Ways of Working:
A Three-tiered Interpretive Model for Video Research

Donald W. Wortham , Sharon J. Derry, Department of Educational Psychology, University of Wisconsin-Madison
1025 W. Johnson St., Madison WI 53706
Email: dwortham@gmail.com, derry@education.wisc.edu

Abstract: In this "ways of working" paper, we describe a three-tier interpretive model for analyzing video data of social interaction. Since members of a research team often work from different theoretical perspectives, individual members' analyses of "what happened and why" are not always congruent. The model and process described here require making explicit methodological (and technological) links between layers of interpretation. Individual group members benefit because the connections between their analyses and data can be effectively warranted; the larger group benefits because the location of interpretive differences is more precisely exposed.

Introduction

The method described in this paper may point to the solution of a problem that confronts much current research on social interaction, especially research that captures this interaction using rich media. Conceptually, a consumer of video research should be able to trace a path from data highlighted by analysis, down through any interim representations (such as transcripts), to the raw data. This capacity to trace a path from highest level of analysis down to raw data would allow the consumer to better understand both nomothetic generalizations and ideographic particulars of the social interaction being studied. In practice, these connections are almost always absent, and our field's current reliance on print media for publication passively discourages these connections. The situation is exacerbated when video research investigates subjects' taken meanings (emics) rather than structural features of social interaction (etics). Since researchers regardless of methodology agree that an important goal of social science is the uncovering of subjective meaning (c.f., Angelillo, Rogoff & Chavajay, in press; Nystrand, Wu, Gamoran, Zeiser, and Long, 2003), this opacity regarding the connections between data and analysis is worrisome.

In our own lab, we have been confronted by these same problems, since a variety of theoretical perspectives are represented among our group's members. Working from a particular theoretical viewpoint, one of us will "see" something in data and construct an analysis based on this apperception. To build a case for this perspectival viewpoint, and more importantly to help others see what they might otherwise not, we have begun to explore a number of interrelated practices. First, we have put forth a model of interpretation, and come to basic agreement about what can be known (and what cannot) from each of the three layers in our model. Second, we make explicit linkages between our analyses and the various forms of data we study. These linkages are both conceptual and technical, so that we can now trace an interpretive path from highest layer of analysis, down through interpretive representations such as transcripts, to our raw video data. Since we believe that these efforts, while still provisional and in an early stage of development, may be instructive for other researchers working with video, we connect our discussion to a broader discourse within the learning-science community about the establishment of standards for video research (e.g., Derry, in press).

The Three-Layer Interpretation Model

Let us begin with a very simple visual analogy: our Three-Layer Interpretation Model is like a three-ring toddler's toy. The toy consists of a wooden base, a spindle, and three rings of decreasing size. Our model consists of a base (the event stream…especially social interaction captured in rich media), and three interpretive levels: a large layer (raw data), a medium layer (observed events), and a small layer (analytic statements). Each of these layers has certain epistemological and technical constraints, and results in specific work products.

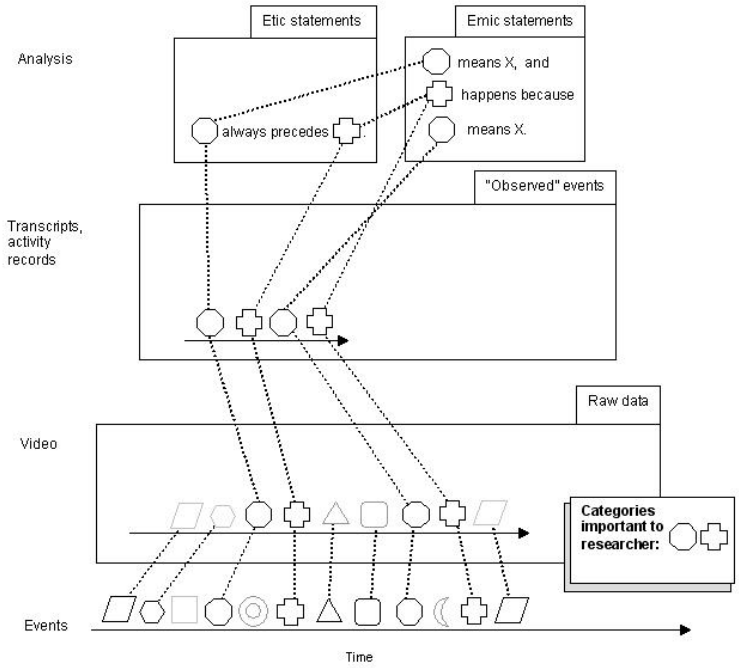


Figure 1. Three-layer interpretive model.

