Clement, J. (2000) Analysis of clinical interviews: Foundations and model viability. In Lesh, R. and Kelly, A., Handbook of research methodologies for science and mathematics education (pp. 341-385). Hillsdale, NJ: Lawrence Erlbaum.

This file is the last draft before page proof corrections.

ANALYSIS OF CLINICAL INTERVIEWS: FOUNDATIONS & MODEL VIABILITY*

John Clement

*The research reported in this study was supported by the National Science foundation under Grant RED-9453084. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

The clinical interview is a technique pioneered by Piaget (1975) to study the form of knowledge structures and reasoning processes. In the last twenty-five years, it has evolved into a variety of methods, including open ended interviews and think-aloud problem-solving protocols. These techniques have played key roles in seminal studies in science and mathematics education as well as developmental psychology. Their strengths, in comparison to nonclinical, data-gathering techniques, include the ability to collect and analyze data on mental processes at the level of a subject's authentic ideas and meanings, and to expose hidden structures and processes in the subject's thinking that could not be detected by less open-ended techniques. These abilities are especially important because of Piaget's discovery that people have many interesting knowledge structures and reasoning processes that are not the same as academic ones--they have alternative conceptions and use nonformal reasoning and learning processes. Mapping this "hidden world" of indigenous thinking is crucial for the success of instructional design. Students cannot help but use their own prior conceptions and reasoning processes during instruction, and they have strong effects on the course of instruction. Since tests are almost always written from the point of view of the academic and are designed to detect standard forms of academic knowledge, they can fail to detect key elements in students' thinking. Clinical interviews, on the other hand, can be designed to elicit and document naturalistic forms of thinking. In some exploratory varieties of clinical interviewing, the investigator can also react responsively to data as they are collected by asking new questions in order to clarify and extend the investigation. Even where the detection of academic knowledge is sought, clinical

interviews can give more information on depth of conceptual understanding, since oral and graphical explanations can be collected, and clarifications can be sought where appropriate. However, the analysis of interviews can be difficult and time consuming, always involving a degree of interpretation on the part of the researcher. One purpose of this chapter is to discuss the scientific foundations that provide a basis for sound analysis.

Partly because of their open-endedness and qualitative and interpretive nature, some individuals have implied that clinical interview studies may be "unscientific" (Nisbet & Wilson, 1977); witness the recently advertised international handbook of *Educational Research, Methodology and Measurement* (Keeves, 1997), in which research methodology is divided into two camps: humanistic (including interviewing for clinical research) and scientific analysis (including quantitative methods.) The present chapter argues that carefully done clinical interview studies are an essential and irreplaceable part of the scientific enterprise of investigating a student's mental processes and that, in the context of the history of science, such simplistic associations between the scientific and the quantitative are grossly misplaced.

There are also controversies at a more detailed level. A variety of protocol analysis methods are available; see, for example, Clement (1979, 1989b), Driver (1973), Easley (1974), Hayes, and Flower (1978), and Newell and Simon (1972). In fact, there are some fairly diverse methods of clinical interviewing; these vary from the most *convergent* approach involving the detailed coding of individual clauses in transcripts by multiple independent coders to the most *generative* approach involving the open interpretation of large episodes by an individual analyst. This raises several questions: (1) One can ask whether the two approaches above are in opposition or are complementary (I will claim the latter); (2) Generative/Interpretive analyses focus on constructing new observation concepts and theoretical models. Because they concentrate on discovering new constructs rather than attaining high reliability with established constructs, they have been criticized by some as being less scientific. So a second question is whether generative studies in particular can be placed on firmer methodological ground as scientific studies in a more explicit way than in the past. (3) In this regard, I want to examine first whether some new trends in the history and philosophy of science focusing on mental models and naturalistic studies of reasoning in scientists can be used as a foundation for analysis methodology.

Thus, I will begin by stepping back to consider the bigger picture of the scientific research enterprise in general. I believe that there is some very good news here. Recent views of the nature of theory generation and evaluation in science lend substantial support to the scientific value of protocol analysis in general and of generative methods in particular. I will develop this conclusion from the view that the core of what an investigator is doing in analyzing a clinical interview is constructing a model of hidden mental structures and processes that are grounded in detailed observations from protocols. Making sure that these models are viable is a central goal. Therefore, there are two main topics in this chapter:

- the foundations of techniques for constructing models-- from history of science;

- a description of different types of clinical interview methods used for different purposes, concentrating on generative methods that foster the production of viable models of students' mental processes.

Foundations: Findings From Recent History of Science Studies on Processes for Constructing Models in Science

The work of a group of scholars in the history and philosophy of science, (Campbell, 1920/1957; Harre, 1961; Hesse, 1966; Nagel, 1961) provides the following view of the types of knowledge involved in science. They suggest that scientists often think in terms of theoretical *explanatory models,* such as molecules, waves, fields, and black holes, that are a separate kind of hypothesis from empirical laws. Such models are not merely condensed summaries of empirical observations but, rather, are inventions that contribute new mechanisms and concepts that are part of the scientist's view of the world and that are not "given" in the data. I first wish to put forward the view that one of the most important needs in basic research on students' learning processes is the need for insightful explanatory models of these processes. To do this I first need to examine the different types of knowledge used in science.

Four Levels of Knowledge in Science

As shown on the left-hand side of Table 1, the authors mentioned in the previous paragraph describe science as having at least four levels of knowledge. In particular, they see a distinction between an empirical law hypothesis (at level two in Table 1) summarizing an observed regularity and what I will call an explanatory model hypothesis (at level three). Campbell's (1920/1957) oft-cited example is that merely being able to make predictions from the empirical gas law stating that pressure times volume is proportional to temperature is not equivalent to understanding the explanation for the behavior of gas in terms of an imageable explanatory model of billiard-ball-like molecules in motion. The model provides a description of a hidden process or mechanism that explains how the gas works and answers "why" questions about where observable changes in temperature and pressure come from. On its own, the empirical law PV=KT does none of these things. Causal relationships are often central in such models. The model not only adds significant explanatory power to one's knowledge but also heuristic power, which stimulates the future growth of the theory.

These models do not consist merely of formulas and statements of abstract principles; rather, they consist of visualizable physical models, such as the elastic particle model for gases, which underlie the comprehension of formulas. Explanatory models are often iconic and analog in nature, being built up from more primitive and familiar notions. In this view, the visualizable model is a major locus of meaning for a scientific theory. (Summaries of these views are given in Harre, 1967; Hesse, 1967; and a related

hierarchy was proposed in Easley, 1978). Thus, the two lower rows of Table 1 constitute the observational level of science and the two upper rows constitute the theoretical level. These distinctions can be applied to studies of cognitive processes as well. For example, the second column in Table 1 shows four types of knowledge for the psychological construct of disequilibrium following a prediction. Disequilibrium as an explanatory model of a mental process is distinguished from expressions of surprise as an observed behavior pattern. The upshot of these considerations is that a central goal of science is to develop explanatory models that give satisfying explanations for patterns in observations.

The Need for Generative Case Studies

Unfortunately, in general within the learning sciences, there has been a significant shortage of explanatory models for higher-order processes. For example, there has been a notable lack of research on theories of content-rich, higher-level, cognitive processing in proportion to other areas of human thought. This is not surprising because psychology favors studies that include controlled experiments comparing behavior patterns at level two in Table 1 and because doing controlled experiments on higher-order processes is extremely difficult. This is the case because, first, some of the processes are so poorly understood that what the most important variables are is not at all clear. Second, there are likely to be many difficult choices open to a human problem-solver, and many possibilities for feedback loops, processes that complete or coordinate with each other, and recursion. For such cases, these non linear properties make the use of a controlled experiment designed to identify simple linear relationships between variables very difficult to apply at best. This perspective gives us a way to see why quite different methods might be required for studies within the learning sciences (in Table 1), as well as why special methods might be required to generate new hypotheses at level three (explanatory models of mental processes) for complex, higher-order thinking.

An Example Illustrating the Role of Generative Case Studies

Here and throughout this chapter, I will cite examples from one line of research that depended heavily on generative case studies. It concerned solutions to algebra word problems and the nature of students' difficulties with "reversal errors" (Clement, 1982). The problem shown in Table 2 was given on a forty-five minute written test to 150 freshmen engineering students.

Insert Table 2 about here

Fully 37% of the engineering majors, most of whom were taking calculus, solved this very basic algebra problem incorrectly. With a similar problem involving a fractional rather than an integral ratio, 73% failed to solve it. At first, it was thought that the errors on such simple problems must be due

primarily to carelessness. However, there was a strong pattern in the errors. Sixty-eight percent (68%) of the errors were reversal errors: 6S = P (or an algebraically equivalent statement) instead of S = 6P. Placing this example in Table 1 at level two (behavior patterns), one might propose *empirical* hypotheses concerning correlations between the incidence of reversal errors and subject variables such as age and mathematical background. These could be tested purely at the observational level without the development of a theory.

On the other hand, at level three (explanatory models), one might try to form *theoretical* hypotheses concerning the alternative reasoning patterns that cause these errors, and the reasoning pattern(s) responsible for correct answers. In order to do this, audio-taped and video- taped clinical interviews were conducted with fifteen freshmen who were asked to think aloud as they worked. From the analysis of transcripts of the subjects making reversal errors, one model of the faulty reasoning that emerged was a "word-order-matching" approach in which the students assume that the order of the key words in the problem statement will map directly into the order of the symbols appearing in the equation. For example, subject S1 immediately wrote 6S = P and said, "Well, the problem states it right off: '6 times students'; so it will be six times S is equal to professors." In this case, some of the protocol data were very suggestive of an explanation. There is no evidence here for any more complicated a strategy than that of mechanically matching the order of the symbols in the equation to the order of the words in the problem statement. This provides a possible explanation for the difficulty. This is a syntactic strategy in the sense that it is based on rules for arranging symbols in an expression that do not depend on the meaning of the expression. (Eventually, we found another source of the error, which will be discussed presently.)

Although the analysis process was unusually easy in this first example, it gives an illustration of constructing a model for a subject's reasoning process--a model that is grounded in the primary-level observations in a transcript. The clinical interviews played an important role here because they furnished a unique and comparatively direct source of information on the source of the error. Also, the model of the faulty, word-order-matching process (at level three in Table 1) goes well beyond the observed behavior pattern of the reversed equation (at level two in Table 1) in providing a description of the mental process underlying the behavior.

Thus, by using generative case studies in these areas, one can propose initial explanatory models of cognitive processes that are grounded in naturalistic observations. Now that the nature of explanatory models has been introduced, the question of how such models are constructed in science in general can be examined.

The Construction of Explanatory Models in Science: Abduction Versus Induction

If generative interviews are to generate grounded explanatory models, this raises the fundamental issue of how the explanatory models at level three of Table 1 are formed in the mind of a scientific

investigator. Much of the recent progress in history and philosophy of science can be seen as a struggle to move away from a simplistic view of how theories are formed in science as being either pure induction upward from observations or pure deduction downward from axioms, followed by testing. Instead, we see movement toward a view that involves both top-down and bottom-up processing in a cycle of conjecture and revision, as shown in Figure 1. Many modern scholars in the history of science and cognitive studies of science now view the process of how models are constructed in science as a cyclical process of hypothesis generation, rational and empirical testing, and modification or rejection. It is difficult to describe so complex a process in a single diagram, but the simplified model in Figure 1, reflecting the work of a number of scholars, will aid in the present analysis. A major change in the basis for such views is that the most recent work has been grounded in systematic studies of scientists rather than in abstract analyses of the nature of science. Theory formation and assessment cycles of this kind have been discussed by Gruber (1974), Nersessian (1984, 1992), Tweney (1985), Thagard (1992), Giere (1988), and Darden (1991) based on studies in the history of science, and by Dunbar (1994) and Clement (1989b) based on naturalistic and “think-aloud” studies of expert scientists. This work means that major elements of the view shown in Figure 1 now have empirical grounding in studies of scientists.

