

Assessing Reasoning Skills

How can a person's reasoning skills be assessed? For example, how can a prospective employer assess an applicant's ability to integrate complex sets of information into meaningful interpretations and identify the implications? In responding to a student's concern about a course, how can an adviser ferret out the assumptions that may be guiding the comment? How can professors who aspire to teach the higher-order thinking skills associated with reflective thinking document their success in doing so? In hiring a resident adviser, how can a hall director evaluate a student's ability to analyze adequately the complex and ambiguous problems that typically arise in the residence hall? How can institutions show prospective parents or legislators that they teach students to think critically? Questions such as these drive the collective need to consider the assessment of reflective thinking.

There are many approaches to assessing "good thinking," whether defined as reasoning skills, logical reasoning, formal reasoning, reflective judgment, problem solving, or critical thinking. These approaches typically represent the specific concepts, elements, or strategies that are central to the particular model of good thinking that a given instrument purports to measure.

The attributes of good thinking measured by these instruments vary widely, from the ability to reason deductively to proficiency in identifying underlying assumptions in an argument. Since the definition of good thinking usually determines (or at least strongly influences) which attributes are measured, it is important for a prospective user of any test of critical thinking to consider carefully whether the definition underlying the measure is consistent with the user's evaluative or educational needs.

In addition to definitional differences, existing measures of critical or reflective thinking also differ in the extent to which they focus on well- or ill-structured problems. (Problems that can be defined completely and solved with certainty are referred to as well structured. See Chapter One for a discussion of these terms.) Some problems focus on strictly logical, deductive reasoning. (We do not mean to imply that all logic problems are well structured. This definition applies to logic problems based on propositional calculus, which is decidable and complete, but not — as Church and Gödel have shown [Rundle, 1972] — to those based on the predicate calculus, which is undecidable and incomplete.) Such problems provide a description of a situation that includes all relevant information necessary for correctly answering the question posed, such as the following: All military men tell the truth. John, a military man, says he was engaged by the president to sell guns to terrorists. The president has been accused of conspiring to sell guns to terrorist nations. On the basis of the statements given above, decide whether the president committed this act.

When the basic elements of the problem are provided (even if finding them requires careful reading) and the problem can be solved with certainty, the problem qualifies as a well-structured one. In well-structured problems, logical reasoning skills such as assumption identification, induction, and deduction are emphasized.

A second kind of well-structured problem involves sorting through and correctly applying complex algorithms. For example: A ship leaves port at noon, sailing north at thirty knots. A second ship leaves port an hour later, sailing east at twenty-five knots. At 3 P.M., how fast is the distance between them changing?

Other kinds of well-structured problems may involve such reading and analysis skills as identifying generalizations that follow from stated evidence and detecting fallacious arguments. It is important to note that well-structured problems may range from very simple to very complex and that they can require the application of skills and knowledge from a variety of disciplines. The unifying principle underlying a well-structured problem is that, given the parameters of the problem, a single answer can be determined to be correct. When the educational goal is to teach this type of reasoning, then a critical thinking test that relies heavily or exclusively on well-structured problems would be appropriate and useful.

However, many major national reports have argued for instructional goals that require a very different type of reasoning. For example, the National Commission on Excellence in Education (1983, p. 7) notes that "For our country to function, citizens must be able to reach some common understandings on complex issues, often on short notice and *on the basis of conflicting or incomplete evidence*" [italics added]. More recently, the Association of American Colleges (1991, p. 14) has called for students to be able "to state why a question or argument is significant and for whom; what the difference is between developing and justifying a position and merely asserting one; and how to develop and provide warrants for their own interpretations and judgments." The skills of deduction and induction are not sufficient for the types of analysis and synthesis called for in these reports, and assessments of this type of reasoning cannot rely on questions that assume a single correct answer. By contrast, assessment of this type of reasoning requires measures that include ill-structured problems in which, by definition, the evidence is conflicting or incomplete.

In this chapter, we first propose a list of desirable features of a measure of reflective thinking that may be used to guide the consumer's evaluation of existing measures or the researcher's development of new measures of reflective thinking. We then examine a range of available measures of critical thinking (broadly defined) and evaluate the extent to which they provide information about the kind of reasoning described in the

national reports cited above, that is, how people reason about ill-structured problems. The Reflective Judgment Interview, which was expressly designed to measure reasoning about ill-structured problems, is discussed separately in Chapter Five.

Desirable Features of a Measure of Reflective Thinking

How should reflective thinking be measured? What kind of information do educators, institutional researchers, administrators, personnel officers, and others interested in conducting formal assessments of reflective thinking need? Do existing instruments adequately measure reflective thinking? One way to answer the first of these questions is to attempt to generate a list of desirable features of an instrument designed to measure reflective thinking.

For better or worse, the practical realities of conducting assessments and research are often given more consideration than the conceptual soundness and demonstrated reliability of potential instruments. As a result, the first feature many look for is that the instrument be inexpensive, amenable to large-scale administration, and easy to score. While these are indeed attractive characteristics, they should not be used as the most important criteria in selecting or evaluating a measure of reflective thinking. The risk of using "quick and dirty" instruments is that they may quickly yield superficial, irrelevant, unreliable, or uninterpretable data. If the assessment of reflective thinking is really important, then it must be measured carefully. While doing so may take longer and cost more, the resulting data are more likely to be appropriate for answering the kinds of important questions listed at the beginning of this chapter.

Several criteria for evaluating a measure of reflective thinking, as defined in Chapter One, are suggested below. (The need for an instrument with adequate reliability and validity will be assumed rather than discussed here.) The following section describes other desirable features to consider when evaluating or constructing a measure of reflective thinking.

1. *Ill-structured problems should be the focus of the assessment task.*

Dewey (1933) made the astute observation that only when

people are faced with **real problems** do they **engage in true reflective thinking**. These problems involve doubt about the adequacy of one's current understanding of the issue. Using this definition, well-structured problems will not tap a person's ability to think reflectively and thus are not appropriate. Consequently, **one critical component in a measure of reflective thinking is the degree of problem structure in the questions it poses.**

2. *The assessment should elicit information about the rationale for a response as well as its content.* This criterion is based on the understanding that different people can arrive at the same answer (that the Egyptians built the pyramids) for a variety of reasons. **The reasons offered may indicate the sophistication of the respondent's thinking.** For example, some of these reasons may relate directly to the evidence (the presence of ancient Egyptian artifacts in the pyramids) or to the source of information (a popular magazine, friend, or parent). In other words, a measure of the quality of reflective thinking will need to attend to the nature of the justification for the judgment as well as to the judgment itself. As a result, the response format will need to allow the respondents to choose from a variety of justifiable answers (a point also made by McPeck, 1981) or to generate their own answers.

Ennis and Norris (1990) discuss the effects of various test formats in terms of the need to understand individuals' justifications for their decision. For example, they note that machine-scorable tests "ask for judgments but do not also seek the reasoning behind the judgments" (p. 25). Essay tests, by contrast, "can easily ask a student to justify the judgments that the student makes. Informal interviews, though very expensive, are even better, because they enable the tester to seek more justification whenever anything seems unclear. Essay tests only invite justification, but do not provide for further interaction when the student's justification itself is in need of clarification or justification. Interviewing may be the best approach if high quality information is needed on a small number of individuals" (p. 27).

3. *Thinking should be sampled across a variety of issues.* This allows for the evaluation of the consistencies and inconsistencies in the participants' thinking across content areas. Sampling across issues also helps minimize the effects of prior familiarity with one topic, discipline, or area of study.
4. *The content should be generally familiar to a wide range of individuals.* This allows for testing across diverse samples of people who have diverse educational experiences. It also allows participants to respond to issues about which they have some working knowledge rather than issues with which they are unfamiliar. This may have the added benefit of increasing their comfort with the task. (This suggestion does not apply to measures that are specifically developed for administration to those who are expected to have expertise in a given area.)
5. *The content should not constrain the instrument's use to individuals in formal educational settings* (such as college students). To tie the content or scoring of a measure to issues or problems that are unique to the college setting or to specific collegiate events would disallow or confound analyses aimed at determining the effects of college attendance or other life experiences on reflective thinking.
6. *The reading level should be such that it can be used with a wide range of potential test takers.* When a test of reflective thinking also requires high-level reading skills, a low score could indicate poor reading skills, poor thinking skills, or both. Taking reading level into account also allows for the examination of the effects of educational experiences that may stimulate reflective thinking but that do not require complex reading skills (such as movies, lectures, or conversations). It also allows for the evaluation of the development of reflective thinking between groups that may differ in reading level, such as high school and college students or those with and without experience in postsecondary education.
7. *If the measure is based on a model that purports to describe the development of reflective thinking, the theoretical model on which it is based should be validated.* While many models of adult cognition have been proposed during the last decade, few have

been extensively researched and validated (Brabeck, 1984; Commons and others, 1990). The process of validation of developmental claims involves many steps and must include longitudinal data. Models in which validation is lacking or is in the very early stages should be considered preliminary conjectures about development.

This list of desirable features of a measure of reflective thinking is designed to assist researchers, program evaluators, and administrators who must identify or recommend instruments designed to measure reflective thinking. Considering these features may help ensure that selected measures will yield high-quality information about students' *reflective thinking*, not just about other important types of thinking, such as deductive reasoning.

Do Tests of Critical Thinking Measure Reflective Thinking?

To what degree are the skills associated with reflective thinking assessed by existing measures of critical thinking? In addressing this question, we will examine whether these measures include ill-structured problems and if so, how these skills are defined and measured. Two major features of each measure will be considered in this analysis. The first of these is the *definition of thinking* in which the assessment is grounded: does the definition reflect well-structured problems, ill-structured problems, or both? The import of this step is emphasized by Ennis and Norris (1990, p. 11): "It is important to base critical thinking testing and evaluation on an appropriate conception of critical thinking. We cannot just pick any conception, since there is so much resting on the result." Second, we will consider *the nature of the task demands* presented to respondents as an indication of the underlying problem structure of the measure. That is, we will evaluate what respondents are asked to do and how much information they are given. For example, if all necessary information is provided and the respondent's task is to identify the correct answer, then we would conclude that this measure reflects a well-structured orientation.

The organization and choice of measures of reasoning that are included in this analysis follow the general framework used by Pascarella and Terenzini (1991) in their comprehensive review of the effects of college. Their review of students' cognitive skills and intellectual growth includes major sections on Piagetian formal operational reasoning, critical thinking, and postformal reasoning. These three approaches, as well as the measures that are associated with each approach, are discussed in order below.

Assessing Formal Operational Reasoning

Learning to reason abstractly and deductively is a hallmark of Piagetian formal operations. Measures of formal reasoning typically focus on scientific reasoning and involve such elements as hypothetico-deductive reasoning (reasoning from the possible to the actual by disconfirming predetermined hypotheses), propositional thinking (exploring logical relationships between propositions), combinational analysis (identifying relevant variables, both alone and in combination with others), isolation and control of variables, and probability (Flavell, 1963). Many examples of problems that can be solved using formal operational skills are presented and discussed by Inhelder and Piaget (1958). One of their popular tests of the isolation of variables involves examining the oscillation of a pendulum. The respondent's task is to identify which of several variables (length of the string, weight of objects attached to the string, height of the dropping point, and force of the push) accounts for the speed with which the pendulum oscillates. This requires the individual to vary each factor while holding the other three constant, thus isolating the causal variable (the length of the string) and excluding the others. Other tests focus on other aspects of formal reasoning.

In these tests, all necessary information can be obtained by manipulating the objects provided in the experiments, the results are falsifiable and replicable, and they ultimately lead to a single correct answer. Therefore, they measure reasoning about well-structured problems. (For a review of college students' success in solving such problems on formal operations tasks, see P. M. King, 1986.)

Assessing Critical Thinking

The term *critical thinking* has been used extensively to refer to a wide variety of thinking skills (see McMillan, 1987; McPeck, 1981; Paul, 1990). According to Pascarella and Terenzini (1991, p. 118), "critical thinking appears to stress the individual's ability to interpret, evaluate, and make informed judgments about the adequacy of arguments, data, and conclusions." The two measures of critical thinking that have been used most extensively will be reviewed here. These are the *Watson-Glaser Critical Thinking Appraisal* (CTA) (Watson and Glaser, 1964), and the *Cornell Critical Thinking Test* (CCTT), developed by Ennis and Millman (1971). The technical and conceptual merits and limitations of the CTA and the CCTT have been extensively reviewed elsewhere (Berger, 1985; Ennis and Norris, 1990; Helmstadter, 1985; King, Wood, and Mines, 1990; McPeck, 198; Modjeski and Michael, 1983). The following discussion focuses only on the question of problem structure within the CTA and the CCTT.

The Watson-Glaser Critical Thinking Appraisal. McPeck (1981) cites Watson and Glaser's CTA as the single most widely used standardized test of critical thinking. In fact, when McMillan (1987) reviewed twenty-seven programs designed to improve the critical thinking of college students several years later, he noted that only four of these studies did *not* use the CTA.

Watson and Glaser (1964) view critical thinking as a composite of three elements: "(1) attitudes of inquiry that involve an ability to recognize the existence of problems and an acceptance of the general need for evidence in support of what is asserted to be true; (2) knowledge of the nature of valid inferences, abstractions, and generalizations in which the weight or accuracy of different kinds of evidence are logically determined; and (3) skills in employing and applying the above attitudes and knowledge" (p. 10). This definition by itself does not indicate which problem structure is appropriate for a measure of critical thinking because the three elements could apply to reasoning about problems of either type.

Researchers who have examined the problem structure of the CTA have offered varying evaluations. For example, Brabeck and Wood (1990) point to the procedure used in developing the test as informative on this point: a jury of experts "agreed unanimously that the key answers . . . are logically correct" (Watson and Glaser, 1952, p. 8). Brabeck and Wood note that the claim of logical correctness is consistent with the definition of well-structured problems. They argue further that, "The problems posed by the WGCTA [CTA] may be considered well structured in that the decision maker (i.e., test taker) is presented with the basic elements of the problem situation, and a single verifiable and correct answer may then be arrived at through logical inference and probabilistic prediction" (p. 136). Using an approach that focused on the CTA subtests (for example, deduction and interpretation), King, Wood, and Mines (1990, p. 169) concluded that the CTA "reflects a combination of well- and ill-structured problems."

In the current analysis, the specific reasoning tasks given to respondents were used to indicate the problem structure of the test. Table 4.1 summarizes the reasoning tasks (what respondents are asked to do) for each subtest of the CTA, along with the corresponding response options (the alternative responses from which the respondents may select an answer). The description of the tasks is taken directly from the instructions for each section of the CTA, Form Zm.

