Level 3: structured observation

All researchers want results and new researchers probably want them more desperately than older hands, who can rest easy in the faith engendered by earlier successes. But the desire for tangible outcomes, convertible currency in the professional and academic market, can distort the research process, turning it away from the search for understanding and towards an obsession with the meaningless accumulation of detail. One of the most important decisions you may need to make at this level of research is what part, if any, the systematic use of predetermined categories will play in your approach. Writers in the field of qualitative inquiry usually distinguish between *participant observation* and *structured observation*, the subject of this level. Structured observation is sometimes also referred to as systematic observation, but this misleadingly implies that participant observation is not systematic, so I shall avoid the term.

I actually prefer to think in terms of a continuum between *open observation*, which might characterise the early stages of participant observation where the observer tries to get a general sense of the setting and the activities associated with it, and *closed observation*, where the observer is strictly coding behaviour on a low-inference schedule, or instrument (the former is the term usually employed to describe what is actually used). This less categorical approach recognises the range of decisions and sensitivities that the qualitative researcher can respond to and does not imply a simple contrast that invites the assignment of evaluative labels. My emphasis in this level will therefore be not so much on the relative advantages and disadvantages of the two standard approaches but on how to reach a decision on the appropriate balance and how to approach structured observation in a way that will contribute to a qualitative project. I will begin, though, by highlighting the dangers of settling too quickly for a more closed approach.

Good decisions are made in the light of evolving needs and the context of the project as a whole, and it helps to have an overall picture of how participant observation compares with structured observation. Box 3.10 provides a brief overview of the main lines of difference. Participant observation, where 'the human instrument is a most sensitive and perceptive data-gathering tool' (Fetterman 1991:92), lends itself to the narrative representation of events recorded retrospectively. Its open orientation means that coding takes place later, as part of the analytic process, though categories may emerge as the process develops. In the case of structured observation they must be determined in advance as part of a clear coding system designed for cotemporaneous application. The limitations of this approach usually make it unsuitable as a primary data collection method in qualitative research, though the existence of a clear system means that observations can be replicated by other researchers in a way that is not possible in the much more open orientation of participant observation. Properly used where the situation allows it, the two approaches can make a potent combination.

Box 3.10 Participant and structured observation

	Participant observation	*Structured observation*
Orientation	Open	Closed
Foundation	Event-based	Category-based
Form	Narrative	Descriptive
Observer status	Observer-as-instrument	Observer-through-instrument
Coding	Post-observation	Pre-observation
Recording	Retrospective	Cotemporaneous
Format	Notebook	Observation schedule
Replicability	Non-replicable	Replicable
QI status	Main or supplementary method	Supplementary method

Deciding whether to use structured observation

The progression from 'What do I want to find out?' to 'What schedule do I use?' is not as straightforward as it may at first appear. We view new situations in the light of our current knowledge, and it often takes creative insight of a very high order to free us from the domination of taken-for-granted perceptions. We can't look at what goes on in a classroom without carrying as part of our mental baggage assumptions about the ways in which events there are organised; in fact, on the most basic level, we couldn't make any sense of things at all if we didn't have some framework for organisation. But a consequence of this is that one way of seeing can easily become the *standard* way of seeing, which is why preliminary participant observation is so important. We can use techniques associated with that to generate new ways of seeing that will in turn prompt fresh questions, a vitally important preliminary because the value of the schedule we eventually devise will depend on the quality of the questions we ask. The process looks something like that outlined in Figure 3.1.

The sorts of questions that can best be answered by structured observation are those related to particular behaviours about which we need specific information, such as the following:

- How often do students initiate interaction with the teacher?
- What's the distribution of teacher initiations to male and female students?
- How much lesson time is taken up with teacher talk?

Working up a schedule

If the decision to use structured observation has arisen from a genuine research need, the researcher should be in an excellent position to respond to the questions that arise in designing an effective observation schedule. Quantitative research tends to rely on a deductive approach, starting with a particular theoretical perspective, which may be behavioural, as exemplified by the FIAC system, or linguistic, as with the Sinclair and Coulthard system (see Chapter 4, Levels 1 and 2). However,

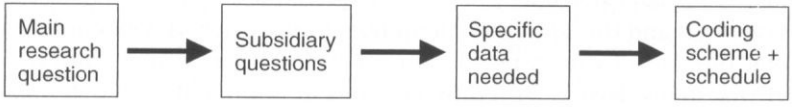

Figure 3.1 Contribution of an observation schedule

qualitative inquiry generally prefers inductive reasoning, which is why structured observation has to be considered in the light of the wider discovery process. Its design represents a creative challenge but also offers the prospect of a flexible and evolving system that allows for substantial revision at the trialling stage. Box 3.12 summarises some purely practical considerations proposed by Evertson and Green (1986:165) and the rest of this section will focus on the design issues that should inform the process of decision-making outlined there.

