Techniques for Assessing Comprehension Monitoring

Fengzhi Ling

ABSTRACT Over the last two decades, comprehension monitoring has attracted increasing interest from cognitive researchers. This review examines comprehension monitoring from the perspective of its research methodology or experimental technique. It reviews the comprehension monitoring mechanism and procedures for its explicit measurement. It analyses four of the main experimental techniques used to investigate the nature of this process. Strengths and weaknesses of each technique are discussed.

INTRODUCTION

Reading is an information processing activity in which readers process a text at a number of levels, integrate the outcomes and form an overall impression or 'macrostructure' representation of the text (Kinnunen & Vauras, 1995). This macrostructure represents the key ideas readers encountered during reading and ideas they might expect. It guides the subsequent reading activity. As readers read, they process portions of the text at a time. Each portion processed is represented by a microstructure. One aspect of the reading activity involves cross referencing the most recent microstructure representation of the 'just read' text with the macrostructure representation. This indicates whether the microstructure is consistent with, contradicts or modifies the macrostructure. This cross-referencing process is comprehension monitoring. It is the process by which readers continually review their comprehension of the text by matching their 'moment by moment' comprehension with their macrostructural understanding of it.

This process directs reading comprehension (Wagoner, 1983). Perceiving a mismatch between the macrostructure and microstructure representations at any time during reading cues readers to modify their pattern of attention allocation and re-evaluate the outputs from some of the levels that led to the particular microstructure representation.
Comprehension monitoring provides an insight into reading comprehension. Although poor comprehension monitoring proficiency can occur at all reading ability levels across a variety of ages (Baker, 1985a; Pressley et al., 1992), it may be "a very sensitive 'leading indicator' of reading ability" (Weaver et al., 1995, p. 190), differentiating good from poor readers (Baumann et al, 1993; Paris & Myers, 1981; Weaver et al., 1995).

WHAT IS COMPREHENSION MONITORING?

Comprehension monitoring is the application of cognitive monitoring to reading comprehension (Baker & Brown, 1984). As such, it is an aspect of metacognition (Baker, 1985; Erickson, 1985; Fujiki & Brinton, 1993; Wagoner, 1983). Metacognition comprises at least two components: an awareness of the role of skills, strategies and resources in effective task performance and the ability to use self-regulatory actions such as checking, revising and remediating for successful performance. The self-regulatory function is cognitive monitoring (Baker and Brown, 1984).

Comprehension monitoring comprises two components: evaluating the on-going match between the macro- and micro- levels of reading comprehension (that is, "knowing about comprehending") and taking remedial actions when comprehension fails (that is, "knowing how to comprehend") (Baker, 1986b; Holbrook, 1986; Wagoner, 1983; Zabrucky & Ratner, 1986, 1989, 1990).

The relationship between these two components is complex. Evaluation is essential to comprehension monitoring (Wagoner, 1983) and can occur without remedial actions, that is, regulation. Regulation, on the other hand, cannot emerge without evaluation (Baker, 1985a). Successful evaluation may accompany poor regulation, but good regulation is unlikely with poor evaluation (Bossert & Schwantes, 1995-96; Zabrucky & Ratner, 1986, 1989, 1990; Zabrucky et al., 1987). Further, regulation develops later than evaluation. The two components are mediated by different cognitive processes and develop at different rates. Regulation not only requires more prerequisite abilities such as identifying the nature of comprehension failure, knowing how to make up failure, but also puts more loads on the working memory (Kinnunen & Vauras, 1995).

It has been said that comprehension monitoring generally proceeds unconsciously (Holbrook, 1986; Pitta, 1983). Even mature readers typically engage in comprehension monitoring seldom consciously. Characteristics of both a reader and a text affect whether the reader comprehends the text subconsciously or consciously (Pitts, 1983). As a consequence, monitoring ability develops slowly and unnecessarily with age because of the critical importance of experience and expertise (Baker & Brown, 1984; Pitts,1983).
THE REFERENTIAL COMMUNICATION TECHNIQUE

The referential communication paradigm, also named the communication adequacy paradigm, is the first technique to be used in the area of comprehension monitoring (Pressley et al., 1990). It is used to examine young children’s ability to beware and detect inadequate message while listening.

The procedure of the referential communication technique

The referential communication paradigm or the communication adequacy paradigm measures comprehension monitoring of spoken language using direction-following tasks (Fujiki & Brinton, 1993; Robinson & Whittaker, 1985). Listeners are typically asked to perform a set of instructions, for example, to construct a block building or select one item from an array. Some of the directions are referentially ambiguous, that is, describing more than one block or item. They are then asked to decide whether they understood the directions. Listeners are permitted to ask questions about any confusion (Markman, 1981). Comprehension monitoring is defined here as the expressed awareness of an inadequate message (Wagoner, 1983). This paradigm is often employed to test children’s ability to monitor comprehension (Bonitatibus, 1988).

