

18

WHICH ELABORATIVE INFERENCES ARE DRAWN DURING READING? A QUESTION OF METHODOLOGIES

Janice M. Keenan, George R. Potts,
University of Denver

Jonathan M. Golding,
University of Kentucky

Tracy M. Jennings
University of Denver

INTRODUCTION

It is well known that speakers and writers cannot make explicit everything that they want to communicate. Nor do they need to. Instead, they rely on their listeners or readers to fill in whatever gaps may exist in the message; in other words, they rely on their audience to make inferences. The question confronting researchers of language comprehension is: Which inferences can people be counted on to reliably draw? This question is particularly important for reading research because the reader, unlike most listeners, is not in a position to ask for assistance in gap-filling from the author. If we can answer this question of which inferences are likely to be drawn, not only will writers (like us) be able to write more comprehensible texts, but we will be well on our way to a general theory of language comprehension. And it is only when this question is answered that we will be able to solve such practical problems as designing machines that we can converse with.

The question of which inferences are drawn during reading usually refers to a class of inferences called elaborative inferences (or forward

inferences). Elaborative inferences are contrasted with bridging inferences (or backward inferences). The distinction is not based on the type of information inferred but rather what motivates the inference. If the inference is drawn in order to establish coherence between the present piece of text and the preceding text, then it is a bridging inference. If an inference is not needed for coherence, but is simply drawn to embellish the textual information, then it is an elaborative inference. There are many possible elaborations one can make in a text. Consequently, research is needed to determine which, if any, of these many possible elaborative inferences are drawn.

Research on inferencing is relatively recent in the history of reading and psychological research. But despite the infancy of this research, or perhaps because of it, there are a large number of paradigms in use to detect inferences. Not only are there a variety of paradigms, but there are also a variety of times during which inferencing is tested. These times range from testing while a person is reading, to testing them immediately after reading, to testing them long after they have read.

Until quite recently, researchers rarely provided a rationale for either their choice of paradigm or their choice of testing time. It was generally believed that the main question to be asked is which inferences are drawn. The methods used to answer that question were thought to be comparable, or at least sufficiently comparable that one did not need to worry about them until after the more important question of which inferences are drawn was answered.

This may have seemed a reasonable approach to take because the field was in its infancy. Furthermore, if the different methods had yielded similar answers, the approach would have been lauded for providing converging evidence across a wide spectrum of methodologies. As it turns out, however, different methods often result in different answers to the question of whether an inference is drawn. And with the focus being on the type of inference drawn rather than the methodology, the differences between studies are often incorrectly attributed to the types of inferences drawn rather than to the methods used to detect them.

For example, there is a tendency to believe that people infer the referents of general terms (e.g., infer that the general term *animal* refers to a cow in the statement, *she was milking the animal*) but that they do not make instrument inferences (e.g., infer *knife* when reading *he stabbed the intruder*). The evidence cited to support this often comes from methodologies as diverse as a recognition test given long after the passage is read for the referent inferences (McKoon & Ratcliff, 1989a) and a Stroop task given immediately after reading for the instrument inferences (Doshier & Corbett, 1982). Consequently, it is unclear whether the differences between the studies are due to the type of inference, as is generally believed, or to the different requirements of the tasks, or when the test was presented.

The use of different methodologies is a problem not only in comparing different types of inferences but also within a given inference type. The result is that it is difficult to say for any single inference type

whether readers typically draw that inference, because frequently there are almost as many studies answering the question affirmatively as negatively. This can be seen in Table 18.1, which summarizes the results of studies examining three main types of elaborative inferences: instrument inferences, instantiations of general terms, and inferences about the likely consequences of events (such as inferring from the fact that someone threw a delicate porcelain vase against a wall, that the vase broke).

Many simplifications have been adopted in constructing this table. First, not every inference study is included; rather, we selected representative studies to show the variety of tasks that have been used. Second, for the purpose of easy viewing, the classification of testing time is simplified. Any study that tested for the inference either while reading the inference sentence or immediately after it is classified as an immediate test. Any study that had some intervening material between the inference sentence and the test, be it from the same text or a different text, is classified as a delayed test. Third, the inferences summarized are always dominant exemplars (e.g., *hammer* rather than *mallet*) or the

The point of Table 18.1 is to show the variety of testing procedures used to detect inferences and to show that for each inference type, there is no consensus as to whether the inference is typically drawn. Furthermore, it is hard to know if the different answers are due to the different requirements of the tasks or to the different testing times. Also complicating the picture is the fact that tasks and test times tend to be confounded. For example, cued recall is always used after a long delay, whereas activation measures (lexical decision, Stroop, naming) tend to be used either immediately or after a very short delay.