The Inference subtest is perhaps unique in tests of critical thinking because its response options place true and false on a continuum, using "definitely true" and "definitely false" as the scale anchors, with "probably true," "insufficient data," and "probably false" falling between these extremes. The crux of this subtest is the *certainty* with which one judges that a conclusion properly follows from the given information. As Helmstadter (1985, pp. 1347-1348) notes: "Unfortunately, in many items of this subtest the judgment component would seem to depend more on a personality response set related to how much evidence is required before one is convinced of an argument than on an ability to ascertain whether an inference is a valid one." The "personality response set" to which Helmstadter refers assumes

Table 4.1. Task Demands of the
Watson-Glaser Critical Thinking Appraisal.

<i>Subtest</i>	<i>Reasoning task^a</i>	<i>Response options</i>
Inference	Judge the <i>degree</i> of truth or falsity about a conclusion based on stated facts.	<i>Inference</i> is definitely true, probably true, insufficient data, probably false, or definitely false
Recognition of Assumptions	Decide whether an assumption is made in a given statement.	<i>Assumption</i> is made or not made
Deduction	Judge whether a conclusion necessarily follows from a statement.	<i>Conclusion</i> follows or does not follow
Interpretation	Judge whether a conclusion logically follows beyond a reasonable doubt from given information.	<i>Conclusion</i> follows or does not follow
Evaluation of Arguments	Judge whether an argument is strong (both important and directly related to the question) or weak.	<i>Argument</i> is strong or weak

^aBased on the instructions for each subtest.

Source: Adapted from Watson and Glaser, 1964.

that evidence is required to convince one of an argument, an assumption that we have shown to be as much related to developmental level as to personal preference—a point also made by Kuhn (1989).

McPeck (1981) raises a more fundamental objection in regard to this subtest, arguing that both the description and the instructions reflect a confusion between a proposition, which may be true or false, and an inference, which may be valid or invalid. He notes further: "Normally, we speak of our knowledge or certainty as possessing degrees, but not truth or falsity" (p. 133). From the perspective of the Reflective Judgment Model, this assumption is held only by those who differentiate between well- and ill-structured problems (Stage 4 and beyond).

A related observation is made by Ennis and Norris (1990, p. 30): "People with different levels of sophistication justifiably give different levels of endorsement to a conclusion. Labeling a conclusion *probably true* instead of *true* constitutes a lesser level of endorsement. Saying that the data are insufficient is no endorsement." From the perspective of the Reflective Judgment Model, willingness to judge a response as "definitely true" is most characteristic of those who reason from a pre-reflective stance (Stages 1, 2, and 3). Those who *reason reflectively* (Stages 6 and 7) are more likely to say "probably true" to a convincing argument because they acknowledge the limits of knowing, balancing this limitation against the need to make judgments in light of uncertainty. Despite these collective concerns that could result in individual variability, there is, nevertheless, a correct answer to each question in this subtest (Brabeck and Wood, 1990). While the nature of the task itself in the subtest appears to be ill structured, it is scored using well-structured assumptions, thus requiring a well-structured solution to an ill-structured problem.

For each of the other four subtests listed in Table 4.1, the response options are of an either/or nature: the assumption is made or not made, the conclusion follows or does not follow, the argument is strong or weak. This is problematic for the Interpretation subtest, where the directions make specific reference to the criterion of whether a conclusion "follows beyond a reasonable doubt, (even though it may not follow absolutely and necessarily)" (Watson and Glaser, 1964, p. 6). The type of ambiguity or uncertainty that is acknowledged in these instructions is common to ill-structured problems. But despite the fact that individuals differ in specifying just where "reasonable doubt" begins and ends, the only response options are "follows" and "does not follow." These two options do not allow for ambiguity or qualification in making this judgment, suggesting that the problem may be well structured after all. In other words, just as with the Inference subtest, while the reasoning tasks themselves *appear* to reflect an ill-structured orientation, the given response options reframe the response into one that is well structured. (See Chapter Seven for a discussion of studies that have

investigated the empirical relationships between the CTA and the RJI.)

The Cornell Critical Thinking Test. In Ennis's (1962) seminal paper on critical thinking, he lamented the lack of careful, systematic attention to the identification of the skills required for judging solutions to complex problems. He was particularly critical of Dewey's (1933) description of thinking, which he claimed, "unfortunately suggest[s] that the problem is solved when the solver thinks it is solved, thus providing a psychological instead of a logical criterion for the solution of a problem" (p. 82). Ennis then set out to offer a definition and a detailed description of the concept of critical thinking. "As a root notion, *critical thinking* is taken to be the correct assessing of statements. Since there are various kinds of statements, various relations between statements and their grounds, and various stages in the process of assessment, we can expect that there will be various ways of going wrong when one attempts to think critically. The following list may be looked upon as a list of specific ways to avoid the pitfalls in assessment" (p. 83).

This definition, especially given its emphasis on "correctness" and the explicit acknowledgement of ways of "going wrong," is consistent with a focus on reasoning about well-structured problems. The list to which Ennis referred describes twelve aspects of critical thinking, such as judging whether a conclusion follows necessarily and judging whether an inductive conclusion is warranted. It also includes **three dimensions of critical thinking: logical, criterial, and pragmatic**. The reasoning described by the dimensions may "require the thinker to go beyond the data given" (McPeck, 1981, p. 145), such as deciding whether there is enough evidence to support a judgment. As such, these dimensions appear less well structured. More recently, Ennis (1987, p. 10) defined critical thinking as "reasonable, reflective thinking that is focused on deciding what to believe or do" and expanded his description to include twelve abilities and fourteen dispositions.

How are these dimensions of thinking measured? The Cornell Critical Thinking Test was developed by Ennis and

Millman (1971) on the basis of Ennis's original concept of critical thinking. In the test manual, they define critical thinking as "the reasonable assessing of statements" (p. 1). Level Z (which is appropriate for college students) consists of seven sections: (a) deduction (determining whether a statement follows from premises in material that is emotionally loaded), (b) detecting fallaciously ambiguous arguments (such as circularity, nonsupporting emotive language, oversimplification of alternatives), (c) judging the reliability of information and authenticity of sources, (d) judging whether or not a hypothesis or generalization is warranted, (e) choosing useful hypothesis-testing predictions when planning experiments, (f) assumption-finding, identifying a definition that best expresses another person's usage of a term, and (g) assumption-finding, identifying a statement that fills a gap in a deductive argument.

What type of problem structure is reflected in this test? Again, the emphasis is on deductive problems that are defined with a high degree of completeness and characterized by a high degree of problem structure. As McPeck (1981, p. 145) has noted, the test's format (short questions and answers, with multiple-choice response options) "does not permit the comprehensive or circumspect judgments that are required by the concept of critical thinking. . . . The resulting content of the questions is thus indistinguishable from what might be found in any beginning-logic test." Later in his critique of this test (pp. 145-146), McPeck comments on the difficulty of assessing reasoning about ill-structured problems:

When the questions [on the CCTT] turn from deductive inferences to questions requiring judgments that are more characteristic of critical thinking (for example, those that involve several of Ennis's twelve "aspects" in conjunction with the "dimensional criteria") the items become problematic, in that they admit of more than one correct answer. . . . Valid questions requiring inductive reasoning are difficult to construct in the best circumstances because no matter how much information is provided,

it is usually possible to interpret those data in several reasonable patterns. Moreover, a standardized multiple-choice test is a particularly inept vehicle for posing such questions.

A further clue to the problem structure of this test lies in the amount of information that is provided to the thinker for use in solving the problems. Even in the Inference subtest, a full array of necessary information is provided, making the level of inference quite low. This feature led McPeck (1981) to conclude that the questions measure reading comprehension rather than critical thinking. In summary, each of these factors suggests that the CCTT is a test of reasoning about well-structured problems.

Assessing Postformal Reasoning

Models of postformal reasoning are distinctive in that they attempt to describe how adults reason about problems that are not amenable to solution when only skills of Piagetian formal reasoning are used. As a result, these models are more likely to address ill-structured problem solving. Piagetian formal reasoning (Inhelder and Piaget, 1958) is defined as the ability to reason about the logical structure of propositions and their deductive and inductive relationships. By contrast, postformal reasoning refers to models that embed formal operations in new systems or offer an alternative to formal operations as an end point of cognitive development. While the emergence of such models has generated a great deal of interest among cognitive psychologists (note, for example, the series of books on postformal reasoning edited by Commons and his associates: Commons, Richards, and Armon, 1984; Commons, Sinnott, Richards, and Armon, 1989; Commons and others, 1990), few of these models have been validated on an extensive research base (Brabeck, 1984), and few of the researchers have actively engaged in test construction and refinement. Nevertheless, several are worth considering for their potential to inform our understanding of the assessment of critical thinking. In this section,

we will discuss four models that describe aspects of reasoning that have been referred to as postformal (Basseches, 1980; Commons and Richards, 1984; Perkins, 1985; Perry, 1970) and offer our evaluation of the problem structure of the tests designed to assess the aspects(s) of critical thinking that each test emphasizes.

Perry's Scheme of Intellectual and Ethical Development. The model of intellectual development that is probably best known in higher education circles is Perry's (1970) scheme of intellectual and ethical development. (Although Pascarella and Terenzini, 1991, classify Perry's model as postformal, we are not convinced that it meets the criteria given above and view this as a topic for future consideration. See Chapter Two for a discussion of this point.) As noted in Chapter Two, Perry and his associates at Harvard University identified a sequence of nine "positions" or vantage points from which students understand the world, especially how they "construe the nature and origins of knowledge, of value, and of responsibility" (1970, p. 1). According to this sequence, students' thinking evolves from a "dualistic" approach (positions 1 and 2) where discrete, concrete, and absolute categories are used to understand people, knowledge, and values. The next level of development is called "multiplicity" (positions 3 and 4). From this vantage point, multiple perspectives on a given topic are recognized, and all points of view are seen as equally valid. With the third level (positions 5 and 6), "relativism," comes the recognition that knowledge is contextual and relative and that the various perspectives are part of a larger whole, the broader context in which the topic can now be seen. Moreover, with position 6 comes a recognition of the need to make one's own choices in the world. The last three positions (7 through 9) address issues of personal commitment, "commitment in relativism," that indicate individuals' affirmations and responsibilities in a pluralistic world, choices (about marriage, a career, and so on) that help them establish their sense of identity in the process. It is from these last three positions that the issues of ethical development noted in the title of Perry's (1970) book are addressed.

Researchers and others interested in assessing students'

developmental status on the Perry scheme have had to select among a variety of "Perry measures" since there is no standard assessment procedure for the scheme. Such measures include those developed by Baxter Magolda and Porterfield (1988); Clinchy and Zimmerman (1982); Erwin (1983); Knefelkamp (1974); Kurfiss (1988); Mentkowski, Moeser, and Strait (1983); Moore (1982, 1989); and Widick (1975). (For reviews of research on the Perry scheme and its measurement, see Baxter Magolda, 1987, 1992; Baxter Magolda and Porterfield, 1988; Moore, 1982; King, 1978; and Rodgers, 1980). Among these instruments, the two that have been used most frequently are the Measure of Intellectual Development (Knefelkamp, 1974; Widick, 1975) and the Measure of Epistemological Reflection (Baxter Magolda and Porterfield, 1988). Each is discussed below.

The foundation for the current version of the Measure of Intellectual Development (MID) was developed by Knefelkamp (1974) and Widick (1975). According to Moore (1982, p. 1), this version is designed to measure "the intellectual aspect of the Perry model (positions 1-5)" and hence is only scored through position 5. (Earlier versions had been scored across the full range of positions.) The MID consists of a series of short open-ended essays about such topics as decision making, classroom learning, and careers. Respondents are asked to provide a written description of an experience in a specified area (such as best class taken in college, a recent major decision) and to comment upon the experience, including how they feel about it. Trained raters match responses to position descriptions. The questions the MID poses are clearly ill structured and require the respondent to generate a description of his or her own way of thinking about each issue. However, the content of the measure renders it useful only for those in collegiate settings.

The Measure of Epistemological Reflection (MER) was developed by Baxter Magolda and Porterfield (1985, 1988), who used concepts and scoring cues from the Perry scheme in its construction; the MER has subsequently been revised on the basis of extensive information from interviews with college women as well as men. This instrument consists of a series of open-ended questions in six domains: decision making in an

educational context, role of the learner in the learning process, role of the instructor in the learning process, role of peers in the learning process, role of evaluation in the learning process, and nature of knowledge, truth, and reality. For each domain, students are asked to provide written answers to a specific series of questions. Like the MID, the MER is scored through position 5, relativism.

The first domain, decision making, is introduced as follows (Baxter Magolda and Porterfield, 1988, p. 103): "Think about the last time you had to make a major decision about your education in which you had a number of alternatives (e.g., which college to attend, college major, career choice, etc.). What was the nature of the decision?" The follow-up questions are as follows (p. 103): "What alternatives were available to you? How did you feel about these alternatives? How did you go about choosing from the alternatives? What things were the most important considerations in your choice? Please give details."

For the sixth domain, which has the most explicit epistemological focus, the following questions are posed (p. 220): "Sometimes different instructors give different explanations for historical events or scientific phenomena. When two instructors explain the same thing differently, can one be more correct than the other? When two explanations are given for the same situation, how would you go about deciding which explanation to believe? Please give details and examples. Can one ever be sure of which explanation to believe? If so, how? If one can't be sure of which explanation to believe, why not?"

The problems posed in the MER are clearly ill structured, including both personal and intellectual problems. Scoring of the students' written responses is completed by trained raters; this process has been greatly enhanced by the availability of a detailed technical and training manual (Baxter Magolda and Porterfield, 1988). The reliability and validity of the MER are also documented in this publication, providing the necessary information for researchers seeking to evaluate this measure.

Basseches's Model of Dialectical Reasoning. Basseches's model (1980, 1984a, 1984b) is clearly characterized as postformal: Bas-

seches argues that perspective dialectical reasoning builds on and is more powerful than formal operations. Basseches (1989, p. 45) notes, as we have, that formal operations are limited to what he calls "closed-system" problems. He also notes that many of the problems faced in the adult world require thinking that takes into account the changing nature of the world and call for a synthesis of different and sometimes competing perspectives. He characterizes the new kind of thinking that emerges as dialectical, which he defines as "thinking which looks for and recognizes instances of dialectic — developmental transformation occurring via constitutive and interactive relationships" (p. 55).

