Developing and Modifying Behavioral Coding Schemes in Pediatric Psychology: A Practical Guide

Jill MacLaren Chorney,[1,2,3] PhD, C. Meghan McMurtry,[4] PhD, Christine T. Chambers,[2,5] PhD, and Roger Bakeman,[6] PhD

[1]*Department of Pediatric Anesthesia,* [2]*Centre for Pediatric Pain Research, IWK Health Centre,* [3]*Departments of Anesthesia, Pain Management and Perioperative Medicine, Psychology and Neuroscience, and Surgery, Dalhousie University,* [4]*Department of Psychology, Guelph University, Children's Health Research Institute,* [5]*Departments of Pediatrics and Psychology and Neuroscience, Dalhousie University, and* [6]*Department of Psychology, Georgia State University*

All correspondence concerning this article should be addressed to Jill MacLaren Chorney, PhD, Centre for Pediatric Pain Research (South), 8th Floor, IWK Health Centre, 5850/5980 University Avenue, Halifax, NS, B3K 6R8, Canada. E-mail. Jill.chorney@iwk.nshealth.ca

Received July 27, 2014; revisions received October 17, 2014; accepted October 20, 2014

Objectives To provide a concise and practical guide to the development, modification, and use of behavioral coding schemes for observational data in pediatric psychology. **Methods** This article provides a review of relevant literature and experience in developing and refining behavioral coding schemes. **Results** A step-by-step guide to developing and/or modifying behavioral coding schemes is provided. Major steps include refining a research question, developing or refining the coding manual, piloting and refining the coding manual, and implementing the coding scheme. Major tasks within each step are discussed, and pediatric psychology examples are provided throughout. **Conclusions** Behavioral coding can be a complex and time-intensive process, but the approach is invaluable in allowing researchers to address clinically relevant research questions in ways that would not otherwise be possible.

Key words behavioral coding; observational data; research methods.

The field of pediatric psychology offers a wealth of opportunities for directly observing behavior. As evidenced by this special issue of the *Journal of Pediatric Psychology*, the collection, coding, and analysis of behavioral observations have become common across a variety of populations (e.g., children with cancer [Dunn et al., 2011], chronic pain [Reid, McGrath, & Lang, 2005], spina bifida [Kaugars et al., 2011]) and research contexts (e.g., parent–child interactions [Borrego, Timmer, Urquiza, & Follette, 2004], mealtime communication [Patton, Dolan, & Powers, 2008], lab-based studies [Chambers, Craig, & Bennett, 2002]). Direct observation of behavior is the method of choice when overt behavior is central to a research question (e.g., interactions between children and parents), when self-report is not possible (e.g., in infants), or when self-report is not practical (e.g., during an ongoing interaction). Even if self-report is available, direct observation may be a preferable or complementary method because reporting on one's own behavior may not always be accurate or fully represent actual behavior (e.g., Cohen, Manimala, & Blount, 2000).

The utility of studying behavior in pediatric psychology depends on the ability to measure it. In the same way that a ruler could be used to measure length, or a questionnaire used to measure anxiety, coding schemes can be used to measure overt behavior. There are several books available on coding schemes and observational methods generally (Bakeman & Quera, 2011; Yoder & Symons, 2010) and overviews of behavioral coding as a methodology (Cone, 1999; Hartmann & Wood, 1990), but there is currently no brief and easily accessible review for those wishing to orient themselves to major issues and key decisions. Thus, this article is intended to be a practical guide to both the selection

Journal of Pediatric Psychology 40(1) pp. 154–164, 2015
doi:10.1093/jpepsy/jsu099
Advance Access publication November 21, 2014

and modification of new behavioral coding schemes and the development and refinement of existing ones.

Steps in Developing a New Coding Scheme or Modifying an Existing Coding Scheme

As with any other design, research using behavioral coding begins by defining a research question and determining a measurement and analytic plan. In some cases, an appropriate behavior coding scheme may be available, and if so, the researcher is in a fortunate position. It is not uncommon, however, for researchers to find that there is no available coding scheme to address their research question, or that an existing coding scheme requires modification to fit their context. This article addresses these two issues: When a new coding scheme needs to be developed or when an existing coding scheme requires modification.

Table I outlines the tasks involved in developing or modifying a coding scheme. Although these steps are structured progressively, it is important to note that the actual process of developing or modifying coding schemes is iterative. Decisions made at each step have implications *both* downstream and upstream and sometimes require revisiting previous decisions.

Refine the Research Question

Refining a research question is an important first step in any research study, but the importance of this task cannot be overstated in behavioral coding research. Here, we provide a series of considerations to hone the research question for observational coding.

Determine Who to Code

When refining a research question, researchers must first determine *whose* behavior is relevant to the research question. Participants in pediatric psychology research often include the target child and some combination of relevant others (parents, health care providers, peers).