The task used varies among available studies: children have been asked to identify described pictures (either geometric figures (Patterson et al., 1981; Bonitatibus & Flavell, 1985; Bonitatibus et al., 1988), people (Reid, 1990; Bonitatibus, 1988; Robinson & Whittaker, 1985), or common objects (Beal & Belgrad, 1990)); to learn how to play a game by following specified directions which are presented directly or indirectly (Markman, 1977; Robinson & Robinson, 1976, cited by Wagoner, 1983); or to identify inadequate messages in story context (Patterson et al., 1981). Reid (1990), for example, showed kindergarteners and first graders a set of clown face cards. The faces varied systematically in three two-valued attributes: mouth (happy or sad), hat (round or square), cheeks (spotted with a red spot on each cheek or freckled). One set of cards comprised a happy clown with spotted cheeks and a square hat, a sad clown with freckled cheeks and a square hat, and a happy clown with spotted cheeks and a round hat. Four types of directions (two informative, two noninformative) were used to describe each display. Each direction contained two attributes. Each child was required to select the picture specified in each direction described.

At least three experimental formats have been used. In two, the subjects act as listeners and the experimenter works either as the speaker (Patterson et al., 1981; Bonitatibus et al., 1988; Fujiki & Brinton, 1993; Beal & Belgrad, 1990; Robinson & Whittaker, 1985) or as a questioner who directs a fictitious speaker (Reid, 1990; Bonitatibus & Flavell, 1985). A third format uses both a fictitious speaker and a fictitious listener (Bonitatibus, 1988). The subjects judge message adequacy and the
experimenter acts as a questioner. Bonitatibus (1988), for example, manipulated two small stuffed animals Bear and Penguin in a communication game for first graders who answered questions about what Bear and Penguin were doing and saying.

Monitoring has been examined both when students have been cued to expect that some instructions may not provide clear information (Beal & Belgrad, 1990; Bonitatibus, 1988; Bonitatibus & Flavell, 1985; Markman, 1977; Robinson & Whittaker, 1985) and without this advance cuing (Fujiki & Brinton, 1993; Patterson et al., 1981; Reid, 1990). These two conditions are referred to as spontaneous and directed communication conditions respectively.

A brief summary of previous research
Young children tend to overestimate message adequacy (Beal & Belgrad, 1990). Those aged up to 7 years have difficulty identifying referentially ambiguous messages (Markman, 1979, 1981; Patterson & Kister, 1981; Patterson et al., 1981, cited by Reid, 1990), interpret them as adequate (Bonitatibus, 1988; Robinson & Whittaker, 1985) and attribute the communicative failure to the listener (Flavell et al., 1981; Patterson et al., 1980, cited by Bonitatibus, 1988; Robinson, 1981). The age at which they realise the responsibility of the speaker for the adequacy of the message continues to be debated. While some investigators have shown that children younger than 6 to 7 years of age can evaluate message ambiguity (Courage, 1989; Pratt & Bates, 1982; Sodian, 1988, cited by Reid, 1990), others (Robinson, 1981) have shown that this is not acquired until 7 years of age. However, there is a widespread agreement on the point that the competence to detect ambiguous message increases with age (Wagoner, 1983).

Explanations for poor performance on direction-following tasks
The determination of the causes of poor performance on direction-following tasks is a vexed question. These causes have been explained in terms of cognitive ability, predisposition and properties of the stimulus message.

Inability to monitor comprehension has been attributed to a relatively inadequate cognitive processing of the relevant messages and stimuli (Markman, 1981; Patterson et al., 1981). Reid (1990) argued that children could evaluate the message but not regulate ambiguous messages. Other investigators (Robinson, 1981; Robinson & Whittaker, 1985) have attributed poor monitoring to children’s ignorance of ambiguous messages.

Some researchers have attributed poor performance on direction-following tasks to students’ inability to differentiate between the literal meaning of a speaker’s direction and his/her intention (Bonitatibus, 1988; Beal & Belgrad, 1990). Young children evaluate message quality using their interpretations of the message rather than its literal meaning (Beal & Belgrad, 1990). An essential difference between poor and successful comprehension monitors is the degree to which they attend to the literal meaning.
meaning of a referential communication (Bonitatibus, 1988). Poor monitors are less able to differentiate and attend to literal meaning. This inability leads to a false sense of quality of the expression, probability of communicative success and state of their own understanding. There is a need to investigate "which factors are responsible for children's propensity to respond to the communicative intention and under what circumstances they are more prone to do so" (Reid, 1990, p. 326).

