CHAPTER 1
INTRODUCTION

The Test of Critical Thinking (TCT) is intended to assess critical thinking in students in grades three to five. The TCT is based theoretically on aspects of the Delphi Report (Facione, 1990a) and especially Paul's (1992) model of reasoning, specifically Paul's eight elements of thought. The TCT consists of ten short stories or text scenarios, each of which is followed by several multiple choice questions that require students to employ critical thinking, rather than reading comprehension skills, to select correct responses. The TCT presents a balanced framework of critical thinking elements within interesting stories that reflect seven important life-domains for children and adolescents (Bracken, 1993, 1996; Wasserman & Bracken, 2003), making it both useful and relevant to the lives of young students. The TCT is a comprehensive, highly reliable instrument with strong initial evidence of validity.

Rationale for the TCT

The TCT was conceptualized as primarily serving as a dependent measure for Project Athena a federally funded critical thinking-based research project for students in grades 3, 4, and 5. The TCT was developed because extant critical thinking tests had limitations of definition, appropriate age levels, required reading levels, administration time requirements, and fit to the Paul (1992) model of critical thinking, which made them less appropriate for the research study. Therefore, the author team decided to create an instrument that would be tailored to the age and ability levels of students participating in the Project Athena research study, and to the model of critical thinking used in the curriculum scaling-up project being evaluated.

Goals for the TCT

During the development of the TCT, the author team created and refined the instrument over a nine-month period with several goals in mind. The primary

goals guiding TCT development and refinement included a desire to accomplish the following:

- To create an instrument that would be theoretically and psychometrically sound.
- To create an instrument that would be brief and easy to administer during a typical academic class period.
- To create an instrument that would assess critical thinking, rather than simply reading comprehension.
- To create stories that would consider and reflect the important life-domains of children and adolescents.
- To create stories that would be interesting to young students.
- To create stories and items that would be sensitive to youth from all racial/ethnic, socioeconomic, and gender backgrounds.
- To write items that would thoroughly sample Paul's model of critical thinking.
- To write items that would be as succinct, clear, and concise as possible.
- To employ a scannable record form that would facilitate test scoring as well as creating and aggregating data files.

Applications of the TCT

The TCT was designed to be a comprehensive research assessment tool for the objective evaluation of critical thinking among children in the third through fifth grades. The TCT appears well-suited for use in academic contexts as a dependent measure because it was theoretically conceived, rigorously developed, and has been shown to be psychometrically sound, as well as brief and easily administered. The TCT can be administered individually or to groups

of students (i.e., entire classes). Also, the TCT can be either hand scored or scanned and scored by machine.

CHAPTER 2
SCALE DEVELOPMENT

The process of TCT development involved four major components. First, test development began with an attempt to define the construct of critical thinking. Second, a blueprint of critical thinking components was developed. Third, brief scenarios or short stories were created, using seven important life-domains for children and adolescents. Fourth, test development continued with the writing of specific items or questions to accompany each scenario, sampling Paul's elements of reasoning. Each of these components is described in more detail below.

Construct Definition

Critical thinking was operationally defined during the development of the TCT as a blend of the expert consensus of central critical thinking skills found in the Delphi Report (Facione, 1990a) and the "elements of reasoning" in Paul's (1992) model of critical thinking. The 46 panelists in the Delphi Method included representatives from the fields of Philosophy, Education, the Social Sciences, and the Physical Sciences who set out to define the central elements of critical thinking expected from students in the first two years of college. Six core cognitive skills were identified, each with two to three subskills. The six core skills include *interpretation, analysis, evaluation, inference, explanation*, and *self-regulation* (Facione, 1990a). These core skills were integrated with the Paul model of critical thinking to develop the combined model, depicted in Figure 2.1, that guided the TCT development.

The emphasis on the Paul model stems from its use in the intervention phase of the Project Athena research study. This scaled up curriculum, previously developed and researched for effectiveness in promoting language

arts learning (see VanTassel-Baska, Zuo, Avery, & Little, 2002), uses the Paul model as a major element of the curriculum framework.

In Paul's model, critical thinking involves eight key elements:

- Issue
- Purpose
- Concept
- Point of View
- Assumptions
- Evidence/Information
- Inferences
- Implications/Consequences

Effective critical thinking as proposed in this model involves reflection on these elements and their interactions within a given context, with reference to standards for thinking including such issues as clarity, logic, fairness, and relevance. Therefore, critical thinking involves both cognitive and metacognitive processes.

In the Project Athena curriculum, the Paul model is used in a variety of ways: to support general discussion of problems and issues, to discuss literature selections using the elements of thought as a scaffold, and as a basis for persuasive writing and issue-based research. The goal of this curriculum is to encourage students to develop skills for effective critical thinking both in academic contexts and in everyday life experiences.

In preparation for the development of the TCT, the project team reviewed literature on the Paul model, other critical thinking frameworks, and extant critical thinking assessments. The bibliography in this examiner's manual provides an overview of the sources explored. One of the more influential sources reviewed, along with the Paul model, was the Delphi report, which represents a consensus of expert opinion on critical thinking, its components, and its assessment (Facione, 1990a). The Paul model and the components of critical thinking identified in the Delphi report share several key features, including a clear reflection of the central position of making inferences or interpretations within