The toddler's toy differs from our model in two important ways: first, rather than being discrete from the layers, the spindle in our model is instead a contiguous set of relations among the layers. Secondly, the top ring is of two parts, corresponding to two different goals for research. In the sections below, we describe our model in brief, working from the bottom up.

The Base: Event Matrix

The toddler's toy has a base, a ground. The base referred to in our three-level model corresponds to an "event matrix" that appears to particular observers as event streams. Since, ontologically, the event matrix exists whether or not we as observers make any inferences about it, the event stream itself is not *per se* part of our inference model.

An event matrix is comprised of the multiple causal chains that transpire in a particular location: in a classroom, or in the vacuum of space, or inside a paramecium. Event streams are available to sensory experience, but in practice unknowable in their entirety. The actions of teachers and students, the movements of planets and stars, or the functions of the paramecium's organelles are indeed occurring, but the full scope of this activity is never available to us as observers. Nonetheless, our research efforts are directed toward investigating aspects of this event matrix and making warrantable statements about it. The belief that we can succeed in this effort is grounded in the notion that video analysis is a social meaning-making process, with an aim towards achieving intersubjectivity among researchers, irrespective of their differences. For certain phenomena, we expect intersubjectivity to be high enough so that at least some kinds of statements about what is happening in the event matrix can be made with a high degree of confidence (often *etic* statements, discussed later). For other sorts of statements, we assume that intersubjectivity will be lower (most often *emic* statements, also discussed later). In fact, the point of the method outlined here is precisely to identify where and why intersubjectivity is low, so that our group's understanding is stretched or even changed through alternative interpretations of events.

Layer 1: Raw Data Layer

The first *interpretive* layer of our model is the "raw data" layer, which most often takes the form of digital video recordings of social interaction. This pseudo-tangibility of digital video is a fantastic asset. We can move digital video around, share it, take it apart, observe discontinuous segments contiguously, and so forth, yet it is also

has the properties of being fixed and permanent, unlike our recollections about what happened, which are subject to both degradation and importations of meaning. This "captured" aspect of videotaped social interaction allows us to later make public what was transitive and ephemeral in the event stream.

Even though we can regard the material stuff of the raw data layer to be a high-fidelity, fixed accounting of social interaction, it must be regarded as limited in two important ways. First, the raw data layer is itself interpretive, and not an exact rendering of social interaction; something is always left out. When we videotape a classroom, we capture only certain aspects of social interaction. We further acknowledge that interactions off-screen, and cognitions "in the head" of participants have an impact on the very things we are filming. The upshot is that raw data always has a point of view, and furthermore that our point of view is never privileged, and is always imperfect.

Sometimes this is obvious, as when one participant's movement occludes our camera-eye view of a second participant. But there are more subtle ways in which we leave things out. When our cameras focus on one part of an interacting whole, we miss out-of-view interactions that may dynamically change the very interactions that are our focus. Our situation is no different from other scientists who observe phenomena *in situ*: when astronomers monitor the position of stars in our expanding universe, they focus on a small sector of the sky, all the while knowing that the movements of other heavenly bodies are both "out of the picture" and affecting the movements of those objects "within the picture." For us, this means we must always regard our statements about *what is happening* in our video…the descriptions that become our "observed data layer," discussed at length below…as tentative, and as circumscribed by the fact that we simply have not captured everything that was happening. This is necessarily the case even if levels of intersubjectivity are high regarding what is occurring in the event stream.

This recognition allows researchers to carry out a set of research actions that are fair game with respect to the raw data layer. These include selecting, rejecting, expanding, and focusing raw data. These actions are very close to their common-sense meanings: *selecting* raw data is the action of noticing something interesting and setting it aside for purposes of further analysis; *rejecting* means eliminating raw data from further consideration (most often because the video has been spoiled in some way, or records "dead time"); *expanding* means bringing multiple data sources together to understand a particular sequence of events, such as multiple video views, stills of particularly interesting actions or artifacts, and written artifacts; *focusing* means limiting our field of vision on the raw data to particular aspects of the social interaction (or other behavior) occurring on screen.

