DOCUMENT RESUME

ED 303 767 CS 009 502

AUTHOR Phillips, Linda M.
TITLE Developing and Validating Assessments of Inference Ability in Reading Comprehension. Technical Report No. 452.
INSTITUTION Bolt, Beranek and Newman, Inc., Cambridge, Mass.; Illinois Univ., Urbana. Center for the Study of Reading.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Feb 89
CONTRACT OEG-0087-C1001
NOTE 61p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Foreign Countries; Grade 6; Grade 7; Grade 8; *Inferences; Intermediate Grades; Junior High Schools; Middle Schools; Models; Protocol Analysis; *Reading Comprehension; Reading Skills; *Reading Tests; Theory Practice Relationship
IDENTIFIERS Canada

ABSTRACT

This report describes the development and validation of the Test of Inference Ability in Reading Comprehension: a scaled-answer, multiple-choice test intended for use in Grades 6, 7, and 8. The report discusses the need for and conceptualization of assessment of inference ability; proposes standards and principles of inference appraisal; and discusses test design issues, specifically audience, kinds of discourse, topic familiarity, readability, test format, test length, and passage and item development. Five pilot studies are then presented to show test evolution, providing details of the modifications at each phase of test development. The fifth pilot study (focusing on test validation) is discussed, involving verbal reports of students' thinking as they worked through the test, and providing a reading score for the answer selected and a corresponding thinking score for a student's explanation of why that answer was chosen. Results of this fifth pilot study indicate that for 94% of the items good thinking was significantly correlated with good inference-making and poor thinking to poor inference-making. The report concludes by presenting the final data collection, analyses, results, and directions for future research. (Thirteen tables of data are included, and 98 references are attached. Two appendixes providing the reading and thinking rating scales for the test conclude the report.) (SR)

CENTER FOR THE STUDY OF READING

Technical Report No. 452

DEVELOPING AND VALIDATING ASSESSMENTS OF INFERENCE ABILITY IN READING COMPREHENSION

Linda M. Phillips
Institute for Educational Research and Development
Memorial University of Newfoundland
and
Center for the Study of Reading
University of Illinois at Urbana-Champaign

February 1989

University of Illinois at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

The work upon which this publication was based was supported by a grant from the Social Sciences and Humanities Research Council of Canada. This report was prepared while the author was on sabbatical leave as a Visiting Scholar at the Center for the Study of Reading. The views expressed herein are those of the author and not the funding agency.

Abstract

This report describes the development and validation of assessments of inference ability in reading comprehension in the middle grades. The need for and conceptualization of assessments of inference ability are raised and discussed within a contemporary theoretical framework. Standards and principles of inference appraisal are proposed. The specifications of the assessment address issues such as audience, kinds of discourse, topic familiarity, readability, test format, test length, and passage and item development. The 5 pilot studies are presented to show test evolution. The fifth pilot study focussed on the test validation. Verbal reports of students' thinking as they worked through the test were collected to provide an independent index of the quality of their thinking as they selected their responses. Thus, for each item there was a reading score for the answer selected and a corresponding thinking score for a student's explanation of why that answer was chosen. For 94% of the items good thinking was significantly correlated with good inference-making and poor thinking to poor inference-making. The final data collection, analyses and results are presented followed by a discussion of directions for future research.

DEVELOPING AND VALIDATING ASSESSMENTS OF INFERENCE ABILITY IN READING COMPREHENSION

This report describes a model for the development and validation of a test of inference ability in reading comprehension. The assumption that the ability to make inferences is necessary to reading comprehension is widely accepted by reading theorists and researchers. This observation, coupled with the fact that no satisfactory procedures exist for determining the extent to which children make good inferences when attempting to understand text, motivated this project. This report is built around four sections. The first section will offer a contemporary theoretical framework for a test of inference ability in reading comprehension. The design, item development, and test development iterations are described in the three subsequent sections.

The Need for and Conceptualization of a Test of Inference Ability in Reading Comprehension

Evidence abounds to suggest that poor reasoning is prevalent in our students. The National Assessment of Educational Progress (1984) reported large decreases in inferencing responses of 13- and 17-year-olds over a 10-year period. Furthermore, the Nation's Report Card (Applebee, Langer, & Mullis, 1987) reports that only small percentages of students can reason effectively as they read and write. Such findings suggest that students are not being guided to perform reasoning activities which require analysis and interpretation.

These findings are not surprising when coupled with research on teaching practices. In the late 1970's Dolores Durkin (1978-79) reported that schools do not teach comprehension, and precious little time was devoted to having students explain and substantiate their interpretations. The authors of the Report of the Commission on Reading (Anderson, Hiebert, Scott, & Wilkinson, 1985) lamented that there is very little direct comprehension instruction in most American classrooms. The increased evidence of poor reasoning has led to claims about deficiencies in school programs and to appeals for action, such as more testing. Yet, most available tests are general and vague about the nature of reading comprehension and do not support instructional improvement. It would make sense that a test designed to specifically measure inference ability and to support instructional improvement would be an important place to start.

In my study of the literature on testing for reading comprehension, and particularly on testing for inference ability, standardized tests of reading were found problematic because it is not easy to make any decision as to what the tests measure. To highlight some of the problems, it is necessary to recognize at the start that prominent researchers in the field declare that reading assessment bears a number of substantial flaws. To elaborate on a couple of pieces of research would serve to illustrate some of the flaws.

Tuinman (1973-74) investigated five widely used standardized tests according to the extent to which questions on the tests could be answered without reading the passages upon which those questions were based. Two points emerged. First, although Tuinman was cautious about his conclusions he found significant reason to believe that it was general knowledge and not reading comprehension which was being measured. He suggested that more exploration was needed to discover the extent of this test invalidity. Second, these five major tests did not provide any technical information on the extent to which items could be answered on the basis of information other than that present in the passage. That is, the *major* tests have failed to address this significant construct validity issue.

The question of validity is a major concern because the predominant approach to construct validity in standardized reading tests, that of correlations with other tests, has a significant weakness. It is based on rather circular reasoning because two tests may be *intended* to measure the same ability, *be* highly correlated, and still *fail* to measure what they were intended to measure. That is, they can possess what

Embretson (Whitely) (1983) has called "nomothetic span," but still fail to be representative of the construct they are proposed to measure.

Many of the problems with standardized reading tests remain unresolved and by continuing to use poorly-produced tests we are not recognizing the concerns raised. Anderson (1972) reported that educational researchers have not yet learned to develop achievement tests that meet the primitive first requirement for a system of measurement, namely that there is a clear and consistent definition of the things being counted. A search for definitions of reading comprehension in the major established reading tests (many of which were developed prior to the 1970's) reveals that for all intents and purposes none exist. Generally, each of the manuals says nothing more than that the test items measure specific skills related to the comprehension of what is stated explicitly in the material, the judgment of what is implied, and the drawing of inferences with reference to other situations. The tests do not identify which specific skills are being measured by particular items, nor do the tests report separate scores for specific skills. Rather, a composite score of comprehension, vaguely defined, is reported.

Much work is required to give rise to orderly, sensible data in the evaluation of reading comprehension in general and inference ability in particular. It is my opinion that as a start much can be learned and applied to reading from researchers in the field of critical thinking, who have attempted to elaborate and clearly define the nature of inference (Govier, 1985; Hitchcock, 1983; Norris, 1984; Salmon, 1984; Scriven, 1976). In particular, I refer to the extensive work of Robert Ennis (1962, 1969, 1981, 1985). His efforts in characterizing rational thinkers has been, and continues to be an invaluable source of ideas for the project, as were available critical thinking tests (Ennis & Millman, 1985; Watson & Glaser, 1980) and a test of induction currently under development (Norris & Ryan, 1987) which include sections testing for inference-making ability.

Defining Reading Comprehension

As might be expected, reading comprehension has been defined in many ways. Articles and books (Anderson & Pearson, 1984; Carroll, 1964; Farnham, 1905; Goodman, 1968; Gray, 1940; Huey, 1908; Plato, 1917; Richards, 1938; Thorndike, 1917) have been written on and about reading comprehension. While each article and book contributes to a more thorough understanding of reading comprehension, each is incomplete. It is beyond the province of this project to attempt to provide a complete and comprehensive definition of such a complex act as reading comprehension. For the purposes of this project, I preferred to remain tentative and prepared to alter my working definition of reading comprehension with subsequent information.

Reading comprehension is believed to be a collection of processes such as predicting, inferring, synthesizing, generalizing, and monitoring, which have been identified and labelled in various ways by different writers in the field (Collins, Brown, & Larkin, 1980; Fagan, 1987; Henry, 1974; Phillips, 1988; Smith, 1971). It is widely accepted that reading comprehension involves more than knowing the correct pronunciation of the words, knowing their individual meanings, and being able to locate information in printed material (Norris & Phillips, 1987; Phillips & Norris, 1987; Spiro, 1977; Tuinman, 1986). Current reading theory defines reading comprehension, more or less, as meaning constructed by a reader through strategic and principled integration of the textual information and background knowledge.

Since an explanation of the intricacies of reading comprehension remains elusive, and since it is agreed that reading comprehension is a complex behavior which continues to be perplexing, then one cannot set about assessing it in its entirety. Thus, it seems to make sense to study specific aspects of the process as a means of seeking advancements in the assessment of the complete process of reading comprehension. It is with the process of inference as an aspect of reading comprehension that I am most concerned.

At a general level, inference is a cognitive process used to construct meaning. Inferring in reading comprehension is a constructive thinking process, because a reader expands knowledge by proposing and evaluating hypotheses about the meaning of text.

Good inference-making in reading comprehension requires the thoughtful use of strategies (Collins, Brown, & Larkin, 1980; Phillips, 1985, 1987, 1988; van Dijk & Kintsch, 1983) and evaluative criteria. Inferences in reading comprehension tend to be good to the extent that readers integrate relevant text information and background knowledge to construct complete interpretations that are consistent with both the text information and background knowledge.

At a specific level, inference requires intelligent human judgment (Ennis, 1973), and necessitates the use of relevant text information and background knowledge. This dependence on background knowledge is important for at least three reasons. First, an inference in reading comprehension is the interaction of relevant information provided in the text and background knowledge. In other words, neither textual information nor background knowledge alone is sufficient to make good inferences. Second, background knowledge enables the generation of alternative hypotheses in inferring. Inference is the basis of understanding which often involves transforming, extending, and relating information (Markman, 1981). Third, without background knowledge one cannot evaluate the strength of inferences to generalizations and explanations (Govier, 1985), thereby making background knowledge a necessary part of inferential reasoning.

The Objectives of the Test

Having implied the complexity of the reading comprehension process in general, and having described the process of inference as one aspect of comprehension in particular, it is important to reiterate that comprehension is a complicated cognitive process. Indeed, there may be considerable overlap and interdependence among inferring and the other comprehension processes of attending, analyzing, associating, predicting, synthesizing, generalizing, and monitoring. A general test of comprehension ability would focus on each of the processes, whereas the Test of Inference Ability in Reading Comprehension (TIA) (Phillips & Patterson, 1987)[1] focuses specifically on the process of inference-making.

TIA is designed to appraise the inference ability of middle grade students on the basis of full length passages representative of the three kinds of discourse commonly found at the middle grade levels and of topics characteristic of classroom reading materials. TIA is designed to inform teachers about students' inference ability and to provide diagnostic information for instructional decision-making purposes.

A Principle of Inference Appraisal

The general guideline of ability test validation that directed this research was that the test would be valid to the extent that good inference-making leads to good performance on the test and that poor inference-making leads to poor performance. To be in a position to distinguish good inference-making from poor inference-making implies that there must be standards for making such distinctions. Reading educators should not accept just any inference merely because it reflects some level of the reader's cognitive competence. When we judge someone's inference to be normatively good we are comparing it to what we take to be some standard of expert competence. So, it is important that the best interpretations are inferences in accord with the best available principles. To be in a position to improve reasoning means to be in a position to distinguish good reasoning from bad. To do so, implies that there must be principles and standards.

To apply a set of standards to the quality of inference-making in reading comprehension certain assumptions about the reader, the task, and the text must be made in order to get off the ground. These presuppositions or necessary conditions are stated as follows:

I. A reader must:

1. be competent with the difficulty level of the text,
2. understand the demands of the task; and
3. intend to understand the text.

II. A text must:

1. be written coherently;
2. adhere to conventions of communication by being:

   a. as informative as is required for the situation;
   b. accurate or complete with adequate evidence for asserted information;
   c. relevant to the ongoing situation; and
   d. unambiguous and clear.

If these conditions are not met, poor performance on the inference task may be explained through failure to satisfy one, several, or all of these conditions, rather than as a lack of inference ability.

The satisfaction of Condition I and II is necessary for the application of the following principle of inference appraisal to judgments of readers' inference ability in reading comprehension:

> Inferences in reading comprehension tend to be good to the extent that a reader integrates relevant text information and relevant background knowledge to construct interpretations that more completely and more consistently explain the meaning of the text than alternative interpretations.

Completeness and consistency are thus the two criteria for judging interpretations. Neither criterion by itself is sufficient; they must be used in tandem. The criteria must also be used comparatively in situations where there are competing interpretations. We must ask which interpretation is more complete, and more consistent, because often neither interpretation will be fully complete and fully consistent (Norris & Phillips, 1987; Phillips & Norris, 1987). Thus, the expression "tends to be good to the extent that" is an important part of the principle. The expression is a qualifier which signifies the limitations of the principle and emphasizes that it is not an absolute principle.

Justification of the Principle

The work of researchers in four fields provides evidence for both the derivation and justification of the principle of inference appraisal presented in this research: critical thinking (Ennis, 1969, 1981, 1985); philosophy and philosophy of science (Harman, 1986; Thagard, 1978, 1982), cognitive psychology (Holland, Holyoak, Nisbett, & Thagard, 1986; McCloskey, 1983; Nisbett & Ross, 1980; Stich & Nisbett, 1980) and reading (Collins, Brown & Larkin, 1980; Markman, 1981; Mason, 1984; Norris & Phillips, 1987; Phillips & Norris, 1987).

The most extensive work done on inference criteria known to me is that of Robert Ennis (1969, 1981). He uses the expression "material inferences" and divides these into two categories: those which generalize the evidence which is offered, and those which derive their support from their power to explain the evidence. The latter category is most representative of the kinds of inferences made in reading comprehension and is thus the focus of this discussion. Ennis presents criteria for judging inferences to explanations. The inferences are justified to the extent that: (a) they explain a bulk and variety of reliable data; (b) they are themselves explained by a satisfactory system of knowledge; (c) they *are not* inconsistent with evidence; (d) their competitors *are* inconsistent with evidence; and (e) they are simpler than their competitors.