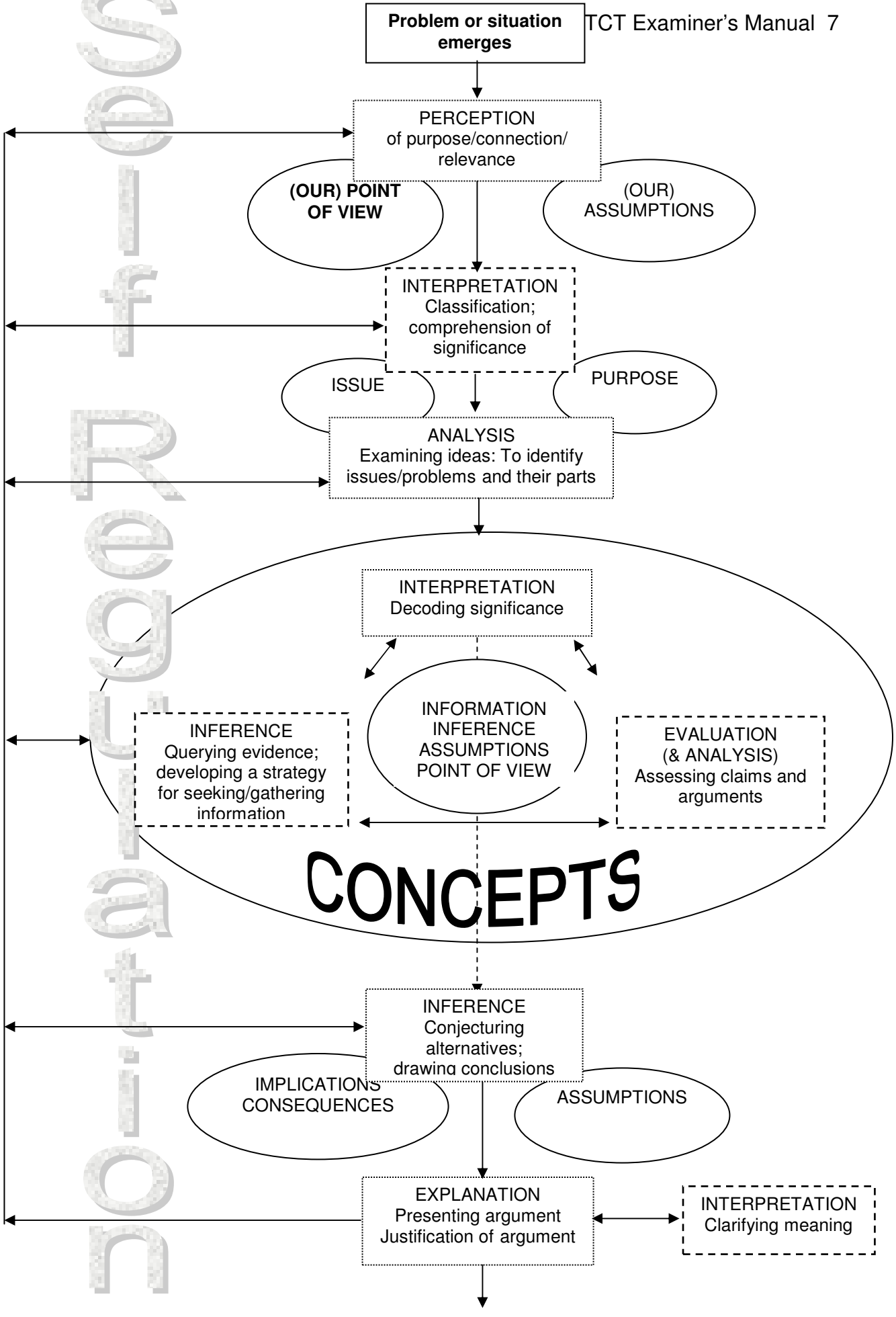
critical thinking. One distinction between the two models, however, is that although the Paul model places emphasis on the key elements *about which* critical thinking is done, the Delphi model emphasizes the key processes that take place during the stages of thinking. Thus, in some respects, the former model is more holistic and the latter is more linear, although recursive, in nature.

Based on this focused literature review, the project team developed an integrative model of the construct (see Figure 2.1) that blended the Paul model with the key components of critical thinking as defined in the Delphi report (Facione, 1990). According to this definitional structure, the TCT authors view critical thinking as the process of making reasoned judgments or inferences about issues or problems based on the evidence available, with recognition of the influence of point of view, assumptions, and context. The process involves an ongoing review of the critical elements and a metacognitive awareness of the interpretations being made along the way, as well as their implications.

In the figure, Paul's elements of thought appear in the ovals, sometimes repeated because of the need to consider specific elements throughout the process and in combination. The critical thinking components identified in the Delphi report appear in rectangles, representing the processes the thinker enacts upon the elements. The overall figure is linear yet recursive, demonstrating that a critical thinker recognizes a problem situation and its relevance, and then performs certain thinking actions related to the problem situation in order to reach conclusions or judgments that drive decisions, actions, and arguments. The central portion of the figure represents the spiraling process of reviewing information, making inferences, considering whether additional information is needed, and asking additional questions. The arrows throughout the figure, with the overlay of self-regulation, demonstrate that an effective critical thinker emphasizes metacognition, while constantly reviewing the process and determining which steps require revisiting.

On the TCT, students are posed questions (i.e., test items) that address each of the elements present in this model. Some questions reference perception of problems and their relevance, others require assessment of evidence or of

inferences made, while still others require identification of the concepts within which the thinking process is embedded. Given that the intent of the TCT is to assess critical thinking or the process of making reasoned judgments, all items on the TCT require inferential thinking, generally grounded within the central portion of Figure 2.1. However, items are identified in the blueprint based on the element of thought *about which* students are expected to make inferences.