There are, of course, other differences among the studies in Table 18.1 besides the tasks and test times. Most notable are differences in materials, especially the context in which the inference is presented. For example, some studies used single sentences, whereas others used short paragraphs. Some paragraphs had previous mention of the inference concept, and some had highly constraining contexts. When making comparisons across studies, these text differences certainly may play a role in accounting for the discrepancy in answers to the question of whether the inference is drawn (cf. Lucas, Tanenhaus, & Carlson, 1987; O'Brien, Shank, Myers, & Rayner, 1988; Whitney, 1986). However, a number of the entries in Table 18.1 show that even within the domain of a single study, when materials are controlled, differences in the domain and testing time yield different results.

Consider, for example, the instrument inference studies in Table 18.1. Although the studies listed are almost evenly split on the question of whether these inferences are typically drawn, the predominant view is that instrument inferences are not drawn. This view is based on disshowing that the cued recall data as flawed and summarizing the others as showing that instruments are not inferred unless the instrument has been previously mentioned in the text, as in the McKoon and Ratcliff

TABLE 18.1

Inference Studies Classified by Type of Task or Measure Used to Detect the Inference, Time of Testing, and Type of Inference Drawn.

	Task or Measure	Test Time	Is Inference Drawn?
Instrument Inferences			
(e.g., infer spoon from stir the soup)			
Dosher & Corbett (1982)	Stroop	Immediate	NO
Lucas et al. (1987)	Lexical decision	Immediate	DC*
	Naming	Immediate	NO
McKoon & Ratcliff (1981)	Recognition	Immediate	YES
Paris & Lindauer (1976)	Cued recall	Delay	YES
Singer (1979)	Reading time	Immediate	NO
Instantiations of General Terms			
(e.g., infer cow from milk the animal)			
Anderson & Ortony (1975)	Cued recall	Delay	YES
Garrod & Sanford (1977)	Reading time	Immediate	NO
McKoon (1988)	Lexical decision	Immediate	NO
	Recognition	Delay	YES
O'Brien et al. (1988)	Fixation duration	Immediate	DC*
Whitney (1986)	Stroop	Immediate	YES
Likely Consequences of Events			
(e.g., infer cut from stepped on a jagged piece of glass)			
Duffy (1986)	Continuation judgment	Immediate	YES
McKoon & Ratcliff (1986)	Recognition	Immediate	NO
	Cued recall	Delay	YES
		Delay	NO
		Delay	YES
McKoon & Ratcliff (1988)	Recognition	Immediate	NO
		Delay	YES
Potts et al. (1988)	Lexical decision	Delay	YES
	Naming	Delay	NO
Singer & Ferreira (1983)	Verification	Immediate	NO

DC* = Depends on context.

(1981) study. However, a recent study by Lucas et al. (1987) showed that the situation is not that simple; methodological factors are paramount. In their study, all texts had previous mention of the instrument and all tests were administered immediately, yet the answer to the question of whether the inference was drawn depended on the particular task used to assess it.

Much of the research on instantiations of general terms suggests that when the inference is highly predictable, it is likely to be drawn. Yet here too, methodology ultimately determines the answer. Consider McKoon's (1988) study. Her immediate lexical decision task suggested the inferences were not drawn whereas her delayed recognition test suggested they were. Of course, in this case it is hard to know whether the controlling factor was task or delay, because the two were confounded. Nonetheless, it is clear that methodological factors are important in determining whether one concludes that these inferences are drawn or not.

Studies of inferences concerning the likely consequences of events are evenly divided on the question of whether they are typically drawn. Even within the same paradigm, one arrives at different conclusions depending on when the test is administered (McKoon & Ratcliff, 1989b) or whether the test item is primed by another item from the text (McKoon & Ratcliff, 1986). And even when delay and priming are controlled, the answer to the question of whether consequent inferences are drawn depends on the task (Duffy, 1986; Potts, Keenan, & Golding, 1988).

In sum, Table 18.1 provides clear evidence that the procedures used to assess inferences seem to be at least as important as the type of inference in determining whether inferences are drawn. Why don't the various tasks used to detect inferences yield comparable results? What are the differences between the tasks that are responsible for the different results? What effect does the time of test have on the results? The purpose of this chapter is to try and answer these questions.

MEASURES OF INFERENCING AND THE DEBATE OVER WHAT CONSTITUTES AN INFERENCE

Measures of inferencing other than reading time measures can be divided into two classes: memory measures and activation measures. Memory measures require subjects to access their representation of the text to see if the inferential information is part of the text's representation. They include cued recall, sentence verification, question answering, and recognition measures. Activation measures, on the other hand, do not require subjects to evaluate their representation of the text. Instead, they detect inferences by seeing whether an inference concept is primed, that is, if it is more activated after reading an inference version of a text versus a no inference control version. The notion is that if an inference has been drawn, then that inference is more activated.

as part of constructing the text's representation. Tasks used to assess priming include naming, lexical decision, and the modified Stroop task.