Determine What Behaviors Are of Interest

Researchers must then determine *what* behaviors are of interest. Behaviors may be relevant for one of two reasons: (1) the behavior matters for its own sake, or (2) the behavior matters because it is representative of, or being used to define, another construct (''methodological operationalism''; Yoder & Symons, 2010). In the first case, behavior matters simply because it is desirable or undesirable for some reason. For example, a child sitting still may be a behavior that matters for its own sake because sitting still is important to accomplish a medical procedure safely. On the other hand, behaviors such as crying, screaming, or facial movements may be of interest because they are being used as indications of pain.

Determine When to Observe

Next, the researcher must identify *when* to observe. In pediatric psychology, the beginning and end of observation sessions are often tied to medical events or other tasks (e.g., mealtime), and many researchers observe for a period of time before, during, and after this event. Alternatively, researchers may choose to use experimental paradigms (e.g., cold pressor task) rather than clinical observations because of the degree of control and internal validity they allow. How long to observe is also important. Depending on the behavior, longer observation times may be required to secure accurate estimates of the frequency, duration or rate of the behavior (Riley-Tillman, Christ, Chafouleas, Boice-Mallach, & Briesch, 2011); thus, times used in the existing literature often provide the best guide.

Determine How to Record

Researchers should next consider *how* to collect their observations. Although in vivo observations can be used for simple coding schemes, audio and video recording are the most common forms of data collection. Reviews on the use of audio and video recording are available (Clemente, 2008; Frank, Juslin, & Harrigan, 2005) and highlight the utility of and considerations in using these methods. Regardless of observation method, researchers should be aware of the issue of reactivity, that is, a change in behavior as a result of the observation itself (Harris & Lahey, 1982). Social desirability effects can be induced by the presence of an observer or the instructions provided (Kazdin, 1982); thus, researchers should attempt to make observations as unobtrusive as possible. Reactivity can also be reduced by extending the observation, thereby allowing participants more time to adapt (Kazdin, 1982).

Determine How Behaviors Will Be Scored

The researcher should then determine how behaviors will be scored. Coding systems may involve either nominal codes (i.e., categories) or rating scales (i.e., ordinal numbers). Nominal codes may be dichotomous (present/absent) or grouped (e.g., not playing/playing beside another child/playing with another child), and can represent intensity with progressively stronger categories (e.g., whimper, cry, scream). Rating scales (e.g., none, minimal, moderate, or maximal) are useful if intensity is of central interest.

Consider the Analytic Plan

Once behaviors of interest are established, researchers should consider their data analysis plan (including sample size justification). Coded data can be analyzed

Table I. *Steps in Developing and Modifying Coding Schemes*

Refine the research question	
Determine who to code	• Whose behaviors are of interest? • Will participants be coded separately or as a class? (e.g., nurse, physician, parent, adult)
Determine what behaviors are of interest	• Based on theory, past research, or clinical experience, what behaviors are relevant to the research question?
Determine when and how to observe	• What defines the beginning and end of the observation period or session? • Are there distinct phases of the observation?
Determine how to record	• Will observations be conducted in vivo or collected by audio or videotape? • How will reactivity be minimized?
Determine how behaviors will be scored	• Is coding (applying nominal codes) or rating (applying ordinal scores) most appropriate? • Do frequency, duration, or timing of behaviors matter?
Consider analytic plan	• What data analytic strategies are most appropriate to answer the research question or questions? • What impact will this have on how you develop and set up your coding?
Consider resource constraints	• What resources are available for coding? • Is software available and familiar? • How many coder hours are available?
Develop or refine the coding manual	
Develop a list of codes	• What codes/labels will be used to represent behaviors of interest? – Physically vs. socially based – Granularity (micro vs. macro)
Develop or refine operational definitions of codes	• How can behaviors be defined based on their observable characteristics? – Consult existing coding schemes, modify to fit context if needed – Conduct field/pilot observations – Consult experts • Using data from the current context, what are some examples of actions that would fit and not fit within each code? • Can mutually exclusive and exhaustive code groups be identified? (especially for continuous sampling)
Determine sampling strategy	• Is instantaneous, interval, or continuous sampling most appropriate? • What metrics would be most relevant (e.g., frequency, duration, order, latency)?
Provide instructions on implementation of the coding scheme	• Will observations be coded from audio, video, transcripts? • How should observations/transcripts be parsed? • How many passes should observers use? • Will coding be completed by hand or using software? • What materials are required?
Pilot and refine the coding manual	
Apply coding scheme to sample observations	• How well do at least two independent coders agree on application of codes? • How can disagreements on code application be resolved? – Do operational definitions need to be refined? – Do new examples, nonexamples, decision rules need to be established?
Consider resource constraints again	• Accounting for increased coding efficiency after training, how long is it likely to take to apply coding scheme for each observation?
Implement the coding scheme	
Define coder requirements and train coders	• What qualifications (professional, technical) do coders require to apply coding scheme with fidelity? • Have coders reached agreement on sample observations coded as a ''gold standard''?