Problem or situation emerges
PERCEPTION of purpose/connection/ relevance
(OUR) POINT OF VIEW
(OUR) ASSUMPTIONS
INTERPRETATION Classification; comprehension of significance
ISSUE
PURPOSE
ANALYSIS Examining ideas: To identify issues/problems and their parts
INTERPRETATION Decoding significance
INFERENCE Querying evidence; developing a strategy for seeking/gathering information
INFORMATION INFERENCE ASSUMPTIONS POINT OF VIEW
EVALUATION (& ANALYSIS) Assessing claims and arguments
CONCEPTS
INFERENCE Conjecturing alternatives; drawing conclusions
IMPLICATIONS CONSEQUENCES
ASSUMPTIONS
EXPLANATION Presenting argument Justification of argument
INTERPRETATION Clarifying meaning

Test Blueprint

The final form of the TCT consists of 45 items, arranged across ten brief scenarios. Each scenario is followed by 3 to 6 critical thinking items, each of which is classified according to one of Paul's (1992) elements of thought. Reflecting the consideration that critical thinking is relevant to everyday life, not just academic contexts, the TCT scenarios were developed across Bracken's (1993, 1996) seven important life-domains: *social, affect, competence, environmental, family, physical,* and *academic*. Each scenario is identified by primary life-domain of emphasis; several reflect both primary and secondary life-domains of emphasis. Table 2.1 illustrates the distribution of scenarios and life-domains sampled. In the table, **P** represents a "Primary" domain of emphasis, and **S** represents a "Secondary" domain of emphasis.

Table 2.1

Scenarios by Primary and Secondary Domain of Emphasis

	Social	Affect	Competence	Environmental	Family	Physical	Academic
STORY 1						**P**	
STORY 2	**S**			**P**	**S**		
STORY 3		**S**	**P**				
STORY 4	**S**				**P**		
STORY 5	**P**						
STORY 6			**P**		**S**		
STORY 7							**P**
STORY 8	**S**	**P**					
STORY 9	**P**						
STORY 10				**P**			

Scenario Development

All scenarios are presented in a consistent format; that is, a brief narration that features two to four characters that are faced with unique situations, sources of tension or conflict, or different points of view. Guidelines for scenario development included the following:

- Scenarios should require only limited background or previous knowledge on behalf of the reader.
- Sufficient information should be presented in each scenario to enable a critical thinker to draw conclusions and make inferences; however, each scenario should also include sufficient ambiguity to encourage consideration of multiple interpretations.
- Scenarios should be written to minimize the reading competence required of students and should reflect relatively low reading levels.
- Within the life-domains sampled, the scenarios should be relevant to the experiences of elementary grade students.

In total, more than 20 scenarios were drafted, edited, and identified by primary life-domain emphasized. Each scenario was reviewed extensively by members of the project team. A final set of ten scenarios was selected for piloting based on the quality of each individual scenario and on a consensus of the group, with attention paid to representation and balance of life-domains, gender emphases, and cultural diversity depicted within the story.

Item Development

Test items were developed to accompany each scenario. A multiple-choice item format was used for all scenarios, with four options presented per item (i.e., one keyed response and three foils). Each item assessed one of the specific

elements of thought identified in Paul's reasoning model (1992). Guidelines for item development included the following considerations:

- Items should require critical thinking rather than merely direct recall or simple literal comprehension of the scenario. All items were required to reflect inferential thinking.
- Item stems and options should be brief and easily read.
- Negative items (e.g., which of these is *NOT*…) should be used only on a limited basis.
- Lowest possible reading levels should be maintained for all items.
- Key terms should be used consistently to indicate elements of thought addressed and how students should assess the information. For example:
    - Items requesting students to review given evidence with great care should include the phrase, "Based on the story..."
    - Items requesting students to choose the best option from several plausible choices should include the phrase "MOST LIKELY" (or LEAST LIKELY).

Although the authors' initial intent was to develop items relating to each of the eight critical thinking elements for every scenario, it was found that each scenario lent itself more readily to questions involving some elements than others. Therefore, the team made a determination to address each element equivalently across the test, rather than across each scenario, as outlined in the Blueprint section.

Items were extensively reviewed and edited by the author team, with considerable discussion of the wording of and rationale for each item stem and options, based on the guidelines listed above and general issues of clarity.

Item Tryouts

Following the team's review of the complete set of scenarios and items, seven adults were asked to take a preliminary version of the test, respond to each item and provide comments regarding their reactions to the scenarios and questions. In addition, several individuals were asked to respond to the items *without* having the opportunity to read the scenarios first, to test for content dependency. Responses from both groups were then reviewed in detail by the project team, with the following major points of discussion:

- whether items answered correctly by most or all of the adult group would likely be too easy for children also
- whether items answered incorrectly by several adults were problematic (i.e., vague, confusing) or appropriately difficult, reflecting several plausible interpretations
- general discussion of comments from adults.

Based on the results of the team review and the responses from the other adults, the test was revised and reduced in length. A pilot version of the TCT containing 59 items across ten scenarios was compiled for try-out with students participating in a summer gifted enrichment program. Details of this pilot, which resulted in the 45-item TCT, are discussed in Chapter 4.

Scenarios were presented in the tryout version of the test in an order intended to provide content variety and therefore to sustain examinees' interest. Also, because it was deemed unlikely that all students would complete all 59 items, two forms that presented the same scenarios and items in reverse order were used to ensure adequate content coverage on the scale tryout. In the tryout and final versions of the instrument, no two scenarios with the same primary domain emphasis appear consecutively, and scenarios with a greater number of items are interspersed with scenarios containing fewer items.

All TCT test items correspond to one of eight elements of thought included in the Paul model of critical thinking. A description of the types of questions classified under each of the elements follows.

Issue: Issue items ask students to identify the central problem, question, or issue reflected in the overall scenario or in a specified portion of the scenario.