With this form of thinking, individuals comprehend that change is a fundamental characteristic of the world and that in the face of change they must often think beyond current stereotypes of solutions to develop new, more inclusive conceptions of a problem as well as solutions for the problem. Basseches has identified twenty-four cognitive abilities that comprise dialectical operations; these include an awareness of whole systems, the essential properties of those systems, and their changing nature; an awareness of internal relationships within a system; and an awareness of the fact that relationships are sometimes essential characteristics of objects or persons (for example, being a member of a family is an essential characteristic of a person's identity). Finally, dialectical thinking can reflect upon itself and demonstrates an awareness that qualitative changes can emerge from quantitative ones.

Basseches uses an example of a mother and daughter who are caught in a conflict over differing values to illustrate dialectical reasoning. The mother reflects on the situation, noting that she, too, had values that were different from her mother's and that her own values have changed over time. The mother begins to understand that her daughter's values reflect an integration of the values she tried to share with her daughter as well as of the experiences her daughter has had in the world, experiences the mother has never had. Thus, the daughter learned and transformed the mother's values to make them relevant to her own life. This conception both acknowledges change and constancy and forms a solution for the mother that transcends the conflict.

Basseches (1980, 1984a, 1984b) has used open-ended relatively unstructured interview procedures to measure dialectical reasoning. In one procedure, he asks interviewees to respond to a series of questions about education, such as "How would you go about formulating a conception of the nature of education?" (1980, p. 224). In a second procedure, he asks participants to read twelve pairs of arguments from cards, one pair at a time. On each card are two arguments about a different topic, such as the family. One of the arguments is dialectical; the other is not. The participant's task is to explain each of the paragraphs in his or her own words. Paragraphs are scored for whether participants comprehend the dialectical and nondialectical passages.

The first method clearly poses an ill-structured problem since there is no single correct answer to questions about the nature of education. Although the second format does not directly pose a problem to the participants, it does involve them in explaining reasoning about ill-structured issues; thus, it measures participants' comprehension of arguments about ill-structured problems. Basseches's method and results are promising (older, more educated participants are more likely to use and comprehend dialectical reasoning); however, the method is time-consuming to administer (two hours) and score (Basseches, 1984a, 1984b; Benack and Basseches, 1989; Irwin and Sheese, 1989). Further, reliability and validity data are minimal since the measures have been used with very limited samples.

Commons and Richards's General Stage Model. The General Stage Model (Commons and Richards, 1984; Commons, Richards, and Kuhn, 1982; Richards, 1990; Richards and Commons, 1984) reflects the authors' attempt to describe formally the general structure of three postformal stages, the systematic, meta-systematic, and cross-paradigmatic, which they claim are hierarchically constructed after formal operations. With systematic reasoning, individuals can operate on whole classes and relations and understand how they generate systems. In meta-systematic reasoning, cognitions about multiple systems and their relations can be generated, and in cross-paradigmatic

reasoning, the relations between systems of systems can be generated. The authors argue that these forms of thinking provide the basis of reasoning about relationships between systems and systems of systems in many fields, including moral philosophy, music, mathematics, and science. (This model is among those for which the data base on the development of the three post-formal stages is not extensive.)

Richards and Commons (1984) use a set of four complex seriation tasks to measure systematic and meta-systematic reasoning. In each of the tasks, different criteria are used to evaluate the order of given elements. For example, one story uses economic value and another uses weight as a basis for a complex story that involves serial relationships. The task for the participant is to compare the ordering systems in the four tasks and to identify which of the six possible pairs of tasks are the most similar. Respondents are asked to judge on a scale of 0 to 9 how similar or different the orderings are in the stories (Richards, 1990). The problem appears to be ill structured since there are many ways in which the tasks are similar or different. However, from a mathematical perspective, Richards (1990, p. 190) notes that an "ideal performer" acting within a meta-systematic perspective would judge two of the stories as more similar than any other pair of stories. Presumably, the judgment of this ideal performer indicates the correct performance on the tasks, implying that the problem, though difficult, ultimately belongs on the well-structured end of the continuum. This "single correct answer" orientation is also reflected in the emphasis on the solution per se rather than on the rationale for the solution or the process used to arrive at the solution. Thus, while many of the models classified as postformal address ill-structured problem solving, one cannot assume that because the model is postformal, the assessment will necessarily focus on ill-structured problems: Commons and Richards' model exemplifies a postformal model where the assessment implies a well-structured orientation.

Perkins's Model of Informal Reasoning. The concept of informal reasoning developed by Perkins and his colleagues (Perkins 1985; Perkins, Allen, and Hafner, 1983) also explicitly addresses

reasoning about ill-structured problems. Even though the research base for this concept is relatively small, it merits inclusion here because of the similarity between the aspects of reasoning it attempts to describe and measure and those of the Reflective Judgment Model. According to Perkins (1985, p. 526): "In informal reasoning, reasons typically occur on both sides of the case, no one line of argument settles the truth of the claim, and no computational procedure assigns the claim a numerical probability. The reasoner must weigh and synthesize to best judge the soundness of the claim."

Informal reasoning is also referred to as everyday reasoning in that it occurs "when a person is deciding which car to buy, which political candidate to vote for, or how to persuade a colleague to adopt a particular position with respect to some burning issue" (Voss, Perkins, and Segal, 1991, p. xiii). These characteristics are clearly consistent with reasoning about ill-structured problems as described by the Reflective Judgment Model.

Perkins (1985) has measured informal reasoning through the use of a ninety-minute interview in which the respondents are presented with a series of "genuinely vexed issues" (p. 564), such as whether violence on television significantly increases the likelihood of violence in real life. In this interview, respondents are asked for an initial judgment on the issue, how confident they are of this judgment, how interested they are in the issue, and how much time they have previously spent thinking about the issue. They are asked to explain their arguments and their reasons for holding them, with several opportunities to elaborate on their statements. In addition, the respondents are asked to state reasons on both sides of the case and to assess the importance of these reasons (Perkins, Farady, and Bushey, 1991, p. 88).

Responses are scored on six scales: (1) number of sentences (a measure of elaboration), (2) number of lines of argument, (3) how many objections were raised to the respondent's own position (a measure of the extent to which each respondent considered the other side of the case), (4) number of times the interviewer prompted the respondent to stay focused on the

issue, (5) adequacy of explanation (how well a reason supported a conclusion), and (6) overall quality. The last two scales use a 1 to 5 rating.

As with the definition of thinking, the interview format used by Perkins and his colleagues also appears to be consistent with reasoning about ill-structured problems. Whether it should also be considered a model of postformal reasoning is questionable since it has not been presented as an integrated psychological theory and its developmental basis is not clear. Nevertheless, this measure holds promise for adding to our understanding of reasoning about ill-structured problems.

Summary

How does an instructor, researcher, administrator, or other decision maker evaluate the potential utility of given measures of critical thinking and decide which to use for a specific purpose? Our analysis has suggested that the astute consumer of such measures should begin by carefully examining how critical thinking is defined and measured in any instrument being considered. Whether a given instrument will suffice will probably depend largely on the consistency between the aspects of thinking the individual or institution desires to measure and the operational definition of thinking used in the instrument (a point also noted by Helmstadter, 1985, and Jacoby, Astin, and Ayala, 1987), as well as how appropriate it is for use with the population being assessed. Further, this analysis has also pointed out that many measures of critical thinking actually measure reasoning about well-structured rather than ill-structured problems and that this cannot always be correctly inferred from the type of thinking the test purports to measure.

This chapter has addressed a variety of issues related to the assessment of reflective thinking. Since the attributes of "good thinking" are defined and measured in such a variety of ways, we began with a consideration of what characteristics a measure of reflective thinking and ill-structured problem solving would have in order to provide useful information to teachers, campus administrators, researchers studying educational out-

comes, and others. Elements of this list provided a foundation for examining existing measures of critical thinking for the purpose of evaluating their potential as measures of reflective thinking. In particular, we examined the structure of the problems that are provided in each instrument and found that several popular measures of critical thinking focus on reasoning about well-structured rather than ill-structured problems. Thus, they do not serve as suitable measures of reflective thinking.

This discussion of assessment of reflective thinking has focused exclusively on formal assessment instruments because they are typically selected for large-scale assessment projects. Nevertheless, we recognize the value of informal assessment, such as occurs when educators must decide how to respond to individual students. This approach may also yield valuable information about a person's skills in thinking critically and reflectively.

Research on the Reflective Judgment Model

with Phillip K. Wood

Being able to think critically is one of the major hallmarks of an educated person and an attribute that commonly appears in statements of educational goals at both the secondary and college levels. Further, educational environments provide many opportunities to assist students in learning how to improve their thinking skills, including reasoning about problems that do not have single, verifiable answers. This is particularly true at the college level, where students are exposed to contrasting perspectives on a multiplicity of ill-structured problems in both academic and social environments and where there are (or should be) many opportunities to learn to reason more complexly and effectively.

The consistency between reflective thinking as a goal of education and the type of thinking described by the Reflective Judgment Model leads us to the major questions of this chapter: (1) Does reflective judgment develop between late adolescence and middle adulthood? (2) How do high school, college, and graduate students reason about ill-structured problems? (3) Does students' reasoning improve with additional exposure to and involvement in higher education? (4) Do adult learners differ from traditional-age students in their reflective thinking? (5) How

does the reasoning of adults who have not participated in higher education compare to the reasoning of those who have earned a college degree? (6) Are there gender or cross-cultural differences in the development of reflective thinking? In this chapter, we turn to the research that has been conducted on the Reflective Judgment Model, presenting the results of both longitudinal and cross-sectional research that collectively inform our understanding of these questions.

Research on changes in thinking that are associated with age and educational experiences, including development in reflective thinking, is of interest to different groups of people for different reasons. For example, such research is important to developmental psychologists who seek to understand the nature and the mechanisms of cognitive development through the adult years. Those with institutional studies responsibilities are interested in such research because one index of an institution's effectiveness in promoting reflective thinking is whether its graduating seniors consistently show more proficiency in critical thinking than they did as entering freshmen. College faculty and student affairs staff look to research on the development of reflective thinking because of the insights it can offer for helping them better educate students to value and demonstrate good thinking skills. We believe the following data will be of interest to all three groups.

Research on changes that are associated with age, such as those described by the Reflective Judgment Model, requires comparisons of individuals at different ages. However, age alone would not be expected to lead to higher scores: there is no reason to believe that the passage of time alone should influence the development of reflective thinking unless it corresponds to some other intervening event (such as neurological maturation or age-related involvement in an intellectually stimulating environment). Since educational institutions attempt to involve students in the acquisition, analysis, and utilization of knowledge, improvement in students' ability to make more reflective judgments can reasonably be expected, and the failure to find such changes would be disturbing.

To date, more than 1,700 individuals representing a wide

variety of student and nonstudent subgroups, have participated in the Reflective Judgment Interview (RJI). This number includes almost 200 people who have taken part in longitudinal studies, over 150 high school students, 1,100 college students (both traditional-age and adult samples), and about 200 graduate students. In addition, more than 150 nonstudent adults have been tested. The results help us understand how students reason about ill-structured problems, whether their reasoning changes over time, and whether improvement in the ability to make reflective judgments is associated with participation in higher education. In this chapter, we examine the patterns of scores that have emerged from both longitudinal and cross-sectional studies as they pertain to the development of reflective thinking, particularly in the context of higher education. In considering these results, readers are reminded that the RJI measures how students typically function, not how they might function under optimal learning and testing conditions. (The RJI and its properties are described in detail in Chapter Five.)

We begin our description of research on the Reflective Judgment Model by looking at those studies that have attempted to ascertain whether the ability to make reflective judgments develops over time, particularly between late adolescence and middle adulthood. Charting the changes in individuals' reasoning requires longitudinal data. Such data are essential for the validation of any developmental model because they indicate whether people change in a manner consistent with the description and sequence proposed by the model. Further, they allow a stricter test of the impact of educational experiences than do cross-sectional studies (at least among the particular cohorts tested). Despite the importance of longitudinal research for the study of intellectual development, most of the existing models of adult cognitive development lack a longitudinal data base (Brabeck, 1984), an observation that, unfortunately, is still true today. The Reflective Judgment Model is an exception to this pattern.

We first examine the results of a ten-year longitudinal study of the development of reflective judgment that was an outgrowth of the original Kitchener and King (1981) study. This study covers the longest span of years and the most testings of

any group of individuals to date, providing the strongest basis for evaluating the developmental properties of the model. Following these, we describe additional longitudinal studies of shorter duration that used the Reflective Judgment Model. In the second half of this chapter, we review the cross-sectional studies; these are discussed separately for each of the following groups: high school students, traditional-age college students, nontraditional-age college students, and nonstudent adults. We conclude this chapter by reviewing the research on gender differences and cross-cultural differences in reflective judgment. Because of the length of this chapter and because some readers may wish to be selective in their reading of various portions of the research summarized here, the concluding summary is more detailed than that offered for other chapters.

Ten-Year Longitudinal Study of Reflective Judgment

The ten-year longitudinal study noted above traced the development of the reflective thinking of eighty individuals from 1977 to 1987. During this decade, these individuals generously contributed their time (up to four hours per testing) and candidly shared with us the means by which they arrived at conclusions about vexing problems. As a result, we now have a picture of their development from the psychologist's equivalent of a zoom lens, taking periodic snapshots of their reasoning every few years. Since these individuals ranged between sixteen and twenty-eight years of age at the first testing, reviewing their progress provides an overview of the development in reflective thinking during early adulthood. In addition, since most of the participants in the sample were engaged in some aspect of formal education at the same time, this study offers insights into their intellectual journeys during this time and provides an initial assessment of whether involvement in higher education leads to higher levels of reflective judgment. The results document changes in participants' reflective thinking based on their assumptions about knowledge, the certainty with which they said they knew, and how they justified their beliefs.

The study began with a cross-sectional investigation of

three groups of students who differed by age and educational level (King, 1977; Kitchener, 1977-78; Kitchener and King, 1981). The original sample included twenty high school juniors, forty college juniors, and twenty third-year doctoral students. The rationale here was that the different ways of reasoning described by the Reflective Judgment Model should be apparent when tested with students from a wide range of educational levels and that a finding of no differences across these groups would raise serious questions about the model and/or its measurement. The high school students were enrolled in two school districts in communities in or near the Minneapolis metropolitan area; the college students were enrolled in either the College of Liberal Arts or the College of Agriculture at the University of Minnesota; and the doctoral students were liberal arts majors at the same university. The younger two groups were matched to the graduate sample by gender (half female, half male) and by academic aptitude (based on scores from the Minnesota Scholastic Aptitude Test). This matching was designed to serve as a check against the possibility that any obtained group differences in reflective judgment might be a function of differences in academic aptitude.