The first criterion is covered in the principle of inference appraisal, where it is stated that a reader integrates relevant text information and relevant background knowledge to construct interpretations that are complete, that is, interpretations that explain all relevant information. Ennis' second criterion (inferences are themselves explained by a satisfactory system of knowledge) and his third (inferences are not inconsistent with available evidence) are incorporated into the principle of inference appraisal where it says interpretations that more completely and more consistently explain the meaning of the text than alternative interpretations. Competing interpretations that are inconsistent with available evidence would be judged to be poor, given the principle of inference as it is stated, thereby automatically incorporating Ennis' criterion 4 into the principle. Ennis' fifth criterion (inferences are justified to the extent that they are simpler than their competitors) is embedded in that part of the principle where it is stated that a reader integrates relevant text information and relevant background knowledge. Irrelevant information can lead to a convoluted interpretation, rather than a straightforward one based on relevant information. (For a detailed discussion of these ideas, see Norris and Phillips, 1987.)

A second source of support for the principle derives from research on failures in everyday reasoning. According to Nisbett and Ross (1980), shortcomings in human inference-making reflect peoples' failure to use normative principles and, instead, to apply simplistic inferential strategies beyond their appropriate limits. They caution that human inference is prone to several major sources of error including, to mention two, an over-reliance on just the information which happens to be available, and an inappropriate weighing of data relevance. Evidence of these two errors has particular bearing for a principle of inference appraisal in reading comprehension. In the case of the first error, readers often place greater reliance on the text information. In the second case, readers may place too great a reliance on some of the textual information or on their background knowledge, thereby failing to properly integrate relevant information from both. The point is that any principle of inference appraisal in reading comprehension must emphasize the necessity of using *both* relevant text information and relevant background knowledge and of properly weighing the relevance of each.

Nisbett and Ross (1980) also present evidence that more vivid or salient information is more likely to enter inferential processes than is less vivid information. Salient information may influence unduly a person's inference-making. Other research has illustrated the tendency for ideas, once formulated or adopted, to persist despite evidence which might be disconfirmatory (Holland, Holyoak, Nisbett, & Thagard, 1986; McCloskey, 1983). It seems people will point to scant positive evidence to sustain their original interpretation, even though substantial negative evidence exists to suggest otherwise. Thus, when some people read and are faced with counterevidence, they will either tend to ignore or misconstrue the evidence to advantage (Phillips, 1987). It seems that a workable principle of inference appraisal must provide a standard against which readers can monitor whether their interpretations are the best explanations; that is, are more consistent and more complete than alternative explanations, or are unduly influenced by one or all of the factors mentioned above.

A third source of support for the principle of inference appraisal is garnered from work in the philosophy of science on inference to the best explanation. Inference to the best explanation entails accepting an hypothesis on the grounds that it provides a better explanation of the evidence than is provided by alternative hypotheses. Three important criteria are proposed by Thagard (1978) for determining the best explanation: consilience, simplicity, and analogy. An explanation is more *consilient* than another if it explains more of the evidence than the other by unifying and systematizing the information, while at the same time being informative. A *simple*, consilient explanation not only explains all that is necessary, but does so without making a host of assumptions with narrow application, merely derived for the moment. The first two of Thagard's criteria, consilience and simplicity, thus offer support for the standards of "completeness" and "consistency" defined in the principle of inference appraisal. *Analogy* also plays a vital role in good inference-making in that it *supports* the posited hypothesis by improving the explanation that the hypothesis is used to give.

Another source of support for the principle of inference appraisal rests within the reading field. Ellen Markman (1981), in her work on comprehension monitoring, acknowledged that distinguishing a good inference from a poor one is complex and closely tied to distinguishing better explanations or better theories. She posits the question of how readers decide whether or not they have understood. Markman shows how theories of comprehension inform theories of comprehension monitoring by describing two fundamental aspects of comprehension. She argues that well organized or tightly structured information is essential to reading comprehension, that comprehension often promotes the making of inferences, and that the two are interrelated. I propose the following points based on Markman's work on comprehension monitoring: (a) good inferences are highly constrained by the context (text and background knowledge); (b) good inferences are based on warranted assumptions and are progressive in that they subsume previous ideas from the context; (c) good inferences are judgments confirmed by subsequent information from the context; and (d) good inferences are judgments having elegance and parsimony within the context. The constraints imposed by context (text and background knowledge), in the four points above, are embedded in the principle of inference appraisal, thereby indicating that context both provides the subject matter (relevant text information and relevant background knowledge) as well as the parameters (. . . to construct alternative interpretations) of the interpretation. Warranted assumptions (point b), and inferences that have elegance and parsimony (point d) are integrated into the principle of inference appraisal in reading comprehension through use of the words "complete" and "consistent" as discussed earlier in this section.

A further elaboration and confirmation of the above four points is found in the work of Collins, Brown, and Larkin (1980), where adult subjects applied at least four different tests in evaluating the plausibility of the interpretations they constructed. The four tests include: (a) the plausibility of the assumptions and consequences of the model (when a default assumption or a consequence of the interpretation seems implausible, then subjects tend to reject the interpretation); (b) the completeness of the model (interpretations are evaluated in terms of how well the assumptions and consequences of the model answer all the different questions that arise); (c) the interconnectedness of the model (the assumptions or consequences of an interpretation are weighted with respect to how they fit together with other aspects of the model); and (d) the match of the model to the text (occasionally, readers seem to weigh the model in terms of how well its assumptions or consequences match certain surface aspects of the text). Within Collins, Brown, and Larkin's model the integration of text information and background knowledge in the construction of interpretations is explicitly stated, as well as criteria used by adults to test the "fit" of their interpretations.

Section Summary

The principle of inference appraisal proposed is representative of what is currently known about inference and provides a framework within which to better understand the process of inference-making in reading comprehension. The principle is intended to be neither canonical nor comprehensive, but rather to be an advance toward a set of principles. The principle of inference appraisal must be considered tentative and alterable in the light of both further understanding and empirical evidence. However, as shown, it is supported by researchers in the critical thinking, philosophy and philosophy of science, cognitive psychology, and reading fields. There is a remarkable compatibility and overlap in the work, as can be seen by the notions of completeness, consistency, and clarity which are all considered to be criteria of sound inferences.

Specifications of the Test of Inference Ability in Reading Comprehension

It is difficult to separate the design and development of a test; however, since somewhat distinct decisions were made about each, I have opted to devote a separate section to each. This section will provide the specifications on the design of TIA.

Test Development Framework

Audience

The intended audience for TIA is students in the middle grades. Students in Grades 6, 7, and 8 were selected for both theoretical and practical reasons.

Some of the basic tenets of reading development guided my theoretical decisions. While it is generally agreed that reading development is continuous, it is also agreed that there are stages of development. By the time students have advanced to the middle grades, they have read graded materials, content area subjects, and have generally achieved some degree of independence in the reading process. These facts make it more manageable to separate out inference ability problems, should they exist, from other problems such as vocabulary, syntax, or other failures (Anderson & Pearson, 1984; Gentner, 1983; Vosniadou & Ortony, 1983).

There is research which suggests that there are developmental differences in story comprehension (McConaughy, 1980). It seems that even grade 5 students focus more on the literal aspects than on the interpretive aspects. Another reason for selecting middle grade students is they have had some instruction in making inferences.

Kinds of Discourse

Discourse is typically classified as one of four types: (a) exposition, (b) narration, (c) description, and (d) argument (Bock & Brewer, 1985; Brewer, 1980; Spiro & Taylor, 1987). *Exposition* answers real or imaginary questions. It presents facts or explains why something is important, how something works or what a thing means. *Narration* informs readers of what is happening; it is an account of events or action and includes characters, plot, theme, and style. *Description* is a discourse form used to appeal to the senses of the reader and is generally about the appearance of an object, a person, or an event. *Argument* is a form of discourse in which there is an attempt to convince or persuade through appeals to reason, emotions, or to both. Exposition, narration, description, and argument often overlap so this global classification omits much of the complexity of discourse. In practice, clear-cut classifications are not always possible.

Three of the four kinds of discourse are more familiar to students in the middle grades; argument is less familiar. The three full-length stories[2] in TIA represent the common discourse forms found at the middle grade levels, thereby providing a more thorough appraisal of students' inference ability across a variety of reading materials than tests which assess performance on either isolated passages and questions or on one discourse form. Narrative, expository, and descriptive texts make distinct demands upon readers, readers' knowledge, and expectations about a task and have important consequences for cognitive processing and learning (Anderson & Armbruster, 1984; Brewer, 1980; Spiro & Taylor, 1987). For instance, narrative text is often argued to be easier to understand than expository text for both adults and children (Bereiter & Scardamalia, 1982; Bock & Brewer, 1985) possibly because readers are less familiar with how expository texts are organized. Since the three discourse forms are an integral part of programs which students at the middle grade levels are expected to learn, then differences in comprehensibility between narrative, descriptive, and expository texts must be taken into account for diagnostic purposes.

Topic Familiarity

The role played by background knowledge in reading comprehension has attained such widespread acceptance that it no longer requires a justification. The prior or background knowledge that a person brings to a text is said to be one of the most important factors in understanding, remembering, and interpreting text information (Anderson, Spiro, & Anderson, 1978; Ausubel, 1963; Holmes, 1983; Johnston, 1984; Pearson, Hansen & Gordon, 1979). Furthermore, while topic familiarity or possession

of requisite domain knowledge does not necessarily guarantee interest, it does affect the readability and comprehension of text (Phillips, 1987; Walker, 1987). Topic familiarity is seen to be necessary, but not sufficient for comprehension.

Background knowledge alone is not sufficient for reading comprehension because a reader must know *how* to use that knowledge and *want* to use that knowledge. This is a particularly relevant point in the appraisal of inference ability because a reader must seek a complete interpretation that is consistent with both the text information and background knowledge in order to make good inferences. Since readers must integrate background knowledge with the text information to infer, then to try to separate background knowledge from the text information is to deny the role of background knowledge in reading comprehension, and of course, once you read something, it becomes part of your background knowledge. It is not clear what readers' performance on such a test would mean. Therefore, it is necessary to assess what prior knowledge students have in order to make accurate appraisals of their inference ability in reading comprehension.

Furthermore, since it is an objective of TIA to serve as a diagnostic tool, then it must be realized that readers do not always integrate completely text information and background knowledge. Sometimes readers integrate only some of the relevant text information and background knowledge; other times, readers will select relevant text information and background knowledge, but fail to integrate the two. There are occasions where readers fail to do any of the above and as a consequence fail to make an inference, go off course in their interpretation, or make unwarranted assumptions.

A multiplicity of approaches were undertaken during the development of TIA to establish accurate estimates of middle grade readers' topic familiarity. A more thorough discussion of the procedures will be presented in a subsequent section, "Selection of Topics." At this point, it is sufficient to say that several groups were involved in determining topic familiarity and interest: (a) a group of graduate students were asked to list 10 topics which they felt their students were interested in; (b) 130 teachers attending a workshop were asked to list 10 topics that they felt their students were interested in; (c) 300 middle grade students were asked to list 10 topics that they thought they could write about without difficulty; and (d) 12 middle grade classes were selected to discuss some of the topics and to write about them. From these information sources three topics were seen to be common areas of interest and within the background knowledge of the intended audience of the TIA test.

Readability

The readability of text is generally assumed to refer to its legibility, ease of reading, and ease of understanding. Many readability formulae have been developed over the years, but they have not been without criticism. Traditional readability formulae have been criticized for having no point of reference (Manzo, 1970), for neglecting the importance of the structure, texture, and informational density of text (Amiran & Jones, 1982), and for lacking face validity (Coupland, 1978).

Alternative ways of estimating readability have been proposed, including the subjective text difficulty approach by Tamor (1981) and others (Carver, 1975-76; Singer, 1975), the psycholinguistic approach by Holland (1981), and the conceptual approach by Rubin (1981). Tamor combines text-based information (readability estimates) and performance-based information (recall scores) to come up with a subjective text difficulty level for individual readers. Holland's psycholinguistic alternative focuses on assessing the meaning-making demands placed upon readers by the language and structure of the text. Rubin's conceptual approach focuses on the concepts conveyed by the text: how arguments are presented, the role of examples in a text, and how characters' interactions are developed and described.

I weighed and balanced the available evidence and decided not to use a readability formula in the traditional sense. Rather, I chose to use what may be described as a composite of both the traditional and alternate approaches to readability. I chose to adhere to the principles of good story writing and to ask students who piloted the test to assess the readability of the stories.

The TIA stories were wri. ı on three topics identified to be familiar to middle grade students. In the initial pilot studies, students were asked to read the stories aloud and to point out areas of difficulty. When the areas identified to be problematic were revised, the text was judged to be appropriate for the intended audience. In accord with Conditions I and II given in the previous section, it was important that the stories and inference questions be written coherently and adhere to conventions of good communication; it was also essential that a reader be competent with the difficulty level of the text, understand the demands of the task, and intend to understand the text. Otherwise, readers' poor performance on the inference task may be accounted for by a failure to meet these conditions, rather than as a lack of inference ability.

Test Format

TIA contains three full-length stories (average of 465 words per story): a narration, an exposition, and a description. Stories consist of four to five paragraphs and 12 scaled-answer, multiple-choice questions. Questions follow each paragraph in the stories. Each question has four answers provided. To answer the questions students are to use information given in the story and information that could only come from their background knowledge. Students are given an example which is thoroughly worked through so that they will see that they are to consider all possible answers before deciding which answer they think is the "best" one.

The challenge in changing reading assessment is to come up with new means to evaluate our current conceptualization of reading and to diagnose areas where instruction is needed. Reading comprehension admits of degrees. However, credit has generally been given on most tests of reading comprehension for one and only one correct answer. There has been no allowance for partially correct responses, that is, for evidence that a student may be capable of selecting relevant information without quite knowing what to do with it.

The challenge in the design of TIA was to provide diagnostic information about students' performance and to use that information to support instructional improvement. To achieve this end, TIA represents a creative melding of the old and new. The old format of selecting an answer is there with the new advantage of giving credit for answers that are not completely correct. TIA may be described as a "scaled-answer multiple-choice" test, since it attempts to account for variations in understanding. The four answer choices represent a range in values (0-3) assigned according to the quality of the answer selected. An answer that is consistent with both the text information and background knowledge is worth 3 points; a partially-correct answer is worth 2 points; a text-based answer is worth 1 point; and an erroneous answer is worth 0 points. (A complete copy of the scoring guide is provided in Appendix A.)

Test Length

The current version of TIA may be administered in a class period (50 minutes). This allows time for giving instructions and the actual test-taking time. It is intended to be a power test, so students are given a reasonable amount of time to complete all items. From teacher reports, it appears that the average test-taking time is 30 minutes so this may be used as a rough guide if teachers wished to use it as a speed test.