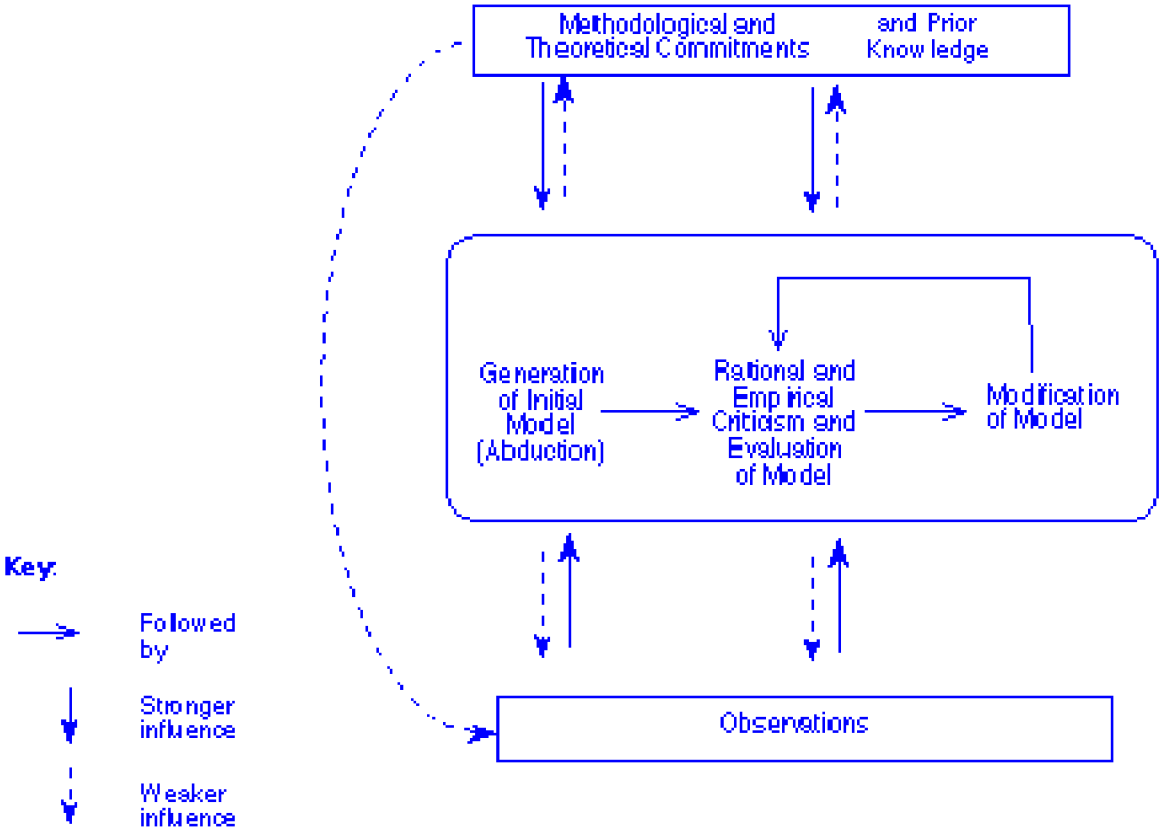

Figure 1

Basic Model Construction Cycle
With Top-Down and Bottom-Up Influences

Essentially, the scientist aims to construct or piece together a theoretical model in the form of a conjectured story or a picture of a hidden structure or process that explains why the phenomenon occurred. Peirce (1958) and Hanson (1958) used the term “abduction” to describe the process of formulating a hypothesis that, if it were true, would account for the phenomenon in question. The initial hypothesis for a hidden mechanism is often more like an abduction than a recognized pattern or induction, and it can be a creative invention as long as it accounts for the observations collected so far.

However, it also should be a very educated invention, reflecting constraints in the scientist's prior knowledge about what might be the most plausible mechanisms involved. Thus, an initial hypothesis for an explanatory model often can be formed by an abductive process, possibly for just a single instance of the phenomenon. Then, the initial model is evaluated and revised in response to criticisms. This can involve evaluations by comparisons with new data, or it can involve evaluations via rational criteria such as simplicity and consistency. By such a process of successive refinements, we cannot arrive at absolute certainties, but a viable and successful explanatory model may be formed. As von Glasersfeld and Steffe

(1991) put it: "The most one can hope for is that the model fits whatever observations one has made and, more importantly, that it remains viable in the face of new observations" (p. 98). And as Hanson (1958) and Kuhn (1962) have emphasized, observations are theory-laden in that they can be shaped and influenced by the investigator's prior conceptions and theories. This is shown by the downward arrows leading to the observations in Figure 1, which indicate that observation concepts and categories can change and develop along with the theoretical model. (See Lakatos, 1978, for a parallel view from mathematics.)

Maintaining The Distinction Between Observation And Theory

In all of the above, however, there is an effort to maintain the distinction between observation and theory. Although these two processes interact, they still can be seen as separate, semiautonomous processes rather than a single process. When they conflict, it is seen as essential to assign greater, but not overwhelming, weight, in the long run, to documented, publicly agreed-upon observations over prior theoretical commitments and newly invented theories. It is also desirable, where possible, to describe explicitly relations of support between observations and theory. Although intermediate examples can be raised that suggest viewing the theory-observation distinction as more of a continuum, the effort can be made still to progress toward separating these two modes of thinking for doing science so that they constrain and stimulate each other. The important goal is to progress toward observation processes that may not be fully "objective" but that are still independent enough from one's theories to yield surprises and anomalies that can motivate revisions in, or even eventual rejections of, the theories.

Using the Interpretive Analysis Of Generative Clinical Interviews to Construct Models

The above perspectives from the history of science provide a methodological grounding for analyzing protocols in generative studies in education. In this method, analysts construct, criticize, and revise hypothesized models of mental structures and processes repeatedly while using them to explain as much of the data in a protocol or a set of protocols as possible. According to the framework developed here, this process is one of abduction, repeated criticism (from both rational and empirical sources), and responsive revision.

Returning to our algebra study example, we gradually became more and more dissatisfied with the "word-order-matching" hypothesis as the only source of the reversal error. Sections of transcript such as the following one from a parallel problem served to discount this hypothesis as the only explanation. The problem is to write an algebraic expression for the statement: “There are eight times as many people in China as there are in England.”

The student wrote: 8C = 1E.

Transcript line 9: Student: ”for every eight Chinamen, but then again, saying for every eight Chinese, yeah, that’s right, for every eight Chinese, there’s one Englishman.” [The interviewer asks the student to explain his equation.]

12 Student: “All right, it means that there is a larger number (points to 8C) of Chinese for every Englishman” (points to 1E).

13 Interviewer: “So you’re pointing to the 8C and the 1E?”

14 Student: “Yeah, and there is a larger number of Chinese than there are Englishmen; therefore, the number of Chinese to Englishmen should be larger - 8C = 1E.”

In line 12, subject S2 indicated clearly that he had comprehended the relative sizes of the two groups in the problem--that there are more people in China. This suggests that he had gone beyond a syntactic, word-order-matching approach and was using a semantic approach dependent on the meaning of the problem. At this point, the "word-order-matching" model was no longer a viable one for these parts of the protocol. The student's intuitions about how to symbolize this relationship were to place the multiplier (8) next to the letter associated with the larger group. This approach is a very literal attempt to compare the relative sizes of the two groups in a static manner, and we detected it in a significant number of students. Therefore we labeled this the “static comparison” approach.

This appears to be an incorrect, but meaningful, way for this group of students to symbolize the relative sizes of two groups--indicating the “base ratio” of eight people in China associated with each person in England. At first, we were blind to this thought pattern because we had come to take for granted the idea that the students were using word-order matching. But an intensive case study of a single student led to generating (abducting) the hypothesis that a meaningful but incorrect symbolization strategy different from word-order matching was being used in these cases. Once we had formulated this hypothesis for one student, *it sensitized us to new observations to look for*, such as the relative- size references. We ultimately found evidence for the static comparison approach in a majority of the subjects making reversal errors on such problems, even though we had not "seen" the evidence in our initial look at the transcripts. This second source of the reversal error was harder to detect, requiring us to go through several of the cycles shown in Figure 1. Thus, by criticizing and revising one's models of the student’s thought process repeatedly, one can develop progressively more adequate models and observation concepts through a model construction process that is used by scientists in general.

An Initial Dichotomy of Clinical Interview Methods: Generative Purposes Lead to Interpretive Analyses Whereas Convergent Purposes Lead to Coded Analyses

The previous examples prepare us to distinguish two major purposes of studies conducted in educational research: *generative* and *convergent.* I will begin by painting these contrasting purposes as a dichotomy in order to introduce the issues, then broaden it later into a spectrum with intermediate points, in order to suggest some of the options available for different research purposes.

Purposes of Clinical Interview Studies:

- *Generative* purposes usually lead to an *interpretive* analysis. Such a study can deal with behaviors that are quite unfamiliar, for which there is very little in the way of existing theory. It attempts to generate new elements of a theoretical model in the form of descriptions of mental structures or processes that can explain the data. This method can deal with larger and richer sections of interview data involving more complex processing. It entails higher levels of inference on the part of the researcher.

- *Convergent* purposes usually lead to a *coded* analysis of interviews that attempts to provide reliable, comparable, empirical findings that can be used to determine frequencies, sample means, and sometimes experimental comparisons for testing a hypothesis.

An interpretive analysis in a generative study tends to present a relatively large section of a transcript, followed by the author's inferences concerning the thought processes of the subject. In contrast to a coded analysis (see below), observation categories are not fixed ahead of time. Analysis can generate new observation categories and models of mental processes giving plausible explanations for the observed behavior. (For example, all of the models of reasoning modes that produce reversal errors were generated during intensive interpretive analysis of individual case studies done for generative purposes.)

There are a number of reasons for the importance of generative studies:

- They are a primary source of grounded theoretical models for learning processes.
- They are a primary source of key observation concepts.
- They are not restricted to collecting immediately codeable observations that fit into existing categories of description; they *allow investigators to develop new categories for description.*
- They provide a foundation for the design of convergent studies.

Interpretive Versus Coded Analysis

This contrasts with a coded analysis in a convergent study which focuses on observations that are assigned to predefined categories by a coder, usually from relatively small segments of a transcript. A transcript is coded when the analyst formulates criteria for recognizing a phenomenon and then lists the

places where that phenomenon occurs in the transcript. The conclusions may be at the level of observation patterns alone, or they can be used as data to support or reject theoretical hypotheses that may have been generated by other means. Two examples of research questions in this type of study are:

- Among a set of previously characterized behaviors, which do these subjects exhibit ?
- Among a set of previously characterized mental concepts or processes with previously hypothesized behavioral indicators, which do these subjects possess?

By using the names 'coded' and 'interpretive' I do not intend to deny that coding of a transcript in natural language that is ungrammatical and incomplete is also an act of interpretation. The intent is only to signify a greater degree of interpretation in the analysis of generative studies.

The Viability Of Theoretical Models In The Analysis Of Clinical Interviews

Viability Versus Validity

Generally, in this chapter, the issue of the viability of a model speaks to our interest in attaining models that are useful to us, that have support in the data, that fit or interact productively with our larger theoretical framework, and that give us a sense of understanding by providing satisfying explanations about hidden processes underlying the phenomena in an area. The literature on validity is tangled with multiple meanings and interpretations for the term. Validity of what? Widely differing referents occur including the validity of an observation, of a theoretical model, or of the relation between an observation and a model. For the sake of avoiding confusion with these past uses, the best strategy, to be used in this chapter, is to put aside the term “validity” altogether in this context and to define a new methodological term-- “viability” of a model-- based on the relation between models and observation in science as described by modern historians of science.