Before leaving this discussion, we must note that issues at the raw data layer span both the epistemological and the technological: certain technical processes can positively affect epistemological problems. If what Jordan and Henderson (1995) say is true…that "video recordings replace the bias of the researcher with the bias of the machine," then making improvements to mechanical aspects of video capture is likely to lessen this unintentional but still real bias. For instance, the knowledge that our raw data cannot capture everything nonetheless inspires us to capture in our raw data as much of the event matrix as is practically possible. Our model is in essence a trickle-up model, so information excluded from raw data is likewise excluded from all subsequent interpretation. Like many other research labs in the learning sciences, ours has, over time, developed strategies and techniques to mitigate some of practical concerns related to videography. For instance, we know through experience that we get more complete renderings of the event matrix when we use multiple cameras. However, our focus in this paper is not the mechanics of raw data capture, nor the placement of recording devices, nor the storing of digital information, nor even the techniques for making links between the raw data layer and other layers of interpretation. Rather, our goal here is to share our model and outline the processes we are exploring, so we will say little more about purely technical considerations, even while acknowledging their importance.

Layer 2: Observed Events Layer

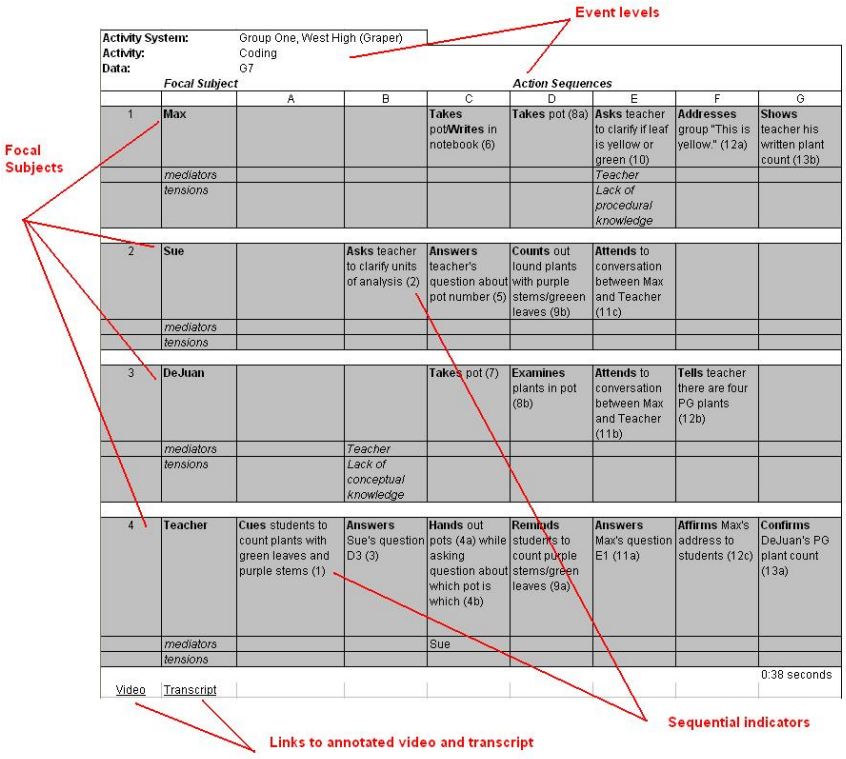
The focus of this section is the second data layer of our three-tier model, what we call the observed events layer. This layer depends upon the existence of raw data, and the previously-mentioned allowable operations performed on it…selecting, rejecting, expanding, or focusing.