Purpose: Purpose items generally ask students the purpose of a character's behavior, to analyze why a character chose to take a certain action or perform a specific behavior.

Concept*:* Concept items generally address the major underlying ideas of a scenario. Concept items often reflect the primary or secondary domain in which the scenario is classified (e.g., a concept question in a Competence domain scenario might focus on aspects of a character's ability to achieve a task *competently*).

Point of View: Point of view items ask students to determine a character's perspective related to key issues in a scenario or to assess the differences in points of view between characters.

Assumptions*:* Assumption items emphasize the underlying beliefs that shape characters' decisions, speech, or actions. Some assumption items call upon students to identify the assumptions they themselves are making that allow certain interpretations of the scenarios.

Evidence*:* Evidence questions focus sharply on the specific information provided (or not provided) within the scenario. Some of these questions ask students to identify evidence that best supports a given conclusion, or to recognize that insufficient evidence is provided to support certain conclusions. Other evidence questions ask students to assess the influence of additional information, provided within the item, on their interpretation of the scenario.

Inference: All TCT items require some aspect of interpretation; as such they are all inherently inferential in nature. However, items classified as inference items ask students to infer conclusions based on the evidence presented in the scenario; many of these items ask students to determine which of the given options is "most likely" or "least likely" to be true, based on the story they have read.

Implication*:* Implication items ask students to determine what is likely to happen next *after* the given scenario, or to trace character decisions back to the conclusions that led to them. Some implication questions also ask students to determine the likely outcomes of possible events not presented in the scenario; some of these address events that might yet happen, while others ask students what *would have occurred* if something in the scenario had happened differently.

The final 45 TCT items were distributed across element categories such that five of Paul's elements are presented in six items each (i.e., Issue, Point of View, Assumption, Evidence, Inference) and three elements have five dedicated items each (i.e., Purpose, Concept, Implication). Table 2.2 illustrates the distribution of items across the scenarios and elements. Numbers in Table 2.2 refer to item numbers as they are presented on the TCT.

Table 2.2 ***Items in Scenarios by Element***

STORY	Issue	Purpose	Concept	POV	Assumption	Evidence	Inference	Implication
1	2			3			1	
2	4, 5	8	7, 9				6	
3		13	14	12		11	10	
4				16	15		17	
5	20				19	18		21
6	22	23	26		27		24	25
7		28	31	30		29		32
8	34			33		35		
9		40		39	36, 41	38		37
10					44	43	42	45
TOTALS	**6**	**5**	**5**	**6**	**6**	**6**	**6**	**5**

As noted previously, reading difficulty was a principal concern during the test development process, reflecting the authors' goal of assessing critical thinking rather than simply reading comprehension. The intent in the test development process was to establish an overall reading level near the lower end of the target population (i.e., third grade). Throughout the development process, the test was regularly reviewed and revised with word substitutions and reduction of compound sentences to simple sentences to ensure a lower reading level. The final version of the TCT was analyzed for reading level using the Flesch-Kincaid Grade Level readability formula. An analysis of the test yielded an overall reading level of 3.7. Analyses also were conducted for each segment of the test (i.e., a scenario and its accompanying items). These analyses yielded a range of reading levels from 2.4 to 5.3, as shown in Table 2.3.

Table 2.3 Reading Levels of TCT segments

Scenario	Reading Level
1	5.3
2	3.6
3	2.9
4	4.0
5	3.6
6	2.7
7	3.7
8	2.4
9	4.5
10	4.8

Chapter 3
Administration and Scoring

This chapter presents the materials, procedures, and directions for completing the TCT record form and administering the TCT. The TCT can be administered to one individual at a time, small groups of students (e.g., 5 to 10 students), or entire classes of students. Regardless of the administration option, the instrument is administered in the same manner.

Materials Needed: To complete the TCT, students will need a #2 pencil, the TCT Scantron Record Form, and a TCT Test Booklet. If the test user is not interested in machine scoring, the student can either respond in the test booklet or on a separate answer sheet. The examiner will need a watch or clock to time the administration.

Time: Students are allowed 45 minutes to complete the test. Timing begins ***after*** the examiner has read all instructions aloud to the students, and the two sample items have been presented, responded to, and discussed.

Administration:

Before reading directions, examiners should distribute the TCT Scantron Record Forms, pencils as necessary, and the TCT Test Booklets to students. Students should be arranged so they cannot easily look onto their neighbors' Record Forms.

The examiner will instruct the students on how to complete each of the demographic information elements at the top of the TCT Record Form. The students must complete all elements carefully and fully, including their names, student identification numbers, race/ethnicity, and gender, as well as their teacher's name and the current date (i.e., year and month).

The directions that are to be read to the students by the examiner are written below in ***bold italics.*** The students should read along with the examiner in their

own test booklets as the directions are read to them. Please keep in mind that it is essential that all instructions are to be read to the students **exactly as they are written** in this manual.

Oral directions - - *Say:*

> ***Please open your test booklets to page 1. Read the instructions silently as I read them aloud.***

After ensuring that the students have correctly opened their Test Booklets to the proper page, continue with the oral directions.

> ***Today, you are going to take a test called The Test of Critical Thinking. How well you do on this test will not affect your grade in this class. During the next 45 minutes, you will read some short stories. After you read each story carefully, you will answer some questions. Think carefully about each possible answer and choose the best one. Bubble in the answer on the answer sheet. Some questions ask you about what happened in the stories and some ask you what might happen. The stories and questions are like the sample question that we will do together. Let's look at the example on the next page. You may turn the page. Read the SAMPLE story silently while I read it aloud.***

Before continuing with the sample story and questions, be sure all students are prepared to continue and are on the correct page. Begin reading the story aloud:

> ***Nathan and Sean were in the same math class. Their teacher returned the tests she had graded. When they saw their grades, Nathan smiled, but Sean looked unhappy. The teacher said that many students had received low grades, and she hoped they would study more for the next test.***

Once you have completed reading the story, pause for a moment to allow time for the students to "absorb" what they have just read along with you.

