THE JOURNAL OF THE LEARNING SCIENCES, 6(3), 271–315

Quantifying Qualitative Analyses of Verbal Data: A Practical Guide

Michelene T. H. Chi
Learning Research and Development Center
University of Pittsburgh

This article provides one example of a method of analyzing qualitative data in an objective and quantifiable way. Although the application of the method is illustrated in the context of verbal data such as explanations, interviews, problem-solving protocols, and retrospective reports, in principle, the mechanics of the method can be adapted for coding other types of qualitative data, such as gestures and videotapes. The mechanics of the method are outlined in 8 concrete steps. Although verbal analyses can be used for many purposes, the main goal of the analyses discussed here is to formulate an understanding of the representation of the knowledge used in cognitive performances and how that representation changes with learning. This can be contrasted with another method of analyzing verbal protocols, the goal of which is to validate the cognitive processes of human performance, often as embodied in a computational model.

For a variety of reasons, there has been an increasing need in cognitive science and educational research to collect and analyze "messy" data. Messy data refer to such things as verbal explanations, observations, and videotapings, as well as gestures. One reason for the need to collect this kind of data is the trend toward studying complex activities in practice or in the context in which they occur. So, for example, to understand how an apprentice learns a trade, one might have to observe the learner in context. Likewise, it is becoming increasingly clear that the performance of experts (such as industrial and software designers) relies on the use of external aids and tools, such as notes and drawings (Norman, 1988). Thus, to capture a complete understanding of their skill, ideally one should incorporate in the analysis not only their verbal transcripts but also their drawings, pointings, and gesturings

Requests for reprints should be sent to Michelene T. H. Chi, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260. E-mail: chi@vms.cis.pitt.edu

(Tang, 1989). Of course, both verbal data and observational data have been used widely for some time, in cognitive simulation research for the former case and in anthropological studies for the latter. However, it has been discouraging for novice students of cognitive science and education to adopt these methods for various reasons, such as the restricted applicability of the protocol analysis method (see Ericsson & Simon, 1984), the subjectiveness of the observational methods (see Schofield & Anderson, 1987), and the time-consumingness of both of these methods. The goal of this article is to attempt to provide guidance for how one can approach an analysis of verbal data more generally, involving a method that integrates elements of qualitative and quantitative analyses so that the interpretation of the results is less subjective. Unfortunately, the time-consuming nature of qualitative-based quantitative analysis will remain, even with an explicit guide. (This article does not, however, address the analysis of videotape data because that topic is already covered by Jordan & Henderson, 1995, and it does not cover the analysis of gestures, which is discussed in Goldin-Meadow, Alibali, & Breckinridge Church, 1993).

There are basically two ways to write such a guide. One way is to survey the literature, identify all those studies that have used some kind of qualitative analysis of verbal data, then describe, analyze, and synthesize all the various methods. Such a survey would have to include not only studies that use some kind of qualitative analysis of verbal data (such as those of Patel & Groen, 1986; Ranney, 1994; Trabasso & Suh, 1993; Voss, Tyler, & Yengo, 1983) but also to include research that focuses explicitly on comparing different methods of collecting and analyzing data, such as the work of Geiwitz, Klatsky, and McCloskey (1988) and Hoffman (1987). This approach is not taken here for two reasons. First, an analysis, synthesis, and comparison type of exhaustive review of the literature focusing on different aspects of verbal analyses has already been carried out (Ericsson & Simon, 1984, 1993; Olson & Biolsi, 1991). Second, the goal of this article is to provide a practical guide; combining a guide along with an analysis and synthesis of the literature would be beyond the scope of this article.

An alternative way to write such a guide, taken in this article, is to describe and deconstruct the methods that I have developed and used in my own research over the past decade or so. Reflecting on these case studies can reveal the techniques and assumptions that underlie the research method, which will henceforth be called *verbal analysis.* Thus, no attempt is made to relate aspects of the verbal analysis method (in terms of degree of overlap or whether the verbal analysis method complements some other methods) to those others have used, with the exception of one very obvious case. The exception is the long-standing work on protocol analysis, initiated by the information-processing approach (Ericsson & Simon, 1984, 1993; Newell & Simon, 1972). Contrasting the protocol analysis method with the verbal analysis method is necessary because the two methods, up to a point, share a number of surface similarities in the mechanics of the coding. There are

numerous other equally well-thought-out methods in the literature using verbal analysis, such as the proposition–discourse analysis of Patel and Groen (1986), conversation analysis of Fox (1991), inferential flow analysis of Salter (1983) and others. Thus, the method described here is but one approach to analyzing verbal data and no claim is being made about its value in comparison to other related methods.

Why is a practical guide on my personal method of any use? First, as lucidly pointed out by Schofield and Anderson (1987), both qualitative and quantitative analyses have shortcomings and strengths, thus some kind of method that can integrate elements of both methods seems desirable, especially for answering complex questions such as learning in context. Thus, laying out and sharing one such method seems useful, even if it is not the best one. Second, because analyses of verbal data are often complex, the techniques for doing them are usually opaque. Thus, an explicit guide seems useful at least as a starting point to overcome that opaqueness, even if this guide illustrates a special case.

This article has basically three sections. The first section raises three introductory issues: my theoretical bias, comparison of this method with the protocol analysis method, and different approaches to integrating quantitative and qualitative analyses. The second section describes the mechanics of the method, which has been decomposed into eight steps (henceforth, the specific steps will be referred to as the *technique*, and the general method is referred to as the *verbal analysis method*). The third section provides additional recommendations and technical details and addresses some remaining questions and caveats. The article closes with a concluding summary.

INTRODUCTION TO VERBAL ANALYSIS

Verbal analysis is a methodology for quantifying the subjective or qualitative coding of the *contents* of verbal utterances. In verbal analysis, one tabulates, counts, and draws relations between the occurrences of different kinds of utterances to reduce the subjectiveness of qualitative coding. Verbal analysis has been used, for example, to code explanations of what one understands as one reads a text sentence-by-sentence, to see whether an explanation is an inference, a monitoring statement, or some other irrelevant comment (Chi, de Leeuw, Chiu, & LaVancher, 1994). Such quantification of qualitative coding is not the same as direct counting methods whereby a researcher picks out aspects of the qualitative data that can be quantified directly, such as counting the occurrence of a given word in a newspaper article (Weber, 1985). Verbal analysis is also to be differentiated from methods whereby a researcher undertakes qualitative observations in a messy context, but then analyzes only the quantitative data from that messy situation. For example, suppose one observes how the introduction of new technological equipment effects the operation in a trauma unit, but then ultimately analyzes the mortality rate or the

cost of treatment per patient as a consequence of having that new technology. In both of these cases, either no subjective qualitative coding was entailed or no qualitative data were actually used. Although such methods sometimes do allow analysis of the content (e.g., one has to read the content of the newspaper article to decide which word should be coded), verbal analysis, in addition, allows for coding that does not require a direct correspondence between the content word(s) uttered and the coding category.

The verbal analysis method is embedded in research that tries to understand cognition, and in particular, the kind of knowledge one gains from learning. However, the mechanics of the method may be adapted to the study of various noncognitive issues (such as social, motivational, and behavioral) and may also be used with observational and video data, rather than strictly with verbal data. However, every approach, as used by a specific investigator, has a theoretical bias that is built into the method. Let me describe that bias first.

Theoretical Bias

The research embodying the verbal analysis method focuses on learning. More specifically, its goal (perhaps not yet achieved) is to capture the representation of knowledge that a learner has and how that representation changes with acquisition. Secondary questions might include contrasting the knowledge of an expert or a more advanced learner with that of a novice. Knowledge representation per se is an active and extensive subfield of artificial intelligence. In that subfield, people have developed sophisticated computational techniques for representing knowledge, using various means such as rule-based systems, semantic and Bayesian networks, and graphical models. The contribution of the method here, however, is not in the representational techniques nor does the method employ sophisticated techniques. The key difference between what knowledge representation researchers do and the method described here is that they are generally concerned with representing "ideal" knowledge that can be interpreted by a computer problem solver or reasoner to draw correct inferences. That is, suppose a researcher figures out (or is given) an optimum solution procedure for an algebra problem. The steps of such a solution procedure can then be represented by some kind of formalism, such as production rules. Thus, their contribution is in their analyses and representation of the ideal knowledge in a way that can be interpreted by a computer problem solver or reasoner.

Instead of representing the ideal knowledge, the goal of the method here is to attempt to figure out what a learner knows (on the basis of what a learner says, does, or manifests in some way, such as pointing or gesturing) and how that knowledge influences the way the learner reasons and solves problems, whether correctly or incorrectly. Thus, the trick is to analyze the learner's utterances (in the case of verbal

data) to capture the knowledge that might underlie those utterances and do so in a way that is not subjective; therefore, it needs to be quantifiable in some ways.

Of course, one can identify and capture what a learner knows from a variety of more traditional dependent measures such as response times and errors. For example, the classic study of Collins and Quillian (1969) used response times to infer the nodes and hierarchical nature of semantic memory. Likewise, errors in classification has been used by Chi, Feltovich, and Glaser (1981) to signify the kind of entities that a student attended to in representing a physics problem. Although response times and errors can uncover the representation of knowledge, analyzing verbal data can provide a much richer, more detailed, and perhaps more accurate representation, so that one can ultimately use such a representation to devise instruction to revise what the student has misconceived or add to a student's missing knowledge.

To uncover what a learner knows requires an analysis of the content of the verbal utterances (i.e., what the student said), along with a procedure to organize the content in some way (i.e., relate what is said) so that one can assess its overall structure. That is, one can determine "what" the student said (the content) by listing it as a set of propositions, a set of concepts, a set of goals, or a set of rules. However, to explore the overall structure, one must then assess the relations among such a set.

One example to illustrate the issue of structure has to do with the debate in the science education literature about the nature of naive, but robust and false, beliefs that students hold about a number of science concepts. The majority of these findings simply illustrate the kind of misconceptions students hold. For example, students may think that heat is a kind of substance that flows from a hot room to a cold room; they may further believe that heat can escape more readily from a poorly insulated room than a well-insulated room. These beliefs reflect the content of what students think. To address the issue of structure, one needs to further say whether such piecemeal beliefs are truly fragmented and unrelated pieces of knowledge (diSessa, 1993) or whether they are theory-like in that they can be captured by a few principles (McCloskey, 1983) or whether they have some other kind of intermediate level structure, such as that the beliefs are based on properties of an ontological class (Chi, in press-a). Thus, the term *structure* simply refers to the relations embedded in the content knowledge.

Contrast to Protocol Analysis

Perhaps the most frequent and systematic use of verbal data is in the context of *protocol analysis,* as espoused by Newell and Simon (1972). Therefore, it is important to point out the underlying assumptions that differentiate the protocol analysis method and the verbal analysis method, especially given that the two methods share some surface similarities in the techniques. There are five key

differences: the instruction, the goal or focus, the analysis, the validation, and the conclusion.

The first difference focuses on the way the verbal data is collected. In the protocol analysis method as described in Ericsson and Simon (1984, 1993), to obtain *think-aloud* protocols, they instructed the subjects to verbalize the information they attended to while solving a problem. They emphasized very much the notion that the utterances refer to the heeded information, almost as if the talking is analogous to a physical pointer that points to the external elements of a problem, if these elements could be displayed. For example, they said that

> a subject given the task of mentally multiplying 24 by 36 while thinking aloud might verbalize: 36 times 24, 4 times 6, 24, 4, carry the 2, 12, 14, 144, and so on. It is important to note that subjects verbalizing their thoughts while performing a task do *not* describe or explain what they are doing—they simply verbalize the information they attend to while generating the answer. (Ericsson & Simon, 1993, p. xiii)

Of course, talking is much richer than merely pointing to the numbers 4 and 6, because the verbalization also reports the intermediate products (e.g., 24, 12, 14, 144) that have no external referents, as well as revealing the goals and subgoals. But the spirit of the instruction is analogous to pointing or can be likened to tracking eye movements, using eye fixations to indicate what information is being heeded at any moment in time. The instruction discourages any description or explanation of what subjects are doing. Thus, the kind of verbalizations that Ericsson and Simon (1993) refer to as explanations, descriptions, justifications, and rationalizations are exactly the kind of verbalizations (henceforth referred to as *explaining*) that this article addresses. Ericsson and Simon further differentiate the two types, the think-aloud type and the explaining type, not only by the instruction but also by the outcome. Think-aloud instruction should not effect the performance of the primary task, whereas explaining does, in that it improves the subjects' performance, such as learning (Chi, de Leeuw, et al., 1994).