Because the job of generating a framework at this level is such a basic one, it makes sense to use a broad definition of viability: the view that the viability of a model should be no less than an estimate of its usefulness or strength as a scientific theory compared to other theories. The bad news here is that strength is a complex thing to measure, and the criteria for it involve human estimates and judgments. The good news is that modern work in the history and philosophy of science has shown that this is a problem common to all of the developing sciences and has made considerable progress on describing these criteria in more realistic terms. Criteria for evaluating scientific theories have been discussed by Darden (1991) and Kuhn (1977) among others and I will use a condensed version of their standards in what follows.

Determinants of Viability

The "explanatory power and usefulness" of a model correspond roughly to what we mean by the "viability" of a model. Given a set of observations, the major factors that affect the viability of a theory or a model that explains them are its plausibility, empirical support, rational (nonempirical) support (such as its coherence with previously established models), and external viability (or "tests over time"), such as its extendability to new contexts. These factors are discussed separately below and summarized in Table 3.

Insert Table 3 about here

1. Plausibility: Explanatory Adequacy and Internal Coherence

The most basic requirement for a model of a person's cognitive structures and processes is that it give a plausible explanation for the observed behavior. We cannot provide merely an informal mixed metaphor as the model. It must be a description of a thought process that we can easily imagine taking place in the subjects, that explains their behavior, and that is internally coherent. This last criterion refers to whether the model is internally consistent, both logically and pragmatically, in terms of the story it tells about a mental mechanism (e.g., it does not speak of conscious processing on unconscious material).

2. Empirical Support

Interviews are subject to what Hayes (1978) has called the porpoise effect: we only see a part of the porpoise a part of the time, and we only derive partial and indirect information on intermittent parts of mental processes by viewing the data in interview transcripts or tapes. Mental processes are by nature hidden processes. Thus, our concluding hypotheses about models of processes in a report will be stronger or weaker depending partly on how much support they derive from empirical observations; and that depends on how prevalent and how pertinent the data are. The strength of empirical support depends primarily on the three factors listed below:

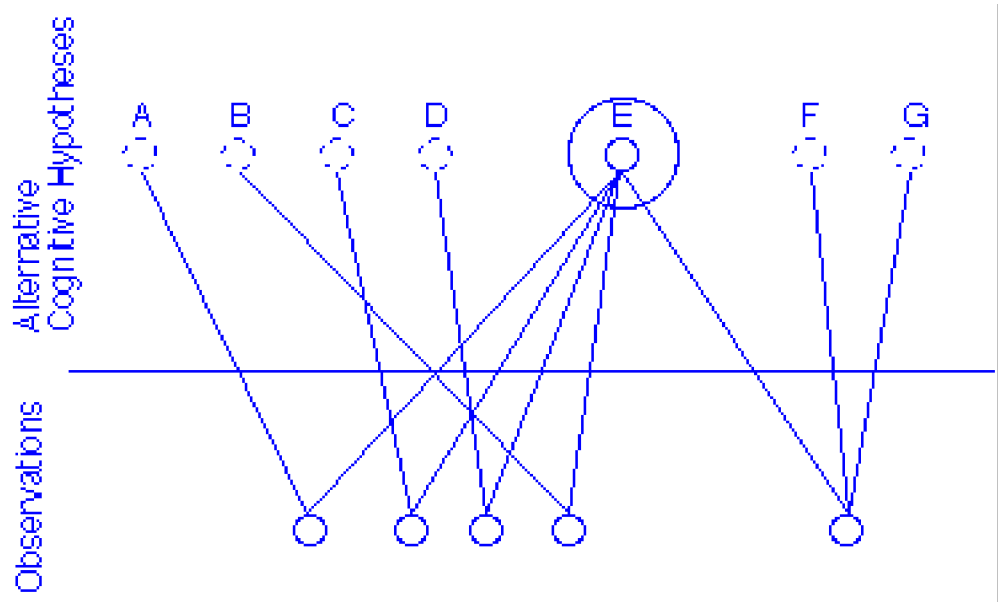

Figure 2

Triangulation: Hypothesis E
With Multiple Sources of Support

a . Explanatory adequacy and scope by means of triangulation and number of supporting observations. The empirical support that a model has rises with the number of supporting observations that it has; in other words, the number of connections to data. When a model gives a plausible explanation for an observation, it, in turn, receives a degree of empirical support from the observation. When a model provides an explanation for more than one observation or aspect of a transcript and derives support from all of them, we say that we can triangulate from the observations to the model. In this way, model construction is responsive to the multiple constraints provided by the transcript. Inferring new models from evidence in protocols is an inherently difficult and creative construction process. when one can triangulate by explaining multiple observations with the same hypothesized model, shown schematically as hypothesis E in Figure 2, that gives one a stronger degree of support for the model.

For example, in the case of the static comparison approach as an origin of reversal errors discussed earlier, a central assumption of this model was that the students comprehended and heeded the relative sizes of the two groups (implying they thought that they were symbolizing that relationship). This assumption distinguishes the static comparison approach from a purely syntactic one of word-order matching. There are three different indicators of this in the excerpts from the transcript above, in lines 9, 12, and 14. Thus, we can triangulate from these statements

to the hypothesized model that a correct, relative- size idea is a part of subject S2's comprehension of the problem.

(Note: Here, "Support" usually means "corroborates" rather than "strongly implicates." On its own, each observation may not provide substantial evidence. Furthermore, a single observation can have explanations other than the one proposed for it, as symbolized by the pairs of lines reaching upward from each observation symbol in Figure 2. But when one hypothesis fits more observations than any of the other hypotheses, this can support the hypothesis as the most plausible explanation. Thus the relation between observations and hypothesized models used here is not one of unique implication but, rather, of collective abduction and support.)

Insert figure 2 here

b. The strength of each connection of support between an observation and a hypothesis--that is, the directness and quality of each supporting observation (e.g., if a subject solving a mathematics problem mentions the name of each step of a standard algorithm as it is used, that observation has a strong connection of support to the hypothesis that he or she is using an algorithm). Figure 3 shows a simplified case where theoretical aspect T1 is supported by observations O1-O3 and T2 by O3, O4 and O5. The T's stand for theoretical models of a cognitive process (such as the static comparison approach). The P's stand for segments of transcript that constitute relatively raw data used in the analysis. (Even here, the transcribing of slurred words or incomplete phrases will involve some top-down interpretation, but, in general, the transcript level will involve the least amount of interpretive inference.) The O's stand for observation concepts (behavior patterns such as reversed equations and stated, relative-size comparisons) that have been identified as relevant. One particular observation (e.g., O1 in Figure 3) may provide more direct evidence for (have fewer layers of interpretation leading to) a theoretical model (shown as a heavier line of connection to the theory) than another observation. (Some would call this the "validity" of using that observation to infer that hypothesis. Of the concepts in Figure 3 that have been discussed, this is the closest one to "content validity" in the testing literature.)

Insert figure 3 here

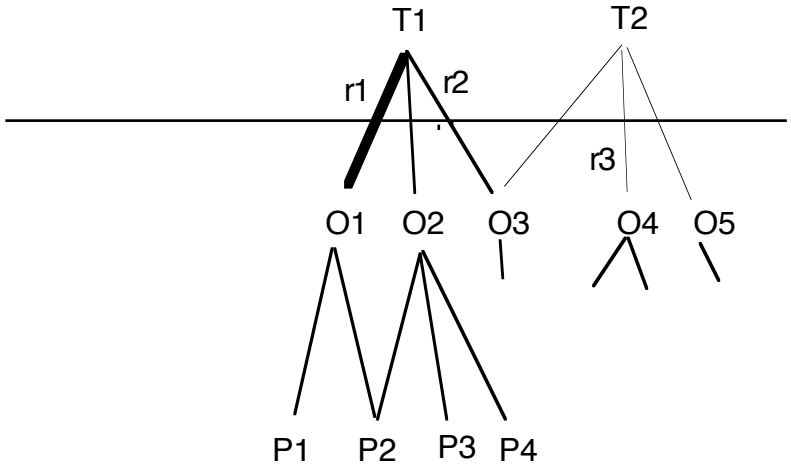

Figure 3

Levels of Analysis in a Clinical Interview

c. Absence of conflicting evidence or anomalies that are inconsistent with the model. In any study of rich behavior, there will be some observations that are not explained by the current theoretical model, and perhaps some that are in conflict with it; the latter are called anomalies with respect to that model. If there are too many anomalies, a model may lose its viability in comparison to another competing model.

Together, factors a, b, and c determine the strength of the empirical support for a hypothesized model and are one component of its viability, as shown in Table 3. Wide triangulation from multiple instances is desirable when available, but it is not the only criterion, especially in initial studies of an area where "existence proofs" (read "arguments from even a single observed instance for instituting a new category of behavior or category of mental processing") are needed.

3. Nonempirical Criteria

Other criteria given in Darden (1991) and Kuhn (1977) for evaluating scientific theories are nonempirical. I have already discussed two of these above under the heading of Plausibility: explanatory adequacy and internal coherence. Others are:

a. Clarity: whether the model is clearly described and comprehensible, allowing it to be used as a tool for thinking.

b & c. Simplicity and lack of ad hocness: whether the model introduces arbitrary elements in order to "kluge" (contrive) an explanation. Other factors being equal, a simpler model is preferred.

d. External coherence: whether the model is coherent with other accepted theories in one's larger theoretical framework, again both logically and pragmatically. For example, the theory that students can tend to use an overly static approach to algebraic symbolization becomes more plausible when one sees related behaviors in a very different domain such as Cartesian graphs. There, one can observe some students symbolizing the speed of a bike coasting over a hill with a speed versus time graph that shows the same hill-like shape! This error has the same character of an overly literal correspondence between the symbol and the symbolized (Clement, 1989a).

4. External Viability of a Model Across Contexts and Populations

Further criteria from Kuhn (1977) and Darden (1991) concern the extent to which the model has explanatory power in the sense that it can be applied to contexts outside the realm of the original study. By other contexts, I mean a different task, setting, or population for the study.

a. Generalizability (of a model): Can the existing model be applied to other contexts in order to explain existing data? For example, the static comparison approach helps to explain why reversal errors are also observed in problems where students translate from pictures (rather than sentences) to an equation. (These errors could not be explained by the word-order-matching model.)

b. Predictiveness (predictive validity): Does the model lead to new predictions for other contexts that turn out to be correct? For example, Clement (1982) discusses how the model above led to a prediction for fewer reversal errors in a computer programming context.

c. Extendability: Can the model be modified or extended to account for even more widely varying contexts?

d. Fruitfulness: whether the model is ultimately fruitful as a heuristic for promoting the identification of further interesting observations and hypotheses.

The last three criteria are essentially tests of a model over time, indicating whether a model leads to further productivity in the field.

A Constructivist Approach to Investigation

The above guidelines constitute a far more extensive set than usually are mentioned in clinical interview studies, including my own. I believe that we need to be more explicit about them in preparing and reviewing reports, especially larger reports that describe a program of research.

In this view, a model of mental processes is generated initially from a protocol by a process of constrained inventive abduction and refined by means of an evaluation and improvement cycle. As argued in the first part of this chapter, the model need not necessarily predict observations deductively nor does it need to summarize a strong statistical pattern of observations; it need provide only a viable plausible explanation for a set of observations. It should have internal coherence, empirical support, and usually, external coherence with an existing framework. In such a system, one can argue for the viability of a theoretical model both from above and below in Figure 1. In this constructivist approach, one can claim to have a powerfully viable model that is firmly grounded in an empirical foundation, without overreaching by claiming to have found certain truth.