Inability to monitor comprehension on these tasks has also been attributed to children's "tendency to focus on the elocutionary force of utterances rather than on their propositional content" (Ackerman, 1981, cited by Robinson & Whittaker, 1985, p. 448), that is, their 'performative bias'. This leads to all utterances being interpreted as requests for action. Bonitatibus et al. (1988) found that children's social cognition about speaker's intentions significantly influenced their performance on referential communication tasks.

A third cause of inability to monitor comprehension on direction-following tasks has been attributed to both complexity of the stimulus array and degree of message ambiguity (Fujiki & Brinton, 1993; Patterson et al., 1981). Patterson and others (1981) found that preschoolers and first graders demonstrated considerable skills in comprehension monitoring under the simplified stimulus conditions.

In the present authors' opinion, within a referential communication paradigm, children receive all messages as orders to do some tasks. As a consequence, their attention mainly focuses on a message's communicative function instead of its clearness or adequacy. Whether informative or uninformative, each message has a communicative function. Thus, children's ignorance of ambiguous referents makes them judge and pick one item on the basis of their own understanding of it. If a task direction is strictly specified to constrain its communicative function, children may attend to literal expression and detect more ambiguous messages. Future investigation is needed to test this hypothesis.

**Strengths and weaknesses of the referential communication technique:**

The referential communication paradigm makes it possible to investigate objectively whether young children who have developed basic listening comprehension ability can monitor explicitly their comprehending. Some researchers have worked with children as young as 2 years old who begin to display comprehension monitoring (Revelle et al., 1985, cited by Fujiki & Brinton, 1993). The procedure can be operated such that a range of complexities in stimuli can be examined. As well, it allows both evaluation and regulation to be assessed through subjects' questions and reactions to ambiguous parts of directions respectively.

However, various limitations have been noted for the referential communication technique. The first concern relates to what the technique has been used to assess. Essentially, the technique tests listening comprehension monitoring rather than
reading comprehension monitoring. It usually assesses awareness of how well verbal input has been understood (Bonitatibus et al., 1988). In terms of Baker's (1985a, 1985b) classification of the evaluation standards, it examines only one evaluation standard, namely informational clarity and completeness. In addition, the technique tests metalinguistic ability extremely stringently in that children are required to "solve" a "communication problem" in a non-naturalistic setting (Flavell, 1977, cited by Bonitatibus, 1988). Moreover, its relevance to reading comprehension monitoring has not been established, for example, whether subjects find it easier to detect referentially ambiguous information by reading than by listening. More problematic messages are detected when both written and spoken instruction are given than only spoken direction is provided (Bonitatibus & Flavell, 1985).

The validity of indicators of comprehension monitoring is a second major concern. It is doubtful that non-questioning and nonselection of a referent made by listeners to ambiguous referents are sensitive enough to demonstrate deficiency of comprehension monitoring. Children's performance on noninformative messages is usually accompanied by such nonverbal signs as facial expressions and greater reaction time (Bearison & Levey, 1977; Flavell et al., 1981; Ironsmith & Whitehurst, 1978; Patterson et al., 1980, cited by Reid, 1990). Those signs could illustrate a certain degree of incomprehension. Furthermore, detection of referent ambiguity is not checked further by a probe question for ambiguous information similar to that asked in the error detection paradigm. The way in which a referent is made ambiguous affects listeners' performance on a task. Listeners would report information confusion when they realise that they cannot follow ambiguous information to do something.

A third concern relates to the age range and characteristics of subjects for which it is appropriate. The procedure has been used mainly with younger children (preschool, kindergarten, and elementary school aged children). As well it would seem to be inappropriate for subjects who have hearing problems.

THE ERROR DETECTION TECHNIQUE

The error detection technique has been used widely to investigate the nature of comprehension monitoring. This procedure defines the evaluation component of comprehension monitoring operationally as the ability to identify successfully errors introduced intentionally into an intact text and the regulation component as the ability to deploy remedial strategies for comprehension failure (Baker, 1985b; Walczyk & Hall, 1989; Zabrucky & Ratner, 1989, 1990, 1992). In other words, the detection and correction of erroneous information planted purposely in a complete text are the operational indices of comprehension monitoring.
The evaluation standard
Comprehension monitoring is multifaceted (Baker, 1985a, 1985b; Zabrucky & Ratner, 1986) and can be evaluated at a number of levels of text processing. These are referred to as 'evaluation standards' (Baker, 1985a, 1985b). Three basic types of standards are used to evaluate ongoing comprehension: lexical, operating at the level of individual words; syntactic, requiring sensitivity to the grammatical constraints of the language and semantic, requiring consideration of the meanings of individual sentences and the text as a whole. The semantic standards are divided further into five categories: *(1)* propositional cohesiveness, checking that the ideas expressed in adjacent propositions can be successfully integrated; *(2)* structural cohesiveness, checking that the ideas expressed throughout the text are thematically compatible; *(3)* external consistency, checking that the ideas in the text are consistent with what one already knows; *(4)* internal consistency, checking that the ideas expressed in the text are consistent with one another; and *(5)* informational clarity and completeness, checking that the text clearly states all of the information necessary to achieve a specific goal* *(1985b, p. 156-157).* Baker's classification covers a full range of evaluation standards varying from low to high levels.