The second difference between think-aloud and explaining has to do with the focus of the research. In protocol analysis, the focus is to capture the *processes* of solving a problem or making a decision (i.e., doing some task). The processes of problem solving correspond to the *sequence* of problem states that the subject undertakes as she or he applies permissible operators. One or more of the possible sequences of problem states can be enumerated in advance. So, for instance, in a simple task of the Tower of Hanoi, if the initial state consists of having all of the three disks (the large, medium, and small) on the first peg, then a tree of all possible states can be enumerated (such a tree is called the problem space). Each state in a tree is enumerated by applying an operator, such as moving the small disk to another peg. For example, the first level of the tree can be obtained by moving the small disk to either the second or third peg. Therefore, two possible states can be enumerated from the initial state. From these two possible enumerated states, a

number of subsequent states can be derived by applying various move operators, such as moving the medium disk from the second peg to a yet unoccupied peg.

Notice two important features of this focus. First, a detailed analysis of the Tower of Hanoi task is required so that a problem space is derived. This is usually referred to as the cognitive task analysis. This cognitive task analysis is carried out at a fairly fine-grained level so that it enables the construction of a runnable computational model. Second, the goal of protocol analysis then is to identify which sequence of states a particular subject progresses through (i.e., the solution path). Or alternatively, a specific computational model is constructed that assumes that a solver moves through the possible states in a given path. Then, the goal of the protocol analysis is to see whether there is a match between the path that a solver took and the sequence of states that a simulation model generates. If not, then one can tweak the model so that it simulates an actual solver's path.

In contrast, the focus of the verbal analysis method is to capture the representation of the knowledge that a solver has and less on the processes of problem solving. One could, however, say that the simulation model is the representation that underlies the subject's performance. If so, then in what sense is the focus of the protocol analysis method different from the focus of the verbal analysis method? Primarily, the difference is that the protocol analysis method starts with a model of the task, which can be referred to as the ideal template. The goal of the method here, in contrast, is to seek the model that a subject has, without creating an ideal template a priori. Thus, the goal of protocol analysis is mainly to test a model, rather than to uncover what the subject is actually doing.

Protocol and verbal analysis do share many of the mechanical details (such as segmentation and coding) and concerns (such as how much context to use for interpretation). What is different about the analyses of the two methods is the emphasis and the workload. In protocol analysis, coming up with an ideal template, which requires cognitive task analysis, coupled with (but not necessarily) translating it into a computational model, is the majority of the workload. As such, the actual analysis of the protocols is simplified in a number of ways. One simplification is that because the elements and operators in the model are already predefined, this means that the coding of the protocols often involves determining when the elements and operators are used and how they are referenced, without the need to identify what the elements and operators are. This means that the first step of protocol analysis, according to Ericsson and Simon (1984) is to

> extract the vocabulary of objects and relations needed to define the problem space and operators. For example, in the Tower of Hanoi task, a subject might refer in his protocols to "disks" and "pegs," and to "moving Disk X from Peg Y to Peg Z," essentially using the language of the problem instructions. ... The subject may identify the disks by numbers (Disk 1 for the smallest, say), and the pegs by letters (A,B, C, from left to right). (p. 264)

Thus, in protocol analysis, because the elements and operators are defined a priori, the analysis consists of identifying what vocabulary in the protocols is used to refer to these elements and operators. In verbal analysis, in contrast, the referents are unknown, so that there is the added complexity of first determining what the referents are. In the self-explanation data (Chi, de Leeuw, et al., 1994), for example, we had to determine what is it that a subject is verbalizing about (such as an inference, a metastatement, a plan, an inquiry), in addition to worrying about how and when they are referenced.

Fourth, the method of validation is also different for the two methods. In the protocol analysis method, the validation of the protocol analysis is the "degree of match" between the sequence of protocol utterances and the sequence of states generated by the model (although what constitutes a "good enough" match is often unclear). Moreover, because one can always tweak the model to "match" the protocol, it's never clear how the model can ever be invalidated. In the verbal analysis method, however, validation is obtained by either applying statistical tests of the quantified qualitative codings to see if the results support a hypothesis (such as the data showing a significant difference between high self-explainers and the amount learned; Chi, de Leeuw, et al., 1994), or validation is determined by some qualitative analysis of the structure and its correspondence to some other measure. For example, if the structure of one's verbal utterances fits a particular mental model that was derived from the verbal data, then one can predict that one's answers to certain questions would be consistent with that mental model (Chi, in press-b). In other words, the validation consists of comparing internal measures and not comparing it to an external model.

Finally, the conclusions one draws from protocol and verbal analyses are also different. This is primarily due to the different theoretical biases of the two approaches. Protocol analysis is coupled with a particular approach to problem solving and decision making, and the approach is modelled by the sequence of states. Moreover, the sequence of states presumably depicts the use of a particular strategy. Thus, in protocol analysis, one commonly tries to identify the strategy a subject used to solve a problem on the basis of the processes of problem solving, such as whether the subject used an efficient strategy (e.g., subgoaling) or used a means–ends strategy. Thus, one can infer the strategy that the solver used by the sequence of states the subject progressed through (or the path taken). The reason one would focus on the strategy is because that is an element of the model that presumably generalizes to other tasks and domains. In verbal analysis, the theoretical bias is on the knowledge representation, hence, no conclusion is drawn about the strategy of problem solving. Instead, the conclusion reached by the verbal analysis method is often in terms of the representation the solver has and claims that it is the solver's representation that determines the problem-solving processes. An example to illustrate this difference in the conclusion reached is to contrast the findings of Simon and Simon (1978) and Chi et al. (1981). Both sets of researchers

tried to capture the way experts and novices solve physics problems. In the Simon and Simon study, by analyzing the sequence of equations in the protocols, they could conclude that the expert solver was using a forward search strategy (working forward from the problem statement) whereas the novice solver was using a backward (means–ends) search strategy (working backward from the goal). However, to explain why that difference in strategy was obtained, Chi et al. asked subjects to explain the reasons for their categorization of physics problems. From their explanations, one could conclude that the experts and the novices were representing the problems differently, leading perhaps to the different solution strategies. Thus, the different theoretical biases of the two methods convey different information and conclusions about a research question.

The differing conclusions one reaches in the two methods occur even at a more micro level. In the protocol analysis method, the sequencing of the utterances are of utmost importance because it conveys information about the processes of solving problems. Sequencing is of much less relevance in verbal analysis because all utterances are taken to reflect the underlying representation, to some degree, irrespective of when exactly they were uttered. Sequencing becomes important only when one is interested in analyzing an evolving representation, as in the case of learning. Even then, the interpretation of the sequenced utterances remains distinctly different in the two methods.

One caveat that must be noted is that the mechanics of verbal analysis, although not intended necessarily for think-aloud type of protocols, can nevertheless be used to analyze think-aloud protocol data. Examples are illustrated throughout this article.

Integration of Quantitative and Qualitative Methods

Before discussing different ways of integrating quantitative and qualitative methods and what type of integration the verbal analysis method proposed here entails, a few words about what is meant here by quantitative and qualitative methods are needed. Qualitative methods generally refer to research that is conducted in natural settings such as classrooms, a community or neighborhood, a specific culture. It sometimes relies on the researcher as the main observer in both the data gathering (such as field notes taken from observations or interviews) and the analysis, thus making both the data collection and the analyses vulnerable to subjective interpretation. The data can be field notes taken from observations, explanations or conversations, or interviews. It is possible to remove some of the dependency on the selectivity of the observer(s) by video- or audiotaping. For instance, videotaped conversations among the attending physician, the interns, and the resident, when they go on rounds are being analyzed (Chi, Hashem, Ludvigsen, Shalin, & Bertram, 1997). Of course, some selectivity still remains in terms of which rounds are chosen to tape, on which days, and with which patients. Because qualitative research is

generally undertaken in naturalistic settings, it is often impossible to control for numerous variabilities that context provides. Quantitative methods, on the other hand, refer to experimental design that carefully controls and manipulates the variables under study. The variables manipulated reflect the specific hypothesis being tested. The data gathered are usually of a quantitative nature (such as latency, errors, frequency) that can be subjected to precise statistical tests.

There are clearly many advantages and shortcomings of both qualitative and quantitative methods. The main advantage of qualitative research is that it can provide a richer and deeper understanding of a situation. Moreover, as was alluded to earlier, many skills are executed in a very different way in context than in a sterile laboratory environment. However, qualitative methods usually suffer from subjective interpretation and nonreplicability. Quantitative methods, on the other hand, have the advantage of objectivity and replicability, but the shortcoming is that one can only make conclusions about the specific hypothesis at hand. Furthermore, the sterile laboratory environment of experimental studies limits the generalization of the results to a real-world context. Clearly, there is a need to blend the two methods in such a way as to remove each method's shortcomings. The verbal analysis method attempts to satisfy these goals by removing subjectivity and yet maintaining the richness of context.

There are several ways to integrate quantitative and qualitative methods, and the verbal analysis method discussed here is but one. As mentioned earlier, blending qualitative and quantitative research does not mean analyzing the easily quantifiable aspects of qualitative data, such as counting the frequency of occurrence of a given word in the newspaper (Weber, 1985) or the mortality rate of a trauma unit that has introduced new technological equipment. In these cases, the qualitative data are not being analyzed. The four methods to be discussed later in this article all try to analyze the qualitative data to some degree, and the verbal analysis method is the most integrated one.

The most conservative way to integrate quantitative and qualitative analysis is to use the qualitative data to help interpret the quantitative results. This "interpretation approach" is one that I used early on in my own research. For example, one can combine categorization results with explanations of the categorization (Chi et al., 1981). That is, experts and novices were asked to categorize problems into different types, as they wished to define them. Their categorization pattern can be treated as similarity judgment data and quantitatively analyzed using a variety of factor analyses. However, to interpret what each category means to the experts and novices, their explanations were examined. In this case, one uses the qualitative data (the explanations) as an aid in the interpretation and understanding of the quantitative data (the factor analyses), but no claim is made about the qualitative data per se. Hence, the predominant emphasis is still on the quantitative data.

A second and most straightforward way to integrate the two methods is to use some kind of quantitative measures along with the qualitative measures. This

"complement approach" has been used widely, such as collecting scores of the problem-solving success along with the verbalizations of problem solving, or collecting IQ or achievement test scores along with the verbalizations. In the former case, the quantitative data collected can serve as confirmation of the qualitative analyses and vice versa. For example, one should find that the high scorers are the students who generated a greater number of inferences whereas the low scorers generated few inferences (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). Such problem-solving success scores thus confirm the analysis of the inferences in the explanation data because one would predict that successful solvers must have learned the materials better, and to learn the materials better, one needs to generate more inferences. In the latter case of collecting achievement test scores, these types of quantitative measures can sometimes be used as an independent variable. For example, one might be able to compare the inference-generation capability of high ability students with low ability students (see Chi & VanLehn, 1991). Note in this second approach, both the quantitative and qualitative data are treated more or less with equal weights.

A third way of integrating qualitative with quantitative methods is to use the qualitative analysis as a backdrop for generating hypotheses, which are then tested by experimental methods. For example, Schofield (1982) observed in a qualitative study of peer relations that students seemed to react differently to peers' ambiguously aggressive acts, depending on the race of the peer who generated the act. This conjecture was then tested experimentally by showing Black and White children sketches of aggressive behaviors by other Black and White children, and more conclusive interpretation can then be derived from such an experimental manipulation. Because verbal analyses is so time-consuming, it is not always possible for an investigator to follow up the first hypothesis-testing stage of analysis with the second experimental stage. Fortunately, this two-step approach can be decoupled in the sense that researchers in other laboratories can undertake the second step. For example, the self-explanation work (Chi, Bassok, et al., 1989) that basically analyzed only 8 students' explanations has now been followed up by careful experimental and quantitative analyses with a larger sample of 36 subjects (Renkl, 1997). Notice that in this two-step approach, the emphasis is beginning to be placed more on the qualitative analyses. The quantitative analyses now play a confirmatory role.

Last, the method to be introduced in this article is one whereby the researchers rely strictly on the qualitative data, but they quantify the analyses. That is, the qualitative data is examined for impressions and trends, methods of coding are developed to capture those impressions, and the codings can then be analyzed quantitatively. Suppose a researcher's impression is that students who self-explained more frequently (i.e., generate inferences to themselves more often) while reading a passage also learned more than students who self-explained less frequently. How do you quantify that impression? To do so requires an analysis of the

verbalizations of students while they are reading, such as identifying what an inference is, coding how many inferences were generated by each student, and comparing the number of inferences generated by the good learners versus the poor learners (Chi, Bassok, et al., 1989). Thus, this quantitative-based qualitative approach basically operationalizes one's subjective impression by coding the verbal evidence for that impression and comparing the frequencies of the codes quantitatively.