The end-product of this layer is a document that re-renders what we as observers "see" in a selected segment of raw data. As the graphic in Figure 1 above shows, not all aspects of the raw data trace are included in the observed events layer. The point is not that aspects are excluded, but that researchers can, with high degrees of intersubjectivity, agree about the nature of *what is included*. This requires that we have a shared "observed event syntax." The key to developing an observed event syntax is to come to agreement about the level of description. In

practical terms, this may not be as difficult as it sounds; research suggests that humans recognize a "basic" level of events, similar to Roschian perception of basic-level categorizations of objects (Zacks and Tversky, 2001). When building a record of observed events, the focus is on what early ecological psychologists (e.g., Barker and Wright, 1954) considered *molar* actions, those "observable by the naked eye," which conform to basic-level events. We refer to this basic unit of description as the *action*. An action is an observable event that conforms to a subject-mediator-object syntax where the subject is an individual. According to this perspective, "Caroline watered Pot One with a pipette" and "Robert held the sample up to the Munsell chart" are actions.

Our research group typically studies inquiry-learning environments, where multiple subjects interact with each other, with various representations, and with physical objects in the environment. Our model handles this complexity by employing focal-subject observation techniques borrowed from ethology (Lehner, 1996; Lorenz, 1981). When using this method, we "track" each participant in a selected segment, and describe (using short, direct statements) what the participant is doing and saying. These observations are recorded in a contiguous row of cells in a matrix. Once one subject has been observed and his or her actions recorded, a second subject is taken up, until all interacting participants have been followed over the course of the segment.

After all participants have been observed and their actions (including speech acts) recorded, individual records are aligned according to time codes and intersecting actions, the entire record is reviewed against the raw video data to ensure that all observable actions have been noted, and sequence indicators are assigned to each subject's action. This way, the record of social interaction can be followed as if it were a typical one-channel, vertical transcript, even though it is multi-channel and horizontal. Once complete, the resulting product is not unlike a digital animation score, where characters occupy particular horizontal "channels" and the entire sequence can be followed by reading left-to-right, and within a particular column according to a sequence indicator in each cell. Ideally, each segment is directly linked back to raw video, so that researchers can continually check to make sure that this re-rendering of activity captures the common-sense, "observable to the naked eye" view of the reality captured on video. This linkage is part of the "spindle" (discussed more fully below) that connects each layer in our interpretive model. The record in Figure 2 below is an example of one of these products (from the first author's dissertation):



	Focal Subject	A	B	C	D	E	F	G
1	**Max**			**Takes** pot/**Writes** in notebook (6)	**Takes** pot (8a)	**Asks** teacher to clarify if leaf is yellow or green (10)	**Addresses** group "This is yellow." (12a)	**Shows** teacher his written plant count (13b)
	mediators					*Teacher*		
	tensions					*Lack of procedural knowledge*		
2	**Sue**		**Asks** teacher to clarify units of analysis (2)	**Answers** teacher's question about pot number (5)	**Counts** out loud plants with purple stems/greeen leaves (9b)	**Attends** to conversation between Max and Teacher (11c)		
	mediators							
	tensions							
3	**DeJuan**			**Takes** pot (7)	**Examines** plants in pot (8b)	**Attends** to conversation between Max and Teacher (11b)	**Tells** teacher there are four PG plants (12b)	
	mediators		*Teacher*					
	tensions		*Lack of conceptual knowledge*					
4	**Teacher**	**Cues** students to count plants with green leaves and purple stems (1)	**Answers** Sue's question D3 (3)	**Hands** out pots (4a) while asking question about which pot is which (4b)	**Reminds** students to count purple stems/green leaves (9a)	**Answers** Max's question E1 (11a)	**Affirms** Max's address to students (12c)	**Confirms** DeJuan's PG plant count (13a)
	mediators			Sue				
	tensions							

Figure 2. Observed-events record from an inquiry-science classroom

As a practical matter, we should note that the building of observed-event records depends critically on the capacity of digital video to be slowed down, stopped, and re-played, and on the observers' patience. The capacity to "see" what each individual in a social group is doing, and to describe it so that it is intersubjectively available, are critically dependent upon having a syntax of description that specifies the level of observation, being able to slow things down, and having the wherewithal to go over and over short segments of raw data.

Layer 3: Analysis Layer

A product like the record of interaction shown in Figure 2 above serves up working data for the top tier of our three-layer interpretation model. The uppermost interpretive layer is where analytic (if probabilistically hedged) statements are made regarding *what* has happened in the event stream…as it boils up through raw data and into observed events; and *why* or *how* these events have transpired.