> ***Read each sample question and bubble in the best answer on your Scantron answer sheet. Be sure to fill in the bubble completely. Do not circle or X the bubble.***

Allow the students 2 to 3 minutes to answer the two sample questions, and then say:

> ***Now let's look at the sample questions together. Please turn to the next page in your test booklet and follow along silently while I read the explanations aloud.***

Explanation of the Sample Scenarios

After ensuring everyone is prepared to consider the sample items, say:

Based on this story, what is MOST LIKELY to be true? Let's consider the options in Sample Item Number One.

A. Nathan received a better grade on the test than Sean did. This answer is INCORRECT. Nathan seemed happier with his grade than Sean did, but we do not know who actually received a higher grade. If Nathan usually receives C's, he might have received a B and been very happy. If Sean usually receives A's, he might be unhappy with an A-minus.

B. Nathan usually receives better grades than Sean in math. This answer is INCORRECT. We cannot tell from the story what grades these two students usually receive.

C. Sean had expected to do better on the test than he did. This answer is INCORRECT. We know Sean seems to be unhappy about his grade, but we do not know if he expected a better grade. Even if

Sean expected to do badly on the test, he might still have been unhappy with a low grade.

D. Sean did not do as well on the test as he would have liked. This is the CORRECT answer. Sean looked unhappy when he saw his test grade, so we can conclude that he most likely did not do as well as he would have liked.

After the students have considered the explanation, pause briefly and then proceed to Sample Item Number Two, and say:

What does the teacher believe?

A. Studying helps students do well on math tests. This is the CORRECT answer. The teacher said that many students had not done well, and she hoped they would study more for the next test. We can conclude from this statement that the teacher believes studying helps students do well on math tests.

B. Many students did not study for the test. This answer is INCORRECT. The teacher's statement suggests that she believes many students did not study enough, but not that they did not study at all.

C. None of the students studied enough for the test. This answer is INCORRECT. The teacher's statement suggests that she hopes the students who had not done well should study more. She did not say the students who had done well needed to study more.

D. Students cannot do well in math without studying. This answer is INCORRECT. The teacher's statement suggests that she believes studying more would help the students who did not do well to do better on the next test. But she may also believe that some students can do well in math without studying.

After completing and explaining both sample items, ask the students*:*

Are there any questions about the sample questions?

Respond as needed to the students' questions. When the students' questions have all been answered satisfactorily, make sure that everyone is ready to begin testing and say:

We are now ready to begin the test. Remember to read each story and each question carefully. Be sure to fill in the bubble for each answer on the answer sheet. You have 45 minutes to complete the test.

You may begin now.

As soon as you have told the students to begin the test, start timing for the 45-minute test administration. Time the test administration carefully, and after 45 minutes stop all students from responding further and collect all test materials.

As the students take the test, the examiner may provide general assistance to students, such as helping students pronounce or read a word or simply providing common definitions for individual words. It is important, however, that examiners understand that they are not allowed to help students answer test items. *Please, DO NOT help students interpret the stories or questions.* Respond to such student requests for assistance by saying something neutral, such as ***"Use your best judgment," "Do the best you can,"*** or ***"Do what you think is best."***

TCT Scoring

Examiners can hand score the TCT or establish a Scantron scoring protocol and machine score the instrument. Keyed responses are identified in Table 3.1.

Table 3.1

Scoring Key for the TCT

Item Number	Keyed Response	Item Number	Keyed Response
1	A	24	A
2	D	25	D
3	A	26	C
4	A	27	D
5	B	28	C
6	B	29	B
7	A	30	B
8	C	31	D
9	D	32	C
10	D	33	C
11	B	34	B
12	A	35	C
13	C	36	D
14	A	37	A
15	C	38	A
16	B	39	C
17	B	40	C
18	D	41	A
19	B	42	C
20	D	43	D
21	C	44	C
22	C	45	A
23	B		

Chapter 4
Technical Adequacy

The technical characteristics of the TCT were investigated in two phases. The first investigation took place during the piloting of the instrument and the second was conducted during its use as a dependent measure in Project Athena, a large scale curriculum intervention project.

Pilot Analyses

A total of 103 rising second, third, fourth, fifth, sixth, and seventh-grade students were administered the original 59 items of the TCT - - second, sixth and seventh grade students were eliminated from the analyses because Project Athena includes only grades three through five. Data presented in Table 4.1 depict the gender distribution of the 99 third, fourth, and fifth-grade students who participated in the study. The students were all participating in a summer enrichment program at The College of William and Mary in July 2003. Based on the students' performance on the instrument and subsequent data analyses (e.g., difficulty levels, discrimination indices, and contributions to scale reliability), the test was reduced to its final 45-item length.

Table 4.1

Demographic Characteristics of Students Participating in TCT Pilot Analyses

Gender	Grade 3	Grade 4	Grade 5	Totals
Males	16	25	20	61
Females	7	19	12	38
Total per Grade	23	44	32	99

RELIABILITY

Test reliability represents the percentage of variance that results from meaningful variation in test scores, as opposed to score variation that results from error. In group administered instruments, such as the TCT, measurement error can be caused primarily by (a) the undesirable aspects of an instrument (e.g., vague or poorly written items) or (b) unfavorable environmental conditions (e.g., distracting test-taking conditions). The authors of the TCT sought to reduce these deleterious testing effects by ensuring that the TCT was as easy as possible for examiners to administer, as easy as possible for examinees to take, and as brief as possible to reduce the effects of an adverse environmental setting. They also sought to ensure that the test was sufficiently well-written and objectively scored to produce reliable estimates of students' ability. These were the goals the authors hoped to achieve during the development of the TCT.