The Survival of the Most Viable Model

In the long run, the viability of the hypothesized model will be compared to the viability of competing hypotheses, and the one with the highest viability should survive. This will take some discussion in the scientific community because, according to the above criteria, the viability of a theory or model is a complex judgment call. As an estimate of how much confidence one places in a model currently, it is the weighted sum of a variety of factors listed in Table 3. Even the factor of empirical support is itself a weighted sum. This makes writing conclusions about models in a generative, clinical, interview study more difficult because the author should indicate, at least approximately, his or her view of the relative strengths of each portion of the conclusion based on all of these factors. An author will not discuss every item in Table 3 always, but the relevant items should be considered in determining the judgments of how strongly to state each conclusion.

Recommendations For Fostering Model Viability

Recommendations For Increasing Model Viability in Reports on Interpretive Analyses of Generative Studies

The new view of scientific theory evaluation above provides a firmer foundation for suggesting the following general recommendations concerning viability in reporting on analysis work. I begin with generative studies and base the recommendations on the goal of generating viable scientific models. Some of these recommendations(as well as those concerning reliability to be discussed later) will overlap to a certain extent with those discussed by Campbell (1979), Chi (1995), Ginsburg (1983) Goetz and LeCompte (1984), Howe (1985), Howe and Eisenhart (1990), Lincoln and Guba (1985), Patton (1980), and

Smith (1987). Initially, the most basic research questions in this approach are: What might be the important mental states or processes involved? What might be the observation patterns that provide evidence for these processes?

Plausibility and Explanatory Adequacy. The first task is to achieve rudimentary explanatory adequacy--the investigator must construct a plausible model that explains the behavior in question and, at the same time, must refine and sometimes construct the descriptors for the behavior. This is done by means of the cyclical, interpretive analysis cycle of segmenting the protocol; making observations from each segment; formulating a hypothesized model of mental processes that can explain the observations (and suggest others to look for); returning to the data to refine and look for more confirming or disconfirming observations; criticizing and modifying or extending the model; and so on. (For an early description of a similar process called the constant comparison method, see Glaser and Strauss, 1967.) Models will attain progressively more and more viability with each cycle as mental process concepts with supporting observation concepts are formed and stabilized.

Empirical Support and Absence of Anomalies. The analyst interested in developing a viable, robust model must be alert constantly for aspects of a protocol that conflict with the current model. They are what inspire improvements in the model. After the model is as well developed as possible for the current analysis, the "lack of multiple anomalies" criterion for the viability of a model should be examined in the report. It should state whether there are any data elements still in conflict with the model; if they are numerous or important, they will expose limitations in the model. Some of the most valuable observations are those that do not fit any current hypotheses. These can stimulate later construction and revision of a model. Analysts have an obligation to be open to such observations and to report them. Not everything in a generative study has to be explained; new empirical findings can be provocative and valuable in themselves.

Attending to Non-Empirical Factors in Evaluating the Viability Of a Model. Reviewing the list of criteria for evaluating a theory presented earlier, one can recommend that for both generative and convergent studies, authors should take pains to:

- display the external coherence of the model- how it fits with other prior models in the investigator's paradigm. This should include a discussion of the background assumptions from the investigator's prior theoretical framework that are part of the argument leading to the conclusions.
- criticize and revise their models for internal consistency and coherence.
- construct models that are as simple as possible and describe them with as much clarity and transparency as possible. One should be able to *think* and *reason productively* with the models.
- compare modular parts of the favored model to other rival models.

- construct models that are as detailed as possible, but not so detailed that they are hard to support from the available observations. This is the basic tension that determines the appropriate grain size of the model.
- say how the model might be extendible, fruitful, or predictive in other areas to be examined.

Methodological Heuristics and Trade-offs. Darden (1991) points out that the criteria for evaluating models are heuristics, not algorithms. She emphasizes the fact that different criteria can pull the investigator in different directions; that is, there are some difficult trade-offs to be considered when attempting to improve a theory. For example, increasing detail so that the theory is closer to a deterministic computer simulation may increase predictiveness, but it may reduce plausibility, simplicity, transparency, and clarity. If the model becomes too complex for the investigator to think with, that must count as a negative factor for an *explanatory* model. This is a major reason that most scientists prefer to retain visualizable qualitative models even in fields where they have led to the formation of much more complex and detailed mathematical models or simulations. In summary, one cannot simply act to maximize each of the above characteristics because they sometimes lead to conflicting goals--compromises must be chosen that make the model as viable as possible relative to other models.

Other Heuristics for Constructing Viable Models. The above criteria are derived from characteristics of scientific theories as described in recent portrayals of thinking in scientists and may have as much or more to do with evaluating models as with generating them. As such they do not provide very down to earth advice about the difficult initial task itself--the creative act of formulating new theories of cognitive processes. For this reason I will also make some less esoteric recommendations based on personal experience with constructing models of processes from protocol data.

• In an area about which very little is known, a case study can be very powerful. Choosing a transcript that is especially challenging to conventional wisdom or to one's existing theories can be very important because its anomalies will stimulate the growth of a new theory. Once initiated, the theory can be tested and refined by determining whether it can explain the behavior of many other subjects as well.

• When the investigator understands very little about such a case study tape, one may have to view the tape 10 or 20 or more times in order to generate an explanation. Often one needs to "live with" such a tape for a period of time in order to create a viable model.

• Successful scientists appear to engage in model construction cycles in which they alternate between a freely creative, divergent mode of work, and a highly critical, convergent mode (Clement, 1989). It may be helpful for analysts to use these two modes to first generate possible models, then aggressively criticize and eliminate or revise them.

• Having more than one analyst can stimulate the creative mode via brainstorming and the critical mode via argument. Sometimes these modes can be consciously fostered (as in the "no criticisms" rule for brainstorming.)

• A good heuristic is to try to **draw** what happened in the mind of the subject. Mental processes can involve many interconnections and relationships over time. Creative representations in the form of many varieties of diagrams can be very helpful devices for keeping track of such complexity, drawing inferences, seeing new connections, and criticizing and improving one's model. New computerized tools such as graphics packages, Semnet (Fisher, 1990) and Inspiration (Helfgott, et at, 1994) can also be very helpful. The creation of new graphical systems for representing thinking processes can be an important part of a theoretical contribution (Driver, 1973; Clement, 1979; 1989b; Easley, 1974; Newell and Simon, 1972.)

• Creative metaphors can serve as fertile starting points for a theoretical model without committing one to being "stuck" with the metaphor and precluding further development of the model beyond the metaphor. Thinking of conceptions competing or cooperating, or having 'momentum', or autonomy may start as somewhat anthropomorphic metaphors, but may eventually lead to very useful theoretical perspectives.

<u>General Recommendations for Fostering Empirical Support as a Component of Viability in Both Generative and some Convergent Studies</u>

Some convergent studies may simply focus on establishing a pattern of observations in a sample without dealing with theoretical models of internal processes. However in other convergent studies within an advanced area of work, the above criteria for the formation of models have been largely satisfied already before the study begins. In other words, viable models have been developed along with stable observation categories. The task remains to show clear connections of support between observations and models, and this is reflected in the following guidelines applying to such convergent studies as well as generative studies.

In a report, an analyst assesses aspects of the model and supporting observations and makes judgments of the model's viability. The strength of each conclusion should reflect this assessment. Investigators also should provide material that enables readers to make judgments of viability, so that readers can compare theories and data, and make global judgments of plausibility and viability. By displaying the patterns that one has identified, or at least sections of raw transcript, the writer allows readers to make their own judgments of the viability of the theoretical claims made.

Whenever possible, studies should display triangulated connections to an aspect of the model from the observations that support it . Readers will be able to make even better judgments about the viability of the model when this is done explicitly. Triangulation in the broad sense, can occur not only across parts of a transcript but also across observation methods, tasks, and subjects.

As discussed earlier, in any study of relatively rich behavior, an analyst will construct models for different aspects of a protocol that have varying amounts of support in the data. There will be strong support for some models and only initial support for others. Conclusions about models should be stated with different levels of strength to reflect such variations. This suggests that there should be at least two sections in the discussion of the theoretical findings in a report. The first can be written conservatively, with only specific, well-supported conclusions about (models of) the subjects in the experiment. By making explicit the arguments and warrants for each conclusion whenever possible, the author should try to give readers enough information to allow them to make parallel judgments. The availability of transcripts, upon request, to interested readers is helpful in this regard. Then, a section that is described as more speculative can be written about the hypotheses that the investigator feels are reasonable, based on his or her entire past experience. In addition to contributing to existing theory, this can provoke new studies for evaluating these hypotheses.

The Quality of Subjects' Reports

Investigators can try to improve the quality of the subjects' reports of their thinking. Ericsson and Simon (1984) have discussed studies of the extent to which subjects' reports reflect an incomplete, biased, or inconsistent account of their thinking during problem-solving. The quality of these reports affects the quality of the data in the study. They conclude that, under the right conditions, thinking aloud does not create major distortions in a subject's thought processes. They urge that subjects be encouraged to think aloud but not to reflect, at a psychological theory level, on their own thought processes. In this way one circumvents some of the classical objections to introspection raised by others (e.g. Nisbet + Wilson , 1977). Reporting is assumed to be restricted to processes that come into conscious attention. Therefore subjects will be able to report on some processes better than others, and their statements will always be incomplete. Ericsson & Simon argue that thinking that uses nonverbal representations, such as imagery, will be somewhat harder to report in detail because it must be translated into language, but the cognitive load required to give at least partial information on these aspects of thinking is not so heavy as to distort them appreciably.

In this chapter, I cannot do justice to the extensive work in this area, but the reader is referred to Ericsson and Simon (1984) for further discussion of the status of “think-aloud” protocols as data, the quality of subjects’ reports, and associated issues of reliability and validity. Here, I will comment only on one consideration involving the design of interview probes (questions that an interviewer uses to elicit a subject's thoughts within a problem solution.) Some techniques for trying to increase the quality of the

subject's reports are: training a subject to think aloud without theorizing about their thought processes; probing to encourage a sufficient amount of thinking aloud or output from a subject; and probing to request that the subject clarify a report. The controlled application of these three techniques can be important in determining the quality of the reports that subjects give. However, while using probes, the interviewers should try to minimize the interference or influence that they might have on the subjects' thinking. This sets up a major trade-off in the decision of how much probing to do during an interview. Put in the simplest terms, the need for completeness argues for doing more probing, whereas the need for minimizing interference argues for doing less. A compromise appropriate for some purposes is to do only non-intrusive probing early on (e.g., the interviewer can say: "Please think aloud;" or "Say that again, please?") and more intrusive probing only later in an interview. How one resolves this trade-off can also depend on the purpose of the interviews. Researchers trying to describe stable knowledge states (e.g., persistent preconceptions) can allow more intrusive probing than describers of transient reasoning processes (e.g., non routine, problem-solving processes) because there is a smaller interference effect. Also, training subjects to think aloud before starting an interview can sometimes reduce the amount of probing needed. Thus the quality of subjects' reports is another factor that an investigator can try to maximize in both generative and convergent approaches.

Beyond the Generative Vs. Convergent Dichotomy: A Spectrum of Clinical Interview Methods

I began this chapter by describing a dichotomy between generative-interpretive research and convergent-coded research. It is time to expand this dichotomy into a spectrum by adding some intermediate methods. They illustrate the variety of techniques that are needed for a field as complex as educational research.

Constructing New Observation Categories

First, however, I need to expand on the idea of what it means to define an observation construct or category. In even a ten-minute section of videotape, we are faced with a continuous stream of behavior that can be extremely rich. Assuming that we transcribe each statement, should we go beyond this to record each type of gesture, each voice intonation, each action with materials, and each line of a drawing? Or should we use larger meaningful units of activity such as "draws a bisected triangle"? A difficulty with a rich source of data source like a videotape is that there is too much data to analyze in a meaningful way! The investigator must decide what aspects of such a continuous stream of behavior are most relevant to the purpose and context of the study. This is the problem of identifying and describing observation concepts. The investigator must also determine what is relevant depending on the level of the research questions in which he or she is interested. Some of these choices will be obvious. But, for others, investigators must narrow their focus gradually to relevant observations and their descriptions and labels.