(continued)

Table I. *Continued*

Code data and check agreement	• Who will be the primary and secondary (reliability/agreement) coder(s)? • How many observations will be sampled to test observer agreement? • What measure of agreement will you use?
Examine validity	• What steps were taken to ensure face and content validity of the codes? • What measures should be included in the study to examine concurrent, predictive, convergent, and/or discriminant validity?
Analyze data and report results	• How can planned analyses be applied and reported accurately? • What development and/or modification steps were followed that should be reported in the manuscript?

using inferential statistics appropriate for rates and proportions (Fleiss, 1973) or newer sequential analysis strategies (Bakeman & Gottman, 1997; Chorney, Garcia, Berlin, Bakeman, & Kain, 2010). Researchers should be particularly cognizant of how frequently their behaviors of interest are likely to occur. Ideally, all subjects would display all behaviors of interest at least once, with number of occurrences over participants well distributed. In practice, however, it is not uncommon for many participants never to exhibit some behaviors at all. If this is the case, researchers may need to consider nonparametric analyses.

Consider Resource Constraints

Implementing coding schemes can be a time-consuming process, and it is important to be explicit about available resources from the outset. As a rule of thumb, researchers can generally expect that it will take 1–5 or even 10 times the duration of an observation to code it, depending on the complexity of the system (sampling technique, number of subjects and codes, coding passes required). Of note, this estimate does not account for the time required to train coders, or additional coding required for observer agreement.

Researchers should also consider what equipment is available to facilitate coding. Commercial software packages (e.g., Noldus' The Observer www.noldus.com/human-behavior-research/products/the-observer-xt, Mangold's INTERACT www.mangold-international.com/software/interact/what-is-interact.html) can significantly improve the efficiency of this type of research, but these systems are not free. Free software such as Elan (http://www.lat-mpi.eu/tools/elan/) is another option, but its capabilities are more limited. And for any of these computerized systems, there is a learning curve. Low-cost low-tech alternatives—which may be necessary when resources are limited—include use of pencil and paper or entering data directly into computer files (e.g., using Excel).

Develop or Refine the Coding Manual

Coding schemes and their accompanying manuals generally include four components: a list of codes (i.e., labels), operational definitions for each code, a behavioral sampling strategy, and instructions on how to administer the coding scheme. Each is discussed below.

Develop an Initial List of Relevant Codes

Codes are essentially labels that are used to represent behaviors, and they may vary, among other dimensions, in concreteness and granularity (Bakeman & Gottman, 1997). In terms of concreteness, codes may be physically based (e.g., muscle movements, physical positioning) or more socially constructed (e.g., ''reassurance,'' ''sensitivity''). Physically based codes such as facial actions (Ekman & Rosenberg, 1997) may require somewhat less human judgment, and in principle could be computer automated (e.g., NoldusFaceReader^{TM}), whereas socially based codes are constructed concepts and thus require human judgment. Codes may also vary in their granularity. Fine-grained microcoding captures behaviors at their most specific level (e.g., utterance by utterance), whereas macrocoding involves applying codes to a broader sample of behavior (Bell & Bell, 1989). Microcoding allows for more specificity and flexibility later in data analysis (i.e., codes can be analyzed sequentially or combined), but is time-consuming and may oversimplify behaviors (a classic illustration of ''missing the forest for the trees''). Macrocoding, on the other hand, can be completed much more efficiently and in some cases may capture the larger context of interactions. Compared with microcoding, macrocoding requires a greater degree of human judgment and therefore can be more difficult to collect reliably (Margolin et al., 1998). Some authors have used a combination of macrocoding and microcoding (Rodriguez et al., 2013) to capitalize on their respective strengths.

In terms of defining codes, it is generally best practice to create a list that is mutually exclusive (i.e., each behavior can be assigned only one code) and exhaustive (i.e., there is a code for every behavior). This strategy facilitates coding and increases agreement because coders are required to make only one decision (*Which code do I apply?*) rather than multiple decisions (*Do I need to apply a code or not? Are there other codes that I need to apply?*). Most schemes can be made exhaustive by including an ''other'' code. Of note, if the ''other'' code is used often (more than 5%), however, it may obscure meaningful information, and additional codes should probably be defined. Coding schemes may also include multiple mutually exclusive and exhaustive sets (e.g., one set for verbal behaviors, another for nonverbal behaviors).

If published coding schemes contain codes that fit the research question, then using them may be appropriate, but caution is warranted if attempting to apply those codes in a different context. In a different context, codes may not capture all the behaviors of interest or may capture behaviors that were relevant in the old but not the new context. This is especially likely if the new context includes another language or culture (Pedro, Barros, & Moleiro, 2010).

Develop or Refine Operational Definitions

In addition to labels, codes are characterized by operational definitions and specific examples. Operational definitions describe a concept in terms of its observable properties, and more than one coder should be able to apply them consistently (Ribes-Iñesta, 2003). If an existing coding scheme is available, the researcher may be able to use its operational definitions, but refinement (e.g., adding context-specific examples or minor wording changes for clarity) is often required.