SPECIFIC TECHNIQUE FOR VERBAL ANALYSES

The specific technique for analyzing verbal data consists of eight steps, excluding the initial collection and transcribing of the verbal protocols. These eight steps are detailed later, but first we will review the overall process.

The process of verbal analysis assumes that some type of verbal protocols were collected and transcribed. What protocols are to be collected of course depends on the theoretical questions and hypotheses of interest. For example, if one was interested in collaboration, then obviously one would tape conversations among dyads or within a group. More specifically, if a hypothesis is that learning requires the generation of inferences, then one would tape the explanations generated while a student reads a text, or an example within a text, if the specific hypothesis is how one learns from examples.

Once verbal data are collected and transcribed, the first step of the analysis is for the researcher to decide whether to analyze the entire corpus of the protocols or some samples. Once that choice is determined, then the analyzer needs to segment the protocols so that each segment can be coded independently. The reason for segmenting is that one has to determine what constitutes a unit of analysis, analogous to a single trial in an experimental design. (Sometimes one can bypass the segmenting if one is sure of what evidence to search for in the protocols.) Once segmented, each segment can then be coded. This requires that codes be developed according to the hypothesis one wishes to test. To implement the coding, one then needs to decide what utterances in the protocols constitute evidence for a specific code. For explanation data, for example, what would the subject have to say to qualify an explanation as an inference? Once the data are coded, they can be summarized either in a tabular or graphical form. Patterns in the data can then be detected and interpreted. The entire procedure can be repeated if one wishes to test another hypothesis that would require a different type of coding or a more detailed coding of a smaller sample of the protocols.

Although the preceding summary of the technique implies a sequential ordering of the steps in quantifying qualitative analysis of protocol data, that is really not the case. In undertaking an actual analysis, a researcher obviously must look ahead of the sequence of steps to assess whether the result of the decision at a current step

is meaningful or appropriate. For instance, a researcher might have decided on a set of codes (Step 3, see the following list) but then must look ahead to see if these codes can be operationalized in the context of the utterances (Step 4) as well as how much coverage of the protocols the codes can account for. If not, then one has to develop new codes, modify them, or refine them. One could say this forward-and-backward testing process is comparable to piloting the analyses. Now to the specifics.

The method of coding and analyzing verbal data consists of the following eight functional steps:

1. Reducing or sampling the protocols.
2. Segmenting the reduced or sampled protocols (sometimes optional).
3. Developing or choosing a coding scheme or formalism.
4. Operationalizing evidence in the coded protocols that constitutes a mapping to some chosen formalism.
5. Depicting the mapped formalism (optional).
6. Seeking pattern(s) in the mapped formalism.
7. Interpreting the pattern(s).
8. Repeating the whole process, perhaps coding at a different grain size (optional).

Reducing the Protocols

Once verbal data are collected and transcribed, they are ready to be coded. Typically, verbal data tend to be voluminous: 1 hr of tape may take 6 to 10 hr to transcribe, which can result in 15 to 50 pages of text. Many researchers find this amount of data overwhelming and choose to code only a sample of it. There are three general heuristics for data reduction: (a) random sampling, (b) choosing a subset on the basis of some "noncontent" criterion, and (c) doing some preliminary coding on the content of the entire set and then more detailed coding on a selected subset. Random sampling is perhaps the most obvious and simple method of reducing the data and needs no explanation. The second method, selecting a systematic sample according to some noncontent criteria, means that features such as equations, pauses, changes in speaker, changes in activity, a given amount of time or speech, are used. (Note that noncontent criteria sometimes do require having a cursory idea of the content, as in the case of identifying the activity.) These noncontent features can often be easily detected in the protocols without reading the content of the protocols. For example, in examining the physics problem-solving protocols, Simon and Simon (1978) analyzed only the equations that were used by the solvers. Hence, selecting the equation parts of the protocols was a relatively easy sampling criterion. Another example of sampling the data is to examine only

the first few minutes of verbalization, as in the case when one wishes to examine only the planning phase of problem solving (Chi et al., 1981, Experiment 4). A third way to reduce the data is to select a particular activity, such as when students are rereading or referencing the examples in the text (although in this case, some content is required; Chi, Bassok, et al., 1989). Notice that only noncontent criteria are recommended here; otherwise, using the content of the utterances to select a sample would not have saved much time or effort in reducing the protocols. Using content to segment amounts to segmenting the entire set of protocols. One caution to note: Selected samples of this kind may give an incomplete or flawed view of the knowledge. In general, it is preferable to preserve a continuous string of protocols, when possible.

Segmenting the Protocols

Once the corpus (set or subset) of protocols to be coded is decided, then one needs to segment the verbal utterances to identify the unit of analysis. There are four issues to consider in segmenting: (a) the grain size of the segment, (b) the correspondence of the grain size to the questions one is asking, (c) the characteristics in the data used for segmenting, and (d) when it may not be necessary to segment.

Granularity. For verbal data, the defining cut can occur at many points, revealing units of varying grain sizes, such as a proposition, a sentence, an idea, a reasoning chain, a paragraph, an interchange as in conversational dialogue, or an episode (such as an event, or a specific activity). The following analyses, taken from Chi, de Leeuw, et al. (1994), illustrate how data can be coded in two different grain sizes. In that study, students were asked to self-explain what each line of a text passage on the circulatory system means after they have read it. The point of the study was to show that the more inferencing a student engages in while learning, the better the student understands. Consider the explanation (in quotes) that a student gave after reading the sentence (text sentences are in bold):

> **During strenuous exercise, tissues need more oxygen.**
> "During exercise, the tissues, um, are used more, and since they are used more, they need more oxygen and nutrients. And um the blood, blood's transporting it to them."

Because this entire response constituted an explanation from the student's point of view, the researchers initially took the entire explanation as a segment and read its content to see whether there is any inference in it. An inference was defined as any newly asserted information that was not stated in the passage sentence. The one macro inference (italicized) captured from this utterance was:

Blood transports more oxygen and nutrients to the tissues during exercise.

Subsequently, they resegmented the utterances into proposition-sized units, as shown by the slashed lines:

"During exercise, the tissues, um, are used more, //
and since they are used more, they need more oxygen and nutrients. //
And um the blood, blood's transporting it to them."//

From the segmented units, each proposition-sized unit was then systematically considered to see if it constituted an inference. Using this smaller unit size, three micro inferences were captured from this utterance:

Inference 1: *Tissues are used more during exercise.*
Inference 2: *When tissues are used more, they need more oxygen and nutrients.*
Inference 3: *Blood transports oxygen to the tissues.*
(We assumed that the "it" refers to oxygen.)

In tallying the results, it would not be legitimate to equate the number of micro with the number of macro inferences within the same coding scheme. Thus, a decision has to be made a priori about the grain size of the segments.

Because there was no theoretical reason to expect a difference in the results at the two different grain size of codings, both codings did produce the exact same pattern of results, as reported in Chi, de Leeuw, et al. (1994). Note also that codings at two different grain sizes is another way to achieve reliability. The coarser grained coding was used for the remaining analyses, primarily because it made more sense in the following two ways (aside from the practical reason that it is also less time-consuming). First, the coarser grain size seemed to capture the semantics of the inference at a more appropriate level. For example, each of the three micro inferences missed the idea that is embedded in the macro inference, that *blood* transports more oxygen to the tissue *during exercise.* Thus, each of the three micro inferences lacked the important relation captured by the macro inference between blood doing more work and exercising. Second, the finer-grained coding was inadequate in a number of other ways: (a) they sometimes seemed redundant, (b) at other times they seemed to capture inferences at the comprehension level (such as a bridging inference of making the referent explicit) rather than at the knowledge level, (c) sometimes they also seem to capture paraphrases more than an inference (e.g., Inference 2 of the finer coding could be considered one whereby **strenuous exercise** is paraphrased as "tissues ... are used more"), and finally (d) sometimes the micro inferences seemed more like a retrieved piece of knowledge rather than an inference, as in Inference 1. For all of these reasons, the larger grain size was

chosen. Thus, the appropriate choice of a grain size can be a subtle and complex decision.

Coding at a coarser grain size requires, of course, less work. But unfortunately there is a tradeoff sometimes between the grain size and the amount of information one will derive from the data. For example, taking data from problem-solving protocols, one could analyze an entire solution protocol merely for its correctness, whether the solution used a particular strategy (such as analogy or means–end) or whether the solution relied on a worked-out example. An approach utilizing a finer grain size would be to analyze the solution protocol in terms of episodes, providing more sensitive data. For example, instead of analyzing the entire solution protocol as one unit, indicating either correct–incorrect solution or the use of certain strategies, the solution protocols can be divided into episodes, and each episode can be analyzed for the presence of certain type of errors.

The fact that coarser grain sizes tend to be less informative applies to other forms of data as well, such as latency data. For example, the total amount of time it takes to read a sentence is not as sensitive a measure as eye movement data, in which one measures how long a subject is fixating on each word of a sentence. For experimental studies in which latency data are collected, the more densely (or frequently) the measures are taken, the more sensitive they are. Thus, in experimental studies, the sensitivity of the data is determined a priori by the dependent measures that are taken. In verbal analyses, on the other hand, the sensitivity of the data depends on the grain size of the analysis, which can be chosen post hoc, making verbal analyses much more flexible. The number of grain sizes one works with is limited only by the tediousness and length of time required to segment and code the verbal data.

Correspondence. Although the example just presented shows that analyses at two different grain sizes produced the same pattern of results, this is not always the case. Usually, one needs to worry about whether the chosen grain size is appropriate for the questions asked to interpret the results meaningfully. That is, there should be a correspondence between the grain size of analysis and the research question one is asking. For example, if you are asking about how people reason, it makes little sense to look at the number of propositions generated in their arguments (unless your hypothesis is that the more distinct pieces of knowledge you can bring into an argument, the more likely your argument will be persuasive) or at the number of words generated in their arguments (unless your hypothesis is that the most talkative person generally wins the argument). A more appropriate unit might be the reasoning chain, which usually involves a grain size of several sentences. In Chi, Hutchinson, and Robin (1989), for instance, the objective of one of the research questions was to look at the way expert and novice children made attributions about novel dinosaurs—dinosaurs that were unfamiliar and had never been seen by the

children before. The task was simply to ask children to explain what they might know about a novel dinosaur when a picture of it was shown. The verbal data consisted of what they said about each dinosaur. The following is an example of two possible analyses from the same data: The first looked at what attributions children made from their knowledge (i.e., what features they may attribute to a novel dinosaur), and the second analysis looked at how such an attribution was made. For instance, when shown a picture of a novel dinosaur, an expert child (ages varied from 4–7) may say something like:

> He's probably a good swimmer. [Why?] 'Cause duckbills are good swimmers.
> [Why is he a duckbill?] 'Cause it has this [bill].

Whereas a novice child of the same age may say something like:

> He could walk real fast. [Why?] 'Cause he has giant legs.

To answer the question of what attributions the children made, one could simply do a gross proposition analysis to show that the attributions an expert child made are "good swimmer" and "has a bill." From such an attribute analysis, one could say something about the differences in the content of children's knowledge bases, and how such differences could help expert and novice children make different attributions. On the other hand, if the research question is how did a child arrive at these attributions, then a larger unit of analysis (spanning several sentences) is needed to understand what kind of reasoning the child used. In this case, the expert child was characterized as using "hierarchical" reasoning to arrive at the attributes. From her first sentence "He's probably a good swimmer," one could say that she identified a feature (such as the presence of a bill from the third phrase, "'Cause it has this [bill]"), and inferred from the bill that the novel dinosaur must belong to the duckbill family. On the basis of that classification into a family type, she extrapolated an additional feature, swimming ("He's probably a good swimmer"), that the novel dinosaur must also possess (the second sentence of "'Cause duckbills are good swimmers"). Thus, all three phrases are needed to deduce how the child's reasoning might have proceeded. This example merely illustrates the point that the appropriate grain size to choose must correspond to the questions being asked.

Features used for segmenting. Notice that the choice of a grain size has implications for how easily the verbal utterances can be segmented. The boundaries of the units can be identified by either noncontent features or by semantic features, in much of the same way that subsets of protocol data can be selected. Noncontent features, as discussed earlier, can be characterized by either of the following: (a) language-related syntax, such as words, sentences, or sentences with connecting

words such as *because* or *therefore,* or the use of equations; or (b) activity features, such as pauses of a certain duration, turn-taking, a change of activity such as from solving problems to seeking information from the text, or from reading to drawing diagrams, or from reading to writing equations, and so forth.

The obvious advantage of using a noncontent feature for segmentation is that one does not have to read the verbal data carefully to perform segmentation. Thus noncontent segmenting is usually more straightforward and less time-consuming. Also, it is conceivable that this sort of segmentation could be done using a computer-aided system.