Often this happens as they converge on an insightful theoretical model. For example, in the case of the static approach as a source of reversal errors, it was not at all clear in the early case studies that students' references to the relative sizes of the two variables were an important observation category. Only as the theory of a static comparison approach was developed did that observation become important.

This illustrates that it is not always the case that theory emerges from observations. In generative studies, emergence sometimes occurs in the opposite direction or hand in hand, consistent with authors who refer to observations as more or less "theory-laden."

A Spectrum of Clinical Methods and Their Relation to Levels of Development in Observation Concepts

Table 4 shows a spectrum of methods from the generative to the convergent, with intermediate methods between them. There is an increasing level of observational reliability and quantifiability as one moves from approach A to approach D. Approaches A and B correspond to generative studies whereas C and D correspond to convergent studies. An exploratory study on a new subtopic may use method A alone. As theory and especially observation concepts become defined more explicitly, studies can be designed at the higher-lettered approaches of Table 4.

(Insert TABLE 4: About Here)

Although this chapter concentrates on model viability, I will touch briefly as well on the problem of the reliability of observations in the analysis of interviews. Figure 3 can be used to contrast the concepts of reliability and viability as used here: Whereas viability refers essentially to the "strength" or believability of the theoretical models or findings about mental processes (T's) at the top of the figure that explain behaviors, reliability refers to the "strength" or believability of the observation findings (O's) that summarize behaviors. Concerns about observational reliability stem most fundamentally from an interest in establishing agreement between scientists on believable findings at the level of observations (at levels one and two in Table 1). This is a slightly broader and more fundamental use of the term "reliability" than is used in some circles because it is not limited to notions of consistency in the measurement of observables, although it includes that notion. Just as an important determinant of the viability of a model is the empirical support that a theoretical model receives from the multiple observations connected to it, an important component of the reliability of an observable behavior pattern is the combined strengths of the connections to sections of primary-level, protocol data represented as P's in Figure 3.

From one point of view, Table 4 describes a sequence of methods by which observation concepts can achieve high reliability gradually from top to bottom. The upper portion of the spectrum in Table 4 is associated with a concentration on the development of viable new models and observation concepts in areas where that is needed, whereas the bottom portion is associated with a concentration on pursuing greater reliability of observations. As concepts are criticized and refined and as observation concepts

become more replicable, investigators may move downward on the spectrum in progressing from one form of analysis to the next .

Seeking improved reliability does not imply seeking a level where observations become "infallibly true readings of the book of nature." Rather, one seeks to establish a data base of conclusions about summaries of behavior that are derived more directly from that behavior and that will attract as wide a span of agreement as possible. These observations can then be used as a relatively solid foundation for supporting more interpretive theoretical claims in a study.

Studies of each type listed in Table 4 are needed and each is important scientific work. A guiding principle is that work on an area of behavior should take place using the approach most appropriate to that area. In general, an area that is more complex, implicit, unexplored, or "hidden" in some other sense will require work using a lower-lettered approach.

Higher-level thinking of the kind usually studied in educational research is arguably the most complex process of all those studied by the sciences. Thus, in areas where processes are complex and relatively unexplored, it may take several investigators years to move through the lowest-lettered approaches. The generative stage is crucial. As illustrated by Darwin and Faraday, the ability to experiment with interpretations of data and to reorganize and modify critical qualitative concepts is part of a criticism and improvement process that can play an essential role in scientific progress (Gruber, 1974; Tweney, 1985). This task will be facilitated by communication between groups. Therefore, it is important that there be channels for publishing the work that takes place in each of the approaches, including the lower-lettered ones.

Additional Guidelines for Fostering Model Viability in each of the Four Types of Studies

In general, approaches A and B in Table 4 correspond to generative studies in section 1 whereas approaches C and D correspond to convergent studies. Recommendations on maximizing viability from the previous sections of this chapter apply accordingly, with the following additional considerations for each category in Table 4:

A. Exploratory Studies: In method A of Table 4, one is striving for initial explanatory adequacy. Descriptions in a report combine models and observations, so models have rough empirical support in observations, but the connections between them are not spelled out in detail yet . Rough initial plausibility of the "whole description" of an event can be judged by the reader when transcript excerpts are included. Some rational criteria for viability (e.g., clarity) apply also. Extendability, fruitfulness, and the potential generalizability of the model, as defined in the third section of this chapter, are major goals at this stage.

B. Grounded Model Construction Studies: In method B, observation concepts and model elements are being constructed separately. Accounts of empirical support for a model by

means of triangulation from and the strength of connections to observations begin to be possible but may not be explicitly defined everywhere. The focus is on all four groups of criteria for the viability of a model in Table 3 and the attempt to construct and describe the essential elements of the model.

C. Explicit Analysis Studies and D. Independent Coder Studies: In these methods explicit criteria for the tallying of observations and in some cases their connections of support to explicit descriptions of models are developed. These steps make tighter argument structures to conclusions about models possible, based on the criteria shown in parts 2 and 3 of Table 3.

Reliability of Observations: Progressive Levels of Development

Reliability Concerns Observations, Not Theories

In the terms used here in, it is observations rather than theoretical models of thought processes that are examined with respect to reliability. That is, concerns about reliability stem from an interest in establishing agreement between scientists on believable findings at the level of *observations*.. In the broad sense, observational reliability refers to the "strength" or believability of the observation findings that summarize behaviors. Such findings will be more convincing to other members of the scientific community. In contrast to this, *theories* are evaluated with respect to broader criteria (cf. the criteria for evaluating viability described earlier). Thus, we are concerned here primarily with the reliability of observations, not theories.

Varying Needs for Reliability

I will discuss the factors involved in fostering the reliability of observations according to the different approaches in Table 4. It makes sense that there will be emphasis on very different aspects of reliability at the two ends of the spectrum in Table 4. For example, there is no mention of measuring reliability across multiple coders in the lower-lettered approaches; this is possible only after a criticism and improvement cycle has operated long enough to create relevant and stable observation concepts.

We also have the following very central trade-off. The more complex or implicit the mental process being investigated, the more difficult it will be to gather evidence on it. It will be harder to find relevant observables that can be coded and harder to connect them directly to the process in a concluding argument. Related to this is the idea that the longer a mental process takes to complete, the larger the sections of transcript that will refer to it. It will be easier, initially at least, to interpret such a section as a molar unit, rather than breaking it down first into coded fragments. I am not saying that it is impossible to eventually develop relevant and supporting codings that can contribute to evaluating certain models of complex processes, but I am saying that it becomes more difficult as the complexity increases. The more

complex the process, the more difficult it will be to develop coding techniques for multiple coders to use to gather direct evidence for it at very high levels of reliability. It is possible that, for some topics, this level of reliability may be unattainable. Therefore investigators determine whether the purposes and topic of the research justify the expenditure of resources required to document high levels of reliability. Procedures for fostering reliability will vary across the approaches in Table 4, as will be described.

Recommendations for Fostering Reliability

Mechanical Recording Of Data. In all of the approaches in Table 4, audiotape, and, even better, videotape, are a great help because they retain a rich record of behavior that can be reexamined again and again. Of course, differences in what is "heard" and "seen" by investigators can still occur, but an automated record of this kind can raise the level of observer agreement significantly on simple behaviors like speech and actions. Portions of them can be included in reports as transcripts and shared with readers. Two-camera, picture-in-picture recording systems are useful for capturing subjects and a computer screen or written work at the same time.

Procedural Replicability of Interviewing Procedures. This refers to the extent to which the procedure of presenting questions and probes to the subjects is specified explicitly enough to be replicable across different subjects within the study or, possibly, by a different investigator with different subjects. Replicability of the interviewing procedures should be designed into all of the approaches in Table 4 as appropriate to the research objectives. It can help to increase the generality of observational findings across subjects in and outside of a study. Which procedures are standardized will depend on the issues that one wishes to reach conclusions about over groups of subjects. Readers are referred to Goldin's chapter in this volume for an extensive discussion of many key points in this area.

Observational Reliability for Approaches A and B in Table 4. The new studies of scientists described at the beginning of this chapter tell us that the model generation process also involves the construction of new observation concepts. This means that in generative clinical interview approaches, theory and observation concepts are developed together. Observation is highly constrained by the primary level data, but observation categories are still being formulated, and questions will remain about what part of the data will be the focus. These initial observations can still be extremely valuable as suggesters of initial hypotheses for models that can start and fuel a criticism and revision cycle. Because the observation concepts are continuing to change and improve, this lessens the focus on (and the possibility of) formal demonstrations of observational reliability at this stage. The primary focus will be on the ambitious tasks of forming viable models of what the subject is thinking and, at the same time, finding initial observation categories that are relevant to what the subject is thinking.

Nevertheless, one can estimate the relative believability and strength of one's different observations in a report and reflect them in the level of certainty expressed in one's descriptions. For example, one can report high estimated reliability for an observation referring to relative-size relations for a

subject who gives a very clear and full report focusing on the relative sizes of two groups in an algebra solution, as opposed to a subject who gives a very short and muddled report. Readers can make their own judgments whether they would describe patterns in the data in the same way by comparing sections of transcript to the observation summaries of the investigator. Thus, displaying at least sample sections of transcript to illustrate the bases for the investigator's observations can enhance readers' judgments of potential reliability.

In sum, even with generative approaches, there are opportunities to increase certain types of reliability in a useful way. I have identified four major factors to which generative studies can pay attention even though, typically, they do not attempt to replicate observations across samples in any formal way. They are:

- increasing the quality of the subjects' reports (discussed earlier)
- mechanical recording of a permanent data record
- replicability of interviewing procedures for presenting tasks and probes
- helping the reader (as well as the investigator) to estimate the credibility or reliability of the observations that were made from a transcript.

Attention to these factors can help an investigator to more firmly ground the findings of a study in a set of observations.

Approach C: Explicit Analysis Studies. In these studies, the separation of observations from theoretical mental processes makes possible some stronger considerations of the reliability of the observation concepts. Here, there is movement toward precise description of observation procedures in a way that could potentially be done by others as well as the analyst. The analyst attempts to arrive at clear criteria that can be used to assign sections of transcript data to observation categories. In defining observation concepts, there is an attempt to find low-inference descriptors--those that require minimal interpretation--as a way of facilitating agreement between observers. However, seeking low-inference descriptors at the level of observations should not inhibit an analyst from also using high-inference descriptors at the level of theoretical models. When such descriptors explain multiple observations, they can be quite powerful.

As one's definitions of observation concepts approach a level where they can be used independently by others, but prior to any attempts at independent coding, the reader who is given sections of raw data and these observation criteria can use the criteria to make a more accurate assessment of whether he or she would be likely to code the data in the same way.

In addition, in approaches A, B, and C, nonindependent multiple observers, if available, can consider data together and try to come to consensus on observations. Agreements and disagreements are reported. For the former, one can say that the possibility of agreement between observers has been shown.

Approach D: Independent Coder Studies. Here multiple observers are first trained to use explicit coding criteria, usually with practice items. Then they code the protocol data independently, tallying instances of each coding category. Inter-rater reliabilities can be calculated as the percentage of agreements out of the total number of judgments. Space does not permit a complete discussion of coding procedures here but an extensive discussion is given in Ericsson and Simon (1984).