When operational definitions are not available, the researcher is in a more challenging position. Definitions can be generated from the literature, investigator judgment, or expert consensus. A researcher may develop (or refine, as discussed above) operational definitions by watching a sample of observations and asking him/herself ''*What do I see/hear/observe that tells me I should be providing this label right now?*'' and then putting these observations into words. A list of examples of actions that fit the label is also generated. Several other collaborators should then review this definition to evaluate its comprehensibility and applicability, after which definitions may be further refined. Researchers may also use expert consensus to generate operational definitions (Yoder & Symons, 2010). Using this methodology, a researcher shows sample observations to experts (i.e., individuals who have published or practiced in the field) and asks them to rate these sample data as ''low,'' ''medium,'' or ''high'' on the behavior of interest. Ratings are then used as probes to facilitate interviews with experts to determine what observable characteristics they used to make their ratings (e.g., *What did you see in the observation that made you give a ''high'' rating in comparison to the other observation you rated as ''low?''*). These interviews can then be analyzed using content analysis to generate themes that can be used in the operational definitions.

Determine the Sampling/Recording/Rating Method

The researcher must next determine how behavior will be sampled in order to apply codes to available recordings (or live observations). As with all steps in this process, the research question will guide the sampling or recording method. If the order or timing of behaviors is important, then the coding scheme must capture this information. In order of least to most intensive, data can be coded using global ratings, instantaneous sampling, and interval or continuous recording. Figure 1 provides examples of data coded using each method and demonstrates their major differences.

Global Ratings. If macro level impressions of observations are of interest, global ratings can be a time- and cost-efficient option. In this method, numbers represent ordinal ratings of the behavior over a period (often the entire observation duration). Ratings may be based on metrics of frequency, duration, or intensity of behavior. Global ratings are common and useful when outcome (rather than process) is central to the research question (Adamson, Bakeman, Deckner, & Nelson, 2014). Sample publications using global ratings include Kubicek, Riley, Coleman, Miller, & Linder (2013) and Shapiro, McPhee, Abbott, & Sulzbacher (1994).

Instantaneous Sampling. In terms of systematic observation at a micro level, instantaneous sampling requires the least resources but is also least commonly used because of its limitations. Here, the coder captures information only on behaviors that are present at an instant in time, repeats this instant at a sampling rate (e.g., every 10 s), and ignores behavior that occurs outside this instant (Leger, 1977). Instantaneous sampling does not capture information on timing, duration, or frequency, and because it ignores behavior outside the instant, risks underestimating behavior. It can be useful, however, in situations in which coding must be completed in vivo or when other activities must be completed outside the instances. The reportable metric for instantaneous sampling is the proportion of instances in which a code occurred (number of instances coded present/total number of instances). Although publications using this sampling strategy are uncommon, relevant

Data are shown below from a hypothetical observation of a child undergoing an immunization and her parent's behavior during this immunization. Behaviors of interest include child crying and parent's use of reassuring statements (e.g., "you're ok"). The child in this observation cried from 12 seconds through 24 seconds, again from 34 through 35 seconds, and again from 41 through 44 seconds. The adult reassured in this observation once at 17 seconds, again at 27 and 28 seconds, and three more times as 35, 36, and 37 seconds.

Sec	10					15					20					25					30					35					40					45					50
Cry																																								
Reas								X										X	X							X	X	X												

Instantaneous Sampling: Coding of behaviors as present or absent at an instant repeated at a rate of every 10 seconds

Time Instant	Cry	Reassure
10 sec	Absent	Absent
20 sec	Present	Absent
30 sec	Absent	Absent
40 sec	Absent	Absent
50 sec	Absent	Absent

One-Zero Interval Recording: Coding of behaviors as present or absent during 10 second intervals

Time Instant	Cry	Reassure
10-19 sec	Present	Present
20-29 sec	Present	Present
30-39 sec	Present	Present
40-49 sec	Present	Absent

Event Sequential Recording: Order of behaviors coded without time information
Cry, Reassure, Reassure, Reassure, Cry, Reassure, Reassure, Reassure, Cry

Timed-event Sequential Recording: Order and timing of behaviors coded
Cry: 12-24 sec, Reassure: 17 sec, Reassure: 27 sec, Reassure: 28 sec, Cry: 34-35 sec, Reassure: 35, Reassure: 36, Reassure: 37, Cry: 41-44 sec

Figure 1. Instantaneous sampling, interval sampling, and continuous recording.

examples include Pellegrini and Davis (1993) and Walters and Hope (1998).

Interval Sampling. Interval sampling also defines a sampling rate, but in this case, the coder captures behaviors that occur at any time during the sampling interval and the coder is continuously alert. If a target behavior occurs at any point during the interval, the behavior is scored as present (otherwise known as zero-one coding; see Bakeman & Quera, 2011, p. 32). Interval sampling has the benefit of capturing all behavior that occurs, but information on the time of onset, offset, duration, or frequency of behaviors is not captured. Perhaps most notably, interval coding does not differentiate between an ongoing bout of a behavior and one that stops and starts within an interval. In this way, interval sampling can overestimate behavior that occurs for only part of an interval. On the other hand, interval coding can underestimate behavior if multiple instances occur within one interval (and thus are counted only once). The reportable metric for interval sampling is the proportion of intervals in which a code occurred (number of intervals coded present/total number of intervals). Sample publications using interval sampling include Camras et al. (2007) and Cohen, Bernard, McClellan, and MacLaren, (2005).