Selection of Topics

Differences in background knowledge among students and students at different grade levels can be manifested in different ways. These differences may lead to variance in performance on reading comprehension tests and hence to invalid interpretations of students' performance. It is desirable that the world views, or empirical beliefs needed to interpret a story on which test items are based, be ones that most students share (Norris, 1988). If scores on TIA are to be taken as measures of inference ability in reading comprehension, then it is necessary to reduce as much as possible the effects of

background knowledge. To minimize differences in performance which might be due to differences in background knowledge, rather than to differences in inference ability, item were selected on the basis of their familiarity to students in Grades 6, 7, and 8.

Sensitivity to the issue of background knowledge led to a comprehensive study of topics for potential selection for item development. The six stages of the study are described next.

Stage one: graduate students. The first stage involved eight graduate students with a diversity of teaching experiences. Each student was asked to list 10 topics which he or she felt students in Grades 6, 7, and 8 would be interested in and have knowledge of, and to provide a justification for their choices. These are the 10 topics which the graduates identified most frequently: travel, space, videos, sports, animals, money, friends, future, styles, and science fiction.

Stage two: teachers. In the second stage, 130 middle grade teachers attending a professional development workshop were asked to list 10 topics which would interest their students and to justify their list. The teachers identified the following topics most frequently: cars, space, money, science fiction, holidays, music, mystery, computers, sports, and hobbies. The graduate students' and teachers' lists overlapped on the following topics: space, sports, money, future, and science fiction.

Stage three: students (topic identification). In the third stage, the ideas of middle grade students were sought. Three hundred students at Grades 5, 6, 7, 8, and 9 were asked to list 10 topics they thought they could write about without difficulty. Grades 5 and 9 were taken in addition to Grades 6, 7, and 8 to account for potential differences at the upper and lower limits of reading ability in the intended test group. The most preferred choices of the students are the following: money, space things, sports, pets, getting out of school, holidays, movies, space friends, war, and travel. Overlap in topic choices among all three groups indicates that the most popular choices are money, space or space-related topics, sports, getting out of school, holidays, and pets.

Stage four: students (unassigned written essays). In the fourth stage, middle grade students were asked to write on a topic of their choice. This was done to distinguish topics which students would choose to write about from those that might sound exciting but about which they would be unlikely or unable to write. Twelve classes of students in Grades 6, 7, and 8 were asked to choose from the most common topics identified up to this point (money, space or space-related things, sports, pets, getting out of school, holidays) or any other topic and to write an essay.

The essays were generally about space, money, and pets in one way or another. Specific differences existed in the general topic, for instance, essays about pets varied from the time it takes to care for them to how pets are wonderful friends. Bearing in mind that each story on TIA was to be representative of the reading materials at the middle grade levels, then from the most popular student topics three topics were selected: UFOs, Money, and a Newspaper Mystery.

Stage five: students (assigned written essays). In the fifth stage, 65 students in Grades 5 through 9 were asked to write a story about UFOs, Money, and a Newspaper Mystery. These essays were studied for vocabulary-choice, sentence and idea complexity, and form.

Stage six: final topic selection. The sixth and final stage of topic selection went through three phases involving free recall and word associations, recognition, structured and unstructured questions and discussions on each of the three topics. These phases represent a synthesis of research on assessing background knowledge (Adams & Bruce, 1980; Anderson, Spiro, & Anderson, 1978; Holmes, 1983, 1987; Holmes & Roser, 1987; Pearson, Hansen, & Gordon, 1979; Spilich, Vesonder, Chiesi, & Voss, 1979; Walker & Yekovich, 1984) and are taken to be some of the best available ways to assess background knowledge. It took approximately five class periods to establish students' background knowledge of each topic.

In Phase 1, students were asked to free recall or to brainstorm on each of the topics. They were directed to think of all the information they would expect to find in a story about UFOs, a story about Money, and a Newspaper Mystery. Also, students were asked to come up with associations for UFO-related words like heavenly bodies, evidence, and explanations; for Money-related words like uses, forms, characteristics, and changes; and for Newspaper-related words like responsibilities, carrier, weather, and newspaper-related confusions.

The second phase involved recognition activities to identify any misunderstanding which students might have about each of the three topics. These activities were developed from the Phase 1 discussions. For example, it became evident that some students thought that scientists know what UFOs are and that UFOs are meteors and "stuff like that" in the sky. Students were asked to identify from a prepared sheet dealing with these matters several possible correct and incorrect answers to questions such as "What are UFOs?", "Is a meteor a UFO?", "Would there be UFOs if scientists know what they are?" The answers provided to these kinds of questions led into the final phase of topic selection.

Discussions guided by structured and unstructured questions completed the final phase of establishing topic familiarity. A structured question on the Money topic, for instance, was "What is money?" Such questions led to unstructured questions about the topic such as "Do all jungle tribes have money?", and lively discussions were held with the students on each of the three selected topics.

For the purposes of this project, students were assumed to have a sufficient amount of topic familiarity if they were able to speak to each topic according to a general outline as follows:

UFO Outline

I. What UFOs are believed to be
II. What UFOs are reported to look like
III. Where UFOs are reported to come from
IV. Available evidence
V. Why UFOs are studied

Money Outline

I. What money is
II. The characteristics of money
III. How money developed
IV. Functions of money
V. Why the form of money changes

Newspaper Delivery

I. Carrier's responsibilities
II. Knowing the route
III. The importance of time
IV. How to deal with people
V. Potential problems

Furthermore, students were deemed to have sufficient relevant background knowledge if at least 70% of them were able to speak to these outlines. Seventy percent was taken as a satisfactory lower bound of general knowledge on each topic because, assuming randomness, this would mean there was less than a 3% chance that a student would have knowledge of neither topic and less than a 10% chance that a student would have knowledge of fewer than two topics. In the case of the TIA topics, the specific student levels were 80%, 75%, and 90% on the UFO, Money, and Newspapers topics, respectively. Thus, the chances of systematic bias against any student *across all topics* is minimal.

The comprehensive information gathered from the topics identified, the students, written stories' and class visits guided the choice of topics for the stories in the TIA test. Three stories were written for the TIA test: UFOs, Money, and The Wrong Newspapers. The UFOs story was modified from previous research projects (Beebe & Phillips, 1980; Phillips, 1985) and continues to be a winner among students. It is a story about unusual phenomena, telling of different UFO reports, offering plausible explanations for some of the reports, and suggesting that with improved technology we may be able to explain UFOs. The Money story is a description of the everyday use of money, of how it works, as well as its historical development. The third and final story is a mystery entitled "The Wrong Newspapers" which involves a mix-up in newspaper delivery, with the culprit being the neighbor's dog.

Principles of Story Writing

Story grammars have been developed to illustrate underlying text structures. The most common types of text used in the middle grade levels are narrative, descriptive, and expository. Each is organized in a particular way and it is believed that children use the structure, once they have it internalized, to assist them in understanding and recalling information (Stein, 1983; Thorndyke, 1977; van Dijk & Kintsch, 1983).

There is overlap in the classifications of narrative, expository, and descriptive structures since all three may be found in the one story. The Wrong Newspapers story fits more within the narrative classification than either the expository or descriptive. However, the UFOs and Money stories are harder to classify because they overlap considerably with the exposition and descriptive forms.

The principles of story grammar were followed in writing the TIA mystery narrative entitled "The Wrong Newspapers." The principles may be summarized as follows: There should be a setting which introduces the characters and provides the time and place of the story; an initiating event should occur which sets the story in action; there should be a response to that action followed by an attempt to achieve a goal or to respond to an action; the consequences of that attempt woven with a reaction are provided. These principles, coupled with a sensitivity to vocabulary choice, sentence structure, and sentence length, were in our minds during the story writing process. The data obtained when students read the stories was used to make final judgments on topic and story suitability.

Narrative provides a resolution or stopping point and therefore it is easier to identify its underlying structure or grammar than is commonly the case in expository material. Expository material has at least six underlying structures: serial, topic, restriction and illustration, definition, argumentation and comparison-contrast. The serial pattern may be considered the generic basic structure since the others are more general secondary structures. These five structures (topic, restriction and illustration, definition, argumentation and comparison-contrast) are more perplexing than the serial structure from another perspective because within a serial pattern there may be occasions when other structures are used. Consider the case in a social studies text where forms of travel are being studied in predominantly a serial fashion, but for a couple of paragraphs modes of travel are compared and contrasted, followed at the end of the chapter by a generalization about the most efficient means of travel. So, it is common to see much overlap and intermingling among text structures.

The TIA stories on UFOs and Money were written following primarily a serial pattern: A general concept is presented in each story; generalizations combined with examples are stated; a sequence of events unfolds; a conclusion follows. The authors were cautious to ensure that vocabulary choices were either known or explained and that sentences were coherent. The UFOs, Money, and The Wrong Newspapers stories were further subjected to the Anderson and Armbruster (1984) test of understandability; do the stories provide enough relevant information to achieve the author's purpose and to be meaningful to its readers? The evidence from the groups of students in Grades 6, 7, and 8 who read and discussed the stories is that the test of understandability was passed.

The Evolution of the Test of Inference Ability in Reading Comprehension

The present form of TIA represents the results of six phases of evolution in design and development. This section will provide details of the modifications at each phase of the test development as well as a rationale for them. A description of the test validation techniques that guided the design and development of TIA will also be discussed.

Initial Test Version

The Initial Test Version contained three stories (UFOs, Money, and The Wrong Newspapers) and 48 short-answer questions. It was given to a graduate reading class and a colleague who was engaged in the development of a test of inductive reasoning in critical thinking.

On the basis of feedback from these two sources the test was edited and written more concisely. The number of test questions was reduced from 48 to 36 because the test was too lengthy even for these subjects. This task was made simple for two reasons: (a) 5 of the 12 questions were judged to be insufficiently related to the stories to allow for more complete and consistent inferences to be made; and (b) 7 of the 12 questions could not be easily identified as inference questions, so in cases of doubt the questions were dropped from the test.

Pilot Study 1 (Short-Answer Version)

Trial One

Following the completion of the revisions to the Initial Test Version, the first pilot was conducted. A short-answer format, rather than a multiple-choice format, was used to help understand what the questions measured. In addition, this trial administration was done to check on test length, passage difficulty, vocabulary choices, clarity of instructions, and question ambiguities. The effects of story order were studied. As reported in a previous section, research shows that narrative text is generally easier to understand than expository text. If this is so, then differences in performance could be expected if the order of presentation of discourse types was altered. If differences were identified, then story order would be an important consideration in subsequent test development.

Sample and procedure. Sixty-five students in Grades 6, 7, and 8 participated. Test booklets were distributed randomly to students with the three stories (UFOs, Money, and The Wrong Newspapers) collated in the six story combinations. The directions and sample paragraph and inference question were discussed with the students. Students were told that they would have to use their background knowledge and the text information to answer the questions, that they would read three stories, and that each story would have four or five paragraphs and questions on each paragraph. Students were directed to read each paragraph, to write their answer to each of the corresponding questions, and to justify their answers. When all student enquiries were answered, then the test was started.

Results. The pattern of student responses to the inference questions was one of the most significant findings. Students' answers were of four types: an implausible response; a non-inference response; a partially-correct inference response; or a complete inference response. (A more detailed description of these may be found in Appendix A.)

It was found that the test was too long, since it took an average of one and one-quarter hours to complete, not counting the time necessary to give and complete the sample item.

No differences in performance on the basis of story order were found. Students may have acquired new knowledge while taking the test; however, it did not add to nor detract from their performance when story order was altered.

Test revisions. From an examination of students' answers numerous revisions were made: (a) six questions were reworded to make their meaning clearer and one question was deleted because of ambiguity; (b) three questions in the UFOs story (#8, 9, 10) were re-sequenced as #10, 8, 9 to match the text sequence; (c) sentences in the text judged to be too similarly worded to the corresponding inference questions were deleted; (d) some sentences were modified to be more general, and less explicit, thereby making the corresponding inference questions more challenging; and (e) other sentences were changed to clarify meaning.

Upon completion of the revisions based on the results of Trial 1, the number of questions on TIA for the Trial 2 study was 36: 12 for each story.

Trial 2

Trial 2 of Pilot Study 1 (short-answer version of TIA) was completed to determine whether the four types of responses identified in Trial 1 would be replicated with more refined passages and questions. If they were corroborated, then subsequent test development would have to take these response variations into account, if test performance was to be taken as a valid indication of student ability.

Sample and procedure. One hundred students in grades 6, 7, and 8 took the short-answer version of the TIA test. The same procedure was followed as in the previous trial.

Results. Students' written responses and accompanying justifications for their answers were studied. The trend of reader response variations identified in Trial 1 was evident in the responses on this trial. Student responses for each question fell within 1 of 4 response patterns identified in Trial 1. The four variations (an implausible response; a non-inference response; a partially-correct inference response; or a complete inference response) in student performance became a major factor in the future design and development of TIA.

Test revisions. Since a multiple-choice format for TIA was the ultimate aim, the fourth version of TIA involved writing a scaled-answer multiple-choice set for use in the second pilot study. The item form on the short-answer and multiple-choice versions of the test were modified so that they would be identical.

It was presumed that a sentence stem format, where possible, might help to reduce writing time, 90 minutes on the short-answer version, to one class period (50 minutes). Distractors for the multiple-choice version of the test were derived from students' answers on the written short-answer versions from Pilot Study 1. Each set of four possible answers were scaled as follows: an implausible response worth 0, a non-inference response worth 1, a partially-correct inference worth 2, and a complete inference worth 3. This "scaled-answer format" was used to afford students the option of selecting the type of response they would likely make if they were taking the short-answer version of TIA. Multiple-choice items were constructed such that distractors were consistent in grammatical style, vocabulary, and length. Answers were randomly assigned to serial position (A, B, C, or D).

Pilot Study 2 (Short-Answer/Scaled-Answer Multiple-Choice Versions)

The second pilot study was conducted to serve four purposes: (a) to examine the degree of similarity of performance on the short-answer and the scaled-answer multiple-choice formats; (b) to compare completion times required by both test formats; (c) to corroborate whether the four patterns of responses identified in Pilot Study 1 would be displayed by the students in this pilot; and (d) to identify potential item ambiguities, vocabulary difficulties, and other problems.

Sample and Procedure

Eighty-one students in Grades 6, 7, and 8 participated in Pilot 2. Forty students wrote the short-answer version and the remainder wrote the multiple-choice version. The same procedure described in previous pilots was followed, with one exception. The students taking the multiple-choice version were provided with answer choices and cautioned to consider all possible answers before deciding the answer they thought best.

Results of Multiple-Choice Format

Test completion time ranged from 50-75 minutes on the multiple-choice format for classes in Grades 6, 7, and 8. This represented an average reduction of 10 minutes compared to the short-answer format; unfortunately, we expected a greater reduction.

Item analysis, on the basis of correct or incorrect answers, of the multiple-choice format showed a KR-20 reliability of 0.68 and a test mean of 17.5 items correct out of a total possible 36 items, with a standard deviation of 4.73. The test means for Grades 6, 7, and 8 were 14.0, 17.1, and 19.6 respectively. Item/test biserial correlations and item difficulty indexes were computed and are presented in Table 1.