Multiple observers need careful training and discussions to attain a common understanding of the criteria to be used for coding. Trained observers from the same research team are used in all branches of science; one is not required to replicate observational findings using laymen recruited at random from the street. This acknowledges the view that observations can be theory-laden. Successful agreement between independent coders requires subject verbalizations that are relatively clear and prolific, disciplined coders, and highly explicit and clear coding criteria (usually involving low-inference descriptors only).

Some studies may use several levels of coding such as primary-level transcript data, low-inference observations, patterns in those observations, and so forth. A less common and more ambitious goal is for the study to be replicated by a different research team on a new sample of subjects. Such external replicability requires, in addition to the other criterion in this chapter, high procedural replicability and the ability to select a similar sample. Such findings, therefore, are restricted to a context and sample with very similar attributes.

The Choice of Method Depends on the Goal of the Research. In summary, there is a spectrum of methods used in clinical interview studies that vary according to whether the purposes of the study are more generative or more convergent. On this dimension, the aim is to choose the level of analysis appropriate to the topic depending on the extent to which the observation categories either are established already or are being generated and refined. (Note: Some studies may use more than one method in Table 4. However, even in these cases, I find this breakdown useful in helping to separate and define the purposes of different activities in such a "compound" study.)

Two Types of Generalizability

Theoretical versus Observational Generalizability

The ability to replicate observations externally across samples also can be seen as a test of the generalizability of the observations across subjects within a population from which the sample has been drawn. For example, one hopes that the observations of reversal errors in freshmen engineering majors in our study would generalize with only minor differences to other samples in the population of freshmen engineers in general. We can contrast this observational generalizability to a second kind of generalizability--that of the theoretical model. Here, instead of replicating the same set of conditions, the investigator uses the theoretical model to explain behavior under a new set of conditions. For example, in

our study of reversal errors in algebra, we found that students would reverse the placement of coefficients in an equation that they wrote from problems where they translated from sentences to equations, as described earlier. We assumed that this was due to a rather mechanical, word-order-matching algorithm. However, we then discovered by generative protocol analysis that some students were doing meaning-based, nonalgorithmic thinking about images of quantities during solutions and still making the reversal error by placing the larger coefficient next to the letter symbolizing the larger quantity in the equation. From this model of the error source, we formed the prediction that the reversal error should appear in other contexts that could not involve word-order matching, such as problems that started from pictures, data tables, or even graphs. This prediction turned out to be correct. Thus our generative study developed a model of a nonstandard, mathematical-reasoning process with predictive validity since it generalized to different contexts.

Also, by designing slightly more difficult problems, we were able to elicit considerable evidence for this difficulty in very different populations from the original (including a sample of university deans!). This illustrates the ability of a good theoretical model in science to apply adaptively to quite different circumstances from the original experiment, in contrast to the much more narrowly defined generalizability of a set of observations to other subjects in a particular population, using a particular procedure in a particular context. This reflects the fact that models are intended to be more general than observation patterns; that is, the generality or power of the model is strong enough to allow it to be applied to a different context or population where we do not expect particular observations to be replicated. Table 5 contrasts the features of generative and convergent studies that affect generalizability.

Insert Table 5 about here.

Thus, there are two, different, major pathways for generalization to occur: *theoretical generalizability* of a model over contexts and populations; and *observational generalizability* of an established observation pattern over samples. However, the range of situations to which coded observations generalize is rather narrow, being limited to cases using the same interviewing procedure and sample characteristics. The predictive applicability of an observation pattern to a new sample is comparatively easy to determine. On the other hand, greater judgment is involved in deciding when and where to apply a theoretical model. The power or potential scope of its application is much wider, given that models are intended to apply to a wider range than the particular observations that they were created to explain. There are limits here, as well, to how different the situation or population can be for the model to apply. But the limits are much less restricted.

In sum, there is a theoretical route for generalizability that operates through a theoretical model, which, if successful, can be applied more widely to other contexts and populations. This is the main route of generalizability for generative studies. On the other hand, there is an observational route for

generalizability to similar samples that extends from observation patterns in one study to similar observations in another study in which the experimental conditions are very similar. The latter type gives us a means for generalizing over samples in a restricted population; the former gives us a means for generalizing over contexts and populations.

Objectivity

Note that the observation concepts for all of the approaches in Table 4 are theory-laden to some extent. Consequently, we do not reach anything like "perfect objectivity" in approach D. Agreement between independent observers in approach D does not guarantee objectivity because those observers may share the same theories and special schemata that effect those observations. Again, the most important foundation for progress seems to be to work toward empirical constructs that may not be completely "objective", but which are still independent enough from one's theories to have the potential to yield surprises and anomalies that can motivate revisions in, or rejections of, the theories. There is a sense in which observations in the higher-lettered approaches are "less likely to be unique" to an individual investigator. But what is of equal or greater value here is that one's own observations be independent to some extent from (and be capable of conflicting with) one's own theories. That is something that can happen with any of the four approaches in Table 4. This is equivalent to saying that if the analyst is doing his or her job, observations will not be completely theory-determined and will be a semiautonomous process. This means that although researchers cannot help but be influenced by their theories when making observations from a tape or transcript, they do try to minimize this influence and attempt to be as open as possible to new or anomalous phenomena.

In summary, observations are not fixed properties of the world. They are fallible processes of the observer that can be affected by the observer's theories, but, in general, they are recognized as typically more trustworthy than theories, and procedures should be designed so that observations are independent enough to generate criticisms of those theories.

Why Both Generative and Convergent Studies Along All Parts of the Spectrum Are Important

Why Generative Studies Do Not Focus on Formal Measures of Reliability

Returning to the original dichotomy of generative versus convergent studies for the sake of simplicity, I can now discuss why generative studies do not focus on formal measures of reliability. In a previous section I listed a number of recommendations for fostering the viability of models generated in generative studies. The main activity of these studies is to: (1) formulate well grounded, viable models that are central and general enough to transfer to other situations and contexts. In other words, we wish to develop models of the most important structures and processes involved in learning; (2) formulate relevant

observation concepts that are important in the context of what the subject is doing. In generative studies these two activities will interact.

Studies of scientists' thinking tell us that hypothesized models of hidden processes are most often formed by an abductive generation, criticism, and revision process. In the early stages of this process, observation concepts will be generated and modified along with the new theoretical models. This means that in the early investigations in a subfield, seeking traditional measures of reliability and replicability of observations may be quite inappropriate because observation concepts will still be under formation and will not yet be stable. And a preoccupation with formal reliability may inhibit the process by which the concepts are revised and improved. It makes little sense to attempt independent coding with high inter-rater reliability while the relevant observation categories in an area are still being revised and defined. At this stage, the frequent modifications of the definitions of the categories would sabotage any attempts to measure coding reliability. That is possible only after a criticism and improvement cycle has operated long enough to create stable observation concepts.

Rather, in generative studies, model and observation concept formation can take place productively without yet insisting on formal measures of reliability such as agreement ratios between judges. The other, less formal types of reliability and viability discussed as appropriate for generative studies can still provide criteria for fostering productive, high-quality research at this level. Models produced by this process should still have the potential to generalize to other contexts and populations in a manner that goes beyond traditional concepts of the external replicability of observations, as summarized in Table 5.

Higher-Lettered Approaches Allow Enumerative Observations

In approaches C and D in Table 4, once criteria for observation categories have been articulated, observations can be coded and frequency counts on observations can be used as collapsed data. In that case, the approach is *enumerative* in the sense that instances of something can be counted (e.g., there were thirty-one analogies produced by ten expert subjects solving a certain problem [Clement, 1988]) . These data can be used either descriptively (e.g., to support the contention that experts generate analogies) or experimentally (e.g., to compare the number of analogies produced under two different conditions). Thus, observations also become more quantifiable in the higher-lettered approaches.

It also is possible to imagine observation categories from approach D evolving into measurable variables in more traditional, quantitative research using written instruments. These methods do not use clinical interviews and do not appear in the table but can be imagined to appear below the table. In this manner, clinical studies can play another role in developing important, new, observation variables that can then be used in more traditional, quantitative measurement studies.

Study Sequences

One way to view Table 4 is as the downward path of possible development of work on a single topic over a period of time (usually years), as models become more fully developed and observation concepts become more explicit, replicable, and enumerative. However, that is only a part of the story; movement can occur in the other direction as well. Figure 4 illustrates how work at a convergent level (or from a traditional, experimental measurement study) on one subtopic can raise questions that trigger work on new subtopics at a generative level. Therefore, the establishment of a reliable pattern of observation at level D does not necessarily mean that a topic is closed to further investigation. Asking *why* the pattern occurs can reopen the investigation on a deeper plane that will benefit greatly from generative methods. Consequently, the overall picture is not always one of a uniform progression of a large field of study through the various approaches in Table 4. Rather, the progression can be cyclical, and different approaches will be appropriate for different subfields and subquestions at different times.

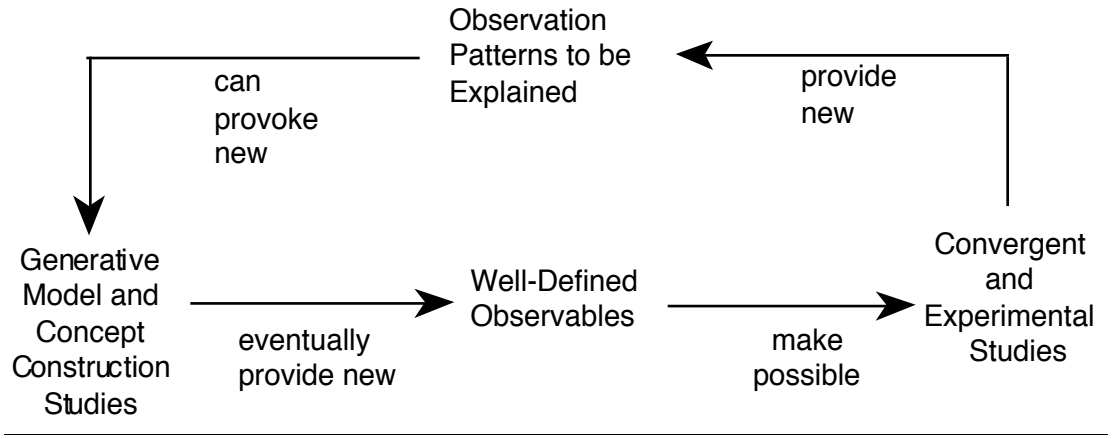

Figure 4. How Work at a Convergent Level Can Initiate Work at a Generative Level

Maintaining a Balance Between Theoretical and Empirical Work

From this point of view, it does not make sense to push ahead for years with developing a very fine-grained model with, say, ten times the number of elements as the number of sections referred to in a transcript, when one does not report on data that are fine-grained enough to support it. Such an exercise may be a worthwhile heuristic in the short run, but empirical analysis should keep pace with theoretical development in the long run. Conversely, it is undesirable to collect large amounts of unanalyzed data because model development should eventually steer the direction for collecting new data.

Comparative Advantages of Generative Versus Convergent Methods

Advantages of Generative Methods. Following the contrasts in Table 5, generative methods are effective for formulating new observation concepts and explanatory models of cognitive processes that are grounded in protocol data. Generative methods can be used with interview data that are very rich in studies of complex mental processes (there has been a shortage of studies that map out such processes). Generative methods are appropriate to the embryonic state of this field. They can deal with behaviors that are unfamiliar and for which there is little in the way of established theory. The weakness of convergent methods is that if they are used too early in an investigation, the investigator can fail to identify key conceptions being used by the subject, key natural-reasoning methods, key processes, and hidden pitfalls in learning. Generative studies are more appropriate for constructing models of these hidden structures and processes.

A successful model can enable generalizations to different contexts and populations. I gave an example of this in the section on generalizability above in which reversal errors are predicted to occur in problems where students translate from *pictures* or data tables to an equation as well as from words to an equation. Thus models from generative studies may be shown to have predictive validity.