Segmentation can also be based on semantic features, such as ideas, argument chains, topics of discussion, or impasses while solving problems. In this approach, the content of the utterances must be read for meaning to determine segment boundaries, such as knowing when an impasse has been reached by the subject or when a topic of discussion has been changed. Although it may be considerably easier to use syntactic boundaries to segment the protocols, it is often psychologically more meaningful to use semantic boundaries. The following excerpt exemplifies the coding of multiple sentences as one episode because the subject was really just trying to resolve, while studying a worked-out physics example, how angle theta was calculated (taken from the Chi, Bassok, et al., 1989, protocol):

Subject:	How they calculated that this angle is theta?
	If this is theta. And this is 90 degrees.
Experimenter:	So you are looking at Figure A. Okay.
Subject:	I'm trying to figure out how they came that this one will be also theta.
	How do I find it? Perhaps ... what did they do here?
	I believe that any time they give the theta that is what it will be, but I want to figure out why.
	So here it would be 90 degrees. (pause)
	Now they move everything. ... They move the axis by theta.
	So if this is a 90 degrees, this one should be theta. Okay. Okay.
Experimenter:	Okay, what?
Subject:	I think I got it.

Thus, this entire episode contains a single activity or idea.

There are a couple of related reasons why a segmentation of this protocol at the episode level was a more appropriate unit of analysis than a sentence. First, an idea might need several sentences to convey, so that coding at the sentence level would overestimate the number of substantive ideas discussed. Second, the same idea could be repeated several times by talkative people, so that counting sentences as the unit of analysis would credit talkative people with more output when in fact they were generating the same idea.

Searching rather than segmenting. It is also conceivable that sometimes no segmentation of the verbal protocols is needed. Instead, one can simply search the protocols for occurrences of the desired activity. In the Chi, Bassok, et al. (1989) self-explanation work, because students generated self-explanations spontaneously, the authors were able to merely search for their occurrences without segmenting the protocols entirely. Once an inference was identified, they then simply ascertained that the same inference was not credited more than once when it was embedded in the same episode. On the other hand, in the Chi, de Leeuw, et al. (1994) self-explanation work, because the students were required to articulate something after reading every sentence, it was necessary that each utterance be segmented and coded for the presence of an explanation inference. Thus, whether or not one can bypass segmenting depends on the specific situation.

Developing or Choosing a Coding Scheme or a Formalism

Once the verbal data are segmented, they are ready to be coded. To do so, codes must be developed to correspond to a formalism which will be used to represent the knowledge. This is probably the most difficult step of the technique to convey because what codes and formalisms are chosen depend entirely on a researcher's theoretical orientation, the hypotheses or questions being asked, the task, and the content domain. Let's start with a simple example of codes that fit a taxonomic categorical scheme. In the physics domain, a set of categories was developed for coding students' explanations generated while studying worked-out examples presented in a text. These categories concerned whether the explanations and elaborations used by the students pertained to physics concepts, principles, systems, or technical knowledge (Chi & VanLehn, 1991). We coded as *concepts* physics entities such as mass, weight, acceleration; as *principles* statements that related entities specified in Newton's laws, such as statements relating mass to acceleration; as *systems* comments about the interaction of two or more objects, such as a block on an inclined plane; and as *technical knowledge* algebraic manipulations such as vectorial decomposition of forces, velocities, and various other concepts. This choice of coding scheme was developed because we wanted to test the hypothesis that students could learn to solve problems correctly without much understanding of the principles or concepts. Besides testing this hypothesis, the analysis also explained why students seem to learn better from and prefer to study worked-out examples, a long-standing dilemma in the literature. We found that the reason students prefer and seem to learn a great deal from worked-out examples in text is simply because examples introduce many concepts, technical procedures, and systems knowledge that are relevant for problem solving but are omitted in the expository part of the text. Thus, these categories were chosen and used to isolate the learning of concepts, principles, technical procedures, and systems knowledge.

The analysis also generated a hypothesis about how understanding of principles can occur.

Because taxonomic categories are simplistic, they do not reveal the formalism that sometimes underlies the development of codes. The codes developed should correspond to elements in a particular formalism. For example, suppose the coded data are ultimately to be represented as a semantic network (Step 5 of the technique); then the codes chosen (Step 3) should correspond to elements in the network, such as nodes and relations (note again the forward and backward nature of the eight steps). For a simplistic example, in the dinosaur study (Chi & Koeske, 1983), a 4-year-old child's knowledge of dinosaurs was elicited. First, the child was asked to retrieve as many dinosaur names as he could; then the child and the experimenter played a game in which the child had to guess the identity of a dinosaur from a description of two of its features by the experimenter, and, in turn, the child also had to generate a couple of features of a dinosaur for the experimenter to identify. For example, the child was asked to identify the dinosaur that the experimenter was thinking of if the dinosaur had the features of sharp teeth and a long tail. Conversely, the child was asked to generate dinosaur features for the experimenter to identify. From such data, adjacently retrieved dinosaur names were coded as being directly linked. For example, if in the first retrieval trial, the child generated a set of dinosaur names in quick succession followed by a pause, then generates another set of names, then links were specified as existing among each set of dinosaurs that were bounded by a longish pause. The underlying assumption was that highly related dinosaur concepts are activated more simultaneously, thus producing a clustering effect in retrieval. Dinosaurs that the child could identify from the named attributes, or dinosaurs' features that the child could generate, were represented as features of each dinosaur that the child knew about and represented them as directly linked to the dinosaur node. In this example, the elements in the protocols (dinosaurs, features) were easily mapped onto the elements in the network (nodes, links).

If one is coding think-aloud protocols in which the search path needs to be identified, then the coding has to capture the sequence of operators that are being used. Thus, to code the think-aloud protocols of writing computer programs, one might code each segment into operators such as "read," "paraphrase," "compare," "evaluate," and so on (C. Fisher, 1987), so that one can characterize the search states or solution path as a consequence of the application of each operator.

There is no clear-cut algorithm for deciding how to choose the appropriate formalism. The general rules of thumb are that procedural tasks tend to be more adaptable to production systems, problem spaces (Newell & Simon, 1972), or flow-chart type of formalisms (Siegler, 1976), whereas tasks that tap declarative and conceptual knowledge would more appropriately be represented by semantic and conceptual networks. Systems knowledge might be more easily represented by mental models (Johnson-Laird, 1983), whereas arguments might be more easily represented by argument chains (Voss et al., 1983). Stories and events can be

represented either as a causal chain (Trabasso & van den Broek, 1985) or a tree of goals and subgoals (Means & Voss, 1985).

In verbal analysis, ideally, a choice of formalism can be developed first in a top-down manner, on the basis of the hypotheses one is testing, in exactly the same way as occurs in designing an experiment. In contrast to traditional experimental design, however, this initial choice can be fine-tuned on the basis of the verbal data, much in the same way one would modify the design of an experiment on the basis of pilot data. For example, in the aforementioned example of coding the writing of computer programs protocols, suppose the protocols indicate that a solver also spends a lot of time mentally simulating the program—a new operator such as mental simulation might be added to the list of codes. Hence, the development of a formalism is an interactive top-down and bottom-up process. It is interactive in the sense that one should be open to additional modifications as one becomes familiar with the verbal data.

Operationalizing Evidence for Coding

Once a formalism is selected, the next step in verbal analysis is to decide what utterances in the verbal data constitute evidence that they belong to a specific category or can be translated into a specific code. In this section, a number of examples are presented to illustrate which units of articulation correspond to the codes that were selected to represent them. Following the examples, two problematic issues that can arise—how to resolve the ambiguity of interpretation and how much context to consider in the interpretation—are presented.

An earlier example described the four categories (concepts, system, principles, technical procedures) the researchers used to capture the knowledge students learn as they study examples in a physics textbook (Chi & VanLehn, 1991). Once these categories are decided on, to implement the coding, one needs to decide what aspect of the content of the explanations constitute evidence for one of these four categories. In this case, an explanation such as "I didn't know the knot can be the body" would be considered an explanation about the concept of "the body" or the concept of "a point mass," whereas comments about decomposing a force into its vectorial components would be classified as a piece of technical knowledge.

Often, much of the ingenuity in verbal analysis stems from finding ways to instantiate the question one is asking by defining what constitutes good evidence. Oftentimes, a researcher's contribution is the evidence that she or he can ferret out of the verbal data. An excellent example is Hutchins and Levin's (1981) analysis of the use of deictic words such as *here* or *there* to reveal the perspective taken by the subject in solving the missionary and cannibals problem ("perspective" in terms of which side of the river the solver was seeing the boat from). In this case, a syntactic analysis—the identification of the use of deictic words—led to an interpretation of the subject's mental model. Similarly, Gobbo and Chi (1986) looked

for evidence of expert children's use of connecting words such as *so* and *because* in their verbal descriptions. The task was simply asking children to describe a dinosaur that they had never seen before, as a picture of it was shown to the child. It was found that expert children were more likely to use connecting words in their description of the novel dinosaurs, whereas novice children were much more likely to list a series of properties that are visibly shown in the picture. The following two quotes illustrate this difference.

Expert child: And he had webbed feet, *so* that he could swim, and his nose was shaped like a duck's bill, *which* gave him his name.

Novice child: He has sharp teeth. He has three fingers. He has sharp fingers, sharp toes, a big tail.

The implication about the structure of the representation was that the expert children had a richly interconnected knowledge base about dinosaurs, so that activating one explicit feature (by describing it in the picture, such as the webbed feet and duck's bill) primed the activation of associated implicit features that was not shown in the picture (such as the dinosaur's ability to swim or its probable name). These associated features were then retrieved in a connected manner by connecting words. Novices, however, had very sparse and disconnected knowledge, so that they could only describe what they explicitly saw in the presented picture of a novel dinosaur. These explicit features were then described in an abbreviated disconnected manner because they did not activate any related implicit associations (Chi, Hutchinson, et al., 1989). Thus, this analysis focused on the use of connecting words as evidence of interconnectedness of a child's knowledge base.

There are occasions when an interpretation of the representation (such as its overall structure) cannot be mapped directly to a particular segment of verbalization. That is, the preceding cases were all instances of the type that a specific segment of data could be pointed to as corresponding to some element of the representation: Specific utterances were mapped directly to a node in the semantic network or to problem-solving operators such as compare; specific words in the verbalization such as *so* or *because* were used to indicate connectedness. However, a way to hypothesize about the representation when the articulations were incomplete or to reserve the interpretation of the representation until the codings could have a cumulative effect, was needed. For example, to capture students' initial mental models of the human circulatory system, each student's initial mental model was determined from numerous *local* statements each student made. Local here means that the student only discussed local anatomical connections, such as the path of blood flow from the atrium to the ventricle. Thus, each student's utterances can be mapped directly to local anatomical connections. A great deal of variation existed among students' models, particularly in the functioning, behavior, and structure of

specific components, as well as in the relations between components. However, six different general types of models were discerned, as shown in Figure 1, ranging from the least accurate No Loop model, to the most accurate Double Loop 2 model (50% of the students started out with the Single Loop model). The summative model that the student was credited with, such as a Single Loop model, was, of course, not necessarily mentioned explicitly by the student. That is, the student may not necessarily have had explicit awareness that his or her model was a Single Loop one as opposed to the alternative Double Loop model; nor was the student necessarily able to describe the Single Loop model. The student was credited with the Single Loop model based on the accumulation of all the utterances about local connections; that is, the analysis revealed the structure of the student's knowledge.

Ambiguity. Up to this point, coding in general has been described as if it is a fairly straightforward procedure, once the codes and formalism are developed. For example, in the physics case in which explanations had to be coded into the four categories of concepts, principles, systems, and technical procedures, this particular coding scheme was straightforward for this content domain, because the interpretation of which category the explanations belong to tends to be unambiguous. Similarly, in C. Fisher's (1987) protocols of programming, it was fairly straightforward to translate the verbal articulations into operators such as *paraphrase, compare, evaluate,* and so forth.

Some types of coding schemes, however, are much more difficult because of the ambiguity in interpreting what was meant. In Chi, de Leeuw, et al. (1994), ambiguity arose in deciding between a minute inference and a bridging inference and between a minute inference and a paraphrase. In the first case, the task was to differentiate between a small, sometimes fragmented, piece of inference versus a substituted referent. That is, a bridging inference (often discussed in the comprehension literature) occurs when an implied referent or agent is specifically articulated and it is not considered a piece of knowledge inference, such as when the student substituted the word *arteries* for the term *large vessels,* as shown:

> This is trying to show that there are, um, valves like the ones in the heart that separate the ventricles from the *arteries.*

When in fact the sentence only stated:

> **Two semi-lunar valves separate the ventricles from the *large vessels* through which blood flows out of the heart.**

So, should the replacement of *arteries* for the words *large vessels* be considered a bridging inference or a minute inference that large vessels are arteries?