[Insert Table 1 about here.]

It can be seen from Table 1 that 3 of the 36 questions had negative biserial correlations (questions 18, 20, and 35). Examination of these three items, coupled with students' short-answer responses, revealed the answer sets for questions 18 and 35 to be ambiguous. Question 20 required students to consider a historical perspective, but it appears that most students answered it from a current events perspective.

Questions 8 and 28 had very low biserial correlations. It was clear, upon examination, that the problems with questions 8 and 28 were vocabulary-related. It seems that many students did not note the relevance of particular examples which were cited. For instance, "meteors" were cited as an example of "astronomical events," but students did not see the relevance in answering item 8 which read "Other kinds of astronomical events that people mistake to be UFOs are." It seems students did not understand "astronomical events," so it was replaced with "heavenly bodies." Revisions were made to all aspects of the test identified to be either definitely or potentially problematic.

The item difficulty levels also pointed to problems with questions 18 and 35 discussed in the preceding paragraph. Question 13 was among the more difficult items; it seems that a word in the question stem was interpreted differently by many students from the test authors. The question read "Money is needed in at least two different ways;" students interpreted the question by focussing on the word "needed" as necessities. The test authors intended the item to get at the idea of commonality of money. Clearly, the problem with the item was with the wording and not with how the students interpreted it. Item 13 was revised to read "Money is a familiar part of our lives because."

Results of Short-Answer Format

Students' responses on the short-answer format were examined and the type of answer identified (implausible response, non-inference response, partially-correct inference, complete inference). Again, the pattern of student responses was consistent with the two previous trials under Pilot 1. This result was taken as evidence that the category scheme could be helpful in constructing distractors that students performing at each level would find plausible.

Student responses on the short-answer format were compared to the multiple-choice key to assess the agreement between the number of responses per item that received full credit. The results are presented in Table 2. It should be pointed out that for purposes of this analysis an item on the short-answer format was not considered correct unless it expressed the same meaning as the answer keyed as

correct on the multiple-choice format, consequently the percentages of agreement between the two are necessarily lowered. For instance, consider item 30 which says "Ann wanted to hand deliver Mr. Jones's newspaper because." The response keyed correct on the multiple-choice format and the one required on the written short-answer format would be "to make sure he got it and to talk to him about the mystery." So, unless students responded on the short-answer version with a compound answer, they were not scored as completely correct, even though they may have been partially correct. A lower percentage of agreement was found on those items that required students to synthesize story information. It seemed that if questions required students to pull together more than one piece of information to formulate a complete response, then they experienced difficulties or did not consider all available relevant information. Question 26, for instance, required the synthesis of three pieces of information; however, the most common response written on the short-answer test and selected on the multiple-choice test was a partially-correct one.

[Insert Table 2 about here.]

The mean on the 36 item short-answer test was 12.97. A recognized restriction of this pilot was that one, and only one, answer was deemed acceptable, which undoubtedly ignores a range of answers which may have been partially correct. Bearing in mind this restriction on acceptable answers, then it seems reasonable to expect that the level of overall performance may have been reduced. The mean on the multiple-choice test was 17.15, which reflects a significantly higher level of performance. Another explanation for the lower performance on the short-answer test could be related to the fact that students had to construct and write an answer, which would seem to be a more demanding task than selecting an answer on the multiple-choice test. Student performance on the multiple-choice test may be a "better" indication of their reading ability than the short-answer test where performance is confounded with students' ability to express their ideas in writing. Also contributing to lower scores was the fact that students tended to leave more items unanswered on the short-answer test than on the multiple-choice test.

Test Revisions

Passage, question, and answer modifications were made to TIA prior to the next pilot. Revisions were made to each of the three stories. For instance, on the UFOs story, it was found that students failed to attend to the word "not" in the sentence, "Many of the older reports are not complete so we need to continue to study UFOs," consequently leading to an erroneous response. The sentence was modified to "Many of the older reports are incomplete so we need to continue to study UFOs."

Some questions were replaced because they did not require students to make inferences, and some answers to other questions were replaced because of ambiguity. Other revisions included substitutions in word-choices and changes in information placement. Further revisions included making answer selections more parallel with one another. For instance, "plastic cards" was replaced with "club cards" in item 23 to make it more parallel with the other options: "trade items," "credit cards," "chocolate bars."

Pilot Study 3 (Verbal Reports as Data)

Pilot Study 3 was conducted using a verbal report methodology. The use of verbal reports in test construction has been suggested by several testing experts (Anastasi, 1988; Cronbach, 1971; Haney & Scott, 1987; Messick, in press). Verbal reports were used as a method to validate whether a complete inference had been made when students selected the answer keyed as correct to help ensure that multiple-choice test questions were functioning effectively as inference questions. In addition, such an approach is particularly useful in test development for revealing potential item ambiguities, vocabulary problems, and hidden cues.

Care was taken to develop interview procedures which would not jeopardize the quality of information to be collected and conclusions to be drawn. Two trial verbal report sessions with six students each were held to ensure that the two interviewers understood the demands and limits of the approach, as well as to determine whether the information needed from the students was being acquired.

Sample and Procedure

Thirty-six students in Grades 6, 7, and 8 participated. Students were each assigned to 1 of the 3 stories on TIA. They were told that to answer each question they would have to use information given in the story and information they already knew. They were told that the story would not directly answer the questions and they would have to use their common sense along with the story information. Students were advised to consider all possible answers before deciding which answer they thought was the best one. Each student worked through a sample item with an interviewer. Once the sample item was completed and students' questions were answered, students were asked to read aloud each paragraph, to read the corresponding test questions, to select 1 of the 4 answers provided, and to tell why they thought that answer was the best.

Interviewers questioned students only if there was a lack of clarity in a response, such as an unspecified pronoun antecedent or an answer so terse or vague that it was too incomplete to follow. At the end of the test interview, general questions were asked about students' interest in the story, about whether the passage vocabulary gave them any problems, and about whether there were other things that were unclear to them. Each verbal report protocol was transcribed and a scheme developed to code the quality of students' responses.

Scoring Responses

In order to reflect the range of responses shown by students in the verbal reports, a scoring system was devised to allow credit for partially correct, as well as complete, inferences. Scores from 0 to 3 were assigned on the basis of the range of completeness of the student responses. See Appendix A for the criteria for grading the test of inference ability.

The following question (Q1) and its possible answers (A, B, C, or D) illustrates the scoring system.

> **Q1 UFOs are sometimes called other names because**
> **(A) people name them according to their shape or probable origin.**

This answer is a complete inference and therefore, is given a score of (3). The relevant textual information was contained in sentence three, "People sometimes call UFOs flying saucers, spaceships from other planets, and extraterrestrial spacecraft." Using background knowledge it can be concluded that the naming criteria for UFOs in this story are based on either shape ("saucers") or probable origin ("other planets" and "extraterrestrial").

The integration of the relevant textual information and background knowledge makes (A) the best inference response for question 1.

> **(D) people don't know what to call them so name them by shape.**

This answer is given a score of (2). It is a partially correct inference for question 1 because it only considers one of the naming criteria, shape, when the textual information supplies two criteria. The criterion of shape was selected for this alternative instead of the criterion of origin because shape was focussed upon in all instances of explanations by students in the verbal reports.

(C) people see an area with many colored lights in the sky

This answer is given a score of (1). It is based on textual information from sentence 5. However, the relevant textual information is contained in another area of the text (sentence 3). Although the textual information selected deals with the appearance of UFOs, it is not the most relevant part of the text.

(B) people know they are unidentified flying objects in the sky

This answer is scored as (0). It is the least correct answer because it makes no sense either in relation to the text, or in relation to background knowledge. People do not know for certain that what they see are unidentified flying objects, and this is not the reason given in the text that synonyms exist for UFOs.

Answer Set Revisions

The process of revising answer sets based on students' verbal reports had two complementary phases. One phase dealt with editing existing answer sets and the other with developing new answer sets which would reflect the range of answers students gave in their verbal reports.

Answer sets were revised where students' explanations of their choice of answer showed either that students made an inference but still selected a less than best answer, or that they used inadvertently placed cues in the answer set to select the best answer. The second phase is discussed in the next section with question revisions.

Vocabulary and Question Revisions

A number of terms which students did not understand became apparent in the verbal report data. Samples of vocabulary revisions include the following substitutions, "scientific equipment" for "technology," and "heavenly bodies" for "astronomical events." Care was taken to maintain the intent of the text and use of precise terms while substituting appropriate vocabulary for students at the Grades 6, 7, and 8 levels. For example, in looking for a substitution for "astronomical events" children's science texts, children's science encyclopedias, and science reference books were consulted.

Eleven questions were deleted from this test version. Five of the 11 questions were judged to be based too heavily on students' background knowledge. The 5 questions did not meet the principle that a good inference question is one that requires a reader to integrate relevant text information and background knowledge to construct complete interpretations that are consistent with both the text information and background knowledge. Four of the 5 questions required students to give answers based on word knowledge. For instance, one of the items stated "The word independence in this story means," which could have been answered without reading the text. In another item, students' lack of background knowledge hampered students in making a complete inference, so the question was deleted. The deleted item read "Money might be more risky to use than credit cards because," but according to students' verbal reports, they did not know that credit cards could be cancelled, and therefore less risky to lose than money.

The remaining six questions were deleted for a variety of reasons. Difficulty level indices from previous pilots indicated that question 13 was one of the most difficult questions (see Table 1). Student verbal reports indicated differences in word interpretations from those intended by the authors. Item 13 says "Money is reused by;" it seems students interpreted "reused," though illogical, to refer to the same money being used over and over or saved by a single individual and not the circulation of money.

Questions requiring students to make time frame shifts were problematic, as evidenced in students' verbal reports and in the item analysis results. For example, students' verbal reports showed that they responded to item 20 which read "Years ago, cows, coffee, and shells did not keep their value as well as money today because" from a current events perspective. A typical response was "they are not wanted

by everyone, whereas money might be because if you traded with people from the city they might not need cows." Item 20 had gone through three revisions and yet students seemed to focus on the current rather than the past, so the item was deleted. The remaining two questions did not function well as inference questions because they were judged to be too text-dependent, so they were deleted.

Students' verbal reports also pointed to item ambiguities (items 9, 18, and 35). For example on item 9, students interpreted "find out" to mean discover new facts, when the intended meaning was "learn." So the item stem "Using the reported information we would find out the most about UFOs by" was changed to "Using available information people learn the most about UFOs by." This modification required students to make the inference that the "available information" was the reports described in the story.

Story Passage Revisions

The final section of test revisions in this pilot deals with the story passages. The major change was with the "Money" story. Due to the fact that the first five inference questions in the "Money" story were deleted, the first two story paragraphs were also deleted. Two new paragraphs and five new inference questions on the functions and characteristics of money were written for the "Money" story.

Minor changes were made to other paragraphs through deletion and addition of sentences. Sentences were added to story paragraphs in instances where more textual information was required for a specific inference question or where a new test question had been added. For example, the sentence "Weather conditions are checked when scientists study available information about UFOs" was added to the second UFOs paragraph to complement the question "Weather conditions affect UFO sightings in the sky because" (UFOs Q5). The sentence in the third paragraph of the "Money" story, "Large animals made trade difficult because there was too much price difference in items," was deleted. There was insufficient story information about trade items for students to answer the corresponding inference questions. Changes of specific vocabulary in story paragraphs were discussed under editing of test vocabulary. Remaining changes were cosmetic in nature.

Pilot Study 4 (Expert Sample)

The revised scaled-answer multiple-choice test was given to two fourth-year college classes in the Faculty of Education at Memorial University of Newfoundland. Sixty-one students participated in this pilot. The purpose was twofold: First, to have an expert adult sample confirm the researchers' rating decisions for young students' responses on the TIA test; and second, to have the experts take the test in order to pinpoint any remaining ambiguities or other problems which might necessitate revision.

So-called experts may be quite unreliable judges of items written for younger students because the adult conception of what is and is not familiar may be quite different from that of younger students. For example recall the item, "Money might be more risky to use than credit cards," required students to know that credit cards may be cancelled. A study of students' verbal reports revealed this to be a piece of information which they did not know. Consequently, while adults consistently made a complete inference on this item, the middle grade students never did. The item was dropped because it did not measure students' inference ability.

For 85% of the items the experts rated the young readers' responses consistent with the rating assigned by the researchers. The remaining 15% were taken to need further revisions. In addition, comments and queries made by the experts were studied and appropriate changes made.

It was assumed that college students would get all 36 items correct on a test intended for middle grade students, so only correct answers were scored. However, the mean number correct for the two college classes was 23.95, out of 36 items, with a standard deviation of 3.68 and a KR-20 reliability of 0.58. There seemed to be distinct divisions among the expert sample about some of the items. For instance,

there were adults who wrote "there is no best answer here " on an item that required them to synthesize two or more pieces of information. It seemed some of the experts would indicate the right information needed for an answer, but would not pull the information together to make a complete and consistent inference. The remaining experts seemed to have little difficulty making complete inferences consistent with those of the researchers. Thus, the majority of the experts were taken to be reliable judges of the best answers.

Pilot Study 5

Test Validation

The fifth pilot study was designed to study the relationship between students' answer selections and their thinking processes in making those selections. One purpose was to determine the quality of students' thinking when they selected their answer for each test item. Understanding students' thinking processes is of fundamental importance because students often arrive at good answers without thinking well and at less than good answers even though they may have thought well. A second purpose was to find out whether the verbal report process affected students' performance, either positively or negatively.

Specifically, four issues motivated the validation procedure; (a) to determine whether students understood the task, that is, that they were to use information from the text and from their background knowledge to answer the inference questions; (b) to determine whether students understood each test item and reasoned well when they picked the best answers; (c) to determine whether students who chose an incorrect answer to an item did so because they did not reason well; and (d) to determine whether verbal reporting affects performance in comparison to writing the test.

For a test of inference ability in reading comprehension to be valid, the test should require that students make a complete inference when they select the best answer for an item (Phillips, 1986). One assumption in multiple-choice test construction is that when students select the best answer for a test item, they do so for the right reason. However, it is possible that students might select the best answer for a test item without fully understanding it. For example, there may be some inadvertent cue prompting students to choose the right answer. A second assumption is that students who choose the incorrect answer do so because they are not reasoning well, yet students might select an incorrect answer for a good reason. For example, there may be an alternate interpretation from that intended by the authors, leading students to choose a less than complete answer even though they reasoned well. Thus, it is important to have students explain their reasoning when they select their answer for each question.