Internal consistency refers to the extent to which items within a test are positively correlated and contribute to the reliable variation of scores. One would expect that items drawn from the same content areas within an instrument would be positively correlated to a moderate degree. Bracken (1987) suggested (Wasserman & Bracken, 2003 elaborated) that tests intended for research applications should minimally be reliable at a level of .70, and preferably .80. This level of reliability (i.e., .70 to .80) was the goal that guided TCT item selection, retention, and test refinement.

Internal Consistency. The principal issue in TCT internal consistency is the extent to which individual items contribute meaningfully to the total scale score. To produce a measure with high internal consistency, all of the TCT items should correlate positively and moderately with each other and their combined total score.

Table 4.2 presents estimates of internal consistency (i.e., coefficient alpha) for the 45 item TCT total test score for the total sample. The total sample reliability is greater than the .70 to .80 criterion set by Bracken (1987) and Wasserman & Bracken (2003) for research instruments. Table 4.2 also presents the total sample reliabilities for grades 3, 4, and 5, with alpha for all grade levels exceeding the proposed criterion. Further data about internal consistency, derived from the scores of the larger sample of Project Athena students who took the test, is presented in later.

Table 4.2 TCT Total Scale Reliability Coefficients for Grades 3, 4, and 5, Boys and Girls, and Total Sample

Group	Internal Consistency (Coefficient Alpha)
Grade 3	.85
Grade 4	.83
Grade 5	.87
Males	.88
Females	.87
Total Sample	.89

Standard Error of Measurement. The standard error of measurement (SEM) of a test is directly proportional to the reliability of the instrument. From its computational formula, it is apparent that as a test's reliability increases, its SEM decreases. This functional relationship further indicates the importance of reliability and its contribution to the confidence an examiner has that an examinee's obtained scores are representative of the individual's theoretical "true

scores." Because the TCT SEMs are quite small (approximately 3 raw score points), estimated true scores lie in fairly tight bands of confidence.

Scoring Reliability. Because the TCT is objectively scored by either hand or Scantron methods, scoring reliability is not an issue that needed to be investigated. Scoring reliability is assured on the TCT, except to the extent that there exists the possibility of occasional clerical or mechanical scoring errors - - minor or occasional scoring errors that exist in all objectively scored tests.

VALIDITY

A test is considered valid to the extent to which it measures what its authors claim it measures. The TCT is a test that its authors propose assesses critical thinking. Given this basic definition of validity and the focus of the TCT, preliminary TCT validity is demonstrated through four general methods - - content validity, item content dependence, age/grade progression, and total test ceilings and floors.

Content Validity. The TCT was based theoretically on Paul's model of critical thinking and the model derived from the Delphi Method (Facione, 1990a). The test blueprint depicted in Chapter 2 illustrates the extent to which the test items were constructed to match Paul's "elements of reasoning." All elements within the model were uniformly assessed across the TCT's ten scenarios. Additionally, to add social relevance and contribute to content validity, each of seven important life elements were represented in the stories throughout the instrument (i.e., academic, affect, competence, environmental, family, physical, and social).

Content Dependence. During the development of test items, several individuals were asked to answer the preliminary test items without having had the benefit of reading the stories. The reason for this exercise was to investigate whether students could answer the questions simply by guessing or by using test-taking strategies. Each of the items used in the final TCT demonstrated sufficient evidence of content dependence to be included in the final instrument.

Age/Grade Progression. Ability tests should demonstrate the differential cognitive development that exists among children of different ages or grades. That is, for the TCT, older students would be expected to answer more items correctly on the TCT than younger students due to the more advanced cognitive development of the older students (assuming neither group had received training in critical thinking). Table 4.3, which provides the means and standard deviations of TCT scores for students in grades 3, 4, and 5, illustrates that older students do in fact obtain higher scores than younger ones, which supports the validity of the instrument. Further, it demonstrates that the TCT standard deviations are quite consistent across the three age levels (i.e., there is comparable variability and similar distributions of scores).

Table 4.3 Means and Standard Deviations of the TCT Total Scale for Boys and Girls, Grades 3, 4, and 5, and Total Sample (Maximum Score = 45)

Group	TCT Mean	TCT SD
Grade 3	16.35	7.60
Grade 4	22.81	7.33
Grade 5	27.72	8.37
Boys	21.18	8.66
Girls	25.73	8.26
Total Sample	22.90	8.75

Total Test Ceilings and Floors. The TCT was piloted on students who were previously identified as intellectually gifted, which allows a reasonable estimate of the strength of the instrument's ceilings and floors for high ability students. Given the means and standard deviations of the total test score across the age levels, it is apparent that the test has a strong floor for third-grade high ability students - - one that is greater than -2Z. Similarly, for high ability fifth-grade students the TCT demonstrates ceilings that exceed +2Z. Given these ceilings and floors, the TCT exhibits sufficient range of difficulty to be an appropriate measure for lower functioning third-grade students and very gifted fifth-grade students.