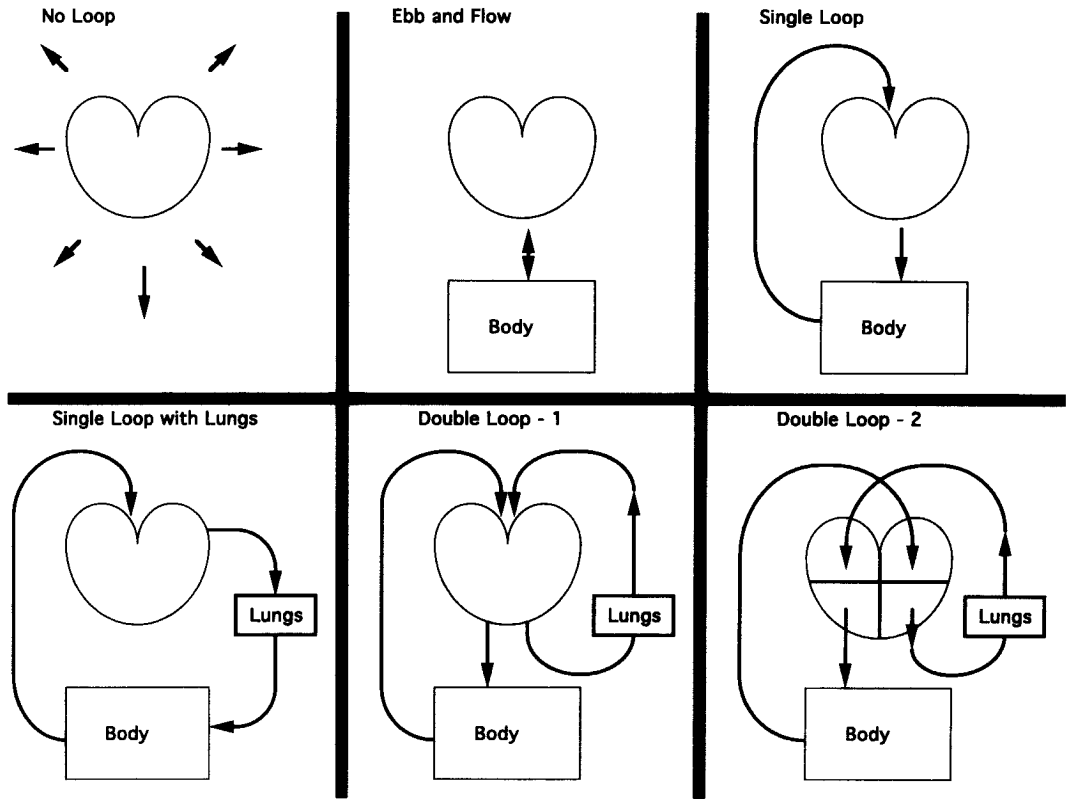

FIGURE 1 Six initial mental models that students hold.

This next example illustrates the second case: ambiguity in discriminating between a minute inference and a paraphrase. After reading the sentence:

The heart is a muscular organ that pumps blood through the body.

The student explains:

> The heart is what causes, is where the power for the blood comes.

Here, should the interpretation by the student that a muscular organ is a powerful source of blood be considered a minute inference or a paraphrase?

Context. A second technical issue has to do with the extent to which context should be considered in coding. Context refers to how many lines of the protocols surrounding (before and after) the current segment the coder should take into consideration when interpreting the meaning of the current segment. That is, a different interpretation sometimes arises if a broader context (several lines of transcribed protocols) is used than if the decision for coding a specific segment of protocol is kept at a very local (one segment) level. One has three obvious choices. The first is to use as much context as necessary to maximize the coder's comprehension of the verbal utterances. The following is an example of resolving understanding of protocols generated over numerous sentences apart. To understand this example, the reader should know that the student's initial mental model of the circulatory system does not involve lungs (Chi, de Leeuw, et al., 1994). Instead, the student's mental model of human circulation is simply that blood leaves the heart, goes to the body, and returns to the heart (the Single Loop model of Figure 1). Sentence 18 of the text that the student read was the first time that the student encountered the information that lungs were a component of the circulatory system:

The right side pumps blood to the lungs, and the left side pumps blood to other parts of the body.

Prior to this sentence, the text was talking about the heart and its chambers. After reading this sentence, the subject explained:

> Just that um, the right side is primarily for the lungs and the left side is to the rest of the body.

One way to interpret this explanation is that the student simply treated lungs as a part of the body, so that she viewed the lungs merely as another destination to which blood had to travel. No special status of lungs as a source of oxygenation was provided here. This conservative interpretation was only substantiated much later when lungs were introduced for a second time, in Sentence 32, which said:

The muscle in the right ventricle contract and force blood through the semilunar valve and into vessels leading to the lungs.

After reading this sentence, the student explained:

(pause) ... Um, the sentence is a little confusing.
Um, blood is just flowing from the ventricle into vessels and going,
um, to the lungs, um, okay.
I guess then I was remembering the thing about the left side is
for rest of the body, and the right side is mostly for the lungs,
but well I guess I don't know. ... I was just thinking that you know,
it was just *providing the lungs with blood* which didn't make much sense,
but that possibly, I don't know, but maybe *it's going there to receive oxygen*
or something. I don't know.

This explanation confirms the preceding interpretation of Sentence 18 that she thought back then about blood going to the lungs for the purpose only of providing the lungs with blood (see the first italicized part) and only now realizes that it's going there to receive oxygen (see the second italicized part). Thus, the use of extensive context helped to resolve the appropriateness of the interpretation when it was ambiguous. The choice of using extensive context might be most appropriate for analyzing single-subject's protocol because one cannot afford to tolerate errors in interpretation.

A second approach is to use minimal context and keep the interpretation at a fairly local level (that is, at the level of the segment being coded). This might be most appropriate for coding data of multiple subjects because multiple-subjects data leave more room for "noise." That is, the resulting trend from the data of multiple subjects can overcome the occasional errors of interpretation from ambiguity. The third way to resolve the issue of context is to code the data twice, once using a strictly local context and once using a broader context. But, of course, this requires more work. In any case, for any given coding, I prefer to allow for the use of context beyond the current segment at minimum for resolving syntactic ambiguity (such as understanding the referents). In general, one should always be consistent in terms of how broad a context to consider throughout a specific coding.

Depicting the Mapped Formalism

Once the data is coded, then the results should be depicted for two reasons. First, it is a way of presenting the data to the audience, just as one would depict quantitative data graphically or in tabular forms. A second reason, again analogous to quantitative analysis, is to see if some patterns can be detected in the depicted data. This second purpose is discussed in this section.

There are multiple ways that coded data can be depicted, depending on the formalism chosen. If one's choice is a taxonomy of categories, then a simple table presenting the means for each category might be adequate. However, for other formalisms, different heuristics could be used to guide the depiction of the data. If one's representation is a sequence of problem states (or problem behavior graph), then there are specific rules for how to depict it in such a way as to maintain the sequence of operators throughout the problem space (Newell & Simon, 1972, p. 173). In depicting coded verbal data into any formalism such as a semantic network, one needs to make assumptions about what codes correspond to a link and what codes correspond to a concept node. In Chi and Koeske (1983), each dinosaur name was designated to be a node in the network, and the relation between successive dinosaur names (without a significant pause) corresponded to a link between two dinosaur nodes. Each property or feature that a child knew about a dinosaur was considered to be a property node that was represented as being associated with the dinosaur node. There are also commercial programs that can lay out such networks, such as SemNet™ (K. Fisher, 1989, 1990).

There do not seem to be any standard techniques for illustrating a representation graphically. Different systems seem to work for different tasks and domains. In depicting an argument chain, for example, Voss et al. (1983) attached a backing node to the warrant node that it was backing. If a given warrant has two backings, then they attached both backing nodes to the same warrant node. In the case of depicting the dinosaur semantic network (Chi & Koeske, 1983), the assumption of no redundant nodes was made, so that all references to the same concept node were shown as linking to that node and not to a duplicate of it. These constraints can sometimes be guided by assumptions in the formalism.

In some sense, it does not matter what format one uses to depict the representation, such as pictorially, in a tree structure, or other types of formal notations such as equations. Conclusions drawn from the analysis depend on the pattern that one can perceive from the format or from the analyses (which will be discussed in the next section), although some errors in interpretation could be made if the data were incorrectly depicted. Errors in depicting the data can be made unintentionally. For example, in depicting the goals, an investigator might lay out the set of goals articulated early on in the protocols on the top of a page and ones mentioned later in the protocols on the bottom of the page and then mistakenly assume that the later-mentioned ones on the bottom of the page were the lower level goals. Although this example sounds ludicrous, it has actually happened.

Seeking Pattern and Coherence in the Depicted Data

Once the coded data are depicted, then one can begin to seek patterns in the results. Seeking patterns in the depicted data is not unlike seeking patterns in other types of dependent measures, such as reaction time plots, in which one looks for linear trends

or U-shaped functions. If the data are coded into taxonomic categories, then the pattern is easy to see because the data can be easily plotted in bar graphs, and statistical analysis can confirm or disconfirm any reliable differences. However, if the coded data are represented in a node-link structure, then patterns of interlinkages should be sought, followed by developing methods to quantify the observed pattern. For example, Chi and Koeske (1983) depicted a child's knowledge of a more familiar set of 20 dinosaurs in contrast to a less familiar set of 20 dinosaurs on the basis of what features of each dinosaur were known to the child, as well as how frequently dinosaur names were mentioned in consecutive sequence (see Figures 2 and 3).

Once depicted, the representation of the more familiar set of dinosaurs seemed more coherent than the representation of the less familiar set of dinosaurs, in the following way: The familiar dinosaurs were clearly segregated into two clusters (see Figure 2). The clusters emerged as a result of the pattern of interlinkages among the dinosaurs, as well as in the way the dinosaurs shared features, and not by the total number of interlinkages or the total number of nodes present in each of the two representations (because exactly 20 dinosaurs and a couple of features of each dinosaur were depicted in each representation). That is, for the more familiar set of dinosaurs, there were numerous links among certain dinosaurs, which then defined them to be in a cluster, and at the same time dinosaurs from distinct clusters had few links among them. Such a pattern of greater within- and fewer between-cluster links was not at all apparent for the set of less familiar dinosaurs. The less familiar set of dinosaurs were divided into five clusters (see Figure 3), whose linking pattern was more diffuse. Hence, the conclusion was made that the structure for the more familiar dinosaurs was better organized and more coherent than that of the less familiar dinosaurs, even though the total number of links and nodes used to represent the two structures were identical. Thus, once depicted, a pattern of interlinkages seemed to have emerged in that the interlinkages among the more familiar set seemed more coherent than the interlinkages of the less familiar set. This difference in the pattern was confirmed by quantifying the number of linkages among dinosaurs within a family cluster versus the number of linkages between family clusters.