Students' thinking ability was examined by having them report verbally why they had selected their answers. Answers were almost always chosen for reasons. The aim was to study those reasons for evidence that good thinking leads to good performance on TIA and poor reasoning leads to poor performance. To do this, there must be a description of the reasoning processes which lead to performance and a way to rate the quality of students' thinking. A study of students' thinking processes was done in an attempt to gain more information on the reasoning underlying their answers than is possible by merely considering the selected responses. Information about the nature of the reasoning process is pertinent to construct validation (Embretson, 1983). Evidence for construct validity of an ability, in this case inference ability, is obtained to the extent that good performance can be explained by students' reasoning well, and to the extent that poor performance can be explained by not reasoning well. Thus, the aim of construct validation is to identify the causes of performance on tests. The aim with TIA was to develop a test such that good inference-making was the cause of good test performance and the construct validation through the use of verbal reports is conducted to determine the extent to which this has been achieved. These verbal report protocols were used in conjunction with the students' answer selections to provide information for test validation. The general principle

followed was that tests would be valid to the extent that good inference-making led to good performance and poor inference-making led to poor performance.

Sample and Procedure

One hundred and eighty-three students in Grades 6, 7, and 8 at three schools participated in this pilot. The students were selected at random from intact classes and assigned randomly to either of two test conditions. There were 95 students tested in the verbal report condition and 88 students in the written test condition.

Students in the written test condition completed the multiple-choice test in their classrooms in a manner similar to any group test. The same administration procedures described in Pilot Study 2 were followed. Students in the verbal report condition were interviewed by two interviewers using the same procedure described in Pilot Study 3. Students in the verbal report cohort were assigned to 1 of the 3 stories on a rotating basis. That is, the first student was assigned story 1, the second assigned story 2, the third did story 3, and the fourth student did story 1, thus starting a repeat of the cycle. The total administrations per story were as follows: 34 students completed the 'UFOs' story, 32 students completed the 'Money' story, and 29 students completed 'The Wrong Newspapers' story.

Coding

Three sets of data were collected: reading scores from the written cohort; and reading and thinking scores from the verbal report cohort. Written cohort selected responses were scored according to criteria developed in Pilot Study 3. (See Criteria for grading TIA in Appendix A.) The reading score for an answer ranged from 0 (implausible) to 3 (complete). Students' total reading scores were the sum of the values assigned to all answers selected by students. The total possible score is 108.

Verbal report explanations of students in the verbal report cohort were assigned thinking scores. The thinking score is a measure of the net supporting evidence for an item available from the verbal reports. The quality of students' explanations for each answer was rated according to specific criteria. (See appendix B for a copy of the Thinking Rating Scale.) Thus, for each item there was a reading score for the answer selected and a corresponding thinking score for a student's explanation of why that answer was chosen.

A trial sample of thinking protocols was selected at random from the three stories and grades. Two raters independently assigned a thinking score to each answer justification. Any inconsistencies between raters' scoring of thinking protocols were studied. The initial rating of this small sample of the verbal report protocols allowed changes in the category descriptions of the thinking rating scale before all the protocols were scored. For instance, it was observed that sometimes students simply repeated the answer they had selected as their explanation. In the initial thinking rating scale there was no provision for such a response. Consequently, a change was necessary and a thinking score of (0) was assigned for such responses.

Both total and individual item thinking scores assigned by the two raters were compared. Inter-rater reliabilities on both comparisons results in correlation coefficients greater than .90. Any explanations assigned different thinking scores by the raters were discussed and re-rated. With a high level of reliability on the rating of students' explanations established, it was concluded that the remaining protocols could be consistently scored. About 25% of the remaining protocols were checked at random and found to have a similarly high level of inter-rater reliability ($r > .91$).

Data Analysis

The data analysis examined six questions: (a) To what extent were students' reading scores and thinking scores on each test item in the verbal report cohort correlated? That is, did students who

reasoned well select the best answer and did students who reasoned poorly select an incorrect answer? (b) To what extent were students' total reading and thinking scores for each story in the verbal report cohort correlated? (c) How did students' reading scores in the verbal report cohort compare with reading scores in the written cohort? (d) How is performance on each item related to overall test performance? (e) Did students' reading scores vary by grade level? and (f) Were there interviewer effects on test performance?

Results and Discussion

Reading and thinking relationships for items. Table 3 presents Pearson's correlation coefficients between students' reading and thinking scores by test item. A positive correlation significant at less than the .05 level between reading and thinking scores was found for 34 of the 36 items with an average correlation of .55. Reading and thinking scores for item 10 were not significantly correlated and for item 17 they were negatively correlated. These two items were examined, but no problems were apparent. The results of the previous pilot studies were examined and no indications of problems with items 10 and 17 were found. The final decision was to leave the items without changes and to examine them in the next trial.

[Insert Table 3 about here.]

For 94% of the items good thinking was significantly correlated with good reading and poor thinking to poor reading performance. This result provides strong evidence that generally when students thought well they selected the best answer and when students reasoned poorly they selected an alternate answer. The significant correlations between reading and thinking scores for items is one piece of evidence that TIA is a valid test of inference ability.

Reading and thinking relationships for stories. The reading and thinking relationship for each item is by necessity related to this relationship for each story. Twelve items accompany each story; therefore items 1-12 accompany story 1 'UFOs,' items 13-24 accompany story 2 'Money,' and items 25-36 accompany story 3 'The Wrong Newspapers.' Table 4 presents Pearson's correlation coefficients between total reading and thinking score for the three stories. The correlation coefficients were similar, high, and significant at the .001 level for the three stories.

[Insert Table 4 about here.]

It is reasonable to conclude that students understood the items and that students who selected the best answers thought well. Thus, the significant reading and thinking relationships for stories is taken to be another piece of evidence that TIA is a valid test of inference ability.

Reading performance relationships between verbal report and written cohorts. Table 5 presents story reading score means by cohort. The maximum reading score for a story would be 36, as each story has 12 test items, with a total possible score of (3) per item. Means for story 1 and story 2 were very similar for the verbal report and written cohorts. Means for story 3 differed by 2.8 in favor of the verbal report cohort. While not significant, it is not clear why a difference occurred. This difference translated into test performance would amount to the verbal report cohort doing better on one item. Across the entire test, the overall mean for the verbal report cohort is 24.7 and 23.3 for the written cohort. A difference between means of 1.4 translates into less than half an item correct in favor of the verbal report cohort. Thus, the difference in means on story 3 was not taken to be large enough to invalidate the verbal report methodology. Asking students to think aloud does not significantly alter their performance.

[Insert Table 5 about here.]

The analyses of variance story by cohort and grade showed no significant effects for cohort for either the UFOs or Money stories. However, cohort showed a significant effect, $F(1, 111) = 6.06$, $p < .05$, for The Wrong Newspapers story. It is not easy to explain why a difference in performance by cohort was found on only The Wrong Newspapers story.

Grade had a significant effect on students' reading scores for story 1 (UFOs), $F(2, 116) = 7.58$, $p < .001$ but was not significant for stories 2 or 3. The discourse type may account for the grade effect found for story 1. Students in Grades 6, 7, and 8 may all be familiar with the descriptive and narrative discourse forms of stories 2 and 3. But, students' reading scores on expository material (story 1) might show an improvement for students in Grades 7 and 8 when compared to Grade 6 students. There was no significant interaction effect between grade and cohort.

In sum, reading performance between the verbal report and written cohorts was taken to be highly related. Assuming that verbal reports are an accurate representation of the thinking that went on during the test-taking and the reports are an accurate representation of the thinking of those in the written cohort, then it can be concluded from the evidence presented that students understood the task and reasoned well when they picked the best answer. In addition, the usefulness of verbal reports in understanding students' reasoning and validating less direct measures of inference is strongly supported.

The relationship of item performance to story performance. Students in the verbal report cohort completed only one story, so item analysis results are presented by story for both the verbal report and written cohorts.[3]

The item/test biserial correlation coefficients were positive for all test items and ranged from a low of .163 (item 17) to a high of .693 (item 3). The correlation coefficients show that generally students' performance on individual test items was positively related to overall test performance.

The difficulty indices were computed as the proportion of students picking the best answer. This is not the best indicator of difficulty for scaled-answer items because it does not take account of partially correct scores (1 or 2). In a subsequent section on final data collection, an index computed as average score on an item is used, but the rough index based on rights and wrongs will suffice here because only the best answer was to be considered. A low difficulty index (.100) would indicate a more difficult test item than a test item with a high difficulty index (.600). The difficulty indexes of the test items ranged from a low of .197 (item 1) to a high of .746 (item 12). The range in the difficulty level of items was expected and is within normally recommended bounds.

The KR-20 reliabilities calculated separately for *each* of the three test stories (12 items each) and only for the correct answer for the combined verbal report and written test cohorts were 0.57, 0.23, and 0.50 for stories 1, 2, and 3. The written test cohort completed all 36 test items with a KR-20 reliability of 0.69.

In sum, the results of the relationship of item performance to overall story performance was taken as evidence that students understood the task and that students reasoned well when they chose the best answer. Thus, it is concluded that TIA requires students to make a complete inference when they select the best answer.

Reading scores by grade. Students' reading scores by grade level were also examined. Since only the written test cohort in the fifth pilot study completed all 36 test items, then only their scores were used in this part of the analysis. As the highest reading score for each test item was (3), the maximum reading score for 36 items was 108. Reading score means and standard deviations are presented in Table 6 for Grades 6, 7, and 8 students.

[Insert Table 6 about here].

A one-way ANOVA was performed with reading score as the dependent variable and grade as independent variable. Grade was found to have a significant effect, $F(2, 87) = 4.98$, $p < .01$. The overall trend in mean performance from Grades 6 to 8 was a desirable result. It is assumed that students would perform better in making inferences with each passing grade. Since TIA is intended as a measure of inference ability in Grades 6, 7, and 8, then a significant difference by grade suggests that TIA is sufficiently discriminating to detect differences in performance, should they exist, at each grade level.

Interviewer effect on test performance. Verbal reports of students' reading and thinking scores were analyzed by story with interviewer as the independent variable. Two separate one-way ANOVAS were performed for each of the three stories. Therefore, for each story there was one analysis with reading score as the dependent variable and a second analysis with thinking score as the dependent variable. No significant effect for interviewer was found at the .05 level for any of the six ANOVAS calculated. Interviewer, therefore, did not seem to affect students' reading or thinking performance in the verbal report cohort. This was an encouraging result which provided support for the usefulness of trial interviews to eliminate potential interview problems which may affect the primary data collection.

Section Summary

The data analysis and test results discussed in the preceding sections show the development and statistical support for TIA as a test with both construct validity and reliability. Each subsequent version of the TIA test was an improvement over each previous version and it was not clear what would be gained from further data collection, so the pilot studies were considered to be complete and the TIA test ready for final data collection.

Final Data Collection: Analysis and Results

This section reports the demographics of the samples, the final data collection procedures, and the basic statistics. It also discusses potential extraneous influences to test performance and presents the reliability estimates of TIA.

Samples and Data Collection

Nine hundred and ninety-nine students in grades 6, 7, and 8 from schools in Alberta, Newfoundland and Labrador, Nova Scotia, and Ontario comprised the samples for the final data collection. Contact was made with educators at schools and school board offices and their cooperation was sought for the administration of the TIA test.

When approval was granted to proceed with the project, the contact persons were forwarded the necessary materials. They either arranged to give the tests themselves or for classroom teachers to administer them during scheduled language arts classes. Each participating teacher was given a copy of the directions (see footnote 3). The original contact person was the facilitator for each province. That person took responsibility for distributing the materials, ensuring their proper administration, collecting the materials, and returning them to The Institute for Educational Research and Development at Memorial University of Newfoundland. The final data collection took place during the winter and spring of 1987.

Students in the Alberta sample were from an urban centre with a population of approximately 60,000. It is a trading centre for an agricultural-based economy. The students were described by their teachers as mostly middle class. The schools range in population from 150-650. Less than 4% of the school population is English as a second language (ESL) or native students. Classes were described as having a few bright students, a majority of average students, and some students requiring additional assistance with instruction through remediation classes or learning assistance programs.

The Newfoundland and Labrador sample of students were from two large rural centres. In one centre the population, including surrounding villages, is approximately 10,000. The area may be described as economically depressed with the majority of families described as low to middle class. The students were from a school with a total population of 430 students. None of the students were ESL students; however, some had been involved in French immersion programs. Classes were described as heterogeneous. The other rural centre has a combined population of approximately 14,000 people in two adjacent towns. It is a mining centre with a high employment rate and, for the most part, middle class families. The students are from schools ranging in size from 350-600 students. There are no ESL students, but French immersion programs are common. The classes were described as heterogeneous.

Students in the Nova Scotia sample were from two rural areas ranging in population from 2,000-5,000 people. The economy is farm-based, with the communities comprising a mixture of lower and middle class families. The populations of the two schools were 200 and 275 with no ESL students. Classes were described as heterogeneous.

Students in the Ontario sample were from an urban centre with a population ranging from 60,000 to 110,000 including the surrounding areas from which children are bussed to the city schools. The students were from a wide range of economic levels, and many from single parent homes. Classes were described as heterogeneous with about 20% of the students requiring remedial instruction. Less than 3% of the school population includes ESL students, and brighter students often go into French immersion programs after Grade 4. The economy in the area is built on service and government institutions.

Analysis and Results

Students recorded their answers to the TIA questions on a standardized answer sheet. Each question has four possible answers (A, B, C, or D). Each answer is worth a value of either a 0, 1, 2, or 3 dependent upon the quality of the selected response. The appropriate value was assigned to each selected response and the assigned values totalled to constitute the test score for each student.

Table 7 also presents mean performance scores for the entire sample by sex, grade, and age. The mean for the entire sample of students for whom data was complete (N = 974) is 73.57 with a standard deviation of 13.63. Table 8 shows the ANOVA main effects on these same variables. Of particular relevance in this report are the main effects of sex, grade, and age on test score.

[Insert Tables 7 & 8 about here.]

In the case of significant sex differences, the females' mean performance is higher than the males'. A comparison of the means presented in Table 7 indicates a difference of about two points. Based on the perplexing welter of research findings on differences between males and females, it seems many questions remain unanswered (Downing, May, & Ollila, 1982). Questions about such matters as the effects of different cultural expectations, genetic factors, and teacher-model differences all seem to point to the necessity of further research prior to the drawing of any conclusions. Differences in this data on the basis of sex are minimal. Furthermore, differences in performance between the sexes are small in comparison to the range of differences in performance within a sex. For these reasons it is not believed that the TIA test is biased in favor of any sex, nor that the data is untrustworthy for generalization purposes, regardless of sex.

Table 8 shows grade and age significant at $p < .05$ level. In the case of significant differences by grade it is important to know whether the differences are between Grades 6 and 7, Grades 7 and 8, and Grades 6 and 8. Sheffes a posteriori comparison of means test was done (Kirk, 1968). While Sheffes S method allows for the calculation of significant differences in means when there are unequal n's, it does set the highest critical statistic of all the multiple comparison tests. The only critical difference in

means on the TIA test was between Grades 6 and 8 where the difference in means (5.70) exceeds the critical value of 3.87. It is likely that had it been appropriate to use a less rigorous test that differences between Grades 6 and 7 and between 7 and 8 would have been found. Nevertheless, the significant differences between Grade 6 and Grade 8 indicate a general tendency for performance to improve with grade level.