Further Assessment of Technical Adequacy Based on Administration of the TCT to a Large Sample of Students in Project Athena, a Curriculum Intervention Study

Project Athena is a language arts curriculum scale up and implementation study for third through fifth grade students in Title I schools. Project Athena is a federally funded (i.e., A Department of Defense, Jacob Javits funded program for gifted education). Beginning in September, the participating school divisions administered several pre-assessment measures, including portions of the *Iowa Test of Basic Skills (ITBS),* portions of the *Cognitive Abilities Test (CoGAT)*, portions of the *Universal Nonverbal Intelligence Test (UNIT),* the *Test of Critical Thinking (TCT)* and performance-based measures of literary analysis and persuasive writing that are embedded in each curriculum unit. Following the pre-test administrations, teachers in experimental classrooms implemented the relevant William and Mary Language Arts Units. Each unit encompasses 24 lessons and required specialized training for fidelity of implementation. During the months in which teachers were either implementing the William and Mary units or their school district's curriculum, classroom observations were conducted using the *William and Mary Classroom Observation Scale-Revised (COS-R)* to determine the effectiveness of teacher and student behaviors. Following the implementation cycle, post-assessment measures were administered, again including portions of the ITBS, the TCT, and the performance measures in the curriculum units.

As noted earlier, the TCT was developed to be used as a pre and post-assessment instrument in Project Athena. The administration of the TCT in this

context provided the following additional information on the characteristics of the instrument.

RELIABILITY

TCT Stability. The stability of a test is an important psychometric consideration, especially when educational interventions are planned. That is, researchers want to be able to show that test scores change as a result of intervention rather than as a result of the instability of the instrument. A test should evidence short-term stability at the same approximate level as its internal consistency, though stability is generally a more difficult standard to achieve and is usually less robust than internal consistency.

Although students in experimental groups are generally considered to not be good subjects upon which to assess stability (i.e., they are experiencing an intervention that is intended to make the instrument less stable by improving students' scores), control students make ideal subjects for stability investigations (i.e., they are not receiving interventions designed to alter their scores). During the pre- and post-testing phases of the TCT for Project Athena, the control students' scores across a semester were contrasted to investigate the approximately 3-month stability of the TCT. It should be recognized, however, that even the control students were experiencing an intervention during this period, albeit their standard language arts classroom instruction.

Table 4.4 presents the TCT mean scores for third, fourth, and fifth grade control group students on the fall pre-test administration and late winter post-test administration. The overall correlation between pre and post-test scores for the control students was .66, which is the stability coefficient for the test. It should be noted that there was significant gain for each grade level in students' critical thinking skills over the approximate six-month test interval, which suggests that both the students' standard curriculum and their natural cognitive development appear to have increased the critical thinking skills in this control sample.

Table 4.4. Comparison of Pre and Post-Test TCT Means by Grade Level

Grade	N	Pre-TCT	Post-TCT	*t-obtained*	Sig.
3rd	142	13.99 (6.17)	18.07 (5.78)	9.55	.000
4th	140	15.84 (6.06)	18.82 (6.67)	5.96	.000
5th	159	19.48 (6.95)	22.38 (6.98)	6.47	.000

Stability coefficient : .66 (N=517)

Internal Consistency: Table 4.5 presents estimates of internal consistency (i.e., coefficient alpha) for the TCT pretest for the total sample, as well as for students in each grade level and for male and female students. The total sample reliability, .83, exceeds the .70 to .80 criterion set by Bracken (1987) and Wasserman & Bracken (2003) for research instruments. Total sample reliabilities for grades 3, 4, and 5 and for male and female students also meet or exceed this criterion. These values suggest that researchers can expect the TCT to provide quite consistent estimates of children's critical thinking. These findings also indicate that across the three grade levels, 80% to 85% of total scale variability is reliable or true, and only 10% to 15% of the variability is due to error. Based on these preliminary total sample estimates of internal consistency, examiners can expect to use the TCT as a dependent measure in research with considerable confidence.

Table 4.5 Total Scale Reliability Coefficients for Grades 3, 4, and 5, Boys and Girls, and Total Sample for TCT Pretest

Group	Sample Size (N)	Internal Consistency (Coefficient Alpha)
Grade 3	379	.80
Grade 4	374	.82
Grade 5	472	.81
Males	377	.83
Females	414	.85
Total Sample	*1259	.83

*Note: The number of students in the total sample does not equal the sum of male plus female students because some students did not indicate their gender. Also, the total N does not equal the sum of the three grade samples because some students were in a combined 3rd through 5th grade class, which was not assigned a grade level.

Effect of Gender on TCT Performance: In order to investigate whether a student's gender is related to his/her score on the TCT, an ANOVA was conducted to compare mean pre-test mean scores for boys and girls across the entire sample of Project Athena. The results, displayed in Table 4.6 highlight a significant gender difference between the groups, with girls scoring significantly higher on the TCT than boys. This gender difference yields a small effect size, however, and is not an important finding (i.e., effect size is approximately .20).

Table 4.6. Comparison of Male and Female Scores on the TCT Pretest

Gender	N	Mean	SD	F	Sig.
Male	551	16.40	6.74	10.69	.001
Female	598	17.77	7.42		

Influence of Ethnicity on TCT Performance: Similarly, to investigate whether a student's race or ethnic heritage is related to his/her score on the TCT, an ANOVA was conducted to compare mean scores of students from the following ethnic groups: African American, American Indian or Alaska Native, Asian (including Native Hawaiian or other Pacific Islander), Hispanic, and White. The results, displayed in Table 4.7, demonstrate that there was a significant effect of ethnicity, and post hoc analyses (i.e., Tukey HSD) revealed the following pattern of significant differences between groups: Whites=Asians>Hispanics=African Americans. There were no significant differences found between mean scores of the American Indian/Alaska Native group and any other group, which is most probably due to the very small number of students in that group.