As shown previously, one can confirm any pattern that one sees in the depicted data by quantitative means without a great deal of difficulty. That is, one can quantify any pattern in a network, an argument chain, or a causal chain, even though they are graphical in nature. There are numerous ways in which a graphical representation, like a taxonomic one, can be quantified and tested for statistical reliability. For instance, Voss et al. (1983) were able to calculate the mean depth of an argument chain, consisting of a claim and a warrant backed by another claim and a warrant, backed by a final claim and a warrant. Such a calculation provided an indication of how extensively arguments were developed. One could also adapt van den Broek's (1989) causal network representation of a story for depicting verbal

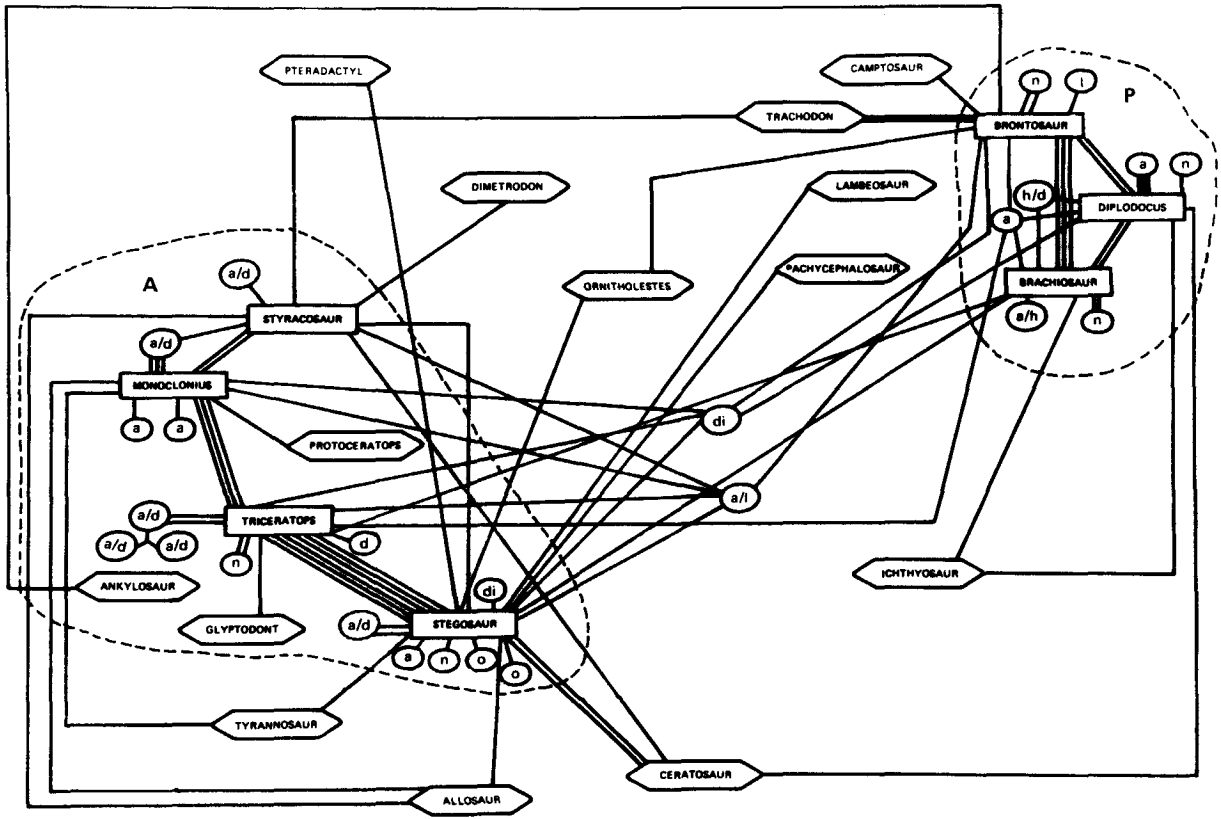

FIGURE 2 A representation of a child's knowledge of 20 more familiar dinosaurs. From "Network Representation of a Child's Dinosaur Knowledge," by M. T. H. Chi and R. D. Koeske, 1983, *Developmental Psychology, 19,* p. 33. Copyright 1983 by the American Psychological Association. Reprinted with permission.

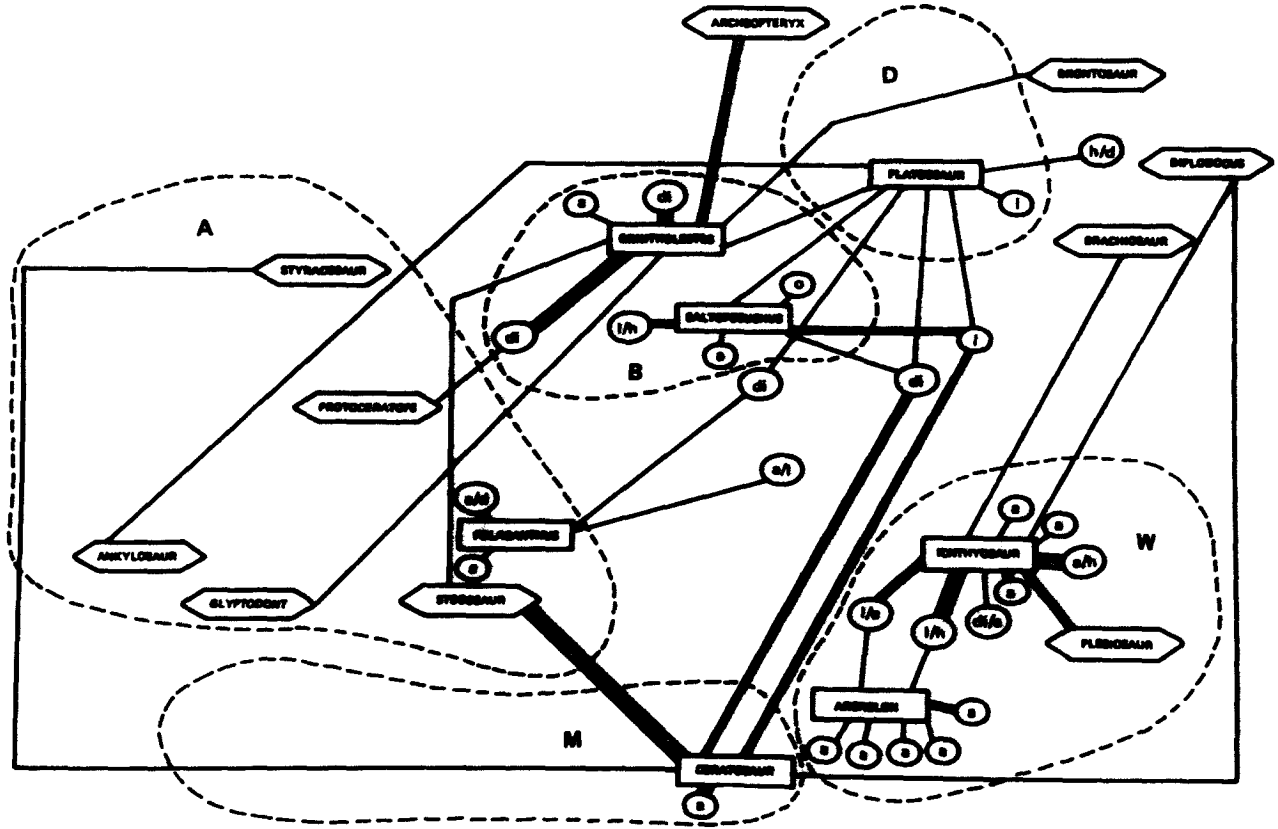

FIGURE 3 A representation of a child's knowledge of 20 less familiar dinosaurs. From "Network Representation of a Child's Dinosaur Knowledge," by M. T. H. Chi and R. D. Koeske, 1983, *Developmental Psychology, 19*, p. 34. Copyright 1983 by the American Psychological Association. Reprinted with permission.

data (de Leeuw, 1993). Once depicted, one can detect patterns of relations such as the number of causal connections or the length of causal chains to the number of inferences, and so forth.

In general, it does not appear difficult to quantify the density of graphical representations of coded data, such as counting the number of nodes attached to the argument chain of one graph compared to another. Thus, one should not have to rely on subjective visual assessment of the structure or graphical representation (such as simply judging that one depicted graph is more hierarchical than another).

Quantifying a perceived pattern in a graphical representation of coded data is not the same thing as quantifying the total number of elements in it. Quantifying the total number of elements merely shows that one representation has more knowledge than another representation, which is usually not a very profound finding, especially in the context of expert–novice research. Instead, what is discussed previously is quantifying the pattern that emerged, which is comparable to capturing the structure of the representation. Two examples will illustrate this type of method. Chi and Koeske (1983) did not quantify the number of dinosaur and feature nodes in the more and less familiar representations; in fact, this number was controlled and made equivalent in the two representations. Instead, the quantification occurred at the level of the pattern. Similarly, Voss et al. (1983) did not merely quantify the number of warrant and backing nodes, but instead quantified the mean depth of the argument chain as an indication of the complexity of the argument. Thus, the quantification occurred at the level of the pattern, serving the purpose of being able to say something about the coherence of the structure, rather than the abundance of the elements in the structure.

It is also possible that coherence and structure in the depicted data can be assessed without a quantitative tallying. Three examples are presented here. If one depicted a causal chain of events for a story, one might capture coherence of the causal chain in terms of whether all the relevant events were part of the causal chain, as did van den Broek (1989), or whether subjects represented the events of a story in a hierarchical or sequential manner. In this case, coherence really refers to the structure of the representation, such as whether all the events of a story are related to the main causal chain of the story. In either case, whether coherence is assessed by some kind of tallying of the pattern of linkages, or whether it is assessed by the nature of the structure it takes (linear versus hierarchical), either mode of assessment seems more informative than the conventional method of relying on a judgment of coherence by two independent raters.

A second way of assessing coherence without quantification can be shown in the analysis that captured students' understanding of the circulatory system (Chi, de Leeuw, et al., 1994). In that analysis, a student's mental model could be coherent whether or not it had many of the correct components and correct connections. That is, a Single Loop model can be more coherent than a Single Loop with Lungs model (see Figure 1) even though the latter model has more components in it (namely, the

addition of lungs). Coherence in this case can be determined by methods such as seeing whether the mental model correctly predicts what answers students would give based on that model or how students handle conflicting information. Thus, a mental model can be coherent in that it is used systematically to generate explanations and answers, even though it may not have as many components and linkages as another more complex but less coherent one.

As a final example, in the physics study wherein experts and novices were asked to categorize physics problems (Chi et al., 1981), there was no reason to expect experts and novices to divide the problems into a differing number of categories (they did in fact divide them into relatively the same number of categories). However, the nature of the categories were significantly different. Thus, the pattern that emerged from the depicted data need not necessarily rely on a quantitative assessment if it can be captured by other means.

Interpreting the Pattern and Its Validity

Interpretation of the perceived pattern in the depicted data, as in the pattern-seeking stage and other stages of the analyses, again depends entirely on the hypotheses being tested, the research questions being asked, and the theoretical orientation of the investigator. One can interpret the data in terms of the strategies and processes, or the structure and content of the knowledge base, or both.

The interpretation of the pattern in the data is often more persuasive if there are other converging evidence or analyses. In a taxonomic representation, the interpretation is fairly persuasive not only because it is fairly straightforward (because the taxonomy was developed to test the hypothesis directly), but more important, one can usually apply statistical analyses to confirm or disconfirm the interpretation. For example, if one's categories are Explanations, Monitoring Statements, and Others, one could readily see whether some subjects generated proportionately more Explanations than Monitoring statements or whether some subjects generated more Explanations than other subjects. Or one could see whether a group of subjects (such as the more-successful solvers) generated proportionately more explanations than another group of subjects (such as the less-successful solvers; Chi, Bassok, et al., 1989).

One way to validate an interpretation is to substantiate it with additional evidence, sort of like the complement approach discussed earlier, in which qualitative analysis is coupled with quantitative measures. Chi (1985) initially used the order of retrieval of classmates' names to indicate how children organized their classmates into seating clusters. The sequential retrieval information was then further substantiated by looking at the duration of interitem retrieval times, such as the rapidity of enumeration of some of the classmates' names, versus those that are separated by a "prolonged" pause (one would have to decide from the data what

constitutes a prolonged pause). The order of retrieval and the clusters as defined by pauses are converging analyses that would constitute a validity check, especially for single-subject analysis, as stated in the preceding paragraph.

Another way to achieve validity is to code the data twice, in something like a two-pass approach. In the mental model analysis of a student's initial understanding of the circulatory system presented earlier, after an initial reading of all the students' protocols, the authors arrived at a preliminary set of six general models, as shown in Figure 1. To validate that each student was correctly credited with the appropriate model, a set of necessary features of each of the six types of models was then specified, as shown in Table 1. The protocols were then reexamined to make sure that evidence for every feature of a model was articulated. For example, in the Single Loop with Lungs model, all five characteristics (listed in Table 1) had to be mentioned at least once by the student to confirm the initial categorization of the student's model. Thus, this second pass through the verbal protocols, looking for the presence of specific features, constituted a validity check for the first more subjective interpretation.

Repeating the Whole Process

Although it seems masochistic, it is often necessary to repeat the entire process over, from Step 1 to Step 7. This need arises often, for example, if one wants to recode the data at a different grain size or if one wants to address a different question. Because a unique asset of verbal data is its flexibility, that is, it can be coded to address a variety of questions, a researcher can err by asking the wrong question initially. The researcher then has the privilege of recoding the data to answer a new question. This can often occur because the data provide a rich source for generating new hypotheses. Controlled experimental studies, on the other hand, do not have this flexibility, because each study is designed to answer an explicit set of questions.

RECOMMENDATIONS, TECHNICAL DETAILS, CAVEATS

There are numerous additional recommendations to provide, pitfalls to avoid, and technical details and caveats to consider. In this catch-all section, these recommendations, pitfalls, and technical details and caveats are discussed, in no particular order.