The grade by age interaction is a result of the fact that scores increase by grade only for students with grade-appropriate ages. Students who are old for their grade tended to do more poorly than students with grade-appropriate ages.

Item Statistics

Item Difficulty

Typically, difficulty level indices reported on a multiple-choice test are given as the proportion of students getting an item right, but on a standard multiple-choice test the only scores are 1 and 0, where 1 is for the correct choice and 0 is for any other choice. Thus, the proportion of people getting the item right, that is the difficulty level, is just the average performance on the item. So by extension, the difficulty level index for the TIA test in which possible scores are 0, 1, 2, or 3 is again the average performance on the item. See Table 9 for the percentage breakdown of students who chose an answer worth 0, 1, 2, or 3 for each test item (total of 36 items). The percentage of students who chose answers other than the best (the most consistent and complete) for each item reflects the variability in performance. Item difficulties are Given in Table 10. When reading this table, note that the higher the difficulty level index the easier the item. As can be seen from Table 10, there is a range of difficulty levels across the test. For instance, many students found item 12 fairly easy whereas item 4 appears to have been more difficult for them. These indices represent a range of challenge for students.

The Relationship of Item Performance to Overall Performance

Typically, the item/test correlation is computed using a biserial correlation coefficient. This is a correlation between dichotomous (0,1) and continuous variables (0-36 range of items). However, on the TIA test item scores are not merely dichotomous variables, but rather interval variables (0, 1, 2, and 3), so the appropriate statistic is the Pearson's r. Table 10 presents the item/test correlations. There was one negative item/test biserial correlation (item 14) and one which was essentially zero (item 17). Item 14, as can be seen from Table 9, was answered by the greater proportion of students as a non-inference question (a score of 1). In other words, students chose a text-based response. Such a response by the majority of students points to a problem with either the wording of the question or a perceived high similarity among answer choices on the part of the students. A reexamination of students' verbal reports from the last pilot study indicated a problem with word-choice with item 14. A revision has been made for future uses of TIA. Item 17, also on the Money story, demanded a high level of understanding of the features of money which may have been unfair to the students. A reexamination of students' verbal reports corroborated the suspicion that in order to choose a 2-point answer, they made a sophisticated, complete, and consistent inference. The scoring key was modified for item 17.

[Insert Tables 9 & 10 about here.]

The Relationship of Story Performance to Overall Performance

TIA is made up of three of the most common discourse forms found in the middle grades. Research indicates that narrative, descriptive, and expository texts make distinct demands upon readers, thus it seems that differences in comprehensibility between narrative and descriptive and expository texts should be expected. Table 11 shows the percentage of all responses obtaining scores of 0, 1, 2, 3 by grade level and story. Seventy-two percent of all responses on The Wrong Newspapers are quality

responses (scores of 2 and 3) compared to 66% on the UFOs story and 68% on the Money story. The Wrong Newspapers story is a narrative, the discourse form taken to be the easier of the three, yet the differences in performance are not as dramatic as expected. This result raises an interesting question.

[Insert Table 11 about here.]

A question which remains to be studied is whether particular types of inference questions, regardless of discourse form, present more of a challenge to students than others. There is circumstantial evidence from the pilot studies and from questions rated as difficult on the final study to support such a suspicion. Logical inference questions on TIA that required the synthesis of several pieces of information were more likely to be answered in an incomplete manner than informational inference questions such as elaboration or setting the context, regardless of the discourse form. A plausible explanation for the minimal performance differences as displayed in Table 11 is that inference questions requiring the synthesis of several pieces of information were asked on all three discourse forms. If students experience difficulty with logical inference questions as suggested, then perhaps the type of question asked is an important variable in addition to the discourse form being studied.

Potential Extraneous Influences

Potential extraneous influences on performance on the TIA test include such factors as test-taking strategies, test-wiseness, and guessing. While these are not mutually exclusive, I will deal with each separately.

Test-Taking Strategies

Care was taken to provide clear, unambiguous directions to all TIA test-takers. Students were informed that they were to use information provided in the story and information they already knew in deciding upon the best answer. They were told to think about which answer out of four they thought was the best one. A sample item was done with the students. (For a copy of the complete instructions see footnote 3.) Special mention was made of the importance for students to consider all possible answers before deciding on the best one. Students were informed of the scoring system.

TIA is a power test, so no strict time limits were set. Students were told that it takes about a class period or so to do the test. Teacher reports of the final data collection indicated that most students finished the test in about 30-35 minutes, excluding time for directions (total time approximately 40-45 minutes). The intent was to allow students time to think and to carefully consider their answer choices without the pressure of a speed component. Test users may use the average completion time of 30 minutes as an indication of how their classes compare with others in time taken to do the test.

During the development of TIA, attention was paid to a host of simple but important rules for test construction which are in harmony with sound established measurement principles (Standards for Educational and Psychological Testing, 1985). Rules such as avoiding items with negative questions, using qualifiers cautiously such as "always" and "usually," and avoiding item stems similar to text information. Other factors which may have contributed to test-taking strategies were considered in the test refinement process and have been reported in a previous section.

Test-Wiseness

There is a sense in which test wiseness has to do with general wiseness or perceptiveness. Students may capitalize upon cues of various sorts which would result in improved performance on a test for reasons other than use of the ability being tested. Students' verbal report protocols were studied in each of the pilot studies for evidence of use of such cues. Test revisions were made if cues were suspected. Test revisions were discussed in a previous section.

Guessing

In the case of a short-answer test students guess only if they construct a response, which reduces the risk of attaining a higher score due to guessing. On the other hand, in the case of a multiple-choice test, the probability of attaining a higher score due to guessing is greater (Slakter, 1967).

The scaled-answer scoring system (scores of 0, 1, 2, 3) used on TIA further complicates the issue of guessing, because the probability of a student getting some positive score for each item on TIA is .75. Contrast this with the standard four-option multiple-choice test with one correct answer where the probability of getting a positive score is only .25 on each item. Given the unusual scoring system used on TIA, it would be instructive to examine the distribution of scores which could be obtained through guessing (Johnson & Kotz, 1977; Larsen, 1974).

Table 12 presents the distribution of total possible test scores (0-108) and their corresponding cumulative probability under the assumption of random guessing on each item. Note that since there is a considerable chance of scoring points through guessing it is virtually impossible to guess and receive less than about 40. Even for a total score of 54 which would be 50%, there is a 47% chance of attaining at least this high by guessing. However, if you look at a total score of 58, only 4 points higher, the chances of attaining a score of 58 or higher are dramatically reduced to 25%.

[Insert Table 12 about here.]

The overall mean score on TIA is 73.48 and as can be seen from Table 12 there is virtually no chance of a student getting this score or higher from guessing. Indeed the chances of getting any score of 60 or higher through guessing are quite low. The pattern of probability distributions displayed in Table 12 indicates that while scores up to about 60 can be expected to reveal very little about inference ability because of the guessing factor, scores above this point are virtually *unattainable* through guessing.

You will recall from the discussion of Pilot Study 5 in a previous section that there were highly significant correlations between reading scores and thinking scores (see Table 3) and that there were no significant differences in performance between the verbal report and the written response cohorts. These two points are worth mentioning here in terms of rounding out this discussion on guessing. The first point provides evidence that generally when students thought well they selected the best answer and that students who reasoned poorly selected an alternate answer. An examination of the verbal report protocols showed that despite the opportunity of having a best-answer option, students generally chose the answer that made most sense to them, the one that they could justify. It would seem then, that there was much more going on than guessing.

The second point that there were no significant differences in performance between the verbal report and written cohorts may point to a uniqueness in the nature of the task. Recall that Pilot Study 2 showed minimal differences in the amount of time taken by students to complete the multiple-choice and short-answer formats, suggesting that the reasoning demands of the task were similar. TIA is a test of inference ability, the ability to integrate relevant textual information and background knowledge and requires reasoning regardless of response format. In other words, determining the best answer on TIA requires making an inference regardless of the format of the test. This argument is very similar to one found in the area of mathematics where it has been argued that tests based on the same content but different formats require equivalent reasoning in test performance (Traub & Fisher, 1977).

Kuder-Richardson 20 Reliability Indices

Table 13 gives means, standard deviations, and KR-20 reliabilities for each story and for the total test. The Kuder-Richardson 20 reliability estimates are conservative; they give a lower bound estimate of reliability on a test. Nevertheless, it would be fair to say that TIA's reliability of .79 is highly satisfactory given the number of test items (36) and the reported reliabilities of similar tests requiring

students to reason well such as the following: Test on Appraising Observations (50 items) .69; Cornell Critical Thinking Test, Level X (71 items) .85; and the Watson-Glaser Critical Thinking Appraisal, Form A (80 items) .80.

[Insert Table 13 about here.]

Summary of Present Efforts and Future Prospects

This report has described the design and development of a test of inference ability in reading comprehension. It is a scaled-answer multiple-choice test intended for use with students in Grades 6, 7, and 8.

The Test of Inference Ability in Reading Comprehension is based upon what is currently known about inference and is in accord with the best available principles and information as described in a previous section. The principle of inference appraisal and the work reported here are not meant in any sense to be definitive, but rather are meant to be a chart in what is an uncharted testing area. It is to be seen as an important starting point open to extension. The objectives, design, and evolution of the test reported in the preceding sections represent a comprehensive methodology aimed at validly appraising inference ability in reading comprehension.

In order to have construct validity in tests of reading comprehension we must seek out the causes of performance on them. Responses on measures of reading comprehension may be correct or incorrect for very different reasons. Correct responses are not sufficient evidence of comprehension because sometimes they are the result of minimal reasoning. Conversely, incorrect responses are not sufficient evidence of a lack of comprehension, because sometimes they are a result of comprehension. Students' verbal reports and written explanations as to why they made their choices are ways to seek out causes of performance.

Central to this work are future prospects. The completion of a manual which will allow for diagnostic information for instructional decision-making purposes is the next immediate project. Diagnostic information will be reported in a manner that describes students' performance in terms of the quality of the inferences they have made, and the variations in inference ability across question types and across discourse forms. Such process-oriented information provides the necessary understanding of where students need instruction (Frederiksen, 1984), and to that end, specific teaching suggestions will be offered. The development of a short-answer form of the test which would allow for a more direct evaluation of the effect of background beliefs and levels of sophistication on students' performance is also planned.

Other prospects include studies to identify the kinds of strategies students' use in attempting to understand the various discourse forms, to measure the effectiveness of those strategies, and to explore the claim that reading well is thinking well by studying the relationships among inference strategy use, good inference-making in text comprehension, overall reading comprehension, and critical thinking performance.

A future prospect of a more collaborative nature is to study the seemingly multiple perspectives on the appraisal of inference ability through an examination of the types of inference demands made by the TIA test compared with those on current but more general comprehension assessment projects (Valencia & Pearson, 1987; Wixson, Peters, Weber, & Roeber, 1987).

References

Adams, M., & Bruce, B. (1980). *Background knowledge and reading comprehension* (Reading Ed. Rep. No. 13). Cambridge, MA: Bolt, Beranek, and Newman.

Amiran, M. R., & Jones, B. F. (1982). Toward a new definition of readability. *Educational Psychologist, 17*, 13-30.

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42*, 145-170.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. (1985). *Becoming a nation of readers.* Champaign, IL: Center for the Study of Reading.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255-292). New York: Longman.

Anderson, R. C., Spiro, R. J., & Anderson, M. C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal, 15*, 433-440.

Anderson, T. H., & Armbruster, B. B. (1984). Content area texts. In R. C. Anderson, J. Osborn, & R. J. Tierney (Eds.), *Learning to read in American schools: Basal readers and content texts* (pp. 193-224). Hillsdale, NJ: Erlbaum.

Applebee, A. N., Langer, J. A., & Mullis, I. V. (1987). The Nation's Report Card. *Learning to be literate in America: Reading, writing, and reasoning.* Princeton, NJ: Educational Testing Service.

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning.* New York: Grune & Stratton.

Beebe, M. J., & Phillips, L. M. (1980). Predictions and inferences in reading. *Elements, 11*(7).

Bereiter, C., & Scardamalia, M. (1982). From conversation to composition: The role of instruction in a developmental process. In R. Glaser (Ed.), *Advances in instructional psychology, volume 2* (pp. 1-64). Hillsdale, NJ: Erlbaum.

Bock, J. K., & Brewer, W. F. (1985). Discourse structure and mental models. In T. H. Carr (Ed.), *The development of reading skills: New directions for child development* (No. 27). San Francisco, CA: Jossey-Bass.

Brewer, W. F. (1980). Literary theory, rhetoric, and stylistics: Implications for psychology. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 221-239). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1964). *Language and thought.* Englewood Cliffs, NJ: Prentice-Hall.

Carver, R. P. (1975-76). Measuring prose difficulty using the Rauding scale. *Reading Research Quarterly, 11*, 660-685.

Cohen, J. L. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences, 4*(3), 317-331.

Collins, A., Brown, J. S., & Larkin, K. M. (1980). Inference in text understanding. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 385-407). Hillsdale, NJ: Erlbaum.

Coupland, N. (1978). Is readability real? *Communication of Scientific and Technical Information.* Washington, D.C.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Downing, J., May, R., & Ollila, L. (1982). Sex differences and cultural expectations in reading. In E. M. Sheridan (Ed.), *Sex stereotypes and reading: Research and strategies* (pp. 17-34). Newark, DE: International Reading Association.

Durkin, D. (1978-79). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly, 14*, 481-533.

Embretson (Whitely) S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Ennis, R. H. (1962). A concept of critical thinking. *Harvard Educational Review, 32*, 81-111.

Ennis, R. H. (1969). *Logic in teaching.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ennis, R. H. (1973). Inference. In H. S. Broudy, R. H. Ennis, & L. I. Krimerman (Eds.), *Philosophy of educational research.* New York: John Wiley & Sons, Inc.

Ennis, R. H. (1981). A conception of deductive logic competence. *Teaching Philosophy, 4*, 337-385.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 44-48.

Ennis, R. H., & Millman, J. (1985). *Cornell Critical Thinking Test, Level X.* Pacific Grove, CA: Midwest Publications.

Fagan, W. T. (1987). *Reading processes in reading comprehension.* Edmonton, Alta: Department of Education.

Farnham, G. L. (1905). *Sentence method of teaching, writing, and spelling.* New York: Harcourt, Brace.

Frederiksen, N. (1984). The real test bias. *American Psychologist, 39*(3), 193-202.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155-170.

Goodman, N. (1983). *Fact, fiction, and forecast.* Cambridge, MA: MIT Press.

Govier, T. (1985). *A practical study of argument.* Belmont, CA: Wadsworth Publishing Company.