Continuous Recording. Continuous recording strategies provide the most comprehensive representation of data, but are also the most time and resource intensive. In continuous sampling, the coder is always alert and records any occurrence of a target code. If the researcher is interested only in frequency, continuous coding can be as simple as a count of the number of times a particular code is recorded. Alternatively, researchers have the option of capturing information on the timing, duration, and order of codes using this method. A thorough review of continuous recording strategies can be found elsewhere (Bakeman & Quera, 2011; Chorney et al., 2010), but we provide a brief overview here.

Event-sequential continuous recording generates a single list of codes that represents the order in which behaviors were observed. In this system, accurate information on how often a behavior occurs is maintained, but timing and duration of behaviors are not captured. Because duration is not accounted for, event sequential recording is mostly used for research questions with relatively brief behaviors or those that can be easily parsed into segments (e.g., content of verbal behavior). Event sequential recording has several reportable metrics including frequency, rate (frequency of target code/observation duration), or

proportion (frequency of target code/frequency of all codes or subset of codes). Sample publications using event sequential coding include Blount et al. (1997) and Chambers et al. (2002).

Timed-event sequential continuous coding is the most time- and resource-intensive coding but also provides the most thorough representation of behavior. This type of coding provides information on frequency and order of behavior but also maintains information on timing. This method allows the collection of the start and stop times for all behaviors with meaningful durations (''event codes''). Alternatively, if a behavior does not have a meaningful duration (e.g., an utterance), the coder can also capture data on the onset of a behavior only (''point codes''). In this way, timing information is still maintained. Timed-event data have the most flexibility for reportable metrics. Point codes and event codes can be reported much like event-sequential codes as frequencies, proportions, or rates. Event codes can also be reported by total duration or by proportion of time (total duration of target code/duration of observation). Because event codes may occur in separate episodes (e.g., a child cries for a period, is quiet, and then cries again), authors can also report on frequency of episodes of a target behavior, or average duration of these episodes (total amount of time in which the code was displayed/# of bouts). Lastly, because time data are available, authors may also report latency (amount of time elapsed between two codes). Sample publications using timed-event sequential coding include Adamson et al. (2014) and Chorney et al. (2009).

Provide Instructions on Implementation of the Coding Scheme

There are a few practical points that should be included in instructions on administering the coding scheme. The manual should be explicit on whether data should be coded directly from observation (in vivo, audio, video) and/or from transcripts. The decision about whether to transcribe data is sometimes a practical one. If coders use unclear audio, there may be problems with agreement, not because coders assigned different codes inappropriately to the same content, but because they assigned different codes appropriately to different content (i.e., they heard different things). When transcriptions are used to facilitate unclear audio, transcripts should be checked by another reviewer to ensure fidelity; this can require multiple listeners.

Because behavior naturally occurs in a stream, the coding manual should provide rules on how to divide (or parse) behaviors, especially for continuous recording. Utterance coding is a common parsing rule used for coding verbal behavior. Using this rule, a code is assigned to the smallest unit of speech that has meaning (even if successive utterances are assigned the same code). The decision on how to parse behavior will affect results, and thus should be consistent with previous research and should be applied consistently across all observations (Margolin et al., 1998). Although not required, transcription is sometimes used to facilitate this task.

Additional practical instructions should also be provided, including the materials required to code (location of observations, data storage) and instructions on coding passes. In some cases, coders may need to watch an observation more than once to capture all behaviors of interest, and there should be consistency in which behaviors are coded in which pass.

Pilot and Refine the Coding Scheme

Apply Coding Scheme to Sample Observations

A great deal of work goes into developing the first iteration of a coding scheme, but (as with many other things in research) a first effort is often imperfect. Thus, before full-scale coding, it is important to pilot the manual with a sample of observations drawn from the study context. When sample observations are not available, a subset of the study data (e.g., first five participants) may be used, but if these data are included in analyses, they should be recoded once the scheme has been finalized.

When pilot testing, at least two individuals (one of whom was not involved in developing the codes) should independently code two to five pilot sessions. The coders then meet with the lead researcher (usually the researcher most involved in developing the research question). The three discuss any disagreements, with the coders explaining why they assigned particular codes. During piloting, disagreements often stem from unclear definitions and coders interpreting observed behaviors in different ways. If this is the case, the coders and lead researcher reach consensus about which of the codes best captures a particular behavior and then refine either the operational definition or the examples that reflect their decision. This process is repeated until, based on the latest pilot observation, no additional changes are made.

There are no hard and fast rules on when piloting a coding scheme is finished, but in our experience once coders consistently agree on two to five observations and less than 5% of behavior is captured as ''other,'' it is likely that the scheme can be applied with relative fidelity and represents the full range of potential behavior.

Consider Your Resource Constraints Again

Once initial piloting has been completed, the researcher can better estimate how long it will take to implement the coding scheme. At this point, researchers should revisit their estimate of coding time (due to increased efficiency, often about 25 to 50% less than during piloting) and ensure that the human resources available will meet this need. If the complexity and time demands of the coding scheme clearly outweigh available resources, researchers need to consider simplifying the coding scheme.