These individuals were contacted in 1977 and again in 1979, when 74 percent of the original sample was retested (the results are reported in King and others, 1983). The third testing occurred in 1983, with a 69 percent retest rate (see King, Kitchener, Wood, and Davison, 1989; Kitchener, King, Wood, and Davison, 1989). The final attempt to contact all eighty of the men and women who originally participated in the 1977 study was undertaken in 1987. At this time, fifty-six members of the original sample (70 percent) were contacted and agreed to continue their participation. Usable RJI data were obtained for fifty-three of these individuals, or 66 percent of the original sample. (Preliminary findings of the ten-year study are reported in King, Kitchener, and Wood, 1990.) Table 6.1 shows the sample sizes and response rates for each subgroup at each year of testing.

Between 1977 and 1987, a large portion of this sample completed either the educational degree they were pursuing in

**Table 6.1. Ten-Year Study:
Sample Size and Response Rate by Group and Year of Testing.**

Year	<i>Educational level in 1977</i>							
	<i>High school juniors</i>		<i>College juniors</i>		<i>Advanced doctoral students</i>		<i>Total</i>	
	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>
1977	20	100	40	100	20	100	80	100
1979	17	85	27	68	15	75	59	74
1983	15 ^a	75	27 ^a	68	13	65	55	69
1987	13	65	26	65	14	70	53	66

Note: Figures are based on those who completed the Reflective Judgment Interview at each year of testing.

^aDue to a coding error, the sample sizes for the 1983 high school and college groups listed here are fewer by one than the number we reported in Kitchener, King, Wood, and Davison (1989). This correction does not affect the patterns of results reported there.

1977 or a subsequent degree. Table 6.2 shows the degrees earned by the participants of this study according to subgroup. By 1987, all but one of the former high school students had earned at least a bachelor's degree. Among the former college juniors, 84 percent had completed at least a bachelor's degree; of these college graduates, almost half had also completed a postbaccalaureate degree. By 1987, 80 percent of the former advanced doctoral students had completed the Ph.D. degrees they had been working on in 1977. As the data show, the sample had a rather high level of educational attainment by 1987.

At each testing, the participants were asked to complete several measures, including the RJI. The four standard RJI problems (see Table 5.1) were used at the first, second, and third testings. At the fourth testing, a new problem was added as a check against the possibility that an increase in scores over time might be a function of practice or familiarity with the issues discussed rather than a reflection of development in reasoning. Each problem was rated blindly at each testing; that is, the raters were not informed about ratings that had been assigned at prior testings, about those assigned to other problems at the current time of testing, or about those assigned by the other rater. Raters

Table 6.2. Ten-Year Study: Educational Attainment by 1987.

Highest degree earned by 1987	Educational level in 1977					
	High school juniors		College juniors		Advanced doctoral students	
	n	%	n	%	n	%
High school diploma	1	9	4	17	—	—
B.A./B.S.	9	82	11	46	1	7
M.A./M.S.	1	9	3	13	2	13
Professional	—	—	6	25	—	—
Ph.D.	—	—	—	—	12	80
Total	11	100	25	100 ^a	15	100

Note: Numbers vary slightly from those in Table 6.1 due to missing data.

^aSum of column figures is 101 due to rounding.

were also blind to gender and other demographic characteristics and to group (high school, college, or graduate student) for the first three testings. (By the time of the fourth testing, group membership at the first testing no longer indicated educational level. Further, students from each group had earned high RJI scores at the third testing, so the need for raters to be blind to this information was no longer compelling.)

The other measures that were also administered were Terman's (1973) Concept Mastery Test, a measure of verbal ability; Rest's (1979) Defining Issues Test, a measure of moral judgment; Loevinger and Wessler's (1970) Sentence Completion Test, a measure of ego development, given at the first, second, and third testings only; the Myers-Briggs Type Indicator, a measure of psychological type, given at the fourth testing only; and a background questionnaire on the participants' educational and occupational experiences. (Discussions of the relationships between the RJI and other measures appear in Chapters Seven and Eight.)

Changes in Reflective Thinking Over Time

Conducting longitudinal research presents a variety of challenges. Many participants move (sometimes many times) over

a ten-year period; thus, locating former participants may take considerable detective work. Difficulties in locating former participants, combined with life events that interfere with their participation, sometimes result in sample attrition. Participants who are missing or unable to participate at one testing may reappear and be eager to participate in subsequent testings. For example, one of the participants in this study was out of the country during the middle two testing periods but did take part in the fourth testing.

Such sampling problems present subsequent problems for data analysis since not every participant has scores for each testing. The question then becomes whether it is better to evaluate change by looking at the scores of everyone in the sample or just of those people who were present at every testing. If everyone in the sample is included (even if some individuals missed one or more testings), there will be gaps in the data. Where scores are missing, we have no knowledge of whether individuals' scores would have increased, decreased, or stayed the same during specified intervals. On the other hand, if the scores of only the participants who were tested every time are included, the sample size is smaller, and generalizations are in turn limited. Consequently, analyses designed to evaluate changes in Reflective Judgment over the ten years of this study were conducted in three different ways. The first used the most conservative test of a developmental sequence by including only the scores of the thirty-eight people who participated in all four testings (1977, 1979, 1983, and 1987). Second, the scores of the thirteen people who had missed only the second or third testing were added to the sample of thirty-eight by estimating their missing scores (missing scores were estimated by averaging the scores that immediately preceded and followed the missing ones); this yielded a sample of fifty-one. Third, the scores of all individuals who had participated in any phase of the study ($n = 80$) were included, without the interpolation of any scores. Since only minor differences in the patterns of results were found in the three types of scores, the following analyses are based on the total sample. (Results based on those with complete data are reported in Resource B, Table B6.1.)

Changes in Group Scores. The three student groups differed in their reflective judgment levels in 1977: doctoral students scored significantly higher than college students, who scored significantly higher than high school students. Additionally, differences between these groups remained in subsequent testings: the former doctoral students scored significantly higher than the other two groups; moreover, while the RJI scores of both of the younger two groups increased over time, the patterns of growth between the two groups was different, and they became more similar.

Consistent patterns of increasing RJI mean scores were found for all three groups at each subsequent testing (with the single exception of the scores for the former doctoral students, which remained stable between 1979 and 1987). In other words, with one minor exception, the RJI mean scores of each group increased with each follow-up testing. These scores are presented in Table 6.3 and plotted in Figure 6.1. As the ranges of scores in Table 6.3 illustrate, not only was higher-stage reasoning more evident over time but lower stage reasoning became less evident. In no case was there a significant decline in scores over time, and in no case did the increase exceed one and a half stages between testings. Taken together, these patterns suggest that the development of reflective thinking as measured by the RJI evolves slowly and steadily over time among individuals engaged in educational programs.

The mean RJI scores for the former high school juniors increased the most dramatically in the ten-year period, from 2.9 to 5.5, a difference of two and a half stages. This increase in the group mean score reflects a change in epistemic assumptions ranging from Stages 2 and 3 (the stages on which the 1977 score was predominantly based) through Stages 5 and 6 (for the 1987 score), a dramatic metamorphosis in thinking toward making reflective judgments. It is perhaps not coincidental that during the ten years of this study, all but one member of this group attended and graduated from college.

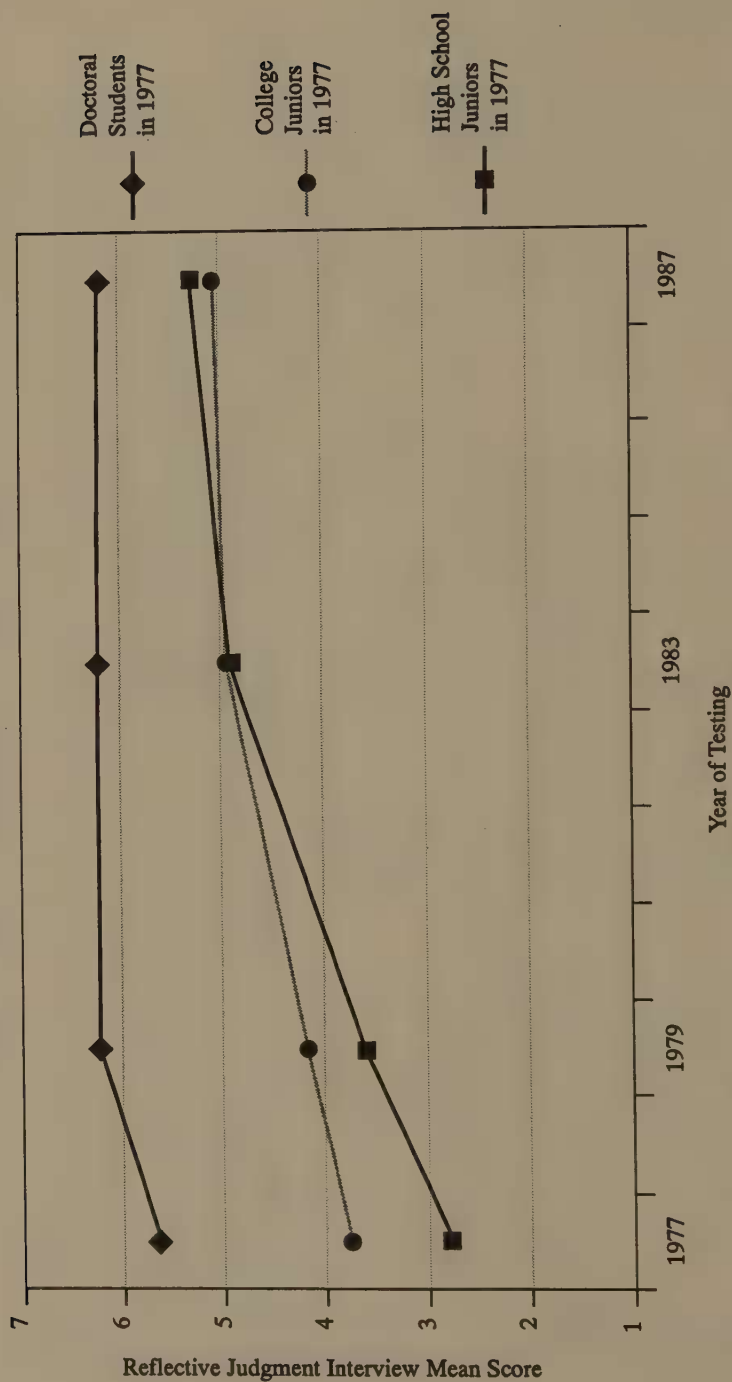
The ten-year change in mean scores for the former college juniors from 3.8 to 5.1 was not as dramatic but still showed a steady pattern of development in reflective thinking. The scores

Table 6.3. Ten-Year Study: Reflective Judgment Scores by Group and Time.

Time	Educational level in 1977								
	High school juniors			College juniors			Doctoral students		
	M	sd	Range of scores	M	sd	Range of scores	M	sd	Range of scores
1977	2.77	.49	2.17-4.04	3.76	.69	2.46-5.58	5.67	.92	3.83-6.92
1979	3.61	.46	2.83-4.50	4.18	.80	3.21-6.38	6.23	.50	4.97-6.96
1983	4.91	.79	3.46-6.00	4.92	.54	4.12-6.17	6.23	.45	5.21-6.88
1987	5.29	.75	4.12-6.59	5.05	.60	3.96-6.44	6.21	.56	5.06-6.94
Change, 1977-1987	+2.52			+1.29			+1.54		

Note: Scores are based on those tested two or more times (1977 plus 1979, 1983, and/or 1987). Sample sizes for each group at each time of testing are given in Table 6.1. The results of a three-way (time $\times$ group $\times$ gender) General Linear Model were as follows:
 Overall model: $F(23,223) = 27.88, p < .001$
 Time main effect: $F(3,223) = 63.49, p < .001$
 Group main effect: $F(2,223) = 164.21, p < .001$
 Gender main effect: $F(1,223) = 4.57, p < .05$
 Time $\times$ group interaction: $F(6,223) = 9.17, p < .001$
 Time $\times$ gender interaction: $F(3,223) = .92, ns$
 Group $\times$ gender interaction: $F(2,223) = 2.84, p = .06$
 Time $\times$ group $\times$ gender interaction: $F(6,223) = .40, ns$

Figure 6.1. Reflective Judgment Interview Scores by Group: 1977-1987.



Note: Scores are based on individuals tested two or more times (1977 plus 1979, 1983, and/or 1987). Sample sizes for each group at each time of testing are given in Table 6.1.

for the former doctoral students are best characterized as stable, changing only from 6.2 to 6.4 over the ten-year period. However, even though this group's mean scores were generally stable, the range of scores shown in Table 6.3 suggests that lower scores were replaced by higher scores, so that by the fourth testing, the entire sample scored close to Stage 6 or 7. The stability in mean scores may reflect a "ceiling effect" on the RJI since the highest possible score, 7.0, can only be earned if all raters consistently rate all responses as Stage 7 over all problems.

Since the 1987 testing included a new problem on the safety of nuclear energy, the correspondence between participants' responses to this problem were compared with their responses to the other problems. If individuals were scoring higher over time as a function of increasing familiarity with the problems, they would be expected to score lower on a less familiar problem. However, differences between the mean score for this problem and the mean scores for the other four problems were negligible, offering no support for the familiarity hypothesis.

Several statistical analyses were run to check for patterns in the mean scores of each group over the ten-year period. First, a three-way General Linear Model (time by group by gender) was run, with the RJI scores used as the dependent variable. (Specific statistical findings for this model are reported in Table 6.3.) Where significant main effects were found, Waller-Duncan K-ratio follow-up *t*-tests were run. The mean RJI scores across samples increased over time ($p < .001$). These increases were significant between 1977 and 1987, but the overall increase between 1983 and 1987 was not significant.

A time by group interaction ($p < .001$) indicated a disordinal pattern in the RJI scores; that is, as illustrated in Figure 6.1, the amount of growth over time differed by group. In 1977 and 1979, the scores of each successive age/educational group were significantly higher. However, in 1983 and 1987, the ordering of the scores of the former high school and college students was reversed: the scores of the high school students had increased more and actually exceeded those of their older counterparts. The scores of the former doctoral students remained significantly higher than the scores of the other two groups. Since

there was such a dramatic increase in the scores of the former high school students, a follow-up ANOVA was conducted only on the two older groups. The time, group, and time by group interactions remained ($p < .01$), indicating that the overall findings of growth over time were not attributable solely to the changes in the former high school students. (Further analyses of group differences, as well as gender differences, are reported below.)