Gray, W. S. (1940). *Reading general education.* Washington, DC: American Council on Education.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test ambiguity. In R. D. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Harman, G. (1986). *Change in view*. Cambridge, MA: MIT Press.

Henry, G. H. (1974). *Teaching reading as concept development*. Newark, DE: International Reading Association.

Hitchcock, D. (1983). *Critical thinking: A guide to evaluating information*. Toronto, CA: Metheun Publications.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Holland, V. M. (1981). *Psycholinguistic alternatives to readability formulas* (Tech. Rep. No. 12, Document Design Project). Washington, DC: American Institutes for Research.

Holmes, B. C. (1983). The effect of prior knowledge on the question answering of good and poor readers. *Journal of Reading Behavior, 15*(1), 1-18.

Holmes B. C. (1987). Children's inferences with print and pictures. *Journal of Educational Psychology, 79*(1), 14-18.

Holmes, B. C., & Roser, N. J. (1987). Five ways to assess readers' prior knowledge. *The Reading Teacher, 40*(7), 646-649.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. New York: Macmillan.

Johnson, N. L., & Kotz, S. (1977). *Urn models and their application*. New York: John Wiley & Sons, Inc.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 19*(2), 219-239.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Wadsworth Publishing Company, Inc.

Larsen, H. J. (1969). *Introduction to probability theory and statistical inference* (2nd edition). New York: John Wiley & Sons, Inc.

Manzo, A. (1970). Readability: A postscript. *Elementary English, 47*, 962-965.

Markman, E. (1981). Comprehension monitoring. In P. Dickson (Ed.), *Children's oral communication skills*. New York: Academic Press.

Mason, J. (1984). A schema-theoretic view of the reading process as a basis for comprehension instruction. In G. G. Duffy, L. R. Roehler, & J. Mason (Eds.), *Comprehension instruction* (pp. 26-38). New York: Longman.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 24*, 122-130.

McConaughy, S. (1980). *Developmental differences in summarizing short stories*. Paper presented at the Fourth Annual Boston University Conference on Language Development, Boston, MA.

Messick, S. (in press). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Norris, S. P. (1984). Defining observational competence. *Science Education, 68*, 129-142.

Norris, S. P. (1988). Controlling for background beliefs when developing multiple-choice critical thinking tests. *Educational Measurement, 7*(3), 5-11.

Norris, S. P., & Phillips, L. M. (1987). Explanations of reading comprehension: Schema theory and critical thinking theory. *Teachers College Record, 89*(2), 281-306.

Norris, S. P., & Ryan, J. (1987). Designing a test of inductive reasoning. In F. van Eemeren, R. Grootendorst, J. A. Blair, & C. A. Willard (Eds.), *Argumentation: Analysis and practices* (pp. 394-403). Dordrecht, The Netherlands: Foris Publications Holland.

Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Reading Behavior, 11*, 201-209.

Phillips, L. M. (1985). Categories of inference strategies. In W. T. Fagan, C. Cooper & J. Jensen (Eds.), *Measures for research and evaluation in the English language arts*, (Vol. 2, pp. 100-102). Urbana, IL: National Council of Teachers of English.

Phillips, L. M. (1986). Is the answer sufficient indication of reading comprehension? *Morning Watch, 14*, 15-20.

Phillips, L. M. (1987). *Inference strategies in reading comprehension* (Tech. Rep. No. 410). Urbana: University of Illinois, Center for the Study of Reading.

Phillips, L. M. (1988). Young readers' inference strategies in reading comprehension. *Cognition and Instruction, 5*(3), 193-222.

Phillips, L. M., & Norris, S. P. (1987). Reading well is thinking well. In N. C. Burbules (Ed.), *Philosophy of education 1986*, (pp. 187-197). Normal, IL: Philosophy of Education Society.

Phillips, L. M., & Patterson, C. C. (1987). *Phillips-Patterson Test of Inference Ability in Reading Comprehension*, St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland.

Plato (1917). *The education of the young in the Republic of Plato* (Translated with notes and introduction by B. Bosanquet). Cambridge, MA: Harvard University Press.

Reading, thinking and writing: Results from the National Assessment of Reading and Literature. Denver, CO: Education Commission of the United States, 1984.

Richards, I. A. (1938). *Interpretations in teaching*. New York: Harcourt Brace.

Rubin, A. (1981). *Conceptual readability: New ways to look at text* (Reading Ed. Rep. No. 31). Cambridge, MA: Bolt, Beranek and Newman.

Salmon, M. (1984). *Logic and critical thinking.* Orlando, FL: Harcourt Brace Jovanovich Publishers.

Scriven, M. (1976). *Reasoning.* New York: McGraw-Hill.

Singer, H. (1975). The SEER technique: A non-computational procedure for quickly estimating readability level. *Journal of Reading Behavior, 7,* 255-267.

Slakter, M. J. (1967). Risk taking on objective examinations. *American Educational Research Journal, 4,* 31-43.

Smith, F. (1971). *Understanding reading.* New York: Holt, Rinehart and Winston.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 275-290.

Spiro, R. J. (1977). Remembering information from text: The "State of Schema" approach. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 178-202). Hillsdale, NJ: Erlbaum.

Spiro, R. J., & Taylor, B. M. (1987). On investigating children's transition from narrative to expository discourse: The multidimensional nature of psychological text classification. In R. J. Tierney, J. Mitchell, & P. Anders (Eds.), *Understanding readers' understanding* (pp. 77-93). Hillsdale, NJ: Erlbaum.

Standards for educational and psychological testing (1985). Washington, DC: The American Psychological Association.

Stein, N. L. (1983). On the goals, functions, and knowledge of reading and writing. *Contemporary Educational Psychologist, 8,* 261-292.

Stich, S., & Nisbett, R. (1980). Justification and the psychology of human reasoning. *Philosophy of Science, 47,* 188-202.

Tamor, L. (1981). Subjective text difficulty: An alternative approach to defining the difficulty level of written text. *Journal of Reading Behavior, 13,* 165-172.

Thagard, P. (1978). The best explanations: Criteria for theory choice. *Journal of Philosophy, 75*(2), 76-92.

Thagard, P. (1982). From the descriptive to the normative in psychology and logic. *Philosophy of Science, 49*(1).

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8,* 323-332.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9,* 77-110.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*(3), 355-369.

Tuinman, J. J. (1973-74). Determining the passage dependency of comprehension questions in five major tests. *Reading Research Quarterly, 9,* 206-223.

Tuinman, J. J. (1986). Reading is recognition when reading is not reasoning. In S. de Castell, A. Luke, & K. Egan (Eds.), *Literacy, society, and schooling* (pp. 196-206). New York: Cambridge University Press.

Valencia, S., & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher, 40*(8), 726-732.

van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Vosniadou, S., & Ortony, A. (1983). The emergence of the literal-metaphorical-anomalous distinction in young children. *Child Development, 54*, 154-161.

Walker, C. H. (1987). Relative importance of domain knowledge and overall aptitude on acquisition of domain-related information. *Cognition and Instruction, 4*(1), 25-42.

Walker, C. H., & Yekovich, F. R. (1984). Script-based inferences: Effects of text and knowledge variables on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 23*, 357-370.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal.* Cleveland, OH: Psychological Corporation.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. D. (1987). New direction in statewide reading assessment. *The Reading Teacher, 40*(8), 740-754.

Author Notes

Special thanks go to Stephen P. Norris for providing insightful and constructive comments at different stages of this work. I thank P. David Pearson for his careful reviewing of an earlier version of the manuscript.

Footnotes

[1]A copy of TIA is available by writing to Linda M. Phillips, Institute for Educational Research and Development, Memorial University of Newfoundland, St. John's, Newfoundland, Canada A1B 3XB or by telephoning (709) 737-8625.

[2]The technical definition of "story" does not strictly apply to all three discourse forms on TIA, but then neither does the alternate term "passages," so I have decided to use story in the generic sense.

[3]Copies of all three tables and appendices are presented in the technical report which may be obtained by contacting the author.

Table 1

Pilot 2, Item Statistics

Item	Item/Total Correlations	Item Difficulty Index	Item	Item/Total Correlations	Item Difficulty Index
1	.448	.341	19	.340	.366
2	.528	.610	20	-.153	.366
3	.667	.317	21	.453	.488
4	.378	.293	22	.544	.561
5	.477	.512	23	.441	.780
6	.710	.512	24	.203	.171
7	.389	.683	25	.306	.732
8	.072	.195	26	.795	.902
9	.414	.488	27	.412	.512
10	.327	.463	28	.101	.537
11	.752	.829	29	.221	.488
12	.844	.805	30	.271	.585
13	.311	.098	31	.608	.488
14	.217	.561	32	.301	.463
15	.217	.561	33	.456	.585
16	.206	.244	34	.560	.463
17	.573	.293	35	-.316	.098
18	-.053	.07[illegible]	36	.505	.683

Table 2

Pilot 2, % Agreement Between Full Credit Short-Answer and Multiple-Choice Responses

Item	%	Item	%
1	51	19	68
2	46	20	11
3	28	21	40
4	32	22	33
5	50	23	75
6	39	24	54
7	16	25	50
8	11	26	10
9	28	27	31
10	59	28	8
11	47	29	42
12	79	30	0
13	18	31	53
14	18	32	49
15	59	33	62
16	27	34	63
17	40	35	11
18	43	36	45

Table 3

Pilot 5, Pearson Correlation Coefficients Between Reading and Thinking Scores by Item

Item	Pearson's r	Item	Pearson's r
1	.82**	19	.54**
2	.62**	20	.45*
3	.69**	21	.51**
4	.77**	22	.39*
5	.62**	23	.45*
6	.80**	24	.48*
7	.72**	25	.66**
8	.54**	26	.45*
9	.94**	27	.42*
10	.09	28	.53*
11	.68**	29	.72**
12	.50**	30	.37*
13	.56**	31	.61**
14	.52**	32	.42*
15	.30*	33	.38*
16	.52**	34	.83**
17	-.12	35	.54**
18	.77**	36	.61*

*p < .05
**p < .001
N = 95

Table 4

Pilot 5, Pearson Correlation Coefficients between Reading Scores and Thinking Scores by Story

Story	Pearson's r
'UFOs' Story 1	.77*
'Money' Story 2	.75*
'The Wrong Newspapers' Story 3	.77*

*p < .001

Table 5

Pilot 5, Story Reading Score Means by Cohort

	Story		
Cohort	1	2	3
Verbal Report	22.3	24.8	27.0
Written	22.2	23.6	24.2

Table 6

Pilot 5, Reading Score Means and Standard Deviations by Grade

Grade	*M*	*SD*
6	65.1	10.8
7	70.3	12.5
8	74.6	11.5
All grades	69.9	12.1

Table 7

Mean Scores by Sex, Grade and Age

Variable	*Mean*	*N*
Sex		
Male	72.57	518
Female	74.46	481
Grade		
Grade 6	70.77	324
Grade 7	72.99	330
Grade 8	76.47	344
Age		
11 Years	71.79	218
12 Years	73.29	299
13 Years	75.14	332
14 Years	72.85	126

Table 8

ANOVA Results on Final Data: Test Score by Sex, Grade and Age

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Sex	576.38	1	576.38	3.73	.054
Grade	2498.39	2	1249.20	8.08	.000
Age	2425.09	3	808.36	5.23	.001
Sex X Grade	1122.91	2	561.45	3.63	.027
Sex X Age	827.75	3	275.92	1.78	.148
Grade X Age	2595.26	4	648.81	4.20	.002
Within	141319.11	914	154.62		

Table 9

Percentage of Students Obtaining each Possible Score per Item

Item	Grade	Scores				Item	Grade	Scores				Item	Grade	Scores			
		0	1	2	3			0	1	2	3			0	1	2	3
1	6*	37	17	27	19	13	6	12	16	13	59	25	6	9	7	9	75
	7**	40	11	24	25		7	13	14	9	64		7	10	8	11	71
	8***	38	10	27	25		8	8	13	10	69		8	6	6	12	76
2	6	6	10	17	67	14	6	5	59	11	25	26	6	23	4	14	59
	7	5	11	12	72		7	5	69	8	18		7	22	5	15	58
	8	5	7	9	79		8	4	68	5	23		8	26	5	11	58
3	6	12	28	10	50	15	6	47	2	5	46	27	6	26	11	20	43
	7	13	28	7	52		7	34	6	8	52		7	27	11	17	45
	8	17	23	7	53		8	29	3	5	63		8	25	8	21	46
4	6	29	34	16	21	16	6	10	14	13	63	28	6	16	9	30	45
	7	26	29	17	28		7	6	10	17	67		7	17	9	28	46
	8	25	31	22	22		8	7	5	18	70		8	13	9	36	42
5	6	4	18	43	35	17	6	12	24	38	26	29	6	11	11	45	33
	7	3	17	33	47		7	9	30	37	23		7	11	8	40	41
	8	2	20	30	48		8	10	22	44	24		8	11	11	33	45
6	6	19	19	11	51	18	6	23	12	11	54	30	6	27	14	10	49
	7	13	18	7	62		7	22	10	9	59		7	24	15	10	51
	8	8	20	5	67		8	18	10	9	63		8	20	11	11	58
7	6	23	19	9	49	19	6	16	14	27	43	31	6	21	4	31	44
	7	15	17	8	60		7	18	14	25	43		7	18	5	31	46
	8	15	17	9	58		8	17	13	27	43		8	10	4	25	61
8	6	15	16	16	53	20	6	15	16	23	46	32	6	35	5	13	47
	7	18	14	16	52		7	11	20	20	49		7	34	5	13	48
	8	17	11	16	56		8	8	19	17	56		8	28	2	14	56
9	6	4	51	6	39	21	6	15	16	15	54	33	6	16	8	20	56
	7	1	43	5	51		7	13	12	11	64		7	20	8	21	51
	8	2	39	4	55		8	7	13	17	63		8	12	12	18	58
10	6	6	29	34	31	22	6	4	17	26	53	34	6	34	8	21	37
	7	6	28	35	31		7	3	17	27	53		7	27	8	22	43
	8	4	22	35	39		8	3	18	22	57		8	19	10	22	49
11	6	3	19	21	57	23	6	25	9	6	60	35	6	21	11	45	23
	7	3	16	19	62		7	22	8	4	66		7	23	12	40	25
	8	3	15	16	66		8	21	10	6	63		8	8	13	12	67
12	6	5	8	8	79	24	6	7	7	48	37	36	6	14	16	11	59
	7	3	9	6	82		7	5	11	47	37		7	19	11	11	59
	8	6	6	7	81		8	5	11	43	41		8	8	13	12	67