Box 3.12 Practical considerations in approaching structured observation

1. **Research question(s)**
   e.g. What is the occurrence of different types of student self-repair and other-repair in teacher-fronted activities and in group work?
2. **Focus**
   e.g. particular group(s), event, strategy
3. **Setting**
   e.g. classroom, staffroom
4. **'Slice of reality'**
   e.g. group work, teacher questioning
5. **Observation instrument(s)**
   e.g. category system, descriptive system
6. **Observation procedures**
   e.g. When? How often? Number of observers?
7. **Analytical procedures**
   e.g. frequency count, event structure
8. **Presenting of findings**
   e.g. written summary, tables, pie charts

Source: based on Evertson and Green 1986:165.

Basic decisions

The fundamental aim of any structured observation is to provide specific information that the researcher needs in order to answer the research questions posed, so the decision-making process should not be a mechanistic one. The most basic decision of all is what behaviours will feature in the observation. It is possible to cover all behaviours (the FIAC system does this) but it may well be that the observation will focus on specific behaviours or events (for example, teacher questions or group work). Box 3.13 on pages 152–3 summarises other options.

Once the focus of structured observation has been established, a number of choices have to be made, four of which are particularly important:

1. The researcher will need to decide whether a *descriptive system* or a *category system* is more appropriate. This will probably depend on the extent to which numerical information is required and whether the relevant behaviour can be broken down into a reasonable number of discrete categories.
2. It may also be necessary to decide whether a *rating scale* is to be used. Rating scales evaluate behaviour according to agreed criteria (for example, 'high, medium, low') and rating systems are particularly suitable for purposes of evaluation (a good example is the RSA Diploma Check List for Practical Tests, reproduced in Malamah-Thomas 1987:73), though rating and non-rating systems can be combined.
3. The degree of inference involved (i.e. the extent to which observers are required to use their own judgement in applying the system) will also need to be assessed. Where the structured observation involves one observer and is essentially descriptive, *high-inference categories* are perfectly acceptable, but if more than one observer is involved *low-inference categories* are preferable.
4 When the categories have been identified, the researcher must decide on how behaviours will be sampled, and there are two options available: *event sampling* (or event-based coding) and *interval sampling* (or time-based coding). In interval sampling behaviours are coded at predetermined intervals (e.g. every three seconds), while event sampling codes behaviours as they occur, regardless of the length of time that elapses during their occurrence.

Box 3.13 Options in structured observation systems

Descriptive Open. May be some (general) preset categories. Entries may extend over a number of paragraph, resembling fieldnotes.	*Category* Closed. Sample specific behaviours and events. Coded concurrently with events and therefore depend on small, precisely defined, easily coded low-inference units.
Rating Relatively open. Involve element of assessment, and therefore retrospective judgement. More	*Non-rating* Scalar rating not needed. Task is to assign behaviour to preset categories.

high inference than non-rating systems.

High inference
Direct attention to important aspects of behaviour and offer the chance to capture a sense of its richness. Descriptive systems usually high inference but might have low inference elements.

Event sampling
Focuses on events and can be used with low frequency behaviours. Decide on events to be observed and ensure all instances are captured. Provides details of events and can address questions such as 'How often does X occur?' and 'How long does X take?'

Low inference
Low-inference approaches tend to be atomistic, artificially dividing up complex events. But can be used by teams and provide precise, detailed information. Category systems should be low-inference.

Interval sampling
Best used with high frequency behaviours. Divide specified period into shorter time segments, then code behaviour at onset of each segment. No record of exact time taken up by particular activities, but gives general picture of time distribution throughout period.

Illustrations

A comparison of extracts from two (hypothetical) approaches, one based on a rating scale and the other on a non-rating scale, will illustrate some of the points made above. In the first example (Extract 3.12), which involves interval sampling but no rating, coding takes place every five seconds, at which point the user enters the relevant code in the box under the time shown, working across in the 5-second units then down through the minutes. The best way of handling the practicalities of such fine tuning is not to rely on a watch, which doubles the observation tasks, but to make a tape on which you record a click every five seconds; this can then be played back quietly through a single earphone and used as a prompt.

Extract 3.12 Four minutes of coding on a 10-category system

	5	10	15	20	25	30	35	40	45	50	55	60
1	2	2	2	6	6	2	2	6	6	6	6	5
2	2	2	10	2	2	6	2	2	2	10	5	5
3	5	3	3	5	3	3	6	6	4	4	4	6
4	2	2	6	6	6	6	6	5	3	6	6	9

If we assume that 2 stands for student talk, 6 for teacher talk and 5 for silence, we can see that in the first minute the talk is divided roughly equally and there is at least one example of silence. Although teacher talk seems to predominate slightly, the system doesn't allow us to state this with certainty (though a clearer picture of unequal distribution emerges over the four minutes), nor does it allow us to be certain about the number of turns because it is possible for a very brief turn to occur with a 5-second segment but not be recorded.