These standards can be categorised into microstructure and macrostructure standards (Baker & Zimlin, 1989). Microstructure standards involve low-level evaluation and include lexical, propositional cohesiveness and external consistency standards, while macrostructure standards, involving high-level evaluation, include structural cohesiveness, internal consistency, and informational completeness standards. The various types are shown in the following table.

Table 1: A Category of Evaluation Standards

<table>
<thead>
<tr>
<th>Basic Types</th>
<th>Microstructure</th>
<th>Macrostructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexical standard</td>
<td>lexical cohesiveness</td>
<td></td>
</tr>
<tr>
<td>syntactic standard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>semantic standard</td>
<td>propositional cohesiveness</td>
<td>structural cohesiveness</td>
</tr>
<tr>
<td></td>
<td>external cohesiveness</td>
<td>internal consistency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>informational completeness</td>
</tr>
</tbody>
</table>

A second classification based on the nature of errors and developed by Englert et al. (1988) proposes three types of errors: reader-based errors refer to information inconsistent with students' knowledge of the world, text-based errors refer to information inconsistent with the specific details of the preceding text and text-structure-based errors refer to information that is consistent with the topic but incompatible with the prevailing text structure.
The procedure of the error detection paradigm
The error detection paradigm involves constructing reading material with contradictory or anomalous information. The various evaluation standards mentioned above are represented by incorporating the corresponding types of errors in the text (Baker, 1985b; Baker & Zimlin, 1989):

- **lexical errors** by replacing one word in a sentence with an orthographically regular nonsense word that varies in length from one to three syllables.
- **external consistency errors** by substituting one word in a sentence with another that makes the information untrue or implausible.
- **propositional cohesiveness errors** by substituting a pronoun or a general noun for a specific noun in a context where the anaphoric referent is ambiguous.
- **structural cohesiveness errors** by inserting a new sentence that is linked semantically to the general theme but provides information irrelevant to the topic.
- **internal consistency errors** by replacing a word by another that conveys a meaning opposite or incompatible with a prior sentence.
- **informational completeness errors** by omitting information about one or more items in a series that prior context suggested would be forthcoming.
- **syntactic errors** by distorting the grammar of a sentence by altering the word order.

In this way investigators can create a variety of types of errors depending on the needs of the specific research.

Readers display their ability to monitor comprehension by detecting the inconsistent information (Walczyk & Hall, 1989). They are asked to report, to underline errors, and to explain the nature of the errors and the reasons for their answers after reading texts silently or loudly (Anderson & Beal, 1995; Vosniadou et al., 1988; Walczyk & Hall, 1989; Zabrucky & Ratner, 1990, 1992). They display regulation by changing the errors and through indicative behaviours such as looking back at texts read earlier (Anderson & Beal, 1995; August et al., 1984; Baker & Anderson, 1982; Beal, 1987, 1990a, 1990b, 1996; Englert et al., 1988; Hacker et al., 1994; Paris & Myers, 1981; Zabrucky & Ratner, 1986, 1989, 1990, 1992).

Evaluation and regulation of errors are elicited through specifically worded questions such as "Does everything make sense? If something doesn't make sense, what change would you make? Make your change directly on the story." and globally worded questions such as "Did the story make sense?" (Englert et al., 1988; Zabrucky & Moore, 1989). Specifically worded questions are preferable (Zabrucky, 1986, 1989, 1990; Zabrucky & Moore, 1989); globally worded questions cannot reflect accurately comprehension monitoring because their interpretation varies with age and reading ability.
Error detection proceeds under two conditions: directed error detection when subjects are told in advance of the existence of errors in texts (Anderson & Beal, 1995; August et al., 1984; Baker, 1985a; Beal, 1990; Chan et al., 1987; Englert et al., 1988; Grabe & Mann, 1984; Hacker et al., 1994; Sousa & Oakhill, 1996; Vosniadou et al., 1988; Zabrucky et al., 1987, 1989) and spontaneous error detection when they are not so forewarned (Kinnunen & Vauras, 1995; Paris & Myers, 1981; Zabrucky & Ratner, 1989, 1990, 1992). It is assumed that the directed error detection condition is designed to assess evaluation ability (Paris & Myers, 1981; Sousa et al., 1996; Zabrucky, 1990). Thus, it is a better indicator of monitoring ability because some readers fail to monitor their comprehension spontaneously but can do it when directed (Paris & Myers, 1981).