Here's an example of how this works in practice. One of our ongoing research projects (the first author's dissertation work) involves studying how students in inquiry science classrooms develop the ways of working of scientists. Borrowing ideas from Charles Goodwin (1994), we are viewing video of students and teachers conducting heredity experiments to see if and where they "create" data by coding aspects of the plants that they are experimentally manipulating. This interpretive process occurs in several steps. To begin, we work with raw data to look for instances of the category of interest, in this case students' creating data by coding. Once these segments are selected (one of the allowable research maneuvers at the raw data level), we can review the raw video data, focus on participants and their actions, and build an observed-events record (Figure 2). This record allows us to probe the sequences, and ask questions such as, "Who is participating?" "What is the pattern of interactions in the segment of interest? We find that questions about "what" and "who" can be answered in a reasonably straightforward manner, if we follow the observed-events syntax discussed above.

Of course we must ask how and why questions as well: "How did the students know what 'mattered' in terms of coding?" "Why did Max ask the question about leaf color that he did?" In accordance with our working model, we explicitly distinguish between questions of the first sort and questions of the second, calling them *etic* and *emic* respectively. This acknowledges epistemological differences in the sorts of statements that we can make from our data. Since this distinction is crucial to our model, we review it in some detail.

More than fifty years ago, Kenneth Pike introduced the terms *etic* and *emic* to social science (Pike, 1954). Pike, a linguist, suggested that the distinction between phon*etics* (physiological patterns resulting in sounds) and phon*emics* (the meanings that humans assign to speech behavior) could be profitably applied to the study of human behavior beyond language. Pike recognized that with *phonetics*, language could be studied and understood at arm's length from the people that speak it. One need not understand what a speaker is saying to distinguish performative aspects of speech, such as differences between an aspirated [p] phonetic and a non-aspirated [b] sound, nor distributional aspects (that, for instance, pathologies such as velopharyngeal frictive speech often co-occur with cleft palate). Pike saw the study of *phonemics* as different, requiring the observer to access the meaning system of the speaker. Pike's critical test for phonemic difference was to determine if exchanging one phoneme in a string changed the meaning for the speaker (Pike, 1954). For instance, substituting *r* for *l* as the first sound in the string //ip changes the meaning of the string (/r/ip versus /l/ip).

Extrapolating from this, Pike's use of the terms etic and emic in social and behavioral research refers to making a distinction between analysis of a behavior stream from a perspective *other than that of the actor* (etic) and a rendering of the same behavior *from the perspective of the actor* (emic). For instance, to describe an action as "the subject used an overhand motion with an inward turn of the wrist to propel a solid sphere toward its target" is to employ an *etic* (if ungainly) description. Etic statements make minimal (if any) reference to a particular cultural meaning system; in principle, they can be understood by anyone with basic communicative competence (c.f. Habermas, 1981). To make the distinction clear, we can experiment with the statement concerning the sphere and its target. If we reformulate it to say "the pitcher threw a slider, it missed the strike zone," we are making an *emic* statement that is only meaningful if one participates in the language of baseball. This is made particularly salient if we change the sentence, so it reads "the pitcher threw a slider at the batter's *head.*" For those knowledgeable about the game, throwing a pitch at a batter's head has a very different meaning than merely failing to deliver a strike…it is seen as an aggressive act.

One of the reasons for making the etic/emic distinction is that it forces us to acknowledge that we are adopting a specific perspectival frame when we make emic statements in an analysis. Leander's (2002) excellent study of "silencing" in classrooms provides a useful example. In this work, Leander investigates four special categories of social interaction in classrooms: students can be active participants, they can be "silent," they can "silent and being silenced," or finally they can be "speaking while being silenced." While the first of two of these categories can be readily observed by anyone with basic communicative competence, the latter two require that the observer participate in the meaning system developed by Leander in his article, and presumably shared by his subjects. These are therefore emic descriptors. By no means does this make them invalid; quite the opposite. To our way of thinking, social science progresses to the degree that new understandings…such as what it means to be "silenced"…develop and diffuse through the community of researchers.