As already noted, the scores of the former high school juniors changed from Stages 2 and 3 in 1977 to Stages 5 and 6 in 1987. The greatest amount of growth for this group occurred between 1979 and 1983; during this four-year period, the average scores changed almost one and a half stages, from 3.9 to 5.3. No other group showed a comparable amount of growth at any point in the study: the average scores of the former college students increased by slightly more than one stage during the ten years of the study (from 3.8 to 5.1), and the mean score of the doctoral students remained stable at Stage 6.

While we can only speculate about the reason for the spurt in the scores of the former high school students, its timing does coincide with their college attendance: 75 percent of the participants in this group entered and completed college between 1979 and 1983. This spurt corresponds to the developmental spurt found among a sample of college students between the ages of eighteen and twenty in the Kitchener, Lynch, Fischer, and Wood (1993) study and is predicted by Fischer's (1980) skill theory model, described in Chapter Two. Thus, the spurt may be due to the influence of being in an educational environment, to maturation, or to both. The premise that the educational environment played a critical role in the development of reflective thinking among the former high school students is supported by an additional finding at the sixth-year testing in 1983. Those who held a bachelor's degree showed significantly more change in their RJI score than did those with only a high school diploma, supporting the contention that there is greater growth in reflective judgment for those who attend and complete college than for those who do not. However, too few of the participants from this group did not attend college for a rigorous test of this hypothesis, so the point remains tentative.

Changes in Individual Scores. Although group data can provide an overview of longitudinal changes, data on individual change over time provide a stronger test. The RJI scores for the entire longitudinal sample in this study ($n = 71$) are listed in Resource B, Table B6.2. These scores are presented in two ways, first as mean scores averaged across all problems and raters. The second index presents each individual's dominant (modal stage score) and subdominant (second most frequently assigned) stage score. The findings presented below on individual changes over time are based on these scores.

The pattern of changes in the RJI scores of the fifty-three individuals who were retested in 1987 gives a full ten-year view of development. (For this purpose, we define change as a difference of at least .1—for example, a mean score of 5.4 compared to 5.5.) Using this criterion, the RJI scores of 92 percent ($n = 49$) of the sample increased over the ten years of the study. The scores of two individuals (both former doctoral students) decreased over this time, although their modal scores remained stable at Stage 7. There was no change in the scores of the remaining two individuals between 1977 and 1987.

A second set of scores, those of the thirty-eight individuals who participated in all four testings, provides a more detailed picture of the development of reflective judgment. Twenty-six of these individuals (68 percent) had scores that either increased or remained stable between each of the three testing intervals (1977–1979, 1979–1983, and 1983–1987), and eleven of the remaining twelve people had only one exception to this pattern (for example, one individual's RJI scores increased between 1977 and 1979, decreased in 1983, and increased in 1987).

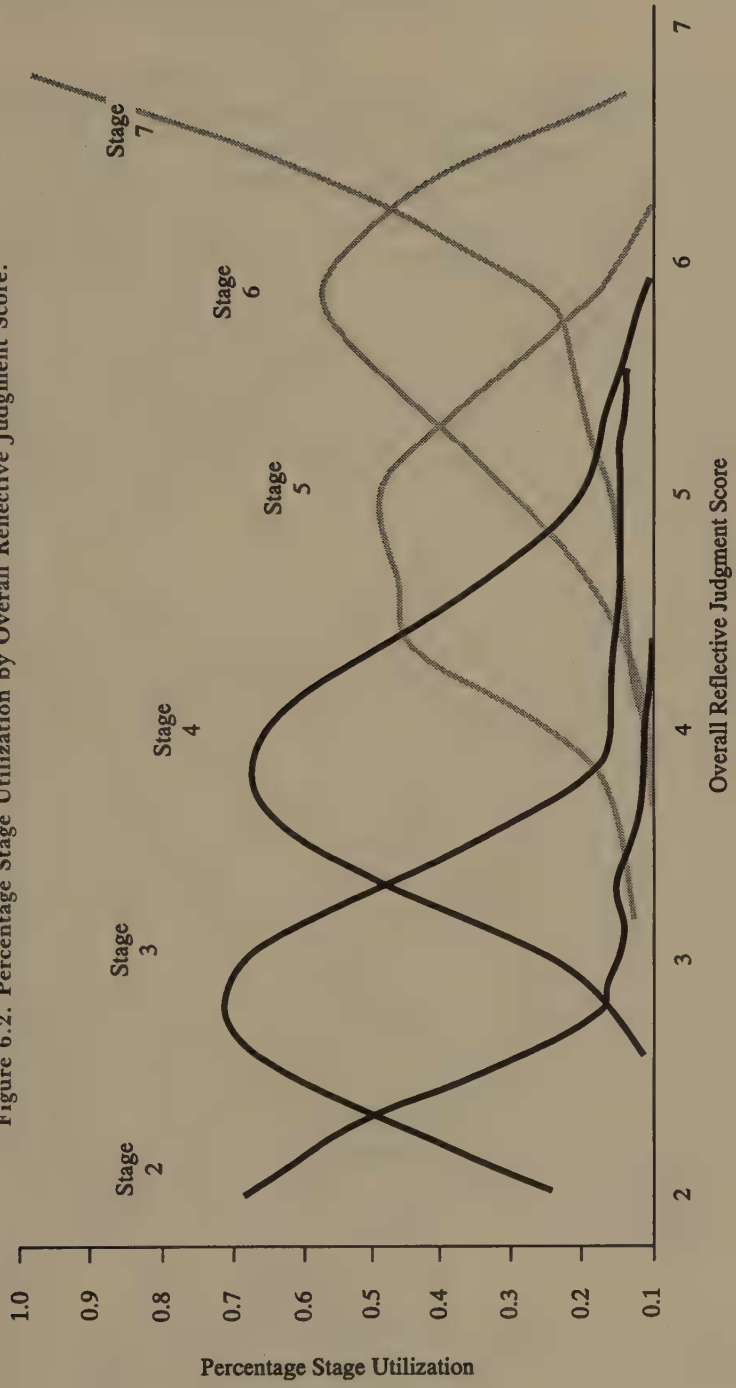
This pervasive pattern of upward change in RJI scores over the ten years of this longitudinal study shows that the vast majority of the individuals who participated in the study improved in their ability to make reflective judgments over time. This suggests that the overall changes in mean scores shown in Table 6.3 cannot be attributed to large increases in one subset of the sample (the former high school juniors). Further, the pattern of increasing RJI scores over time clearly indicates the slow, steady emergence of reflective thinking during

early adulthood. When individual scores are examined for the dominant and subdominant stages at each time of testing, a sequential (not random) ordering of stages is apparent, and there is consistent evidence of lower-stage usage prior to higher-stage usage. Few examples offer evidence to suggest that stages were skipped. For example, in 1977, one female participant scored at a dominant Stage 2 and a subdominant Stage 3 on the RJI; in 1979, she scored at a dominant Stage 4 and a subdominant Stage 3; in 1983, she scored at a dominant Stage 4 and a subdominant Stage 5; and in 1987 (by which time she had graduated from college), she scored at a dominant Stage 5 and a subdominant Stage 4. Not only had her RJI scores changed over two full stages during the ten years of the study, but the changes had occurred sequentially and gradually.

Variability in Scores Across Stages. Do individuals evidence reasoning that is characteristic of more than one stage of the Reflective Judgment Model at a time? If so, this would be consistent with the complex stage model proposed by Rest (1979), which is described in Chapter Two. If not, how can the variability be characterized? The listing of dominant and subdominant stage usage in Resource B, Table B6.2, informs this question: a subdominant stage was assigned for most individuals in the study; this was almost always an adjacent stage. In only two cases were the ratings limited to a single stage score.

Wood (1993a) has examined this question in more detail in the only existing study to systematically analyze variability in stage reasoning across studies that used the RJI. To do this, he conducted a secondary analysis of RJI scores from fifteen studies that used trained interviewers and raters and for which raw data were available for reanalysis ($n = 1,995$). He calculated the proportion of time each stage was represented in an individual's ratings across the four RJI problems and referred to this as the "percent stage utilization score." To indicate the general patterns in stage utilization as a function of the overall RJI score, he then calculated a series of spline regressions (Darlington, 1990), which predicted each stage utilization on the basis of overall reflective judgment score. The spline regressions for each stage utilization are displayed in Figure 6.2. This figure shows

Figure 6.2. Percentage Stage Utilization by Overall Reflective Judgment Score.



that Stage 2 ratings account for about 70 percent of the assigned ratings for individuals whose overall score was Stage 2. By contrast, individuals who scored at higher levels of reflective judgment evidenced reasoning at a wider variety of stages. For example, the line corresponding to an overall score of Stage 3 indicates that, on the average, over two-thirds of the assigned scores for these individuals were at Stage 3, with the remaining scores fairly equally divided between Stages 2 and 4. A similar pattern was found for Stage 4, with the remaining scores being evenly divided between the adjacent stages, or Stages 3 and 5. By contrast, about half the scores for individuals having an overall rating of Stage 5 were at Stage 5, but the remaining scores indicate reasoning characteristic of Stages 3, 4, 6, and 7 as well. Overall, the stage utilization for Stages 5 and 6 cover a wider range of scores than that for Stages 3 and 4. The rather steep curves for the two extreme stages, 2 and 7, may be an artifact of their status as the lowest and highest RJI stages assessed.

The upward progression of overlapping distributions of scores portrayed in Figure 6.2 bears a striking resemblance to Rest's hypothesized graphic presentation of a complex stage model (1979, Figure 6-2.2, p. 66), which he notes depicts "a messier and complicated picture of development" (p. 65). Wood's empirically derived figure offers a more complex picture of the development of concepts of justification than has been described in more traditional cross-group or longitudinal analyses or than would be predicted by a simple stage model of development.

These longitudinal results strongly suggest that the development of reflective thinking as defined by the Reflective Judgment Model unfolds gradually over time, even among those enrolled in educational programs. The sequence of stage usage generally follows the order hypothesized by the model, with higher-stage usage being preceded by lower-stage usage. Regressions appear to be infrequent. Consistent with Figure 6.2, stages are characterized as waves across a mixture of stages, where the peak of a wave is the most commonly used set of assumptions. While there is still an observable pattern to the movement between stages, this developmental movement is better described as the changing shape of the wave rather than as a pattern of uniform steps interspersed with plateaus.

The finding that people commonly use multiple ways of reasoning may give educators a framework for understanding the variability they see within individual students. For example, a student who characteristically approaches problems using Stage 4 assumptions ("all opinions are equally valid as long as they include evidence"; "you can't evaluate opinions") may also surprise an observer with a carefully balanced perspective characteristic of Stage 5 reasoning. As Kroll (1992a) asserts, our educational interventions should be targeted at the leading edge of development, not its central tendency. In other words, educators should teach about and model higher-stage reasoning even when that is not the type of reasoning they may expect students to produce themselves. (This point is elaborated in Chapter Nine.)

Verbatim Examples of Changes over Time. To illustrate the substantive changes in reasoning that the foregoing numbers reflect, we next provide two verbatim examples from the RJI transcripts of participants in this study. As these examples show, changes in reasoning associated with changes in RJI scores do not necessarily reflect a change in the content of a person's basic point of view about an issue (such as whether the Egyptians did or did not build the pyramids or whether news reporting is biased or unbiased). Rather, changes in RJI scores reflect changes in the basis for holding that point of view and the person's assumptions about how she or he could know and understand the topic.

The first example is a response to the creation/evolution problem; it was given by a female high school student who scored at the 76th percentile on a scholastic aptitude test (MSAT). We have included excerpts from her first interview as a high school junior and then from her interview ten years later. (Ratings listed are based on the entire transcript of the problem being discussed, not just the excerpted portions.) This respondent's 1977 rating was dominant Stage 2 and subdominant Stage 3. (As before, *I* represents the interviewer and *R* the respondent.)

I: What do you think about the issue?

R: Like in the Bible, in the beginning God created the earth and everything. I think it's just an idea. How can you describe

to people who don't know many things about life? How can you explain to them that they evolved from a lower animal form? You can't. And so I think they tried to put it in a way that they would understand. He (God) thought that was the way they would understand it best. It doesn't really matter how He says it. It is the truth that He is saying. It's not like He's lying or something. It doesn't matter how He says it. It's just that He says it and tries to describe to the people at that time how things came about, in what order, and stuff like that.

I: 'On what do you base that point of view?

R: A lot of it is our pastor. I used to take a Bible study class with him. There's a lot to it, but I just don't see how people could take such a literal interpretation of the Bible because you just can't take everything literally. Because that's not the way it was meant to be taken.

I: How did you come to hold your point of view?

R: By listening to our pastor and other people and finding mistakes [in the Bible] myself that couldn't always be right.

I: Can you say you know for sure that the literal view presented in the Bible is not correct?

R: Yes, from my science classes and people I've talked to.

I: In this case, can you say that one point of view is right and the other point of view is wrong?

R: I'm tempted to say yes, but I can't because of what people think is truth. I think it's wrong, but you can't tell a great aunt that the Bible is wrong. You just can't say that to them.

I: What does truth mean to you?

R: Truth is like the facts or something like that—what actually took place.

At this point in her life, this young woman held the point of view that the literal creation story was not correct because of what others, such as her pastor, told her. She believed a more literal interpretation of the Bible was wrong but kept her views

to herself because of social convention. She held to a concrete view of the truth in her reliance on the "facts." Ten years later, as a college graduate, she offered a different explanation of her views. At this time, her rating was dominant Stage 6, subdominant Stage 7.

I: What do you think about these statements?

R: I tend to stay on the evolution side of it and believe the arguments on evolution as opposed to the creation of the world as it exists today because I think the evidence suggests evolution more than creation. Looking at things like the continental drift, different animals that evolved on different islands, and looking at the fossil records, I have more confidence in that.

I: Can you say you know for sure that we evolved?

R: Yes, I believe I know that for sure, but whether or not we have all of the facts today about evolution, I don't think we do. I think there's always more things being uncovered every day. I don't think we have the full story on it, but I believe that evolution occurred and that it is still occurring.

A decade after her first interview, this young woman continued to hold a similar view about the subject (she still endorsed the evolutionary view), but her rationale for holding it and her understanding of evidence had changed. This time she pointed to a variety of types of evidence she had considered. Further, she was able to reason to a conclusion while also acknowledging that knowledge about evolution grows and changes and that we do not have "the full story on it."

The second example comes from a young man who was also a high school student at the first testing and was responding to the problem about the building of the Egyptian pyramids. He scored at the 86th percentile on a test of scholastic aptitude (MSAT) when in high school. His 1977 rating was dominant Stage 3, subdominant Stage 2.

I: What do you think about this issue?