Technical Aspects of Collecting Verbal Data

The analysis of qualitative data is only as good as the way the data was collected. There are many technical aspects concerning how verbal data should be collected, such as:

TABLE 1
Necessary Features for Each Type of Mental Model

No loop	1. Blood is pumped from the heart to the body. 2. Blood does not return to the heart.
Ebb and flow	1. Blood is primarily contained in blood vessels. 2. Blood is pumped from the heart to the body. 3. Blood returns to the heart by way of the same blood vessel.
Single loop	1. Blood is primarily contained in blood vessels. 2. Blood is pumped from the heart to the body. 3. Blood returns to the heart from the body.
Single loop with lungs	1. Blood is primarily contained in blood vessels. 2. Heart pumps blood to body or to lungs. 3. Blood returns to heart from body or from lungs. 4. Blood flows from lungs to body or from body to lungs without return to heart in between. 5. Lungs play a role in the oxygenation of blood.
Double loop–1	1. Blood is primarily contained in blood vessels. 2. Heart pumps blood to body. 3. Blood returns to heart from body. 4. Heart pumps blood to lungs. 5. Blood returns to heart from lungs. 6. Lungs play a role in the oxygenation of blood.
Double loop–2	1. All features from Double loop–1 2. Heart has four chambers 3. Septum divides heart lengthwise—sense of preventing mixing of blood. 4. Blood flow through heart is top to bottom. 5. At least three of the following: Blood flows from right ventricle to the lungs. Blood flows from lungs to left atrium. Blood flows from left ventricle to body. Blood flows from body to right atrium.

Note. From "Eliciting Self-Explanations Improves Understanding," by M. T. H. Chi, N. de Leeuw, M. H. Chiu, and C. LaVancher, 1994, *Cognitive Science, 18,* p. 468. Copyright 1994. Reprinted with permission.

1. How the experimenter should be as unintrusive or as uniformly intrusive as possible.
2. How subjects should be given practice trials.
3. How the data should be transcribed.
4. Whether the act of verbalizing changes the cognitive processes.
5. How to control for the fact that some people are more verbose than others.

Ericsson and Simon (1993) have addressed several of these issues, including the theoretical concern that giving verbal explanations affects the cognitive processes

in question (Ericsson & Simon, 1980, 1993; Nisbett & Wilson, 1977; Schooler & Engstler-Schooler, 1990). Here, a few more comments on Points 1 and 5 (see previous list) are added.

The amount and kind of interruption posed by the experimenter while collecting verbal protocols will affect the data in a number of ways. If the protocols are collected in an interview situation, then all of the questions addressed to all of the subjects should be the same. If there are to be follow-up questions based on what the subject said, then there should be an objective guideline about what format the follow-up questions can take. This is to avoid leading the subject in a specific direction. There should also be a decision made a priori about when the protocol answers or explanations should be terminated. For example, a typical error that has been committed in the past is for the experimenter to terminate the solution protocols of a problem (in a problem-solving session) when the experimenter realizes that the solution has been reached even though the student has not. Telling the student that he or she should go on to the next problem amounts to giving the student feedback about the correctness of the solution at that point in time. If this is what one student receives, then all the students in the study should be given the same feedback. Other examples include how frequently a student should be prompted for clarification or under what circumstances they should be prompted. If it is necessary to vary the amount of probing between students, one way to handle nonequivalent amounts of resulting verbal utterances is to analyze the protocols up to the point of probing separately from the protocols generated after probing.

The technical aspects of collecting verbal data have raised the issue of how to control the fact that some people are more verbose than others. Although verbosity seems to be a concern for verbal analysis, such individual differences actually exist for other kinds of dependent measures as well, and thus, are not concerns restricted to verbal data per se. For instance, the fact that there are individual differences in verbosity may be similar to the fact that some people are faster overall in processing information in reaction-time studies. In latency data, such individual differences are overcome by examining the pattern of the linear function relating response times to some other iterative factor, such as the number of items one has to search in short-term memory, rather than the intercept of the linear function (called the Subtractive Technique; Sternberg, 1966). The same approach can be applied to verbal data. One way to factor out verbosity might be to gather some baseline verbosity data, such as how much a subject normally talks about another topic or while engaged in another task.

It is also possible that people who know more about a topic actually talk less. However, even if the knowledgeable subjects are more quiet, they still cannot mask what they choose to talk about, even if it is sparse. What they choose to talk about is an indication of what they think is important, even if they don't talk about everything they know. This is why it is particularly important in verbal analyses to focus on the content of the utterances.

Thus, individual differences in verbosity can be factored out by focusing on what the subjects say rather than how much they talk. This means that one would not count the number of words a person has spoken as an index of the amount of elaboration, for example, but use a more appropriate measure such as the number of independent ideas generated. In such a case, verbosity can be factored out by segmenting the protocols into idea episodes, as illustrated earlier.

More serious than verbosity is inarticulateness. That is, the subject may choose not to utter anything at places where it is mandatory for the researcher to know what the subject is thinking about. It is always dangerous to draw any inference (e.g., about what a subject does not know) based on vacancies in the verbal data. To avoid this "completeness" problem, it is important not to allow the subject to remain silent for a prolonged period of time. However, pressing subjects to speak in a neutral way should not be confounded with unintentionally directive intrusions, such as serendipitously letting the subjects know that they can stop working on a problem because they have reached a solution when in fact the subjects did not themselves know it.

Interrater Reliability

Another thorny technical issue is interrater reliability. There are three problems here. The first concerns how carefully one has to define the categories a priori, before each rater codes the data. The second has to do with whether discrepancies between two coders should always be resolved. If so, the third issue has to do with when discrepancies should be resolved.

With respect to the first issue, my preference has always been to start coding with only a very preliminary discussion of what is being sought (without defining specific cues or features for identification). This is because the success with which two coders can come up with highly correlated results using such a gross and preliminary definition is itself an indication of the tangibleness or robustness of what one is seeking. Once such a preliminary coding is done, one can then identify and specify more precisely the features by which one is categorizing, so that the second coding can be done on the basis of these features. The reason that such a two-pass approach is preferred is because some categories may not be well-defined by a small set of well-defined features, as Rosch's research (Rosch, Mervis, Gray, Johnson, & Boyes-Bream, 1976) has informed us. For instance, a recent coding attempt was to read the explanations students gave about science problems to see whether their responses contained materialistic conceptions (Slotta, Chi, & Joram, 1995). After the preliminary codings were done in deciding which explanations were substance-based and which were not (were process-based), Slotta et al. then proceeded to define which words or features students used in their explanations that conveyed the ideas of substance (or processes). It was found that words such as *used up, drains,* and *burns out* were all used to denote the idea of *consume.* Hence,

in subsequent recodings, they could then clearly specify that any words related to consuming should be coded as substance-based. If there is a great deal of discrepancy between two raters in the first pass (interrater reliability of less than 80%, for instance), then this should caution the researchers to redefine the categories, rather than to concentrate their efforts only on resolving the interrater discrepancies.

The second issue of interrater reliability has to do with whether discrepancies between two coders should always be resolved. There are two kinds of discrepancies. In the one kind, both coders have firm ideas about which code a particular segment of protocols should be assigned. This kind of discrepancy is the kind that is computed in an interrater reliability index. However, a second kind of discrepancy can occur not because the coders disagree with each other, but rather, each coder is unsure which code should be assigned to a segment, because the segment is very ambiguous. In these cases, instead of resolving the discrepancies between the two coders, it may be better just to complete the coding and then count the number of these ambiguous cases as the uncodable portion of the data. For example, if 20% of the segmented protocols were uninterpretable, then it may be more appropriate to simply say that only 80% of the data were coded.

The third issue has to do with when discrepancies of the first kind (i.e., between two coders when the segment is interpretable) should be resolved. My preference has been to not resolve the interpretation of the segment in the midst of coding, because resolving them can actually bias the interpretation of subsequent codings. Rather, one should attempt to code a section and see how much reliability is achieved. If very little reliability is achieved, then perhaps the coding system should be abandoned and a new one should be developed. Note that interrater reliability should be calculated at various steps of the verbal analysis technique: during segmentation into units, categorizing or coding of the units, depicting the coded data, seeking pattern(s) in the depicted data, interpreting the pattern(s), and so forth. Thus, as one can see, verbal analysis is an extremely time-consuming process.

Successive Analyses

One of the key principles of verbal analysis is to analyze the results successively or iteratively, using smaller and smaller grain sizes, or focusing on subsets of the data, to obtain more sensitive results. This is somewhat analogous to the process of statistical analyses whereby one may start with a general analysis of variance, which may then be followed by a test for linear trend, using a more restricted sample of the data. The following example illustrates a successive analysis using behavioral categories. One behavioral category, used in Chi, Bassok, et al. (1989), was to identify the frequency with which students solving a physics problem actually referred back to the example illustrated in the text. To find this out, we needed to read the content of the problem-solving protocols to see how often references were made to the examples in the text. This was a relatively easy analysis that revealed

that all students relied on examples in the text for each problem that they were solving. At this point, we could have stopped the analysis because the results replicated the findings in the literature (e.g., Pirolli & Anderson's, 1985, results), that all students rely on examples. However, additional successive analyses with smaller grain sizes told us something more revealing and furthered our understanding of how examples are used, as discussed next.

Instead of analyzing each problem-solving protocol as one unit (to see if a student referred to an example within that problem solution), each physics problem-solving protocol was further divided into episodes, with each episode corresponding to a major activity (such as trying to draw a free-body diagram, finding values for an equation, etc). It was then found that the successful solvers referred to the examples infrequently (about once per problem solution), whereas the less successful solvers referred to the examples quite frequently (about six times per problem solution). At this point, the analysis could have stopped again because there is an interesting and significant finding. But additional successive analyses examined the number and location of text lines within an example that the students referenced. It turned out that the successful solvers generally referred only to 1 line in the example, and furthermore, they targeted a specific line for their reference, whereas the less successful solvers reread almost the entire example (around 13 lines) and started at the beginning of the example, indicating that they had no idea where to look for the information they needed. Their reference was just a general search for information.

There are two things to point out about this analysis. One feature is its repetitive nature: that is, successively smaller and smaller grain sized units were analyzed. The researchers first looked at the whole solution to see whether examples were referenced, then looked at episodes inside each solution to see how frequently each example was used, and then finally examined how many lines and which lines of the examples were used. The last analysis revealed something about the kind of knowledge the successful solvers gained: Their problem-solving behavior seemed to be dictated by a top-down search so that they knew precisely which equation or pieces of information to seek from the example (i.e., their search of the example was directed to a specific piece of information that they lacked). Less successful solvers, however, were simply seeking any piece of information (usually any equation containing the appropriate unknown variables) that matched the givens. Because they often merely reread the example, they had probably gained little or no knowledge from studying the example earlier.

The second thing to point out about this analysis is what urged the researchers to continue to analyze the data at successively smaller grain sizes. The answer to this is simply that reading the protocols gave us the impression that there was a difference in the way successful and less successful students used examples. The analyses became a problem of finding a way to code the data so that this impression can be captured. Clearly, as suggested earlier, more sensitive data codings are more likely to lead to capturing such differences.

Within-Subject Analysis

Many of the analyses reported in this article are cross-sectional so that multiple subjects' data were averaged together and quantitative analyses were then carried out. It is often more interesting to perform single-subject analyses, because this allows one to capture the knowledge representation more precisely and make explicit predictions about which underlying conceptual representation enables which kind of performance. Because within-subject analysis is usually very labor intensive, the assumption here is that one would not attempt multiple within-subject analyses. Because it is more difficult to support this kind of within-subject analysis using statistics, there are other means of demonstrating validity. One way, for example, is to use a predictive method: That is, after capturing what knowledge the subject might have, predict what questions she or he can answer correctly on the basis of that knowledge or what reasoning errors she or he might make. This can typically be done by analyzing separate task performances administered at different times. For instance, one could analyze the knowledge manifested in a pretest, let's say, and predict performance outcomes based on that knowledge. Alternatively, one could compare performances on the first half of the task with the second half, and so forth. An example of a within-subject analysis of tutoring can be seen in Chi (1996).

Another method of validating a single-subject analysis is to compare the results to the literature at large. In the analyses of example-studying data from physics, Chi and VanLehn (1991) found that both good and poor solvers had acquired a great deal of technical procedural knowledge. In the context of the relevant problem-solving literature, which shows that students can often solve problems without understanding, they used their results to offer the interpretation that because students often acquire a great deal of technical procedures, it is this acquisition which permits the generation of a successful solution without real conceptual understanding. Basically, the best way to achieve validity for single-subject analyses is to devise more than one way of reaching the same conclusion. Even so, it is probably wise to use at least a few subjects rather than just one single subject.