Explanations for error detection failure
For both spontaneous and directed error detection conditions, failure to report errors occurs at all reading ability levels across a wide range of ages (Baker, 1985a; Pressley et al., 1992). Four types of explanations have been proposed to explain poor performance on these tasks:

2. Incorrect beliefs about the writer or the text: readers do not believe that a text can contain errors (Winograd & Johnson, 1982).
3. Self-predisposition to draw a conclusion: children can evaluate better others’ understanding than their own (Ruffman, 1996).
4. Difficulty in the relevant cognitive knowledge and cognitive processing, such as lack of detailed knowledge about the topic (Baker, 1985a; Beal, 1996; Ruffman, 1996; Winograd & Johnston, 1982); misinterpretation of the information (August et al., 1984; Ruffman, 1996; Winograd & Johnston, 1982) or absence of constructive processing (August et al., 1984; Baker, 1985a; Ruffman, 1996; Sousa & Oakhill, 1996; Vosniadou et al., 1988; Winograd & Johnston, 1982; Zabrucky & Moore, 1989).

Strengths and weaknesses of the error-detection technique
The error detection technique has been used widely in comprehension monitoring research. Its operationalised definition of comprehension monitoring allows ready measure of the two components of comprehension monitoring, evaluation and regulation, under a range of circumstances. It has been used with a wide range of ages of readers of different levels of intellectual or reading ability, for example, to examine developmental trends in comprehension monitoring (Anderson & Beal, 1995).

The validity of the technique in naturalistic education settings is questioned. The insertion of artificial errors in reading materials is difficult to implement in regular school reading contexts. Readers may not be concerned about whether or not their texts contain errors. Particularly when self-reading, readers aim to comprehend what
they are reading. They might be expected to deal with comprehension problems in their own ways.

The relationship between the spontaneous and directed error detection conditions requires clarification. These two types of evaluation are different. The directed error detection is postulated to test the ability to evaluate the current comprehension state (Paris & Myers, 1981; Sousa et al., 1996; Zabrucky, 1990). Probably for this reason, a variety of research has focused on the directed error detection condition (Anderson & Beal, 1995; August et al., 1984; Baker, 1985a; Beal, 1990; Chan et al., 1987; Englert et al., 1988; Grabe & Mann, 1984; Hacker et al., 1994; Sousa & Oakhill, 1996; Vosniadou et al., 1988; Zabrucky & Ratner, 1989; Zabrucky et al., 1987). Compared with the above studies, studies under the spontaneous error detection condition have been quite limited (Kinnunen & Vauras, 1995; Paris & Myers, 1981; Zabrucky & Ratner, 1989, 1990, 1992). Given this, it is not clear precisely what the spontaneous error detection is supposed to measure.

In the present authors' opinion, the spontaneous error detection rather than the directed error detection can reflect the ability to monitor comprehension, because it represents self-awareness of the comprehension state. On the contrary, the directed error detection directs readers to be conscious of their comprehension processes. What more errors are detected in this case is within expectation. As a metacognitive skill, monitoring comprehending should function spontaneously at least to a certain degree. Directed error detection can indicate monitoring ability. Still, it is not the better indicator of monitoring ability by comparison with spontaneous error detection.

THE PERFORMANCE JUDGMENT TECHNIQUE

The performance judgment technique is a subjective measure of comprehension monitoring (Wagoner, 1983). It includes two modules: prediction of future performance and judgment of past performance (Maki et al., 1994, 1995). Incorrect prediction and judgement are used to imply inappropriate awareness and evaluation of comprehension. Thus, they reflect the deficiency of comprehension monitoring (Schemer & Surber, 1986; Zabrucky et al., 1987).

Prediction of future performance
The performance prediction technique tests the ability to predict future performance. It has been used widely to examine the relation between self-assessment of comprehension and objective assessment of comprehension. The term "Calibration of Comprehension" has been used to refer to the subjective evaluation of comprehension in experimental studies (Glenberg et al., 1985, 1987; Maki et al., 1990, 1992; Morris, 1990, 1995; Weaver, 1990). It is defined as the correlation between subjective assessment of knowledge gained from reading (that is, readers' confidence in comprehension ratings) and their performance on objective comprehension tests (Glenberg et al., 1987; Morris, 1995).
The procedure of performance prediction
Readers read brief expository texts and use a many-valued confidence rating scale to rate their confidence in ability to answer inference comprehension questions about each text (Glenberg et al., 1985). They then answer these questions in the form of either multiple-choice or yes-no or true-false. The extent of correlation between predicted and actual performance reflects the level of comprehension monitoring (Pressley et al., 1990; Weaver et al., 1995). Assessment of comprehension monitoring can be enhanced by increasing discrimination of the level of comprehension, which can be reached by asking more questions about the topic (Weaver et al., 1995).