R: I am a science fiction buff, and I've read books a lot like

Chariot of the Gods, but what they taught us in World History is enough to believe that the Egyptians had all the mathematics and power that they needed to build them. What really proves that point is the way they had their property worked out. Each year the river Nile flows through there. It would flood every year. Well, they would all mark their property off by using special lengths of rope. . . . Everything would be perfect each year because of the perfectionism of their race. Everything worked out exactly. There's no other proof at all of visitors from outer space at that time.

I: Can you know for sure that the Egyptians built the pyramids without help?

R: I cannot. One of the things about me is I can never have a completely closed mind on a subject. So I'll say again that I'm four-fifths sure that they didn't have help, . . . but there's another interesting thing about this: pyramid power. Pyramids have preserving power or something, so maybe there's something more to the pyramids than we think.

I: Can you say that point of view is right and one point of view is wrong?

R: No, I can't because I don't have enough data on it. I don't know for sure; I wasn't living then. There are many different theories you could have about anything, if you don't know anything about it.

In this excerpt, the young man expressed some awareness that evidence is relevant when solving problems such as this one. Although the evidence he chose was not directly related to the issue, he seemed quite convinced that his "proof" was conclusive. Despite this, he argued that he couldn't know for sure how the pyramids were built because he didn't have enough data. This kind of reasoning (especially the emphasis on the quantity of information without reference to its quality) is typical of individuals whose thinking is in transition between Stages 2 and 3.

While the content of his point of view had not changed

ten years later, his reasoning was quite different. During this decade, he had graduated from college and had entered a graduate program. By this time, his rating was dominant Stage 7 and subdominant Stage 6.

I: What do you think about these statements?

R: I think the Egyptians did it all on their lonesome, but I would not completely rule out any help. I would like to invoke Occam's Razor here and to say that the most simple theory is generally the true one, as evidenced by the Copernican revolution. And it's just easier for me to believe that the Egyptians did this all by themselves rather than having the highly unlikely event of aliens coming to help them.

I: How did you come to hold that point of view?

R: Again, falling back on Occam's Razor, my general education about how to develop theories and how science has worked has led me to believe that generally the simplest explanation is the true explanation. To extrapolate, the Egyptians were a very sophisticated race as evidenced by the cultural artifacts that they have left in addition to the pyramids. We've also got evidence elsewhere that early civilizations had a very sophisticated means of measuring things like winter and summer solstice—that sort of natural phenomena—as evidenced by Stonehenge. So it seems reasonable to me to expect that the Egyptians could do this one on their own. In addition to their sophistication, they had a large population of slaves. From what little I've read about it in *National Geographic*, they apparently have collected at least some evidence suggesting that it was the sort of human effort which had created the pyramids. On the other hand, there is absolutely no evidence that aliens came to help, no hieroglyphics with UFO saucers and things like that. I would need convincing evidence that in fact that did occur. The circumstantial evidence suggested by people like *Chariots of the Gods* is just not enough.

I: OK. In light of all of that, can you say you know for sure that the Egyptians built the pyramids?

R: No, I cannot because there is no conclusive evidence. . . . I can say that the available evidence points toward the Egyptians building this on their own, so as comparing the two theories, I can say that this one fits the facts better than the other and thus, in relative terms, is more right.

In this excerpt, the respondent still strongly holds his same point of view on the topic but qualifies his assertions in ways that show a change in his assumptions about the certainty with which he holds his beliefs. He also shows how he has synthesized evidence from several sources as the basis for his conclusion and critically evaluates competing perspectives in terms of the quality of the evidence offered. Each of these characteristics is consistent with higher-stage reasoning.

At the end of the same interview, this young man commented on his own developmental processes, noting first that he enjoyed discussing these kinds of issues.

R: I think my education has certainly prepared me to deal with these sorts of issues. It's also given me a set of structured means to go about analyzing a gray area, and I can use that set of evaluation tools in a variety of situations. I think also it's just my natural bent towards liking to discuss stuff. I have friends who have similar bents, and we spend time just choosing sides and really fleshing out an idea.

I: Did you do that in course work as well, or was it more informal?

R: It was more from my friends. Most of my course work was them telling me. You really don't learn discourse in undergrad, and really in the first year of grad school, they're just cranking on you so you don't really learn it there either. But you do learn the tools there, but to use them, it's out in the real world.

This characterization of undergraduate course work, "them telling me," and graduate courses, "they're just cranking on you," is not a flattering portrayal of this man's formal education. How fortunate he was to have (and to have taken ad-

vantage of) opportunities to practice using his analytic skills “out in the real world.” Rich and diverse opportunities exist on college campuses in both curricular and cocurricular settings for students to be involved in the practical application of their learning, from cooperative education or internship experiences to student organizations and community service projects. This important aspect of a college education may provide a primary incentive for students to grapple with and make difficult choices about issues they define as important. Such conditions for the development of reflective judgment should not be left to chance.

Examination of Gender Differences in the Ten-Year Study

In the last decade, several researchers have proposed new developmental models that include elements they claim are distinctive to or prevalent among women (Baxter Magolda, 1992; Belenky, Clinchy, Goldberger, and Tarule, 1986; Clinchy and Zimmerman, 1982; Gilligan, 1982). In this section, we report the results of our examination of RJI scores by gender in the ten-year study. While these results are by no means conclusive (for example, they vary by the type of analysis and specific subgroups used), they yield some interesting patterns. We begin by examining the results of the three-way (time by group by gender) repeated measures GLM. Gender differences were not found in 1977 or 1979, but a group by gender interaction approached statistical significance in 1983 for those tested all three times. In 1987, when the same three-way GLM was used, the group by gender interaction also approached significance, and a gender main effect ($p < .05$) was obtained, with men scoring higher than women (see Table 6.3 for specific GLM results). (When only the scores of those who participated all four times were used for this analysis, the group by gender interaction and the gender main effect were nonsignificant; see Resource B, Table B6.1.) To better understand these results that suggest different patterns of development for women and for men, we conducted several follow-up analyses.

First, we ran a two-way (group by gender) GLM to see whether this finding could be explained by gender differences

on the new problem about the safety of nuclear power. However, both the gender main effect and the group by gender interaction were nonsignificant. We next asked whether men and women had scored differently on any of the four standard RJI problems. Based on a three-way (time by group by gender) repeated measures GLM, no significant gender differences were found for the Egyptian pyramids, news reporting, creation/evolution, or nuclear energy problems. However, for the problem on chemical additives in foods, an inconsistent gender effect was found: among the former high school juniors, men scored about one-third of a stage *lower* than women; among the former college juniors, men scored about half a stage *higher* than women; and among the former doctoral students, the scores were virtually identical. The findings of no difference between the scores of the women and the men on four of the five problems and the inconsistent ordering of scores on the fifth problem offer little support for the hypothesis that the obtained gender differences are attributable to differences on one or more of the four specific problems.

Second, in light of the educational-level differences reported above, we reexamined the data on educational attainment reported in Table 6.2 to see if there were different patterns in educational attainment across gender that might have contributed to these differences. Since the advanced doctoral students had already demonstrated a high level of educational attainment at the time of the first testing, we focused on the former high school and college students. We found that almost half of the men in these two groups (47 percent) had earned postbaccalaureate degrees (M.A., M.D., Ph.D.) by 1987; by contrast, only 15 percent of the women had earned advanced degrees. Since we have already suggested that educational attainment may have a strong influence on reflective judgment, the differences in the educational achievement of the men and women in this sample may have contributed to the gender differences in RJI scores at this time of testing. (Inconsistent patterns of RJI scores by gender have been found in other studies and are reported below.)

Age and Reflective Judgment Level

Another way to examine the development of reflective thinking over time is to examine the relationship between age and RJI scores. To do this, the modal RJI scores of all individuals who had been tested one or more times in the ten-year study were pooled and categorized by their age at the time of testing. For example, all four RJI scores of those participants who were sixteen years old at the time of the first testing and who had participated in all three subsequent testings were entered into these analyses for ages sixteen, eighteen, twenty-two, and twenty-six. By contrast, the score of a participant who was tested only once at age twenty would be entered once under this age category. Ages were then collapsed into five-year intervals. The results of this analysis are shown in Table 6.4.

The modal (or bimodal) reflective judgment stages by age category were as follows: sixteen- to twenty-year-olds, Stage 3; twenty-one- to twenty-five-year-olds, Stage 4; twenty-six- to thirty-year-olds and thirty-one- to thirty-five-year-olds, Stage 5; thirty-six- to forty-year-olds, Stage 6; forty-one-year-olds and

Table 6.4. Ten-Year Study:
Frequency Distribution of RJI Scores by Age.

Age in years ^a	RJI stage scores (%)					
	2	3	4	5	6	7
41 +					50	50
36-40				15	57	29
31-35			19	40	26	16
26-30			15	47	3	35
21-25		17	51	17	14	1
16-20	18	46	29	6	2	—

*Note: Numbers in bold print indicate the modal Reflective Judgment Model stage(s) for each age category. Some rows total 101 percent due to rounding. Total number of cases = 254.

^aFor ages sixteen to twenty, $n = 55$; for ages twenty-one to twenty-five, $n = 72$; for ages twenty-six to thirty, $n = 60$; for ages thirty-one to thirty-five, $n = 43$; for ages thirty-six to forty, $n = 14$; for ages forty-one and older, $n = 10$.

older, Stages 6 and 7. These findings suggest a strong linear relationship between age and the stage in the Reflective Judgment Model. The modal scores of virtually all of the participants who were tested at age thirty-six or older were at Stage 6 or 7, the stages associated with reflective thinking. Virtually none of the youngest age group and only a handful of those twenty-five and under fell into this category. The emergence of Stage 4 reasoning in these analyses is noteworthy. Recall that the hallmarks of this stage include the acknowledgment of the uncertainty of knowledge, the ability to differentiate between well- and ill-structured problems, and the ability to work with simple abstract concepts (including "evidence" and "knowledge"). These attributes and skills are commonly emphasized in many types of collegiate programs. Stage 4 was the modal score for individuals between the ages of twenty-one and twenty-five, which are the ages at which many traditional-age college students graduate and at which many adult learners are enrolled. Further, recall that virtually all of the former high school group earned an average RJI score of at least 4.0 by the time they were twenty-two years old (in 1983). In other words, by age twenty-two, even those who scored at the earliest stages of the Reflective Judgment Model six years earlier now acknowledged the uncertainty of knowledge and demonstrated their understanding of simple abstract concepts. The progress toward reflective thinking indicated by a shift from Stage 2 to Stage 4 reasoning should not be underestimated: the quasi-reflective reasoning characteristic of Stage 4 is a marked advance over the authority-based, pre-reflective thinking it replaces and lays the foundation for true reflective thinking in later adulthood.

These data suggest that age may be an important factor in the development of reflective thinking. However, the role of age alone should not be overstated on the basis of this study, since members of the sample were actively engaged in a variety of educational pursuits during the decade (that is, this is not a test of simple maturation) and since the sample itself is characterized by a high level of scholastic aptitude and is therefore not representative of adults in general. Whether this progression in scores is due predominantly to age, to education, or to

the interaction of the two in this sample remains open to speculation.

Conclusions Drawn

The longitudinal testings of individuals over a decade provided the opportunity to observe long-term changes in their reasoning. In 1977, individuals in each of the three age/educational groups were quite different from their matched counterparts in the other groups. Some of these differences remained intact ten years later (the higher scores of the former doctoral students); others no longer exist (the early differences in RJI scores between the former high school and college students). Over the ten years of the study, most of these individuals were actively involved in earning degrees in higher education, both baccalaureate and postbaccalaureate. To date, no other study of reflective judgment has followed a sample for such an extended period of time; accordingly, no other study has followed a group of high school students through their college years, and no other study has followed advanced doctoral students into middle adulthood. The rich data base that accrued from this study allowed us to begin to answer a variety of questions about the nature of the development of reflective judgment and of the effects of age and education on the ability to think reflectively.

First, however, it should be noted that the sample on which the ten-year study was based is distinctive in several respects. This sample was comprised of white adults from the upper Midwest and was therefore neither geographically nor ethnically representative. While RJI scores from cross-sectional studies (reported later in this chapter) have not been found to differ by region of the country or ethnic background, an unknown bias may be embedded in the homogeneity of this sample. Thus the comparability of this group to other groups is unknown, and the consistently high scores of the doctoral students may reflect not only the demands of doctoral study but a special and unknown characteristic of this particular group of people.

Second, the procedure used to select the high school and

college students (matching them on scholastic aptitude to the graduate students) yielded samples of students at both levels who had unusually high academic aptitude scores. The selection of students who had already shown themselves to be academically successful in 1977 may help account for the high rates of participation in college and postbaccalaureate degree programs until 1987. Nevertheless, in spite of the initial matching, significant differences in reflective judgment levels were found in the three groups. Further, these differences remained over time: the former graduate students started out at a high level and remained there; some of the former college juniors started out at the early or middle stages and progressed to the most advanced stages. The pattern was similar among the former high school students, but for this group, most started out at the early stages and showed evidence of successively higher scores at each time of testing.

The analysis of individual scores over time yielded strong evidence for development in reflective thinking: the observed changes in reasoning over the ten years of this study were clearly consistent with the posited sequence of stages of reasoning outlined in the Reflective Judgment Model. Moreover, the data informed our understanding not only about the nature of the development of reflective thinking among those involved in educational programs but also about the relatively slow pace of its development.

This finding, combined with the observation that individuals evidence reasoning from a variety of stages, serves as a reminder that the ability to think reflectively does not emerge fully formed but develops in a sequential fashion, with earlier stages building on prior stages and laying the foundation for subsequent stages. For example, a student learns to recognize the role of evidence before she composes a compare-and-contrast analysis of two strategies and before she uses the selected strategy to construct a debate-winning argument. Educators have long understood the power of breaking a task into its constituent parts. Since these parts are developmentally ordered as well, educators trying to teach students to think reflectively may want to consider the order in which they ask (or expect) students to put

the pieces together. For example, asking students to coordinate a panel discussion so that several points of view on a controversial topic are explained not only presumes that they are aware of (and perhaps appreciate) multiple points of view on the topic but that they understand that people hold different points of view for different reasons. The data presented here suggest that these assumptions should not be taken for granted and, further, that they emerge in a developmentally ordered fashion.

The findings from this study raise new questions about the course and influences of development among college-educated adults. For example, the influence of age, which presumably reflects a range of adult experiences, may have been underestimated in other studies based on shorter retest intervals. While the influence of formal education is consistent with other research (Pascarella and Terenzini, 1991) and may even be gratifying to some college educators, the role of specific collegiate experiences (from participating in debate clubs or departmental honorary societies to taking classes where reflective thinking is an explicit goal to serving in student leadership positions) has not been studied in this regard. Further, the effects of other life experiences (such as changing careers or making decisions about a child's nutritional or medical needs) on reflective thinking are not yet well understood and merit further study.