Although convergent studies place a higher priority on empirical replicability, a study that has achieved high replicability at the observational level is not guaranteed to have identified viable models at the theoretical level. This is reminiscent of the accuracy and reliability with which Ptolemy was able to predict the positions of the stars using a geocentric model with epicycles. Although he lacked a parsimonious model of the universe, he was able to make extremely replicable observations and accurate predictions. This did not prevent his core model of the universe from being rejected and replaced subsequently. Thus, the reliability or replicability of observations achieved by a convergent study does not on its own imply the presence of a viable and powerful model.

Advantages of Convergent Methods. On the other hand, the value of convergent studies leading to potentially replicable observations lies ultimately in the ability to tie one's conclusions to a relatively stable platform of fixed observation categories with a demonstrated level of agreement between observers and across samples. Observations of behaviors that are widespread and that are shared by other scientists in the field provide the strongest possible foundation for the arguments that will be made in the conclusion of such a study. Arguments to conclusions can be less implicit, complex, and interpretative than in generative studies. In addition, enumerative data from convergent studies allow one to count instances and, therefore, to compare frequencies between two types of conditions. This information can be used as a variable for descriptive, comparative, evaluative or experimental purposes.

Thus, Tables 4&5 along with Figure 4 show how **both generative and convergent methods are essential to the scientific enterprises of educational research and the study of higher- level cognition.** Most generative studies have been done in the fields of education and developmental psychology; this illustrates a key role that these fields have played in cognitive science.

Conclusion

Although some have argued that clinical methods are "less scientific" than more traditional methods in educational research, this chapter argues that clinical methods are an essential part of the scientific enterprise of investigating students' mental processes. In confronting the task of giving an explicit description of a methodology for clinical interviewing, it has been somewhat daunting for me to observe the wide disparities of views on this subject within the field. There are those who do clinical work that resist notions of standards of reliability or replicability in their findings. There are others who insist just as adamantly on formal coding procedures during analysis in order to insure reliability. This chapter has proposed a rapprochement between these camps in saying that, depending on the goal of a particular clinical study, reliability will have a higher or lower priority relative to other considerations. It is my view that both generative and convergent clinical approaches are required for different purposes and that investigators need to be aware of the strengths of, and appropriate uses for, each of them, as indicated in Table 5.

In particular, there are generative studies using interpretive analysis techniques where the first priority is generating viable models and relevant observation concepts. Generative studies allow us to formulate new observation concepts and models of mental processes that are grounded in protocol data in order to explain important, but poorly understood, behaviors. On the other hand, convergent studies using coding techniques allow us to frame conclusions that are anchored in more verifiable observations. Hence, generative studies will be done more often at an early ("Darwinistic") stage in a research topic and convergent studies at a later stage (although this cycle can repeat itself at a deeper level on the same topic). The other approaches shown in Table 4 present intermediate options between these two extremes. These considerations should apply to the analysis of learning processes in studies such as tutoring interviews as well. They should also be extendible to the analysis of social phenomena in groups, but that extension is beyond the scope of this chapter.

Recent history of science studies and studies of cognition in scientists can provide foundational principles for designing methodologies in the learning sciences. What is new in modern history of science studies is that they are willing to describe and analyze theory (model) generation processes in addition to theory testing processes. This has given us an important new resource in formulating methodologies for generative clinical research. There is considerable evidence now that explanatory models in science are generated by an abductive generation, criticism, and revision process. In generative studies, model formation can take place productively through this process without formal measures of reliability, such as agreement ratios between judges. A higher priority goal for these studies is the viability of the proposed model--the strength or usefulness of the model as an explanatory mechanism--although some aspects of reliability are still very relevant. The history of science provides guidelines for defining forms of viability and reliability that are appropriate for generative, clinical interview studies and that can provide criteria for fostering productive, high-

quality research. Models of student thinking produced in this way have the potential to generalize to other contexts and populations in a manner that goes beyond traditional concepts of the external replicability of observations. In this view, generative, convergent, and quantitative measurement methods are seen as linked complementary techniques for generating, supporting, and testing models of students' thinking, rather than as rival approaches. Thus, both generative and convergent clinical methods have roles to play as essential elements of a scientific approach to educational research.

References

Campbell, D. (1979). Degrees of freedom and the case study. In T. Cook & C. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills, CA: Sage.

Campbell, N. (1920). Physics: The elements. Cambridge, UK: Cambridge University Press. Republished in 1957 as The foundations of science. New York: Dover.

Chi, M. (1995). Analyzing verbal data to represent knowledge: A practical guide. Journal of the Learning Sciences.

Clement, J. (1979). Mapping a student's casual conceptions from a problem solving protocol. In J. Lochhead & J. Clement (Eds.), Cognitive process instruction. (pp. 133-146) Hillsdale, NJ: Lawrence Erlbaum.

Clement, J. (1982). Algebra word problem solutions: Thought processes underlying a common misconception. Journal for Research in Mathematics Education, 13(1), 16-30.

Clement, J. (1988). Observed methods for generating analogies in scientific problem solving. Cognitive Science, 12, 563-586.

Clement, J. (1989a). The concept of variation and misconceptions in Cartesian graphing. Focus on Learning Problems in Mathematics, 11(2), 77-87.

Clement, J. (1989b). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. Glover, R. Ronning, and C. Reynolds (Eds.), Handbook of creativity: Assessment, theory and research (pp. 341-381). New York: Plenum.

Darden, L. (1991). Theory change in science: Strategies from Mendelian genetics. New York: Oxford University Press.

Driver, R. (1973). The representation of conceptual frameworks in young adolescent science students. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Dunbar, K. (1994). Scientific discovery heuristics: How current day scientists generate new hypotheses and make scientific discoveries. Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society 16, 985-986.

Easley, J. (1978). Symbol manipulation reexamined: An approach to bridging a chasm. In B. Presseisen, D. Goldstein, & M. Appel (Eds.), Topics in cognitive development 2, 99-112. New York, : Plenum.

Easley, J. A., Jr. (1974). The structural paradigm in protocol analysis. Journal of Research in Science Teaching, 2, 281-290.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Fisher, K. (1990). Semantic networking: The new kid on the block. Journal of Research in Science Teaching, 27(10), 1001-1018.

Giere, R. (1988). Explaining science. Chicago: University of Chicago Press.

Ginsburg, H. (1983). Protocol methods in research on mathematical thinking. In H. Ginsburg (Ed.), The development of mathematical thinking. New York: Academic Press.

Glaser, B., & Strauss, A. (1967). The discovery of grounded theory. Chicago: Aldine.

Goetz, J., & LeCompte, M. (1984). Ethnography and qualitative design in educational research. New York: Academic Press.

Gruber, H. (1974). Darwin on man. New York: E. P. Dutton.

Hanson, N. R. (1958). Patterns of discovery. Cambridge, UK: Cambridge University Press.

Harre, R. (1961). Theories and things. London, UK: Newman History and Philosophy of Science Series.

Harre, R. (1967). Philosophy of science: History of. In P. Edwards, (Ed.), The encyclopedia of philosophy, (pp. 289-296). New York Free Press.

Hayes, J.R. (1978). Cognitive psychology: Thinking and creating. Homewood, Ill: Dorsley.

Hayes, J. R., & Flower, L. S. (1978). Protocol analysis of writing processes. Paper presented at the annual meeting of the American Educational Research Association, March.

Helfgott, D., Helfgott, M. and Hoof, B. (1994). Inspiration®, Inspiration Software, Inc.

Hesse, M. (1966). Models and analogies in science. South Bend, IN: Notre Dame University Press.

Hesse, M. (1967). Models and analogies in science. In P. Edwards (Ed.), The encyclopedia of philosophy (pp. 354-359). New York Free Press.

Howe, K. (1985). Two dogmas of educational research. Educational Researcher, 14, 10-18

Howe, K. & Eisennhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. Educational Researcher, 19(4), 2-9.

Keeves, J. (1988, 1997) Educational research, methodology, and measurement: an international handbook. Oxford, England; New York: Pergamon Press.

Kuhn, T. (1962). The structure of scientific revolutions, (1st ed.) Chicago: University of Chicago Press.

Kuhn, T. (1977). Concepts of cause in the development of physics. In T. Kuhn, The essential tension: Selected studies in scientific tradition and change (pp. 21-30). Chicago: University of Chicago Press.

Lakatos, I. (1978). The methodology of scientific research programmes. Philosophical papers Vol. 1. Cambridge: Cambridge University Press.

Lincoln, Y., & Guba, E. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage.

Nagel, E. (1961). The structure of science. New York: Harcourt, Brace, and World.

Nersessian, N. (1984). Faraday to Einstein: Constructing meaning in scientific explanation. Dordrecht, Netherlands: Martinus Nijhoff.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In Cognitive models of science, R. Giere, (Ed.). Minneapolis, MN: U. of Minn. Press.

Newell, A., & Simon, H. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Patton, M. (1980). Qualitative evaluation methods. Beverly Hills, CA: Sage.

Peirce, C. S. (1958). Collected papers (8 Vols.) C. Hartshorne, P. Weiss, & A. Burks. (Eds.). Cambridge, MA: Harvard University Press.

Piaget, J. (1975). The child's conception of the world (Chapter 1). Totowa, NJ: Littlefield, Adams.

Smith, M. L. (1987). Publishing qualitative research. American Educational Research Journal, *24*, (2), 173-183.

Thagard, P. (1992). Conceptual revolutions. Princeton, NJ: Princeton University Press.

Tweney, R. (1985). Faraday's discovery of induction: A cognitive approach. In D. Gooding, & F. James (Eds.), Faraday rediscovered: Essays on the life and work of Michael Faraday, 1791-1867 (pp. 189-209). New York: Stockton Press.

von Glasersfeld, E., & Steffe, L. (1991). Conceptual models in educational research and practice. The Journal of Educational Thought, 25(2), 91-103.

Table 1. Four Levels of Knowledge Used in the Physical and Cognitive Sciences

	Physical Science: Study of Gases	**Cognitive Science: Study of Disequilibrium**	**Cognitive Science: Diagnostic Algebra Research**
THEORIES			
4. Formal Principles and Theoretical Commitments	$P = 1/3 \frac{Nmv^2}{L^2}$ (Refers to theory of molecules)	Psychological need for local coherence	Tendency to symbolize static rather than operational relationships
3. Researcher's Explanatory Models	Colliding elastic particle model	Mental disequilibrium between conceptions and perceptions	Static versus operative conceptions of equations
==========	==========	==========	==========
OBSERVATIONS			
2. Observed Behavior Patterns and Empirical Laws	$Pv = kt$ (refers to observations of measuring apparatus)	Subjects make predictions and express concern or surprise at opposite result	Reversal pattern in equations; references to relative sizes of quantities
1. Primary-Level Data	Measurement of a single pressure change in a heated gas	Individual subject expresses concern or surprise	Individual reversals

Table 2.. An Example of the Reversal Error in an Algebra Word Problem

Test Question (n = 150)	Correct Answer	% Correct	Typical Incorrect Answer
Write an equation using the variables S and P to represent the following statement: "There are six times as many students as professors at this university". Use S for the number of students and P for the number of professors.	S = 6P	63%	6S = P

Table 3. Viability of a Model: Criteria for Evaluating Theories

1. Plausibility
   a. Explanatory adequacy
   b. Internal coherence

2. Empirical Support
   a. Triangulation and number of supporting observations
   b. Strength of connection to each observation
   c. Lack of anomalies

3. Nonempirical Criteria
   a. Clarity
   b. Simplicity
   c. Lack Of "ad hocness"
   d. External coherence

4. External Viability
   a. Generalizability
   b. Predictiveness
   c. Extendability
   d. Fruitfulness

Table 4. Spectrum of Clinical Interview Approaches from Generative to Convergent

More **Generative** and Interpretive Studies use approaches A and B.