An alternative form of interval sampling would involve identifying activities on one axis and time periods on another, as in Extract 3.13, though in practice the time periods in this example would probably be longer and might run down the page, allowing the ten categories to go across the top.

Extract 3.13 Example of non-rating system (interval sampling)

	5	10	15	20	25	30	35	40	45	50	55	60
TT				√	√			√	√	√	√	
ST	√	√	√			√	√					
Sil												√
Oth												

A rating system (Extract 3.14) looks very different from either of these two and although it allows us to make general comments about (in this case) aspects of teacher delivery, it gives us no indication whatsoever about how often this occurs or how long it lasts.

Extract 3.14 Example of rating system (event sampling)

Feature	*Aspects*	*R*	*Comments*
	Clarity	4	*Speed often too fast. Big*
Delivery	Loudness	4	*ability range in class, but T*
	Speed	2	*made no allowances when*
	Appropriateness	2	*speaking to weaker Ss.*

Defining the units

When basic decisions have been made, attention turns to the design of the observation schedule itself. As with all such systems, it is best to think

first about higher order categories and then work out lower order definitions in terms of these. Trialling is absolutely vital if the categories are to do their job, but the considerations below should help you to avoid some initial pitfalls. They can be applied to all structured observation but are designed with category systems particularly in mind.

1. Within category:

- *Make sure the category is as clear as possible.*
  'Teacher doubt' would be a difficult category to code because it's not clear what it refers to in classroom terms.
- *Check that the category is related to observable behaviour.*
  Even if we knew what 'Teacher doubt' referred to, it would be hard to code because we can't see into the mind of the teacher.
- *Be as precise as possible.*
  'Uses a wide range of target language structures' is clear enough but doesn't specify what constitutes 'a wide range'. If this is to be used as a category, there will need to be accompanying specification.
- *Consider the range of each category.*
  'Corrects fellow pupil in group work' covers two complementary categories. It might have been more effective to separate the two, so that 'corrects fellow pupil' can be assigned to whatever category ('group work' would be one option) applies at a particular time.

2. Among categories:

- *Ensure that all definitions are clear and exclusive.*
  Categories should be defined clearly enough to allow the observer to assign classroom events to individual slots. Overlapping categories make effective coding impossible.
- *Check that there are no gaps in the coverage of chosen behaviours.*
  Taken together, the categories must cover all the instances of behaviour that fall within the parameters of the study. This may involve including an 'other' category (notice how the FIAC system in Box 3.11 above, includes 'Silence or *confusion*' as a category), but where too many behaviours are assigned to this category it is a sign that something more specific is called for.
- *The schedule must be practicable.*
  However good the schedule may be on paper, it's useless if coders can't apply it in practice. For example, where real time coding

applies (i.e. when the coding takes place 'live', in the lesson), the observer must be able to apply the system effectively in the time available.

Box 3.14 Summarises the essentials.

Box 3.14 Essential characteristics of an effective category system

- Clearly definable categories related to observable behaviour
- Mutually exclusive categories – no overlap
- The category set is exhaustive
- The system can be operationalised

Some practical problems

I began this level by indicating some of the dangers of closed observation, but there are also practical challenges to be overcome. Four of these are particularly important and these are outlined below.

1. Observer effect: if you have established yourself in the setting to the extent that your presence is almost taken for granted (which may mean setting aside data from earlier observations), observer effect will be minimalised. Nevertheless, you should not underestimate the effect that filling in a form (which is what completing an observation schedule amounts to in the eye of the beholder) can look like. In establishing a vantage point from which you can observe all events in the setting, you should therefore consider whether it is also possible to hide or disguise the writing or coding you are doing. In any case, it is not a good idea to rely on a single observation so you can, if possible, discard early sessions.

2. Expectancy effect: observation schedules are designed with specific purposes in mind and these cannot simply be erased from the memory. They may well create a tendency to 'see' particular features, so in order to reduce the strength of this expectancy effect you should reflect very carefully on your initial assumptions, however slight these may be. The higher the degree of inference involved in the observation, the more important it is to be aware of factors that might influence your coding.

3. *Observer drift*: this arises from the fact that observers become familiar with the schedule they are using and begin to 'see' things in expected ways, which creates a drift away from the original coding. To some extent this is inevitable, but awareness can help to reduce it and where teams of coders are involved periodic checks on inter-rater agreement can direct attention to emerging problems.

4. *Central tendency*: this applies only to rating scales and refers to the tendency for opt for something at or near the middle. Keeping options to a minimum can help to reduce it, as can clear criteria and, as in the case of observer drift, awareness and monitoring.

If you are working as an individual, awareness and reflection are probably your best defence against these problems, but if there is a team of observers an additional check is available in the form of inter-observer agreement. The next section describes a relatively straightforward way of establishing whether the level of this is acceptable.