Implementing a Coding Scheme

Define Coder Requirements and Train Coders

Once researchers have piloted their coding scheme, the next step is to identify and train coders. Although the rigor built into the coding scheme ensures that these data are as ''objective'' as possible, most systems require some level of judgement; thus, coder background, expertise, and training are of utmost importance. When definitions are socially constructed or require judgment about intent, it may become obvious during piloting that coders require a certain level of professional expertise in the topic area to apply the coding scheme with fidelity. Although requiring professional expertise may be feasible for some studies, it limits the pool of potential coders and should be avoided if possible.

The coding manual should also include instructions for training. Training new coders usually begins with orientation to the study context and familiarization with the coding manual (Margolin et al., 1998). Coders are then provided with a sample observation and may code this observation together with a trainer. This process provides an opportunity for early feedback on coders' thought processes and the chance to highlight decision rules in the coding scheme. Trainee coders then code sample observations on their own and compare these ratings with ''gold standards'' set by consensus (developed using a consensus process as in pilot testing). Discrepancies are discussed with the lead researcher, and explanations should be provided to help the new coder reach the same decisions as the gold standard. Typically, new coders are considered reliably trained when they meet an a priori defined criterion (e.g., a kappa coefficient of at least .80 or 80% agreement, although lower values may be acceptable depending on the number of codes in the scheme; see Bakeman & Quera, 2011). It is worth noting that in rare cases, despite good training and multiple attempts at feedback, some individuals have great difficulty becoming reliable coders. In our experience this occurs with about 1 in 10 people and usually becomes clear relatively early in training.

Code Data and Regularly Check Agreement

Once the coding scheme has been developed, piloted, and refined, and coders have been trained, researchers can begin the task of coding study data. Ideally, a primary coder would code all data, but sometimes this is not possible and multiple coders serve as primary coder. In either case, a single second coder should double code a sample of observations to assess observer agreement and ensure a consistent benchmark across primary coders. Although not always possible, coders should be blinded to study hypotheses and intervention groups (if applicable).

Researchers have several options when considering how to monitor and quantify the degree of observer agreement. The most common options for assessing observer agreement are (1) kappa, (2) percent agreement, and (3) intraclass correlations. Each of these approaches has its own strengths and weaknesses, and some may be more appropriate than others. An extensive discussion of options is beyond the scope of this article; instead, the reader is directed to the following resources that provide an excellent summary of these issues (Bakeman & Quera, 2011; Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Fleiss, 1973).

Sample observations for agreement should be selected at random, and the primary coder should not be aware of which observations will be used. Standard practice in the literature has been to code between 10 and 25% of all observations for observer agreement, but if there is marked variability with some observations having much lower values than desired, the selected subset may not be representative (suggesting a larger subsample needs to be assessed for agreement). Alternatively, there may be an issue with the coding scheme or training.

If reliability cannot be achieved despite retraining and if there is a consistent issue with reliability for several coders, the investigators may need to return to the pilot phase and attempt to further refine the manual. If this is the case and substantial changes are made, earlier data will need to be recoded. In rare instances, reliability may still be difficult to achieve, especially for more socially based judgments. In this case, researchers may consider coding all of the data by consensus (two coders and a third coder to resolve disagreements) or researchers may consider double coding all data and using some summary of these ratings or statistical models to estimate, and thereby control, for interobserver variance (Lei, Smith, & Suen, 2007).

The issue of observer drift and regular assessment of reliability throughout coding is an important one. Although raters are trained to criterion before starting to code study data, there is often drift in how they implement this system over time (Kobak et al., 2007). Reliability should be

monitored at regular intervals and intervention applied if drift is apparent (Warshaw, Dyck, Allsworth, Stout, & Keller, 2001). Researchers may also consider using intrarater reliability checks to assess drift by comparing ratings of the same observation coded by the same coder at two different points in time. Intervention could include providing feedback on disagreements, but this feedback should serve only to remind coders of the criterion coding rules rather than developing new rules that might influence coding midstream.

Examine Validity

A behavioral coding scheme must be both reliable and valid. Face validity (i.e., the extent to which a measure appears to assess the construct of interest) and content validity (i.e., evidence that the content of the items reflect the construct or domain of interest; Kazdin, 2003) should be considered during the stage of code generation, and can be accomplished via literature review, expert consultation, and/or pilot work (e.g., Yoder & Symons, 2010). If a researcher is interested in capturing behavior as a representation of another construct, then construct validity is important. The forms of validity that are most salient include (a) concurrent validity (i.e., relations with criterion measures at the same point in time); (b) predictive validity (i.e., relations with criterion measures in the future); (c) convergent validity (i.e., relations with other measures that assess similar or related constructs); and (d) discriminant validity (i.e., relations with measures that assess dissimilar or unrelated constructs; Kazdin, 2003). It is important that researchers consider ways they could examine the validity of their coding scheme while designing the study; indeed, there may be measures that would be helpful for establishing validity that researchers would not have otherwise included.