A. Exploratory Studies: Relatively large sections of transcript are explained by a global interpretation that may contain several elements. The analyst formulates an initial description of the subject's mental structures, goals, and processes that provides an explanation for the behavior exhibited in the transcript. This involves the construction of new descriptive concepts and relationships on a case-by-case basis. Examples of transcript sections are usually exhibited in reports along side the analysts' interpretations. In exploratory studies, sensitivity to subtle observations is important; e.g., investigators may make use of facial expressions, gestures, and voice inflections. Although it may be impossible at this stage to code some of these observations reliably with multiple independent coders, analysts who become sensitive to them may generate key insightful hypotheses that would otherwise be difficult to attain. This generation technique does not prevent one from evaluating and increasing the support for these hypotheses later by other, more reliable means.

B. Grounded Model Construction Studies: Analysts generate descriptions as in A above. In addition, some initial observation (O) concepts are identified that describe patterns of behavior. Investigators analyze smaller segments of transcripts and begin to separate theoretical (T) concepts (partial models or process characteristics) from observations. They also begin to connect theoretical models to specific observations that support them, triangulating where possible (as in Figure 3). Interview procedures are standardized that are needed to provide a stable context for those observations that will be compared across different subjects and episodes.

C. Explicit Analysis Studies: Investigators: criticize and refine observation concepts and theoretical concepts (model elements) on the basis of more detailed analyses of cases; articulate more explicit definitions of observation concepts (definitions of observations should approach independent codeability); code for certain observations over a complete section of transcript according to a fixed definition or criterion; If the study has a theoretical component they will point to sets of observations in a transcript and explain them by means of a model; articulate more explicit descriptions of theoretical models; and describe explicit triangulated lines of support from observations to theoretical models.

D. Independent Coder Studies: Analysts refine concepts as in C above. In addition, coding of observation patterns (O's) is done by independent coders; inter-rater reliabilities are calculated. Note that it is much easier to define rules

or guidelines for coding observable O's than for theoretical unobservable T's in Figure 3. Coding that is restricted in this way still can provide a strong source of support for a constructed model T when coded O's are judged by readers to provide evidence for T. In advanced fields explicit criteria may also be established for the subsequent inferring of T's from the presence of certain O's after O's have been coded.

More **Convergent**, Confirmatory Studies use approaches C and D.

Table 5. Characteristics of Generative vs Convergent Studies

Generative Studies	Convergent Studies
Generate new observation categories and new model elements that can explain relatively unexplored phenomena	Use fixed observation categories and model elements and document where these are present in records of relatively familiar phenomena
Interpretive analysis concentrates on model viability and relevance of focus for observations	Coded analysis concentrates on the reliability of observations and, in advanced studies, the tightness of their connections to models
Some attention to nonformal aspects of reliability	Some attention to viability through explicit connections between observations and models
Major strength is the external transfer of a new model	Major strength is the external replicability of observations
Theoretical generalizability of models across investigators, and across some populations and contexts	Empirical generalizability of observations across samples and investigators, but generalizations are limited to a narrowly defined population and context
Need procedural replicability for conclusions that compare subjects	Need procedural replicability for observational replicability as well as conclusions that compare subjects

Figures:

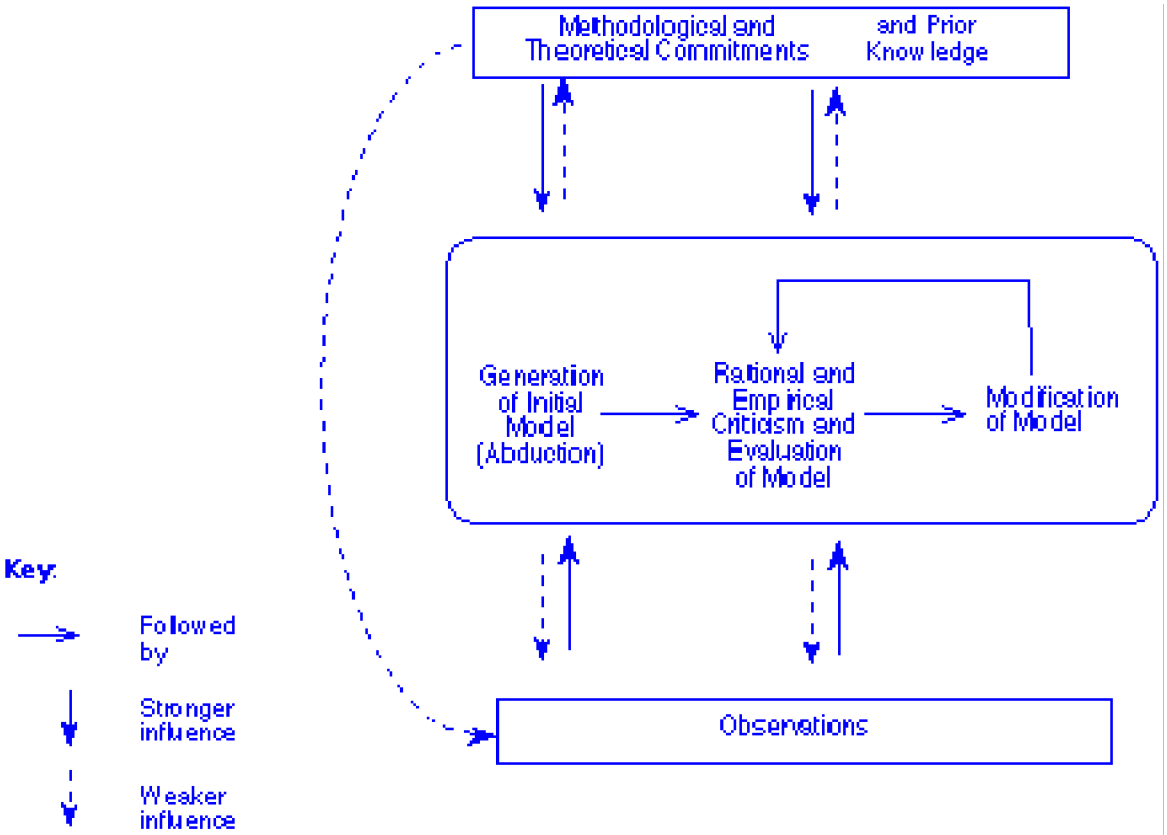

Figure 1

Basic Model Construction Cycle
With Top-Down and Bottom-Up Influences

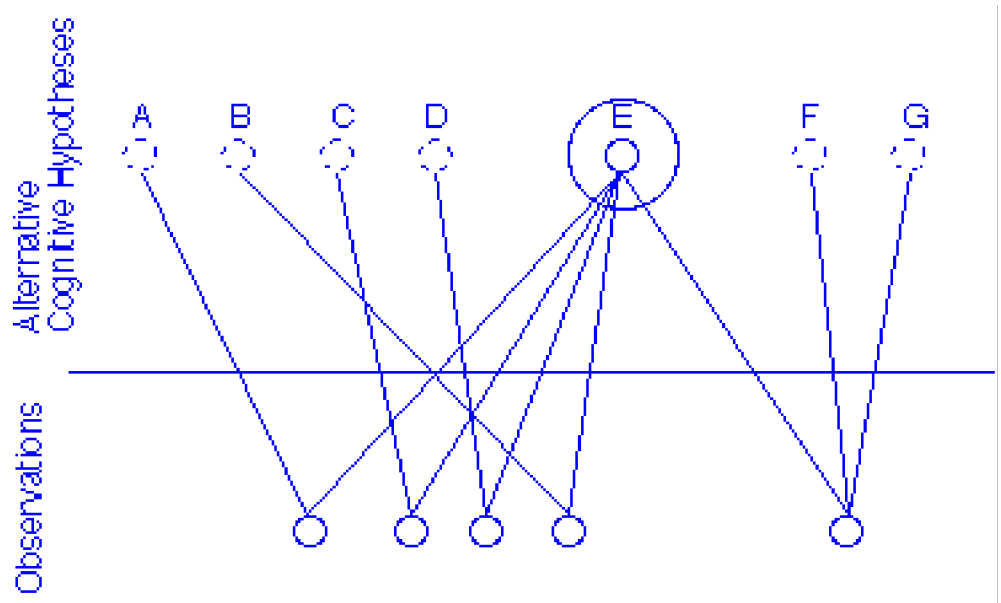

Figure 2

Triangulation: Hypothesis E
With Multiple Sources of Support

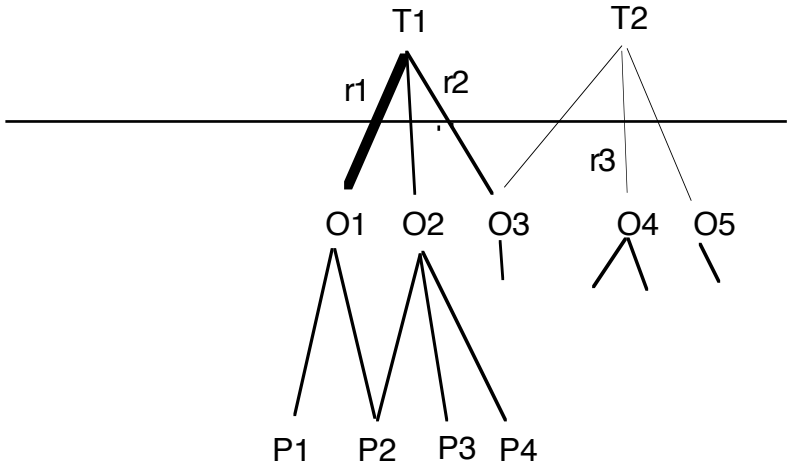

Figure 3. Levels of Analysis in a Clinical Interview

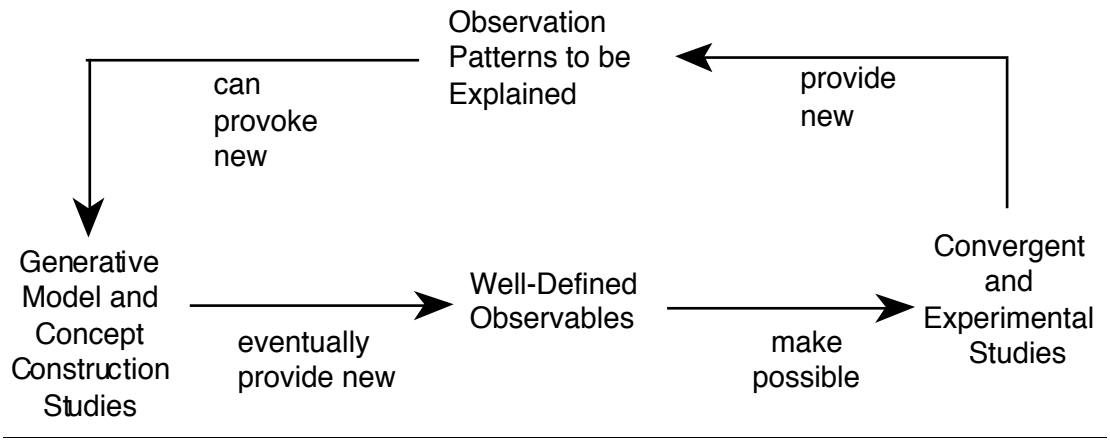

Figure 4. How Work at a Convergent Level Can Initiate Work at a Generative Level

ANALYSIS OF CLINICAL INTERVIEWS: FOUNDATIONS & MODEL VIABILITY

John Clement

School of Education and
Scientific Reasoning Research Institute
University of Massachusetts
Amherst, MA 01003

January 2, 2019

*The research reported in this study was supported by the National Science foundation under Grant RED-9453084. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.