Other Longitudinal Studies of Reflective Judgment

In this section, we examine the patterns of data found across six other studies in which individuals were tested and then retested after a period of time to see whether their RJI scores had increased, decreased, or stayed the same (Brabeck and Wood, 1990; Polkosnik and Winston, 1989; Sakalys, 1984; Schmidt, 1985; Van Tine, 1990; Welfel and Davison, 1986) plus prior testing of the ten-year sample (King and others, 1983; King, Kitchener, and Wood, 1985; Kitchener, King, Wood, and Davison, 1989). These studies include twelve different samples and a total of 241 individuals, most of whom were tested twice, at intervals of from three or four months to four years.

Table 6.5 provides a descriptive overview of the seven

Table 6.5. Longitudinal Studies of Reflective Judgment.

Sample source	Sample size and educational level at first testing	Age at time 1	Time between testings	RJ scores				RJ growth over time
				Time 1		Time 2		
				M	SD	M	SD	
Polkosnik and Winston (1989)	16 college students from U. of Georgia	17-23	3 months	3.42	.44	3.44	.49	.02 (ns)
Sakalys (1984)	25 senior female nursing students in a research course at U. of Colorado	25	4 months	3.68	.58	3.87	.61	.19 (ns)
	25 senior female nursing students (control group) from U. of Colorado	25		3.59	.52	3.64	.45	.05 (ns)
	Overall			3.64	.55	3.76	.54	.10 ^a
Polkosnik and Winston (1989)	14 college students from U. of Georgia		6 months	3.42	.44	3.61	.57	.19 ^b
Brabeck and Wood (1990)	25 high school females, divided between high- and low-level critical thinkers	18-19	1 year	3.40	.40	3.56	.41	.16 ^c
Van Tine (1990)	21 high school freshman from a semirural school district in southern Colorado ^d	14	2 years	2.76	.36	3.26	.29	.50 ^e
King and others (1983) ^f	17 high school juniors from Twin Cities area	18	2 years	2.79	.51	3.61	.46	.82 ^e
	27 juniors from U. of Minn.	22		3.75	.72	4.18	.80	.43 ^e
	15 advanced doctoral students from U. of Minn.	30		6.03	.63	6.26	.48	.23 ^e
Schmidt (1985)	34 traditional-age freshmen from U. of Minn.	18	3 years	3.35	.26	3.56	.29	.20 ^e
	11 adult freshmen from U. of Minn.	21		3.47	.39	3.52	.34	.05 (ns)

Welfel and Davison (1986)		Not reported	4 years	Time 3	
13 engineering freshmen from U. of Minn. 12 social sciences freshmen from U. of Minn. Overall				3.69	.25
				4.20	.65
				3.56	.31
14 high school juniors 22 college juniors 12 advanced doctoral students				4.17	.36
				3.62	.28
				4.19	.52
10 high school juniors 17 college juniors 11 advanced doctoral students				3.62	.28
				4.19	.52
				3.62	.28
King, Kitchener, and Wood (1990) ^b				3.62	.28
				4.19	.52
				3.62	.28

Notes: Data are for seven studies, twelve samples, 241 individuals. The numbers of studies and individuals are based on the first testing only for the Kitchener and King (1981) and the Polkosnik and Winston (1989) samples, even though the samples were retested more than once. Articles are listed in order of length of time between testings.

^a $p < .01$.

^bPolkosnik and Winston (1989) report this as a significant difference ($p < .01$); however, in an independent analysis of the data reported in Polkosnik (1985), we were unable to replicate the time main effect they report (Wood, 1993a).

^c $p < .05$.

^dSample originally tested by McKinney (1985).

^e $p < .001$.

^fThis is a follow-up study of the original Kitchener and King (1981) sample; the scores include only those of individuals who participated two years later.

^gThis is a follow-up study of the original Kitchener and King (1981) sample; the scores include only those of individuals who participated two, four, and six years later.

^hThis is a follow-up study of the original Kitchener and King (1981) sample; the scores include only those of individuals who participated two, four, six, and ten years later.

longitudinal studies, including sample characteristics of each study, the duration of each study, and how RJI scores changed over time. (This number counts only the first retest of the two samples that were tested more than twice. The four testings of the ten-year study described in the preceding section are counted here as one study but listed separately for each testing to illustrate differences by length of time between testings.) The most noteworthy finding from this group of studies is that *in every sample tested, the scores either stayed the same or increased over time*. Further, with two exceptions, *the mean score increased significantly for all groups tested at one- to four-year intervals*. (The exceptions were the adult college freshmen in Schmidt's [1985] sample and the advanced doctoral students in the Kitchener, King, Wood, and Davison [1989] and the King, Kitchener, and Wood [1990] studies.) These findings offer strong support for the claim that the ability to make reflective judgments changes developmentally over time.

Another noteworthy finding shown in this table is that the amount of change over time appears to be strongly related to the duration of time between testings. For example, the two smallest increases were found in studies of only three or four months' duration (Polkosnik and Winston, 1989; Sakalys, 1984). (Schmidt's [1985] adult freshmen, who were retested after three years with a net gain of .05, are an exception to this pattern.) By contrast, the greatest increase was 2.58 (King, Kitchener, and Wood, 1990), which was found in the ten-year retesting of the group of former high school students discussed earlier in this chapter.

Only the three samples tested over a short period of time showed nonsignificant change. The study of shortest duration in which significant changes were reported was that by Polkosnik and Winston (1989) in their second retest. After six months, they found that the students' scores had increased modestly but significantly over their first-testing scores (a main effect that Wood [1993a] was unable to replicate in an independent analysis of these data). By contrast, studies with at least a year between testings consistently reported significant positive change in RJI mean scores (with the two exceptions noted above).

Although longitudinal data used to validate a develop-

mental model typically focus on increases in functioning over time, "regressions," as indicated by decreases in scores, also provide useful information in understanding changes in intellectual performance over time. Information about the incidence of regressions is reported for several of the studies discussed above (Brabeck and Wood, 1990; King and others, 1983; King, Kitchener, and Wood, 1985; Kitchener, King, Wood, and Davison, 1989; Sakalys, 1984; Schmidt, 1985; Welfel and Davison, 1986). In each of these studies, the proportion of the sample for which the RJI mean score either remained constant or increased ranged from 84 to 100 percent, which is consistent with that found in the ten-year study. The finding that stability rather than development characterized the thinking of some participants at the longitudinal testings suggests that the rate of change across the stages varies across individuals. Of course, the test of this observation would require a study of longer duration, with more than two observations of individuals' reasoning over time.

Because of the short-term duration of Sakalys's study, her data on reversals (Sakalys, 1982) may be used as an estimate of measurement error. In her study, raters were blind to time of testing of the responses being rated; that is, they were unaware of whether a given transcript was from the first or second time of testing. Sakalys reported reversals (downward change of more than .1) in 16 percent of her cases. Since the reported incidence of reversals in all the studies of longer duration occurred less frequently than this and because of the overall low incidence of reversals, observed "regressions" may be considered as measurement errors.

Despite the consistency between these data and the description of intellectual development proposed by the Reflective Judgment Model, alternative explanations that might account for or contribute to increases in reflective judgment scores over time still merit consideration. Since questions regarding both test-retest and internal consistency reliability are addressed in Chapter Five, we focus here on the question of whether the obtained increases in RJI scores over time can be attributed to differences between those who continued their participation in

the ten-year longitudinal study and those who were not retested. For example, a person who was keenly aware of reasoning more complexly at the second testing than she was at the first might be more likely to agree to participate in a follow-up interview in order to show her new ways of thinking. By contrast, a person who earned a low RJI score because she did not enjoy explaining or defending her beliefs in detail and found the task too challenging to be enjoyable might be less inclined to continue her participation. Following this logic, any obtained increases in scores might reflect a selection effect on the part of the participants. Several studies have tried to assess whether this selection effect occurred by comparing the scores of participants who dropped out of the study to those of participants who returned for retesting. When such comparisons have been made, the differences between dropouts and ongoing participants were nonsignificant (King and others, 1983; Kitchener, King, Wood, and Davison, 1989; Schmidt, 1983). Another way to approach this question is by looking at the similarity in the patterns of responses between those who participated all four times and those who took part in at least one longitudinal retest. With the few exceptions noted above, these patterns were very similar.

A second alternative explanation is that increases in RJI scores reflect practice or familiarity with the test rather than development in reasoning. It is plausible that those who were interviewed may have given serious thought to one or more of the dilemmas between testings and may even have sought out additional information on the topics, especially if they had already agreed to be retested in the future. To the degree that this happened, taking part in the interview may have indirectly served to stimulate their intellectual development. The only direct evidence of this possibility is reported above from the fourth testing of the sample in ten years. In this study (King, Kitchener, and Wood, 1990), a new problem on the safety of nuclear energy was added to the four standard RJI problems as a test of familiarity. The researchers reasoned that if familiarity with the standard problems accounted for higher scores between testings, then participants would score higher on the standard problems than on the nuclear energy problem. However, no differences were found in participants' scores between problems.

In summary, findings from a number of longitudinal studies offer additional evidence that development in reflective judgment occurs slowly and steadily over time and that the increases in scores are not an artifact of selective participation or practice. Further, these studies show that stability and development are much more common than regression in reflective thinking. Significant increases in RJI scores were reported in all studies of at least a year's duration, predominantly among those involved in collegiate programs. Answers to the questions of whether age or education is the more powerful influence on the ability to make reflective judgments and whether similar growth also occurs among other groups of people will require additional research.

Cross-Sectional Studies of Reflective Judgment

Cross-sectional studies provide an index of levels of reflective thinking across different groups of individuals. We have organized these individuals into ten educational levels that range from early high school to advanced graduate school plus non-student adults. In this section, the results of studies that tested almost fifteen hundred students will be reviewed; these students were from diverse geographical regions of the United States and ranged in age from the teen years to middle adulthood. Data from specific studies are summarized in the text; the interested reader may consult Resource B: Statistical Tables for the findings of each study on which the following summaries are based; there is one table for each major educational level plus one for studies that tested nonstudent adults.

We begin this section with an examination of the RJI scores of high school students, which we then compare with the RJI scores of college students at various levels (such as freshmen and seniors) and with the scores of those who have not only completed college but pursued postbaccalaureate degrees (graduate students). Finally, we contrast the scores of nonstudent adults with the scores from the other groups. Our aim in this section is to provide the reader with a broad overview of how individuals at different age and educational levels think about and resolve ill-structured problems.

Before proceeding with this discussion, we would like to acknowledge that increases in reflective judgment across educational levels should not be attributed solely to the effectiveness of educational institutions; as already noted, age and educational level are frequently confounded in studies that contrast students of different grades or classes. Age brings with it more opportunities for a broad range of life experiences. As students complete more class credits, they also age and face a myriad of life events and personal decisions (such as work experiences, increased family responsibilities, involvement in community organizations, or decisions about health care). Moreover, educational level reflects both a *degree of exposure* to the classes and programs of the institutions and a *degree of success* in mastering the academic demands necessary for advancement to the next educational level. Those who graduate are also the “survivors” of the experience, those who have been successful in negotiating social and institutional systems in support of this goal. Nevertheless, if students at increasingly advanced educational levels do evidence greater skill in reflective thinking, this may also suggest the effectiveness of the educational programs in which the students participated.

Table 6.6 provides a summary of how Reflective Judgment scores differ by educational level. This table was constructed by averaging the RJI mean scores from all available studies* for each educational level to provide a composite view across studies. The mean scores in this table correspond to stages: an RJI mean score of 3.0 indicates that the average stage usage for a given educational level was Stage 3. The table shows that RJI scores increase slowly but steadily from high school to col-

*A series of studies on the Reflective Judgment Model used a “modified” version of the RJI (Barone, 1989; Currier, 1989; Liberto, Kelly, Shapiro, and Currier, 1990; Milman, 1988; Montecinos, 1989; Pape and Kelly, 1991). These studies are not reviewed here because they used a nonstandard version of the RJI and/or a nonstandard scoring procedure and did not use trained raters. As a result, it is impossible to interpret their findings or meaningfully compare them to findings from other studies. An additional study by Kitchener and Wood (1987) is not included because the sample consisted of West German university students whose educational levels were not comparable to those of university students in the United States. This study is discussed later in this chapter under cross-culture differences.

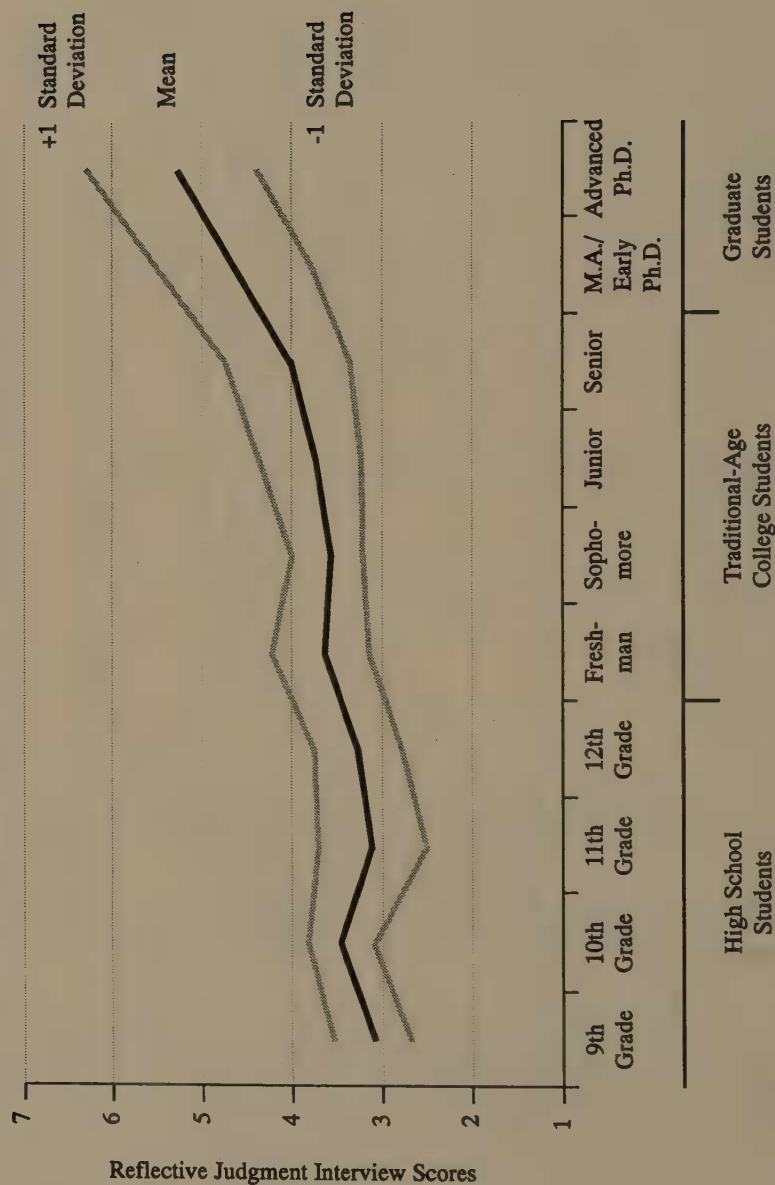
Table 6.6. Reflective Judgment Scores by Educational Level.