A brief summary of previous research
The extent of the relationship between predicted and actual comprehension performance for short expository texts varies between studies from little (Glenberg et al., 1985, 1987) to reliable correlations (Maki et al., 1990, 1992; Morris, 1990; Weave, 1990). Whatever the correlation, subjects ranging in age from young children or college students, are generally overconfident of their comprehension (Maki et al., 1984, 1990, 1992; Glenberg et al., 1985, 1987; Morris, 1995; Pressley et al., 1990, 1992; Weaver et al., 1995). They display low calibration of comprehension.

Explanations for poor performance on prediction tasks
The domain familiarity hypothesis was proposed to explain poor calibration (Glenberg et al., 1987). Comprehension prediction is based on prior knowledge or general familiarity with the topic of a text rather than on knowledge gleaned from reading a particular text. Domain familiarity determines confidence in comprehension. Low calibration of comprehension results from difficulty in assessment of knowledge gained from a particular text.

A second type of explanation derives from the tasks used (Weaver, 1990). The types of texts used (simple, single-paragraph expository texts) and very few items (inference questions) to assess comprehension lead to underestimation of true calibration. Thus, it is shortcomings in the measurement technique that result in the finding of poor calibration.

A third explanation focuses on the influence of availability of information. Morris (1990) did not think that domain familiarity was a major factor in prediction because of its weak influence on calibration. He reported no correlation between prediction and domain familiarity, while domain familiarity was correlated with test performance. He assumed that availability of information in memory affected both confidence ratings and test performance. Also, Morris (1995) rejected the methodological constraint explanation and claimed that poor discourse comprehension monitoring had real substantive causes. He found that subjects used low level information from the text to rate their confidence. Subjects had little access to knowledge and were not able to
distinguish irrelevant and relevant information. Thus, Morris (1995) concluded that both these factors had an influence on calibration.

Unlike Glenberg and Weaver, who concentrated on the role of either domain familiarity or measurement technique in low calibration, Maki and her colleagues (1992, 1995) took both these factors into account simultaneously. Domain familiarity is not the sole basis for performance prediction over text materials. More information about text materials, including the corresponding test questions, accompanies more accurate comprehension monitoring. They stress that sensitivity to the amount of information gained from reading the various texts plays a great part in subjects' comprehension confidence ranking.

**JUDGMENT OF PAST PERFORMANCE**

The technique of judgment of past performance tests the ability to assess past performance. As the other type of comprehension monitoring (Maki et al., 1990, 1994, 1995), the judgment of past performance is less employed to explore the complicated nature of comprehension monitoring by comparison with the prediction of future performance. Moreover, in available articles (Bouffard-Bouchard, 1994; Maki et al., 1990, 1994, 1995; Pressley et al., 1990; Schraw & Roedel, 1994), it is generally accompanied with the performance prediction.

**The procedure of performance judgment**

Readers are first asked to answer comprehension questions about texts they have read and then use a many-valued confidence rating scale to rate their certainty for each question that they answered correctly (Maki et al., 1990). Correlation between certainty ratings and the correctness of the responses of questions is calculated. A mismatch between certainty ratings and accurate answers indicates poor comprehension monitoring (Bouffard-Bouchard, 1994). It is assumed that certainty rating of comprehension constitutes a more sensitive indicator of basic comprehension monitoring ability (Reid, 1990).

**Explanations for performance judgment accuracy**

Findings about posttest confidence judgment are inconsistent. While some studies have shown that college students display accurate confidence judgments (Maki et al., 1994), others (for example, Bouffard-Bouchard, 1994) report that college students often have difficulty in accurately evaluating their comprehension. Confidence judgments depend on memories for specific details (Maki, 1995; Pressley et al., 1990). Task difficulty determines accuracy of confidence in comprehension; easier tasks yield more accurate confidence judgments than more difficult ones and there are few stable individual differences (Maki, 1995). In addition, some investigations report that confidence judgment accuracy is related to verbal comprehension ability and to speed of comprehension (Koriat, 1993; Maki et al., 1994). Inference ability, for example, influences comprehension monitoring measured by confidence judgment
Fengzhi Ling

Techniques for assessing comprehension monitoring

(Bouffard-Bouchard, 1994). Other investigators, for example, Pressley et al. (1988, 1990), report no relationship between posttest confidence judgment and comprehension ability.

**Strengths and weaknesses of the performance judgment technique**

A literature review of this technique by the present authors suggests a recent increase in interest in the performance judgment technique. It can be used easily in both ordinary reading contexts using intact texts and in contexts using the error detection technique. Subjects need to have acquired the capacity to engage in self-reflection. All studies reviewed so far using this technique have involved college students.