Analyze Data and Report Results

As indicated in the planning analysis section, investigators should use analytic methods that are appropriate for their data type and research question. Although outside the scope of this article, the analyses may include descriptive, inferential, or sequential statistics. In terms of reporting results, when using a new or modified coding scheme, researchers should also report details of their development process, including the steps outlined in Table I. Space limitations may preclude full reporting, but many journals now allow supplemental online material or this information should be readily available on request.

Conclusions

Direct behavioral observation methods have led to substantial research and clinical developments in pediatric psychology. These methods allow for the assessment of constructs that would not be available via self-report (e.g., in infants, young children), and even in the presence of self-report, behavioral observations may add complementary information. In pediatric psychology, behavioral observations have advanced the measurement of important outcomes such as pain and distress (e.g., Blount et al., 1997) and have allowed for the study of salient processes such as health care provider–patient communication (e.g., Howells, Davies, Silverman, Archer, & Mellon, 2010) and family interactions (e.g., Dunn et al., 2011). Findings from this research have contributed to our understanding of the scope of problems in pediatric psychology and have led to the development of interventions that have improved child health outcomes (e.g., Wysocki et al., 2000).

This article summarizes the major steps involved in developing and modifying behavioral coding schemes for use in pediatric psychology research, and shares rules of thumb based on our extensive experience with behavioral coding. This approach has applicability to pediatric psychology, and more broadly to other research areas in which direct observation of human behavior is of interest. As is evident from this article, behavioral coding can be a complex and time-intensive process. However, the approach is invaluable in allowing researchers to address clinically relevant research questions in ways that faithfully reflect the behavior of interest.

Acknowledgments

The authors would like to acknowledge the contribution of the collective expertise of their many collaborators, trainees, and coders to this manuscript.

Funding

J.M.C. is a Canadian Institutes of Health Research (CIHR) New Investigator, and her research is funded by CIHR, the Nova Scotia Health Research Foundation (NSHRF), and the Canadian Foundation for Innovation (CFI). CMM's research is funded by the CFI, CIHR, and the University of Guelph. C.T.C. was a Canada Research Chair, and her research is funded by CIHR and CFI.

Conflicts of interest: None declared.

References

Adamson, L. B., Bakeman, R., Deckner, D. F., & Nelson, P. B. (2014). From interactions to conversations: The development of joint engagement during early childhood. *Child Development, 85*, 941–955.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York, NY: Cambridge University Press.

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York, NY: Cambridge University Press.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics, 27*, 3–23.

Bell, D. C., & Bell, L. G. (1989). Micro and macro measurement of family systems concepts. *Journal of Family Psychology, 3*, 137–157.

Blount, R., Cohen, L., Frank, N., Bachanas, P., Smith, A., Manimala, M., & Pate, J. (1997). The child-adult medical procedure interaction scale-revised: An assessment of validity. *Journal of Pediatric Psychology, 22*, 73–88.

Borrego, J. J., Timmer, S. G., Urquiza, A. J., & Follette, W. C. (2004). Physically abusive mothers' responses following episodes of child noncompliance and compliance. *Journal of Consulting and Clinical Psychology, 72*, 897–903.

Camras, L. A., Oster, H., Bakeman, R., Meng, Z., Ujiie, T., & Campos, J. J. (2007). Do infants show distinct negative facial expressions for fear and anger? Emotional expression in 11-month-old European American, Chinese, and Japanese infants. *Infancy, 11*, 131–155.

Chambers, C. T., Craig, K. D., & Bennett, S. M. (2002). The impact of maternal behavior on children's pain experiences: An experimental analysis. *Journal of Pediatric Psychology, 27*, 293–301.

Chorney, J. M., Garcia, A. M., Berlin, K. S., Bakeman, R., & Kain, Z. N. (2010). Time-window sequential analysis: An introduction for pediatric psychologists. *Journal of Pediatric Psychology, 35*, 1061–1070.

Chorney, J. M., Torrey, C., Blount, R., McLaren, C. E., Chen, W. P., & Kain, Z. N. (2009). Healthcare provider and parent behavior and children's coping and distress at anesthesia induction. *Anesthesiology, 111*, 1290–1296.

Clemente, I. (2008). L. Wei, & M. G. Moyer (Eds.), *Recording audio and video* (pp. 177–191). Malden, MA: Blackwell Publishing.

Cohen, L. L., Bernard, R. S., McClellan, C. B., & MacLaren, J. E. (2005). Assessing medical room behavior during infants' painful procedures: The measure of adult and infant soothing and distress (MAISD). *Children's Health Care, 34*, 81–94.

Cohen, L. L., Manimala, R., & Blount, R. L. (2000). Easier said than done: What parents say they do and what they do during children's immunizations. *Children's Health Care, 29*, 79–86.

Cone, J. D. (1999). P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Observational assessment: Measure development and research issues* (pp. 183–223). Hoboken, NJ: John Wiley & Sons Inc.

Dunn, M. J., Rodriguez, E. M., Miller, K. S., Gerhardt, C. A., Vannatta, K., Saylor, M., … Compas, B. E. (2011). Direct observation of mother–child communication in pediatric cancer: Assessment of verbal and non-verbal behavior and emotion. *Journal of Pediatric Psychology, 36*, 565–575.