Table 4.7. Comparison of TCT Pretest Scores of Students from Different Ethnic Groups

Ethnicity	N	Mean	SD
African American	303	15.44	6.53
American Indian or Alaska Native	9	18.11	5.16
Asian	41	18.07	5.69
African American	303	15.44	6.53
Hispanic	169	14.17	5.64
White	540	18.81	7.52
Other	35	15.60	6.68

Relationship of TCT Scores to Scores on Other Cognitive

Instruments: Table 4.8 depicts the correlation coefficients between the TCT pre-test and the following other pre-assessment instruments:

- Universal Nonverbal Intelligence Test (UNIT) - - Abbreviated Battery
- Iowa Test of Basic Skills, Reading (ITBSRD)
- Iowa Test of Basic Skills Language Arts (ITBSLA)
- Cognitive Abilities Test – Verbal (CogAT-V)
- Cognitive Abilities Test – Non-Verbal (CogAT-NV)

Table 4.8. Pairwise Correlation between Student Assessment Measures: Whole Sample (N=732-1171)

	TCT Pre-test
UNIT	**.25****
ITBSRD-Pre	**.63****
ITBSLA-Pre	**.47****
CogAT-V	**.63****
CogAT-NV	**.45****

**** *p* < .01**

The above table demonstrates that the TCT has a moderately low correlation (.25) with the abbreviated version of the UNIT, which is consistent with the fact that the TCT is a test requiring reading comprehension and verbal critical thinking skills while the abbreviated UNIT is totally nonverbal, with no verbal directions, item presentation or item response. The correlation of the TCT is somewhat higher for the CogAT Non-Verbal (.45), which is consistent with the

fact that that all portions of the CogAT includes verbal instructions. The correlation between the TCT pre-test and CogAT Verbal subtest is higher still (.63), reflecting the overlap between that test's verbal directions and skills assessed by that test and the reading comprehension and verbal critical thinking skills assessed by the TCT.

The correlations of the TCT with the Iowa Test of Basic Skills also demonstrate a higher correlation with the Reading subtest (.63) than with the Language Arts subtest (.47). Again, because the TCT requires reading comprehension and verbal critical thinking skills, it is not surprising that it would correlate to a higher degree with a reading subtest than a subtest that also includes mechanics of language.

Conclusion

The TCT, developed as a dependent measure for Project Athena, a "scaling up" curriculum intervention study, assesses verbal critical thinking skills. Administration of the test to a large sample of third through fifth grade students has documented its technical adequacy and relationship to a number of other cognitive assessments. The TCT provides a convenient, reliable, and valid means of conducting individual or group assessments of critical thinking for elementary school students.

REFERENCES

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment. 5,* 313-326.

Bracken, B. A. (1993). *Multidimensional Self Concept Scale*. Austin, TX: Pro-Ed.

Bracken, B. A. (1996) (Ed.). *The handbook of self-concept: Developmental, social, and clinical considerations.* New York: Wiley.

Costa, A. L. (Ed.). (2001). *Developing minds: A resource book for teaching thinking* (3rd ed.). Alexandria, VA: ASCD.

Ennis, R. H. (1996) *Critical thinking*. Upper Saddle River, NJ: Prentice-Hall.

Erwin, T. D. (2000). *The NPEC sourcebook on assessment (Volume 1): Definitions and assessment methods for critical thinking, problem solving, and writing.* Retrieved on 8/7/03 at http://nces.ed.gov/pubs2000/2000195.pdf.

Facione, P. A. (1990a). *Critical Thinking: A statement of expert consensus for purposes of educational assessment and instruction (Executive summary: "The Delphi Report.")* Milbrae, CA: California Academic Press. Retrieved on 8/7/03 at http://www.insightassessment.com/pdf_files/DEXadobe.PDF.

Facione, P. A. (1990b). Thirty great ways to mess up a critical thinking test, *Informal Logic*, 12, 106-111.

Fasko, D. (Ed.). (2003). *Critical thinking and reasoning: Current research, theory, and practice.* Cresskill, NJ: Hampton Press.

Fawkes, D., Adajian, T., Flage, D., Hoeltzel, S., Knorpp, B., O'Meara, B., & Weber, D. (2001). Examining the exam: A critical look at the Watson-Glaser Critical Thinking Appraisal Exam, *Inquiry, 20*, 19-33.

Halpern, D. F. (1997). *Critical thinking across the curriculum: A brief edition of thought and knowledge.* Mahwah, NJ: Erlbaum.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking* (4th ed.). Mahwah, NJ: Erlbaum.

Levy, D. (1997). *Tools of critical thinking: Metathoughts for psychology.* Boston: Allyn & Bacon.

Nosich, G. M. (2001). *Learning to think things through: A guide to critical thinking in the curriculum.* Upper Saddle River, NJ: Prentice Hall.

Paul, R. (1990). *Critical thinking handbook (K-3): A guide for remodelling lesson plans in language arts, social studies & science*. Rohnert Park, CA: Foundation for Critical Thinking.

Paul, R. (1992). *Critical thinking: What every person needs to survive in a rapidly changing world*. Rohnert Park, CA: Center for Critical Thinking and Moral Critique, Sonoma State University.

Paul, R. (1995). *Critical thinking handbook (6th-9th grades): A guide for remodelling lesson plans in language arts, social studies, & science*. Rohnert Park, CA: Foundation for Critical Thinking.

VanTassel-Baska, J., Zuo, L., Avery, L. D., & Little, C. A. (2002). A curriculum study of gifted student learning in the language arts. *Gifted Child Quarterly, 46,* 30-44.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric considerations of assessment procedures. In J. Graham and J. Naglieri (Eds*). Handbook of assessment psychology (pp. 43 – 66).* New York: Wiley.