The Spindle

We now come to the discussion of our three-level interpretive model's spindle: the relations that connect statements at one level with statements at another, like the central spindle of a toddler's toy. The spindle in our model is a conceptual entity that becomes manifest as links between various data documents and our raw digital video. By allowing researchers to traverse these connections, the spindle ultimately grounds interpretation in raw data. This proves particularly valuable when team members' interpretations are novel, or develop out of a theoretical perspective not shared by the group. In a sense, the spindle helps research teams like ours to keep track of "fissures of interpretation."

At the very onset of this paper, we stated that interpretability gaps caused by weak links or even holes between what individuals "see" and what others in a group see are a central problem for research teams. To resolve this problem, we have been experimenting with ways to ensure that theoretical concepts and categories…when applied through analysis…are traceable to actions in transcripts (at the observed events level), and one tier down, to the video of the raw data corpus. We surmised that when connections among levels of interpretation are explicit and direct, they can be followed more easily by others who may not share a particular theoretical perspective, or who are simply less familiar with video data projects (peripheral participants in research, such as practitioners and even scholars-in-training). These connections constitute our spindle.

Other researchers seem to be moving in similar directions; Leander's (2002) study on "silencing" again provides a useful example. In his own musical score-style transcripts (which bear structural resemblance to our own example shown in Figure 2 above), Leander embeds codes derived from his theory of social silencing directly into his transcript. This is invaluable for the reader, since it allows us to "see" exactly *where* Leander "sees" the action of silencing occurring in the classrooms he studies. To embody the three-tier model we discuss here, all that remains is for the author to directly connect this annotated transcript *up* to his analysis, and *down* to the raw video data he captured in the classrooms. Of course, neither of these steps is technically possible in the medium of paper-based print, so we can hardly blame Leander for not making these connections. Indeed, our working model begs many questions about the adequacy of current academic norms for conducting and reporting video research in the learning sciences.

The Problem that the Three-Layer Model Solves

Let us sum up. We believe that the three-layer model and the processes that follow from it hold potential for solving a key problem for research groups: by explicitly naming layers of the model, by concretizing interpretive processes, and by demanding that interpretations must "link down" to raw data, the model can illuminate opaque connections among categories and codes used in research, empirical data, and analysis. In practice, this helps to effectively channel subjectivity—or better, creativity—so that any private or closely-held meanings have a fighting chance of becoming intersubjectivity held. The model and process require making technical (in addition to methodological) links among layers of interpretation: video researchers need methods for explicitly showing connections among analytic statements, records of observed events, and raw data. Using off-the-shelf data-handling tools, we are exploring these practices in our work. To date, they have proven useful for us. As we share them with the larger community, we ask if…perhaps…we should all consider making these connections explicit.

References

Angellio, C. & Rogoff, B. (in press). Examining shared endeavors by abstracting video coding schemes with fidelity to cases. In R. Goldman, R., Pea, B. Barron, & S. Derry (Eds.), *Video Research in the Learning Sciences.* Mahwah, NJ: Erlbaum.

Barker, R.G. & Wright, H.F. (1954). *Midwest and its children.* Evanston, IL: Row Peterson.

Derry, S. J. (in press). Video research in classroom and teacher learning: Searching for standards in a complex terrain. In R. Segall, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video Research in the Learning Sciences.* Mahwah, NJ: Erlbaum.

Goodwin, C.A. (1994). Professional vision. *American Anthropologist* 96(3), 606-633.

Habermas, J. (1981). *The Theory of Communicative Action: Reason and Rationality of Society, Vol. 1.* Boston: Beacon.

Jordan, B. and Henderson, A. (1995). "Interaction analysis: Foundations and practice." *The Journal of the Learning Sciences, 4*(1): 39-103.

Leander, K.M. (2002). Silencing in classroom interaction: Producing and relating social spaces. *Discourse Processes 34*(2), 193-235.

Lehner, P. (1996). *Handbook of Ethological Methods (2nd edition).* Cambridge: Cambridge University Press.

Lorenz, K. (1981). *The Foundations of Ethology.* New York: Springer-Verlag.

Nystrand, M., Wu, L.L., Zeiser, S., and Long, D.A. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes 35*(2), 135-198.

Pike, Kenneth L. (1954). *Language in relation to a unified theory of the structure of human behavior.* Glendale, Calif., Summer Institute of Linguistics.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 12*, 3-21.