Of the two performance judgment procedures, the performance prediction has been used more frequently by itself as a comprehension monitoring procedure than the performance judgment. However, while the outcomes reported using this method are conflicting, the performance judgment technique has demonstrated relatively consistent findings. Additionally, the relationship between these two modules has not been explored.

The performance judgment technique by itself is unlikely to emerge as an effective method to measure comprehension monitoring, since it tests readers' subjective prediction and judgment of reading comprehension performance rather than how they control ongoing understanding or misunderstanding during reading. Children's evaluation of comprehension needs to be inferred indirectly from self-reports of understanding (Morre & Zabrucky, 1989). It is hard to determine whether the subjects' performance on this kind of assessment task can be attributed to their self-evaluation of their actual level of comprehension or to their self-judgment of comprehension ability. Reid (1990) employed the judgment of past performance to test children's ability to evaluate message quality, which, in turn, was questioned by Robinson and Whittaker (1985). In addition, this kind of subjective assessment would be constrained strongly by personal, task and context variables. Consequently, its consistency would be in serious doubt. As a vital component of reading comprehension (Paris & Myers, 1981; Pitts, 1983), independently of domain of reading materials (Schraw et al., 1995), the ability to monitor comprehending should to a great extent remain consistent, and the stable relationship between comprehension monitoring and reading ability should be reflected consistently as well. Nevertheless, this is not the case when comprehension monitoring is assessed by the performance judgment technique. While some research shows a high relationship between comprehension monitoring and reading ability, other research does not (Maki et al., 1994).
THE THINK-ALOUD TECHNIQUE

The thinking aloud technique was developed as a procedure to investigate on-line cognitive processing. It has been used to explore the nature of reading, reading comprehension processes, reading strategies use and reading comprehension instruction (Kucan and Beck, 1997; Thoreson et al., 1997). In the early 1980s, it was used to deal both with problem detection and use of strategies in the area of comprehension monitoring (Kucan & Beck, 1997; Wagoner, 1983). The rationale for using it in comprehension monitoring research is that thinking aloud during reading not only reflects comprehension monitoring but also represents a form of comprehension monitoring itself (Baumann et al., 1993; Garner, 1988).

The procedure of the think-aloud paradigm
"Think-aloud procedures produce concurrent verbalisation about an activity that is temporarily interrupted for provision of the verbal report" (Garner, 1988, p. 65). Readers cease reading periodically to say aloud what they are processing, what they understand and the reading strategies they are using (Baumann et al., 1993). The instruction such as "Tell me what you are thinking and doing while you read this article" elicits this process. The verbalisation, typically recorded on audiotape for analysis, require transcribing and categorizing. Irrelevant information in protocols is discarded (Garner, 1988). The think-aloud paradigm is also named a self-report paradigm on the grounds of its self-processing verbalization (Wagoner, 1983).

A brief summary of previous research
Thinking aloud instruction helps students learn strategies to enhance their understanding of a text and to deal with comprehension difficulties. Baumann et al. (1993) used this paradigm to develop an instructional program to help students learn to monitor their reading comprehension and to use various strategies to deal with comprehension breakdowns. The present authors were unable to locate research that has used thinking aloud directly to test comprehension monitoring ability.

Strengths and weaknesses of the think-aloud technique
The think-aloud technique is a valuable tool for investigating how readers comprehend what they are reading (Trabasso & Magliano, 1996; Whitney & Budd, 1996). It can be used flexibly to examine a range of research questions relating to reading (Baumann et al., 1993; Casanave, 1988; Trabasso & Magliano, 1996; Whitney & Budd, 1996). It can provide information about comprehension monitoring processing that is difficult to obtain using other procedures (Crutcher, 1994; Wagoner, 1983).

Despite its value, the think-aloud technique has been used in comparatively few studies to examine comprehension monitoring during reading, perhaps because of its dependence on highly reliable transcription and categorization of verbal data and its demand on time investment. Most criticism, however, has focused on its validity and
reliability. Although some researchers claim that its validity has been upheld (Kucan & Beck, 1997; Payne, 1994; Wilson, 1994), others (Wilson, 1994) question the usefulness of the data collected. Thinking aloud demonstrates the norms for behaviour rather than a veritable picture of the underlying processes employed in a task. Meanwhile, a task itself and the way to collect verbalization also affect its validity. Some tasks which involve higher level cognitive processes and verbal types of information, are better suitable to thinking aloud than other tasks (Payne, 1994). Performance suffers when the task involved cannot be easily verbalized (Wilson, 1994).