Top-Down and Bottom-Up Processes of Interactive Analysis

At every step of the verbal analysis technique—reducing the data, segmenting them into units, categorizing the units, depicting the coded data, seeking pattern(s) in the depicted data, interpreting the pattern(s)—both top-down and bottom-up processing must be occurring. Top-down means that questions and codes used, for example, are driven by theory; bottom-up means that the codes can be refined on the basis of the protocols, and new hypotheses can be generated from the data. In discussing the technique, this article has described mainly the top-down nature of the deci-

sion-making process (such as deciding the unit size to segment the data). Obviously, in undertaking an actual analysis, many specific decisions are made bottom-up, on the basis of the nature of the data. The example cited earlier about capturing the initial mental model of the circulatory system, followed by identifying the defining characteristic of each mental model, then reconfirming the correct classification of each model, illustrates the bottom-up first, then top-down interactive nature of the analysis. Notice the very different conclusion one might have drawn about this data had the analysis been reversed, namely, top-down first then bottom-up. In this latter case, a researcher would have constructed an ideal template model of the circulatory system (corresponding to the Double Loop 2 model in Figure 1) and then looked for the extent to which each student's model lacked certain components and connections as compared to the ideal model. Such a top-down analysis (sometimes known as the overlay approach in intelligent computer assisted instruction; VanLehn, 1988) would lead to the conclusion that students' models are necessarily incomplete and incoherent, without ever being able to say exactly what the student's misconceived model is. With the analysis that was done (first bottom-up, then top-down), just the opposite conclusion was made, namely that students' mental models, although incorrect, may be completely coherent in that there were no inherent contradictions, there were no incomplete connections, and students could use their mental models to extrapolate predictions and answer questions. Hence, this example suggests that one should not proceed in a top-down manner. Thus, in a sense, the entire research orientation is bottom-up because the goal is to seek the knowledge representation that a learner has, rather than determining a priori a possible set of representations and see which one the learner's representation fits the best.

However, although the orientation of the analysis itself was bottom-up, the more global and theoretical decisions are made top-down (such as the kind of questions to be asked, the nature of the formalism, and the kind of codes to be use). The following example illustrates the top-down nature of verbal analysis. For a number of years, I have been thinking about and developing a theory of conceptual change (Chi, 1992; in press-a; Chi, Slotta, & de Leeuw, 1994). In that theory, I had proposed some fundamental (ontological) differences between entities consisting of physical *substances* (e.g., a carrot, a book) that possess ontological attributes such as *can have color, can have volume, can have weight,* and *can move or be moved,* versus *processes* of interaction, such as the process of equilibrium seeking (e.g., the diffusion of a drop of red ink in a jar of water, heat transfer, natural selection), which do not possess such attributes. To examine students' understanding of situations involving a physical substance (such as gas) versus an equilibrium-seeking process (such as heat), their predictions and explanations for two analogous situations, one involving water and the other involving heat, were examined (Slotta & Chi, 1996; Slotta et al., 1995). The gas situation required the students to predict which of the two balloons is more buoyant after several hours of floating inside a closet, the one

made with an ordinary paper bag or the one made of durable elastic rubber, when both are filled with helium gas and sealed tightly at the opening. Similarly, in the heat situation, the students were asked to predict which cup of coffee is hotter after leaving it on a table for 20 min: the styrofoam cup or the ceramic cup.

The mechanics of the analysis involved coming up with codes (Step 3 of the technique) that could differentiate the kind of explanations students gave, without worrying about either the correctness or incorrectness of their predictions or the accuracy of their explanations. The codes that they decided on using were ontological attributes of substances and processes (such as can have color, or move for substances). For instance, situations involving a gas (a physical substance) could be described by predicates such as *moves* or *is heavy,* but processes such as heat should not be described with these predicates. Hence, the kind of predicates the students used was analyzed within each explanation of each problem situation to conclude how a student was representing the entity: as substance-like or as a process. To ascertain that the students considered heat to be a substance, the students should use words that fit the predicate of *moves.* For example, the student who explained that the coffee in the ceramic mug was hotter than in the styrofoam cup, explained that it's "because the heat in the styrofoam cup is gonna *escape* because a styrofoam cup is not totally sealed ... styrofoam has little holes in it. So it, the heat is gonna *go out, escape,* in the holes" (Slotta et al., 1985, pp. 380). Thus, the words used that have synonymous meanings as *moves,* such as *goes out, escape,* constituted evidence for the code *moves* (Step 4 of the technique). This example illustrates a way of formulating the appropriate codes and searching for evidence in the subject's utterances for the codes that was completely guided by a theory of conceptual change. Thus, verbal analysis involves both top-down and bottom-up process of analyses.

SUMMARY AND CONCLUSIONS

This article discusses a kind of analysis of verbal data that integrates both qualitative and quantitative components. Basically, it's a kind of analysis that quantifies qualitative codings, as opposed to three alternative methods that do not actually integrate quantitative and qualitative methods as much as they use them side-by-side. Verbal analysis is also contrasted with protocol analysis of verbal data, introduced by Newell and Simon (1972). This article then lays out the mechanics of analyzing verbal data into eight steps, and each step is illustrated with examples taken from published work. Hopefully, these steps are sufficiently transparent that the readers can use them as a guide to undertake analyses of their own. Finally, additional concerns, recommendations, and caveats are raised about verbal analysis.

Two aspects are involved in doing both the traditional quantitative type of research involving experimental design and the current qualitative type of research involving verbal analysis. One aspect concerns the generation of the right questions,

which then determines what kind of experimental conditions to test in the quantitative method case and what kind of coding and formalism to use in the qualitative case. This is the theory part: That is, one's theory generally drives the questions, which then drives the analyses. The second aspect of research in both experimental design and verbal analyses is the mechanics part. What this article has laid out is largely the mechanics of doing the analyses. In any research paradigm, there is a theory part and a mechanics part. For some reason, people have often attributed the theory part of qualitative analyses to "art," as if the analyses are based on some mysterious intuitions that cannot be clearly laid out. Hopefully, this article has amended that perception, especially after having laid out in the previous section an example of how a theory of conceptual change contributed to the selection and the formulation of codes for the verbal analysis. In that case, no art was involved at all, unless one considers theory-building as an art. Hence, it would probably make no sense to use this guide unless the researcher is already armed with some questions or a theory, because this guide basically lays out only the mechanics of doing the analyses.

In order not to end on a discouraging note because this guide may still appear opaque, let me recommend a few conservative ways to start. The first time users of verbal analysis might want to take either the complement approach, the interpretation approach, or the two-step approach. Recall that these approaches involve a less aggressive integration of quantitative and qualitative analyses mentioned earlier. Doing so may well enlighten the way to a clearer conception of how verbal analysis is done.

ACKNOWLEDGMENTS

This article was prepared in part while the author was a Fellow at the Center for the Advanced Study in the Behavioral Sciences, 1996–1997. It was supported by Grant 199400132 from the Spencer Foundation. Comments and criticisms from the following people are greatly appreciated: Gitti Jordan, Janet Kolodner, Bernadette Kowalski, Michael Ranney, Peter Reimann, Ross Silberberg, Stephanie Siler, Fritz Staub, Roger Taylor, Kurt VanLehn, and five anonymous reviewers.

REFERENCES

Chi, M. T. H. (1985). Interactive roles of knowledge and strategies in the development of organized sorting and recall. In S. Chipman, J. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Current research and open questions* (Vol. 2, pp. 457–485). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science* (pp. 129–186). Minneapolis: University of Minnesota Press.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10,* 1–17.

Chi, M. T. H. (in press-a). Creativity: Shifting across ontological categories flexibly. In T. B. Ward, S. M. Smith, R. A. Finke, & J. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery and change.*

Chi, M. T. H. (in press-b). Self-explaining: A domain–general learning activity. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Chi, M. T. H., Hashem, A., Ludvigsen, S., Shalin, V., & Bertram, D. (1997). *Stolen knowledge: What is learned in the context of a medical intensive care unit.* Manuscript in preparation.

Chi, M. T. H., Hutchinson, J., & Robin, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill-Palmer Quarterly, 35,* 27–62.

Chi, M. T. H., & Koeske, R. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology, 19,* 29–39.

Chi, M. T. H., Slotta, J. D., & de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction, 4,* 27–43.

Chi, M. T. H., & VanLehn, K. A. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences, 1,* 69–105.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240–247.

de Leeuw, N. (1993). Students' beliefs about the circulatory system: Are misconceptions universal? In W. Kintsch (Ed.), *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 389–393). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

diSessa, A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10,* 105–225.

Ericsson, K. A., & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251.

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

Ericsson, K. A., & Simon, H. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Fisher, C. (1987). Advancing the study of programming with computer-aided protocol analysis. In G. Olson, E. Soloway, & S. Sheppard (Eds.), *Empirical studies of programmers: Second workshop* (pp. 198–216). Norwood, NJ: Ablex.

Fisher, K. (1989). SemNet™ [Computer software]. San Diego, CA: San Diego State University, Center for Research in Mathematics and Science Education.

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

Fox, B. (1991). Cognitive and interactional aspects of correction in tutoring. In P. Goodyear (Ed.), *Teaching knowledge and intelligent tutoring* (pp. 149–172). Norwood, NJ: Ablex.

Geiwitz, J., Klatsky, R. L., & McCloskey, B. P. (1988). *Knowledge acquisition techniques for expert systems: Conceptual and empirical comparisons* (Final report). Fort Monmouth, NJ: U.S. Army Electronics Command.

Gobbo, C., & Chi, M. T. H. (1986). How knowledge is structured and used by expert and novice children. *Cognitive Development, 1,* 221–237.

Goldin-Meadow, S., Alibali, M. W., & Breckinridge Church, R. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100,* 279–298.

Hoffman, R. R. (1987). The problem of extracting the knowledge of experts from the perspective of experimental psychology. *AI Magazine, 8,* 53–67.

Hutchins, E. L., & Levin, J. A. (1981). *Point of view in problem solving* (Rep. No. 105). San Diego: University of California, Center for Human Information Processing.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, MA: Harvard University Press.

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

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Means, M. L., & Voss, J. F. (1985). Star Wars: A developmental study of expert and novice knowledge structures. *Journal of Memory and Language, 24,* 746–757.

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

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

Norman, D. A. (1988). *The psychology of everyday things.* New York: Basic Books.

Olson, J. R., & Biolsi, K. (1991). Techniques for representing expert knowledge. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise* (pp. 240–285). Cambridge, MA: Cambridge University Press.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10,* 91–116.

Pirolli, P., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39,* 240–272.

Ranney, M. (1994). Relative consistency and subjects' "theories" in domains such as naive physics: Common research difficulties illustrated by Cooke and Breedin. *Memory & Cognition, 22,* 494–499.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21.*

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Bream, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 7,* 573–605.

Salter, W. J. (1983). Tacit theories of economics. In *Proceedings of the Fifth Annual Conference of the Cognitive Science Society.* Rochester, NY: Cognitive Science Society.

Schofield, J. W. (1982). *Black and White in school: Trust, tension or tolerance?* New York: Praeger.

Schofield, J. W., & Anderson, K. (1987). Combining quantitative and qualitative components of research on ethnic identity and intergroup relations. In J. S. Phinney & M. J. Rotheram (Eds.), *Children's ethnic socialization: Pluralism and development.* Newbury Park, CA: Sage.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36–71.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481–520.

Simon, D. P., & Simon. H. A. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325–348). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Slotta, J. D., & Chi, M. T. H. (1996). Understanding constraint-based processes: A precursor to conceptual change in physics. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 306–311). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Slotta, J. D., Chi, M. T. H., & Joram, E. (1995). Assessing students' misclassifications of physics concepts: The ontological basis of conceptual change. *Cognition and Instruction, 13,* 373–400.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654.

Tang, J. C. (1989). *Listing, drawing, and gesturing in design: A study of the use of shared workspaces by design teams* (Xerox Tech. Rep. No. P89–00032). Stanford, CA: Stanford University, Department of Mechanical Engineering,

Trabasso, T., & Suh, S. (1993). Understanding text: Achieving explanatory coherence through online inferences and mental operations in working memory. *Discourse Processes, 16,* 3–34.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24,* 612–630.

van den Broek, P. (1989). Causal reasoning and inference making in judging the importance of story statements. *Child Development, 60,* 286–297.

VanLehn, K. (1988). Student modeling. In M. Polson & J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 55–78). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Voss, J. F., Tyler, S. W., & Yengo, L. A. (1983). Individual differences in the solving of social science problems. In R. F. Dillion & R. R. Schmeck (Eds.), *Individual differences in cognition* (pp. 205–232). New York: Academic.

Weber, R. P. (1985). *Basic content analysis.* Newbury Park, CA: Sage.