveness with Assisted Communication Strategies (SPEACS). *Contemporary Clinical Trials*. 2008; 29(5):801–808. [PubMed: 18585481]
- Heacock P, Souder E, Chastain J. Subjects, data, and videotapes. *Nursing Research*. 1996; 45(6):336–338. [PubMed: 8941303]
- Hubbard G, Downs MG, Tester S. Including older people with dementia in research: Challenges and strategies. *Aging and Mental Health*. 2003; 7(5):351–362. [PubMed: 12959804]
- Hulley, SB.; Cummings, SR.; Browner, WS.; Grady, DG.; Newman, TB. *Designing clinical research*. 3rd. Philadelphia: Lippincott, Williams & Wilkins; 2007.
- Hunt AA, Goldman A, Mastroyannopoulou K, Moffat V, Oulton K, Brady M. Clinical validation of the paediatric pain profile. *Developmental Medicine & Child Neurology*. 2004; 46(1):9–18. [PubMed: 14974642]
- Hunt AA, Wisbeach A, Seers K, Goldman A, Crichton N, Perry L, et al. Development of the paediatric pain profile: Role of video analysis and saliva cortisol in validating a tool to assess pain in children with severe neurological disability. *Journal of Pain and Symptom Management*. 2007; 33(3):276–289. [PubMed: 17349497]
- Hurley AC, Volicer BJ, Hanrahan PA, Houde S, Volicer L. Assessment of discomfort in advanced Alzheimer patients. *Research in Nursing & Health*. 1992; 15(5):369–377. [PubMed: 1529121]
- Kaasa T, Wessel J, Darrah J, Bruera E. Inter-rater reliability of formally trained and self-trained raters using the Edmonton Functional Assessment Tool. *Palliative Medicine*. 2000; 14(6):509–517. [PubMed: 11219881]
- Kerlinger, FN.; Lee, HB. *Foundations of behavioral research*. Boston: Cengage Thomson Learning; 2000.
- Kolanowski A, Hoffman L, Hofer S. Concordance of self-report and informant assessment of emotional well-being in nursing home residents with dementia. *The Journal of Gerontology: Psychological Sciences*. 2007; 62B(1):20–27.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33(1):159–174. [PubMed: 843571]
- Laschinger HK. Intraclass correlations as estimates of interrater reliability in nursing research. *West Journal of Nursing Research*. 1992; 14(2):246–251.
- Latvala E, Vuokila-Oikkonen P, Janhonen S. Videotaped recordings as a method of participant observation in psychiatric nursing research. *Journal of Advanced Nursing*. 2000; 31(5):1252–1257. [PubMed: 10840260]
- Le May, AC.; Redfern, SJ. A study of non-verbal communication between nurses and elderly patients. In: Fielding, P., editor. *Research in the nursing care of elderly people*. Location: Wiley; 1987. p. 171–189.
- Light J. Interaction involving individuals using augmentative and alternative communication systems: State of the art and future directions. *Augmentative and Alternative Communication*. 1988; 4(2):66–82.
- Morrison EF, Phillips LR, Chae YM. The development and use of observational measurement scales. *Applied Nursing Research*. 1990; 3(2):73–86. [PubMed: 2357077]
- Morse JM, Beres MA, Spiers JA, Mayan M, Olson K. Identifying signals of suffering by linking verbal and facial cues. *Qualitative Health Research*. 2003; 13(8):1063–1077. [PubMed: 14556419]
- Nelson JE, Meier DE, Oei EJ, Nierman DM, Senzel RS, Manfredi PL, et al. Self-reported symptom experience of critically ill cancer patients receiving intensive care. *Critical Care Medicine*. 2001; 29(2):277–282. [PubMed: 11246306]
- Paterson BL. A framework to identify reactivity in qualitative research. *Western Journal of Nursing Research*. 1994; 16(3):301–316. [PubMed: 8036805]

- Peters JW, Koot HM, Grunau RE, de Boer J, van Druenen MJ, Tibboel D, et al. Neonatal Facial Coding System for assessing postoperative pain in infants: Item reduction is valid and feasible. *The Clinical Journal of Pain*. 2003; 19(6):353–363. [PubMed: 14600535]
- Phillips LR, Morrison EF, Chae YM. The QUALCARE Scale: Developing an instrument to measure quality of home care. *International Journal of Nursing Studies*. 1990; 27(1):61–75. [PubMed: 2312236]
- Rosenstein B. Video use in social science research and program evaluation. *International Journal of Qualitative Methods*. 2002; 1(3):22–43.
- Roter DL, Larson S, Fischer GS, Arnold RM, Tulskey JA. Experts practice what they preach: A descriptive study of best and normative practices in end-of-life discussions. *Archives of Internal Medicine*. 2000; 160(22):3477–3485. [PubMed: 11112242]
- Sloane P, Miller L, Mitchell C, Rader J, Swafford K, Hiatt S. Provision of morning care to nursing home residents with dementia: opportunity for improvement? *American Journal of Alzheimer's Disease and Other Dementias*. 2007; 22(5):369–377.
- Spencer R, Coiera E, Logan P. Variation in communication loads on clinical staff in the Emergency Department. *Annals of Emergency Medicine*. 2003; 44(3):268–273. [PubMed: 15332070]
- Stevens, B. Pain in infants. In: McCaffery, M.; Pasero, C., editors. *Pain: Clinical manual*. St. Louis, MO: Mosby; 1999. p. 626–669.
- Stevens B, Johnston C, Petryshen P, Taddio A. Premature Infant Pain Profile: Development and initial validation. *Clinical Journal of Pain*. 1996; 12(1):13–22. [PubMed: 8722730]
- Topf M. Three estimates of inter-rater reliability for nominal data. *Nursing Research*. 1986; 35(4):253–255. [PubMed: 3636827]
- Treloar C, Laybutt B, Jauncey M, vanBeek I, Lodge M, Malpas G, et al. Broadening discussions of “safe” in hepatitis C prevention: A close up of swabbing in an analysis of video recordings of injecting practice. *International Journal of Drug Policy*. 2008; 19(1):59–65. [PubMed: 18312820]
- Warden V, Hurley AC, Volicer L. Development and psychometric evaluation of the Pain Assessment in Advanced Dementia (PAINAD) scale. *Journal American Medical Directors Association*. 2003; 4(1):9–15.
- Washington C, Moss M. Pragmatic aspects of establishing interrater reliability in research. *Nursing Research*. 1988; 37(3):190–191. [PubMed: 3368362]