The reliability of thinking aloud has also been questioned. Doubts are based on two main aspects. Think-aloud protocols may cannot exactly reflect what is actually thought about during understanding (Trabasso & Magliano, 1996). During thinking aloud, readers may use constructed text representation to "tell a story" about their understanding (Long & Bourg, 1996). Moreover, what students report may depend on their linguistic proficiency. Some students may report more than they really know (Garner, 1988; Wagoner, 1983), while others, with limited language skills, may know more than they can tell (Garner, 1988). In these circumstances, thinking aloud either overestimates or underestimates individuals' actual thinking ability. As well, because think-aloud procedures use at least some of the cognitive resources (Payne, 1994), it may have a negative influence on concurrent cognitive activity (Wilson, 1994). For some readers, this attention distribution may delay or stop their normal thought flow.

CONCLUSION

Comprehension monitoring is the ability to detect and respond to breakdowns in one's understanding of language (Markman, 1981). The four experimental techniques examined in this paper differ on five main attributes: the nature of the method (subjective vs objective method); the function (components of comprehension monitoring measured); the format (multiform vs unitary form); operationality (difficulty vs easiness); and the application of the technique (wide vs narrow range). They have contributed to our understanding of comprehension monitoring.

The more objective methodologies, the referential communication and error detection paradigms focus on the evaluation and regulation of message quality, while the performance judgment paradigm evaluate subjectively performance on the message. Theoretically, the think-aloud paradigm is the most effective way to measure comprehension monitoring, because it provides both message evaluation and performance evaluation simultaneously, explicitly and directly.

The referential communication paradigm has stimulated extensive research on comprehension monitoring. It is the best way to explore how young individuals develop their monitoring ability. Its procedure is similar to processing during problem-solving. By analysing how individuals deal with items that have referential ambiguity,
researchers can identify the strategies used to cope with the ambiguity spontaneously and how personal and task variables influence regulation strategies. With resolution of the limitations noted for this technique, this paradigm might be used more widely in the future to examine comprehension monitoring.

The error detection paradigm is an effective technique for studying comprehension monitoring in experimental contexts. Nonetheless, an objective way to examine comprehension monitoring under a realistic reading setting or in the classroom context is really needed to help both teachers and students to beware of and assess their comprehending processes. As discussed before, the error detection technique builds upon the operational definition of comprehension monitoring. This specific definition does not suit ordinary reading. We need to create one plausible definition for direct assessment of comprehension monitoring during reading in normal conditions. Such definition should reflect the nature of comprehension monitoring.

Unlike the referential communication paradigm and the error detection paradigm, which test directly ongoing comprehension processing under some conditions, the performance judgment paradigm explores comprehension monitoring from the perspective of self-awareness of comprehension processes. It reflects pre-awareness and post-awareness of comprehension performance. Although the conceptual framework of comprehension monitoring adopted in this article does not mention explicitly that awareness is one component of comprehension monitoring, it should be accepted that awareness logically presupposes evaluation and regulation, namely, two stressed components of comprehension monitoring. Wilson's metacognition model (1997) concerns the value of awareness in metacognitive processes. If we focus on the whole process of comprehension monitoring, that is, awareness, evaluation and regulation of comprehension processing, we can see clearly that the performance judgment paradigm plays a role in the study of comprehension monitoring. Its validity as an effective technique to measure comprehension monitoring could be more closely examined if it were accompanied by the error detection paradigm simultaneously in one study.

It seems that no other paradigms have been accompanied by continuous debate more than the think-aloud paradigm since its birth. However, the importance of thinking aloud in cognitive science definitely cannot be denied. It is argued that thinking aloud is an excellent methodology to explore conscious, easily verbalizable cognitive processes (Trabasso & Magliano, 1996; Wilson, 1994). One question pertinent to this argument is what standard can be set to discriminate easily verbalizable cognitive processes from difficultly verbalizable ones. In comparison with other techniques mentioned in this article, the think-aloud paradigm has been seldom used to explore the nature of comprehension monitoring. Given its strengths and weaknesses, it would be better used in combination with other methods (Long & Bourg, 1996; Wagoner, 1983; Whitney & Budd, 1996; Wilson, 1994). This implies the possibility of mixing thinking aloud with the other three techniques or other
different methods whenever suitable to assess comprehension monitoring. However, it is suggested that thinking aloud would be secondary to the performance of the task at hand (Payne, 1994). It is performance of the task not giving a high rate verbalization that remains primary in a study.

Each technique has its advantages and disadvantages. Flexible combination of a variety of different methods or techniques whenever possible are recommended to help us understand how comprehension monitoring is accomplished (Payne, 1994; Whitney & Budd, 1996). The invention of new tools and methodologies for studying comprehension monitoring will inevitably lead to great progress in Cognitive Psychology, Reading Psychology, Educational Psychology and other relevant areas.

REFERENCES