Ekman, P., & Rosenberg, E. L. (1997). P. Ekman, & E. L. Rosenberg (Eds.), *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. New York, NY: Oxford University Press.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. Oxford, England: John Wiley & Sons.

Frank, M. G., Juslin, P. N., & Harrigan, J. A. (2005). J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *Technical issues in recording nonverbal behavior* (pp. 449–470). New York, NY: Oxford University Press.

Harris, F. C., & Lahey, B. B. (1982). Subject reactivity in direct observational assessment: A review and critical analysis. *Clinical Psychology Review, 2*, 523–538.

Hartmann, D. P., & Wood, D. D. (1990). A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *Observational methods* (pp. 107–138). New York, NY: Plenum Press.

Howells, R. J., Davies, H. A., Silverman, J. D., Archer, J. C., & Mellon, A. F. (2010). Assessment of doctors' consultation skills in the paediatric setting: The Paediatric Consultation Assessment Tool. *Archives of Diseases in Childhood, 95*, 323–329.

Kaugars, A. S., Zebracki, K., Kichler, J. C., Fitzgerald, C. J., Greenley, R. N., Alemzadeh, R., & Holmbeck, G. N. (2011). Use of the family interaction macro-coding system with families of adolescents: Psychometric properties among pediatric and healthy populations. *Journal of Pediatric Psychology, 36*, 539–551.

Kazdin, A. E. (1982). Observer effects: Reactivity of direct observation. *New Directions for Methodology of Social and Behavioral Science, 14*, 5–19.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn & Bacon.

Kobak, K. A., Lipsitz, J., Williams, J. B. W., Engelhardt, N., Jeglic, E., & Bellew, K. M. (2007). Are the effects of rater training sustainable? Results from a multicenter clinical trial. *Journal of Clinical Psychopharmacology, 27*, 534–535.

Kubicek, L. F., Riley, K., Coleman, J., Miller, G., & Linder, T. (2013). Assessing the emotional quality of parent–child relationships involving young children with special needs: Applying the constructs of emotional availability and expressed emotion. *Infant Mental Health Journal, 34*, 242–256.

Leger, D. W. (1977). An empirical evaluation of instantaneous and one-zero sampling of chimpanzee behavior. *Primates, 18*, 387–393.

Lei, P., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational research. *Psychology in the Schools, 44*, 433–439.

Margolin, G., Oliver, P. H., Gordis, E. B., O'Hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review, 1*, 195–213.

Patton, S. R., Dolan, L. M., & Powers, S. W. (2008). Differences in family mealtime interactions between young children with type 1 diabetes and controls: Implications for behavioral intervention. *Journal of Pediatric Psychology, 33*, 885–893.

Pedro, H., Barros, L., & Moleiro, C. (2010). Brief report: Parents and nurses' behaviors associated with child distress during routine immunization in a portuguese population. *Journal of Pediatric Psychology, 35*, 602–610.

Pellegrini, A. D., & Davis, P. D. (1993). Relations between children's playground and classroom behaviour. *British Journal of Educational Psychology, 63*, 88–95.

Reid, G. J., McGrath, P. J., & Lang, B. A. (2005). Parent-child interactions among children with juvenile fibromyalgia, arthritis, and healthy controls. *Pain, 113*, 201–210.

Ribes-Iñesta, E. (2003). What is defined in operational definitions? The case of operant psychology. *Behavior and Philosophy, 31*, 111–126.

Riley-Tillman, T., Christ, T. J., Chafouleas, S. M., Boice-Mallach, C., & Briesch, A. (2011). The impact of observation duration on the accuracy of data obtained from direct behavior rating (DBR). *Journal of Positive Behavior Interventions, 13*, 119–128.

Rodriguez, E. M., Dunn, M. J., Zuckerman, T., Hughart, L., Vannatta, K., Gerhardt, C. A., . . . Compas, B. E. (2013). Mother-child communication and maternal depressive symptoms in families of children with cancer: Integrating macro and micro levels of analysis. *Journal of Pediatric Psychology, 38*, 732–743.

Shapiro, E. G., McPhee, J. T., Abbott, A. A., & Sulzbacher, S. I. (1994). Minnesota preschool affect rating scales: Development, reliability, and validity. *Journal of Pediatric Psychology, 19*, 325–345.

Walters, K. S., & Hope, D. A. (1998). Analysis of social behavior in individuals with social phobia and nonanxious participants using a psychobiological model. *Behavior Therapy, 29*, 387–407.

Warshaw, M. G., Dyck, I., Allsworth, J., Stout, R. L., & Keller, M. B. (2001). Maintaining reliability in a long-term psychiatric study: An ongoing inter-rater reliability monitoring program using the longitudinal interval follow-up evaluation. *Journal of Psychiatric Research, 35*, 297–305.

Wysocki, T., Harris, M.A., Greco, P., Bubb, J., Danda, C., Harvey, L., & White, N.H. (2000). Randomized, controlled trial of behavior therapy for families of adolescents with insulin-dependent diabetes mellitus. *Journal of Pediatric Psychology, 25*, 23–33.

Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer Publishing Co.