There are two ways in which an inference concept can be activated. One is through intralexical associations as a result of reading related words in the text; we refer to this as word-based priming. The other is by actually drawing the inference based on one's knowledge about the situation described in the text. Many researchers are not inclined to consider word-based priming as drawing an inference; so, they control for intralexical activations and focus on knowledge-based activations in assessing inferencing (see Keenan, Golding, Potts, Jennings, & Aman, in press; McKoon & Ratcliff, 1986).

The decision to use a memory measure versus an activation measure is often governed by one's theory of what constitutes an inference. At issue is whether one considers simply activating an inference concept as a legitimate case of drawing an inference. Those who reject an activated concept as constituting an inference prefer memory measures to activation measures. There are two reasons for this preference, and these correspond to two dimensions involved in defining an inference.

One dimension is the unit of inferencing. It can range from an activated concept, to a set of concepts constituting a proposition, to a set of propositions constituting a higher order knowledge structure such as a schema. Some researchers' belief that only a proposition or a set of propositions constitutes an inference has led them to reject measures that test the activation level of a single concept. We wish to point out, however, that if a subject constructs a proposition, this process will involve activating concepts. Hence, measures of activation can be sensitive to all units of inferencing.

Another dimension defining an inference is the level to which an inference is processed. This can range from simple activation, to selection for maintenance in working memory, to incorporating the inference into the long-term representation of the text (see Kintsch, 1988). Some researchers believe that what should count as an inference is only what gets incorporated into the final representation of the text. Consequently, these researchers prefer memory measures that require subjects to access their representation of the text over activation measures. We wish to point out, however, that memory measures are also sensitive to activation. So, unless the test is sufficiently delayed to allow for activation to dissipate (and no one yet knows what factors govern the dissipation of activation in text comprehension), a memory test may be measuring activation as well as, or instead of, incorporation.

In sum, although memory measures have sometimes been preferred on the grounds that they measure a higher level of inferencing and can detect incorporation into the text base, these arguments are open to criticisms. In addition, memory measures suffer from another serious problem. Namely, these measures do not allow one to determine whether the inference occurred during comprehension or as a result of the test. That is because whenever the task requires the subject to go back to the text and process it, the subject is given another opportunity to draw

the inference. Consequently, it is difficult to determine whether apparent evidence for the inference was due to processes occurring encoding or processing at the time of test. Note that having the text occur on-line as opposed to delayed (as in cued recall) is not enough to prevent the problem. In other words, testing during comprehension may be preferable to testing long after the text has been read because it minimizes reconstruction. But as long as the measure involves accessing the text, there is no guarantee that the inferences detected were drawn when the text was read. We develop this point more as we discuss each measure.

A CRITICAL ANALYSIS OF TASKS USED TO DETECT INFERENCES

Cued Recall

One of the earliest techniques used to detect the occurrence of elaborative inferences was cued recall. In this paradigm, subjects read sentences in which some information, such as the instrument of the action is either left implicit or made explicit. An example from Corbett and Doshier (1978) is:

Explicit: The athlete cut out an article with scissors for his friend.
Implicit: The athlete cut out an article for his friend.

After reading a set of such sentences, the instruments (e.g., *scissors*) are then given as cues to recall the sentences. Paris and Lindauer (1976) found that an instrument is just as effective a retrieval cue when it is implicit as when it is explicit. Based on this, they concluded that the instrument must have been inferred during encoding.

Singer (1978) challenged this conclusion by providing evidence that suggested that cued recall reflects reconstructive processes occurring during retrieval rather than inferences at encoding. He used sentences containing actions, such as *stir the soup*, for which the possible instruments, ladle and spoon, are asymmetrically related to the action. There is a strong forward association between *stir the soup* and *spoon* but a rather weak backward association from *spoon* to *stir the soup*. On the other hand, there is a weak forward association between *stir the soup* and *ladle*, but a strong backward association. If instruments are inferred during reading, then reading *stir the soup* should lead one to infer *spoon* more often than *ladle* because it is more strongly associated. Therefore, *spoon* should be a more effective retrieval cue than *ladle*. On the other hand, if cued recall reflects reconstructive processes at retrieval, then *ladle* should be a better cue than *spoon* because *ladle* is more strongly associated to *stir the soup*. Singer's results showed *ladle* to be the more effective cue, suggesting that cued recall reflects reconstruction at retrieval. A similar point was made by Corbett and Doshier (1978). As a result of findings such as these, most inference researchers abandoned the cued recall paradigm as a method of detecting inferences.