* 324 students
** 330 students
*** 344 students

Table 10

Item Statistics

Item	Item/Test Correlations	Item Difficulty Index	Item	Item/Test Correlations	Item Difficulty Index
1	.212	1.338	19	.179	1.943
2	.223	2.531	20	.281	2.101
3	.190	1.961	21	.388	2.241
4	.151	1.387	22	.371	2.302
5	.313	2.183	23	.162	2.078
6	.359	2.153	24	.295	2.169
7	.326	2.021	25	.403	2.501
8	.224	2.063	26	.172	1.825
9	.376	1.998	27	.195	1.825
10	.229	1.982	28	.325	2.501
11	.198	2.396	29	.427	2.072
12	.266	2.632	30	.388	1.906
13	.185	2.277	31	.478	2.145
14	-.010	1.469	32	.303	1.816
15	.222	1.766	33	.451	2.132
16	.263	2.414	34	.334	1.806
17	.055	1.781	35	.357	1.736
18	.220	2.053	36	.485	2.199

Table 11

Percentage of All Responses Receiving Scores of 0, 1, 2, and 3 by Grade Level and Story

	UFOs				Money				Newspapers			
Grade	0	1	2	3	0	1	2	3	0	1	2	3
6	14	22	18	46	16	17	19	48	21	9	22	48
7	12	20	16	52	13	18	19	50	21	8	21	50
8	13	20	16	53	12	17	19	52	16	8	23	53

Table 12

Distribution of Total Test Score Under Assumption of Random Response

Total Score	Cum %	Total Score	Cum %	Total Score	Cum %
1	2.12E-20	37	.667	73	99.8
2	1.49E-17	38	1.01	74	99.9
3	1.94E-16	39	1.50	75	99.9
4	1.93E-15	40	2.18	76	100
5	1.58E-14	41	3.10	77	100
6	1.11E-13	42	4.31	78	100
7	6.75E-13	43	5.88	79	100
8	3.68E-12	44	7.85	80	100
9	1.82E-11	45	10.2	81	100
10	8.23E-11	46	13.2	82	100
11	3.44E-10	47	16.7	83	100
12	1.34E-9	48	20.7	84	100
13	4.90E-9	49	25.2	85	100
14	1.68E-8	50	30.2	86	100
15	5.47E-8	51	35.5	87	100
16	1.69E-7	52	41.2	88	100
17	4.96E-7	53	47.0	89	100
18	1.39E-6	54	53.0	90	100
19	3.74E-6	55	58.8	91	100
20	9.64E-6	56	64.5	92	100
21	2.39E-5	57	69.8	93	100
22	5.71E-5	58	74.8	94	100
23	1.31E-4	59	79.3	95	100
24	2.92E-4	60	83.3	96	100
25	6.30E-4	61	87.0	97	100
26	1.31E-3	62	90.0	98	100
27	2.65E-3	63	92.0	99	100
28	5.21E-3	64	94.0	100	100
29	9.92E-3	65	95.7	101	100
30	.018	66	96.9	102	100
31	.033	67	97.8	103	100
32	.058	68	98.4	104	100
33	.100	69	98.9	105	100
34	.166	70	99.3	106	100
35	.271	71	99.6	107	100
36	.430	72	99.7	108	100

Table 13

Means, Standard Deviations, and KR-20 Reliabilities for Story and Total Test

Aspect	Mean	Standard Deviation	KR-20
UFOs (items 1-12)	24.65	5.37	.60
Money (items 13-24)	24.59	4.99	.49
Newspapers (items 25-36)	24.23	7.31	.77
Total Test (items 1-36)	73.48	13.50	.79

Appendix A

Reading Rating Scale for Test of Inference Ability in Reading Comprehension

The reading score is based upon the quality of the interpretation given by students.

TIA contains 36 items. For each item a student is assigned a score of 0, 1, 2, or 3 for a possible total of 108 points if a student made complete inferences on all items. For each item, reading will be rated according to the following scale.

Rating	TIA Reading Evaluation
3	The student integrates relevant text information and relevant background knowledge to construct complete interpretations that are consistent with both the text information and background knowledge. Thus, the student has given a complete inference answer.
2	The student integrates some text information and background knowledge but fails to take into account the available relevant information. The student's answer is consistent with some relevant text information and background knowledge but is incomplete. Thus, the student has given a partially-correct inference answer.
1	The student locates relevant text information but fails to integrate it with relevant background knowledge. Thus, the student has given a non-inference answer.
0	The student makes an inconsistent use of the text information and background knowledge. Thus, the student has given an implausible answer.

Each of the scale graduations are exemplified in the subsequent section. Examples are taken from selected items on TIA. Test item stems are provided with student answers in bold. Evaluator's comments are also provided.

3 points (complete inference)

The student integrates relevant text information and background knowledge to construct complete interpretations that are consistent with the text information and background knowledge. Students substantiate their answers with relevant evidence from both text information and background knowledge in a logical and coherent manner. Examples include the following:

(1) "Using available information people learn the most about UFOs by **examining the information in all the reports.**" In this example, the student recognizes the story says that scientists use three kinds of reports to study UFOs. The most information about UFOs would be attained by pooling all three, thus the preceding answer is consistent and complete with both text information and background knowledge.

(2) "Increased evidence is available today about UFOs than years ago because **we have more scientific equipment to study UFOs.**" In this example, the student uses the text information about potential use of weather cameras in satellites and increased knowledge of the universe to compare the present to the past. This information is then used to further reason that scientific equipment is an important factor in learning more about UFOs, thus the preceding answer is consistent and complete with both text information and background knowledge.

2 points (partially-correct inference)

The student gives an answer that indicates an integration of some text information and background knowledge but fails to take into account available, relevant information. The answer is consistent with some text information and relevant background knowledge, but is incomplete. Examples of partially-complete inference answers follow:

(1) "UFOs are sometimes called other names because **people don't know what to call them so they name them by shape.**" In this example, the student reasoned from some relevant text information but did not provide alternate interpretations. The student did not appear to monitor for consistency and completeness with available text evidence, that is, UFOs are called other names because of their probable origin, thus the preceding answer is only partially-correct.

(2) "Something unidentified in the sky may be called a UFO because **that is what people call it when they jump to conclusions.**" In this example, the student reasoned from some relevant text information but did not remain tentative. The preceding answer is a case where such a statement may be true, but it does not represent a complete interpretation of available, relevant text information pertaining to the question posed.

(3) "It is not known where UFOs come from, it seems they **could be from Earth because we have the materials and people to build such craft.**" In this example, the student reasoned from an unwarranted assumption to a justifiable alternative interpretation. The student has constructed an interpretation, but overlooks the fact that we do not know what UFOs are, so how could we construct them? On the other hand, the student may be thinking that some UFOs are misnamed spacecraft from Earth which makes the answer partially-correct.

1 point (non-inference)

It may be that the student did not understand the task, or that the student is more accustomed to non-inferential questions which often require the mere location of information than to inference questions which require the integration of relevant text information and background knowledge. In the latter case, students give an answer directly related to the text, or an answer which reflects minimal substantiation with the text evidence. Examples of non-inference answers follow:

(1) "UFO stories are very different from each other because **people sometimes think things like weather satellites, clouds, and bright stars are UFOs.**" In this example, the answer given by the student is directly from the UFOs story.

(2) "It is not known where UFOs come from, it seems they **couldn't be from almost anywhere because the story said they came from other planets.**" In this example, the student seemed to forego other possible answers for the sake of specific text information.

0 point (implausible answer)

It may be that the student did not understand the task, or that the student's answer is unsubstantiated. Examples of implausible answers follow:

(1) "UFOs are called other names because **people know they are unidentified flying objects in the sky.**" In this example, the student's answer is circular, it does not answer the item.

(2) "Some people think the study of UFOs should be continued because **some scientists think UFOs are not real.**" In this example, the student makes inappropriate use of text information. The text says,

"In 1969 one group of scientists concluded that there was not enough evidence to prove that UFOs are real and that UFOs are not worth further study." The answer given is implausible.

(3) "Something unidentified in the sky may be called a UFO because **that is the shape of whatever it is in the sky.**" In this example, the student was vague. The answer is unclear because it does not say what the shape is, nor specify what it is that is in the sky.

(4) "UFO stories are very different from each other because **people tend to exaggerate what they see and think of different names for UFOs.**" The student did not take into account available text information which says, "The weather, the time of day, and the number of people watching UFOs may make the stories different" in formulating an answer.

Appendix B

Thinking Rating Scale for Phillips-Patterson Test of Inference Ability in Reading Comprehension

One of the variables derived to appraise the quality of TIA is a thinking score. The score is based upon an analysis of verbal report interviews as students cited why they chose a particular answer as the best answer on the test.

TIA contains 3 stories each with 12 items. Students were asked to think-aloud on 1 of the 3 stories. For each item a student's thinking will be rated between 0 and 3 for a total of 36 points if a student thought well on all 12 items. For each item, thinking will be rated according to the following scale.

Rating	TIA Thinking Evaluation
3	The student cites all relevant textual and background information in the explanation of an answer choice. That is, the student considers the question and the available textual and background information pertinent to it in the formulation of a response which is complete and consistent.
2	The student cites some of the relevant textual and background information in the explanation. That is, the student considers either a part of the question and the available textual and background information pertinent to it, or the student considers the question and part of the available information pertinent to it in the formulation of a response which is consistent but incomplete
1	The student cites insufficient relevant textual and background information in the formulation of a response. That is, the student's response is not sufficient to indicate a clear understanding of either the question or the story. It is in need of elaboration and contains information which is partially correct and partially erroneous. However, it does reflect minimal integration of relevant information.
0	The student cites irrelevant or erroneous or repeats textual, background information, or both in the formulation of a response. That is, the student either misunderstands, misconstrues the story, or repeats the selected answer or textual information with no interpretation.

It is important to be cognizant of at least two precautions in rating students' thinking. The first is that this scale is fundamentally a measure of reasoning ability, not expressive ability. The goal is to focus on the quality of thinking in the verbal report interviews, rather than on the quality of effective speech.

The second precaution concerns the context of student justifications of answer choice on TIA. When students give a justification or what is intended to be a justification they do so having made an answer choice, so the story, the item stem, and the answer selected by the student create the context for the verbal report interview. In other words, when students tell why they selected a particular answer to be the "best" the context of a student response must be used in rating the quality of student thinking.

Each of the scale graduations are exemplified in the subsequent section. Examples are taken from student verbal report interviews for each of the 3 stories on TIA. Test item stems are in bold and student "best" answer choices complete the test item. Student interview comments are then presented, followed by an evaluator's comments.

Item 1 on the UFO story is evaluated as follows:

3 points

UFOs are sometimes called other names because people name them according to their shape or probable origin.

Student says, "I think that's the best answer because we do not really know what they (UFOs) are. UFOs are called 'flying saucers' and 'spaceships from other planets' and other names in the story, so since we don't really know what they are or where they come from, or if they come from anywhere, then people give them names that kinda makes sense according to their shape or where they might be from."

Shape and origin are the factors which must be inferred by a student from the textual information and background information in order to best explain why UFOs might be called other names.

2 points

UFOs are sometimes called other names because people don't know what to call them so name them by shape.

Student says, "I think that's the best answer because people don't know what to call them so they exaggerate about the shape, cause like they call them 'flying saucers'."

Shape was only one of the factors to be inferred about why UFOs might be called other names; the student's reported thinking reflects the use of some relevant textual and background information. The student did not incorporate the textual information "spaceships from other planets" and "extraterrestrial spacecraft" into reasoning for the best answer.

1 point

UFOs are sometimes called other names because people see an area with many coloured lights in the sky.

Student says, "I think that's the best answer cause it says it in the story" (student points to text).

An interviewer says "So, why are UFOs sometimes called other names?" and student responds by reading from the text:

"Stories have been told that UFOs light up an area with coloured lights and that creatures of different sizes and colours have been seen in them."

In this instance, the student failed to make the intended inference in response to the question but does seem to indicate awareness of related textual information. The selected response and reasoning are examples of insufficient use of relevant textual and background information.

0 point

UFOs are sometimes called other names because people know they are unidentified flying objects in the sky.

Student says "It's strange to see unidentified flying objects in the sky and mostly they are UFOs, which is a name they gave them."

In this instance, the student failed to answer the question nor to provide quality reasoning; indeed the student response begs the question.

Item 9 on the Money story is evaluated as follows:

3 points

The money system is different from the trade system because money gave things a standard unit whereas trade did not.

Student says, "I think that's the best answer because in the old days people could never be sure they got the same exchange value for their trade products. Like a cow might not really be the same value as say a piece of land but if one fellow wanted the land and the other fellow the cow, then they exchanged because they needed it even though the trades might not be fair. Money is better because say if a cow costs three hundred dollars, then people know if that is a good price, three hundred dollars is three hundred dollars no matter what you're buying, and with money there's change. Another thing is that cows might be worth different amounts."

To successfully answer this question a student must reason as to the differences between the money and trade systems by using both textual and background information, which the above student has done.

2 points

The money system is different from the trade system because one animal could be worth more in trading than the other.

Student says, "because one cow could be a good healthy cow giving milk and stuff but another cow might be old and sick, so they wouldn't be a fair trade for a piece of land. The cows and the land wouldn't be worth the same."

This response addressed only part of the test item. The student used textual and background information to cite the inequities of the trade system but did not speak to the money system.

1 point

The money system is different from the trade system because it might cost you ten cows for a piece of land in trading.

Student says, "because that's what it says up in the story." Upon further enquiry as to why the two systems are different, the student does not add further clarification.

This student response does not represent an integration of the textual with background information, the student merely offered related information without any indication of having reasoned through the question.

0 point

The money system is different from the trade system because money is what people use all the time so it is newer.

Student says, "Everybody knows what money is and how to use it."

In this particular instance, the student used background information to respond with an irrelevant statement which does not answer the item.

Item 3 on The Wrong Newspapers story is evaluated as follows:

3 points

Yesterday's paper was still laying in the puddle because something odd has happened.

Students says, "it seems from the story that Ann knew Mr. Jones was home and it was peculiar that he didn't pick up his paper. It couldn't have been the rain because Ann was able to deliver her papers, so something must have happened. Anyway the title is a clue that something strange has happened."

In this item, the student must incorporate the textual and background information to answer why the paper might be lying in the puddle.

2 points

Yesterday's paper was still lying in the puddle because the weather was too wet.

Student says, "more than likely it was too wet for Mr. Jones to come out to get it, so it was still in the puddle."

In this example, the student used only some of the relevant information to formulate a justification, and appears to have relied more upon background than upon an incorporation of both textual and background information.

1 point

Yesterday's paper was still lying in the puddle because it was in a plastic bag.

Student says, "It says in the story that's how Ann left the paper."

In this instance, the student did not incorporate the relevant textual and background information but the response indicates an awareness of related textual information.

0 point

Yesterday's paper was still lying in the puddle because it should have been picked up.

Student says, "because she knew that Mr. Jones was away and she knew that there was someone in the house and the paper should have been picked up."

In this case the student offered a subjective response, misread the textual information, and failed to answer the item.