<i>Educational level</i>	<i>Average RJI score</i>	<i>SD</i>	<i>n</i>
High school			
Grade 9	3.08	.41	57
Grade 10	3.46	.35	15
Grade 11	3.12	.61	33
Grade 12	3.27	.51	67
Average	3.19	.50	172
College (traditional-aged students)			
Freshman	3.63	.53	329
Sophomore	3.57	.43	89
Junior	3.74	.59	159
Senior	3.99	.67	369
Average	3.79	.61	946
College (nontraditional-aged students)			
Freshman	3.57	.42	78
Sophomore	4.30	.59	13
Senior	3.98	.74	46
Average	3.78	.61	137
Graduate			
Master's/Early doctoral	4.62	.81	126
Advanced doctoral	5.27	.89	70
Average	4.76	.85	196

lege and from college to graduate school. Among high school students, the average RJI score is 3.2. Among undergraduate students, the overall score was 3.8 for both traditional-age and adult students. By contrast, the average RJI score among the graduate students, 4.8, was a full stage higher than that for the college students. These scores are graphed in Figure 6.3, which shows the mean score for each level (the middle line) and also the variability within educational level, as reflected by the lines that graph the upper and lower limits of the standard deviation for each educational level. The RJI scores of two-thirds of the samples tested fell within the range of scores indicated by the top and bottom lines on this graph.

It is noteworthy that not only do the mean scores of the samples increase across the educational levels but the amount of variability in responses also increases (indicated in Figure 6.3 by

Figure 6.3. Reflective Judgment Interview Scores by Educational Level.



Note: Scores are averaged across all studies reviewed here that reported RJI scores for students ($n = 1,334$) at these educational levels. For specific scores, see Table 6.6

the wider band of scores at the more advanced levels). This suggests a pattern of mixed success in improving reflective thinking (King and Kitchener, 1993). Although the RJI scores increase, even those who are about to complete a college degree do not consistently evidence reflective thinking. We are not suggesting that these students are not able to think reflectively but that they did not appear to do so in their typical manner of reasoning elicited in the context of the RJI. Further, we acknowledge that higher-level thinking is harder to sustain without ongoing practice and contextual support (Kitchener and Fischer, 1990) and that this factor might also have influenced their scores.

While averaging scores in this manner offers a general picture of the trends in reflective thinking across educational levels, it also masks the variability inherent across samples, even those at the same educational level. While we have aggregated scores within educational level, we recognize that there is a great deal of variability between samples (for example, some control for academic ability and some do not) and that comparisons between educational groups thus constituted should be made cautiously. Nevertheless, our purpose here is to provide a broad view of how students' ability to make reflective judgments differs by educational level, not to provide a specific prediction of students' reasoning based solely on their educational level.

Reflective Thinking Among High School Students

We set the context for an examination of changes in reflective thinking across educational levels by examining the scores of high school students; this group provides relevant information about the levels of reflective thinking that exist prior to the students' enrollment in college. We believe that university faculty and staff need to understand the foundation upon which they have the opportunity to build in teaching students to think reflectively, a foundation that may still be in place during the freshman year of college. Five studies have sampled high school students. (Average RJI mean scores for high school students appear in Resource B, Table B6.3.) About half (five of eleven) of the mean scores for the high school samples were below 3.0. There

have been too few studies of tenth and eleventh grade students (one and two, respectively) to be able to draw meaningful conclusions about their reasoning. As a consequence, the fluctuations in scores that appear in Figure 6.3 for tenth and eleventh graders should be interpreted with caution. The mean scores for ninth and twelfth graders are 3.1 and 3.3, respectively, indicating that, on average, high school students tend to hold epistemic assumptions consistent with Stage 3 of the Reflective Judgment Model. Students who hold these assumptions typically maintain that certain truth is ultimately attainable but that in some areas it is not currently available. When faced with uncertainty that they believe is temporary, they are often confused about what to believe. Frequently, they have difficulty relating evidence to their opinion and fall back on a "whimsical" style of reasoning, choosing to simply believe what they want to believe, at least until the truth is known. In other cases, they maintain the belief that authorities have the truth, and they rely on authorities to inform them of what to believe about such dilemmas. (See Chapter Three for a detailed description of reasoning at each stage of the Reflective Judgment Model.)

It should be noted that the scores summarized here are probably not representative of high school students across the United States: of the eleven samples listed in Table B.6.3 in Resource B, seven were from studies that sampled academically talented high school students (Kitchener and King, 1981; Kitchener, Lynch, Fischer, and Wood, 1993;* McKinney, 1985). While such sampling can assist us in comparing high school students with college students by controlling for academic achievement within studies, it may have yielded a pattern of results germane only to academically talented students and not to high school students in general. However, despite the number of studies that attempted to control for the effects of academic ability or achievement, it is noteworthy that the mean score for high

*To avoid confusion, the Kitchener, Lynch, Fischer, and Wood (1993) reference will be cited throughout this chapter even though RJI scores are reported there by age rather than by educational level. RJI scores from this study are reported by educational level in Kitchener (1992) and Wood (1993a).

school seniors was almost half a stage lower ($p < .001$) than the mean score for college freshmen. (See Chapter Seven for a discussion of the effects of intellectual ability on reflective judgment.)

Reflective Thinking Among Traditional-Age College Students

Traditional-age college students have been tested in twenty cross-sectional studies ($n = 946$) that have used the standard RJI format and scoring procedure (Bowen, 1989; Brabeck, 1983a; Evans, 1988; Glatfelter, 1982; Kelton and Griffith, 1986; Hayes, 1981; King, J. W., 1986; King and Parker, 1978; King, Taylor, and Ottinger, 1989; King, Wood, and Mines, 1990; Kitchener and King, 1981; Kitchener, Lynch, Fischer, and Wood, 1993; Mines, King, Hood, and Wood, 1990; Nickerson, 1991; Polkosnik, 1985; Sakalys, 1984; Schmidt, 1985; Shoff, 1979; Strange and King, 1981; and Welfel, 1982b). Counting the RJI scores reported for each subsample tested within each study yielded a total of forty-four samples. (These are listed by RJI score in Resource B, Table B6.4.)

Most studies of traditional-age college students that used the RJI have tested freshmen and seniors (twelve and twenty samples, respectively); fewer have tested sophomores (five samples) and juniors (seven samples). In general, the freshman students in these samples evidenced the use of assumptions of Stages 3 and 4 of the Reflective Judgment Model. Both the overall mean score and the median score across all freshman samples ($n = 329$) were 3.6. Only one sample averaged 4.0 or above, and this was a sample of academically talented students (Kitchener, Lynch, Fischer, and Wood, 1993). The means of the remaining samples of freshmen clustered between 3.2 (the lowest score) and 3.5 and between 3.7 and 3.9. The prevalence of Stage 3 reasoning means that like the high school students, many of the college freshmen articulated the belief that absolute truth is only temporarily inaccessible, that knowing is limited to one's personal impressions about the topic (uninformed by evidence), and that most if not all problems are well structured (defined with a high degree of certainty and completeness). However, as is apparent in Figure 6.3, two-thirds of the freshmen rea-

soned between Stages 3 and 4, suggesting that other freshmen were beginning to accept the concept that uncertainty may be an ongoing characteristic of the knowing process. Consequently, they are beginning to accept the idea that some problems are truly ill structured.

The overall mean score for the college sophomore samples was 3.6 ($n = 89$), the same overall mean as that found for the freshmen. The overall mean score for the juniors ($n = 159$) was slightly higher, 3.7. The highest average mean score among the college student samples was that of the seniors ($n = 369$), 4.0, which is significantly ($p < .001$) higher than the freshman average, by about half a stage. By contrast with the freshman samples, all the senior samples had an RJI mean score of over 3.5, and half had a mean score of 4.0 or higher. In other words, Stage 4 reasoning was used by all the senior samples. The range of scores for the seniors was more variable than that of any other class level—from a low of 3.6 (Glatfelter, 1982) among seniors at a state university in the West to a high of 5.0 (Kelton and Griffith, 1986) among seniors enrolled in a highly selective liberal arts college in the Southeast.

A difference of half a stage between the averaged scores of college freshmen and seniors may not suggest a dramatic metamorphosis in thinking. However, while the numerical difference is small, the development in reasoning that it reflects is nevertheless noteworthy. The Stage 4 reasoning prevalent among the senior samples is clearly more adequate and defensible than the reasoning it replaced. Recall that at Stage 4 comes the acknowledgment that uncertainty is not just a temporary condition of knowing. It is at this stage, too, that students begin to use evidence systematically to support their judgment, a development of no small consequence. As Barry Kroll (1992b) has described this type of development, students are abandoning "ignorant certainty" (characteristic of earlier stage reasoning) in favor of "intelligent confusion." While promoting students' confusion is not a goal typically published in college mission statements, Kroll's observation serves to remind us that intelligent confusion is a developmental advance over ignorant, dogmatic certainty and that it paves the way for more thoughtful, reasoned

judgments that may follow. In this sense, then, the development from Stage 3 to Stage 4 reasoning can be hailed as an important step in the process of learning to make reflective judgments.

Nevertheless, the prevalence of Stage 4 reasoning (most apparent among the senior samples) also indicates that many college students are at a loss when asked to defend their answers to ill-structured problems, for Stage 4 has as a major characteristic the assumption that, because there are many possible answers to every question and no absolutely certain way to adjudicate between competing answers, knowledge claims are simply idiosyncratic to the individual. In other words, an answer to an ill-structured problem is seen as merely an opinion. Further, many college students are not demonstrating an ability to articulate the role of evidence in making interpretations (Stage 5) or to defensibly critique their own judgments or explanations as being in some way better than or preferable to alternative explanations (Stage 7), which is a hallmark of reflective thinking. So while college educators can take solace in the observation that upward movement toward being able to make reflective judgments is identifiable across educational levels during the college years, college seniors do not typically articulate concepts that underlie reflective thinking. However, as noted in Chapter Five, Kitchener, Lynch, Fischer, and Wood (1993) found that under certain circumstances, college seniors can demonstrate an understanding of Stage 6 reasoning. Kroll (1992b) also presents convincing evidence that college freshmen can learn to think reflectively in college classes that are structured to stimulate questioning and in which they are expected and encouraged to make thoughtful, reasoned judgments.

The King, Taylor, and Ottinger (1993) study is noteworthy because it also examined the relationship between RJI scores and students' academic standing and graduation status five years later (these data are reported in King, Taylor, and Ottinger, 1993). The sample for this study consisted of 146 African-American students; five years after these students took the RJI, their GPA and graduation status were examined. By means of a stepwise multiple regression procedure, RJI scores were found to predict students' college GPA ($p < .001$). Further, those who

had graduated scored significantly higher on the RJI than those who were on academic warning, were suspended, had been dismissed, or had withdrawn from the university.

In light of this body of research on the reflective judgment of college students, it may be helpful to reexamine the educational proposals that currently serve as the foundation for collegiate educational reform. For example, the following observation was made in a major report on curricular reform in higher education: "By attending to the knowledge claims of the major over time and by treating increasingly complex matters from multiple points of view, students discover that nothing is self-evident, that nothing is simply 'there,' that questions and answers are chosen and created — not given — and that they always are framed by context; for that reason, they always are contingent" (Association of American Colleges, 1991, p. 13). This document, *The Challenge of Connecting Learning*, persuasively and eloquently presents a vision of the kind of reasoning that should be associated with the goals of undergraduate liberal education and certainly with the completion of a baccalaureate degree. Underlying it are the assumptions that answers are contingent and knowledge is contextual. These assumptions reflect Reflective Judgment Stage 5 reasoning, which, as shown above, is a full stage above the average RJI score of college seniors.

A second example from the same document illustrates a related discrepancy between educational goals and many students' reasoning: "Students cannot be allowed to be content with the notion that issues may be addressed by any number of equally valid formulations among which they cannot choose. They must learn to discriminate by arguing, and they must realize that arguments exist for the purpose of clarifying and making choices" (Association of American Colleges, 1991, p. 14). In this example, the skills associated with learning to discriminate the valid from the invalid (or less valid) formulations are associated with the upper stages of the Reflective Judgment Model (Stages 6 and 7), as is the ability to make discriminating arguments. And while some college students show an appreciation or understanding of the necessity of making informed, thoughtful choices about ill-structured problems, most college students do not typically evidence high-level reflective thinking. As King

(1992a, p. 8) notes: "Teachers who attempt to teach about this process . . . without attending to students' underlying epistemic assumptions will probably be very frustrated, and their students will probably be dissatisfied as well." Educators who understand their students' epistemic assumptions will be able to make more informed decisions when choosing assignments or activities designed to help students learn about, practice, and improve their reflective thinking skills.

Dings (1989) studied faculty members' perceptions of the epistemic assumptions of college freshmen and seniors (described by the Reflective Judgment Model) and their evaluation of their own assumptions. He found that faculty members considered Stage 2 reasoning to be most descriptive of freshman students, Stage 6 reasoning to be most descriptive of senior students, and Stage 7 to describe their own reasoning. These findings suggest that faculty members are underestimating the cognitive complexity of freshmen and may be overestimating the cognitive skills of seniors. Dings's study raises questions about how educators arrive at assumptions about students' reasoning skills, how they translate these assumptions and expectations about student reasoning skills into assignments and grading criteria, how students understand these expectations, and whether and how discrepancies between educators' assumptions and students' skills are addressed. It also raises the question of the degree of variability teachers expect to see in students' responses. For example, Lynch (1989) found that in using a recognition task, college students could understand information about a stage above where they could produce a response on the RJI. Whether discrepancies such as those found by Dings result in frustration and low performance or whether they serve to challenge students to operate closer to their optimal level remains to be seen. Studies conducted to date have not addressed this question; we encourage further research in this area.

Reflective Thinking Among Nontraditional-Age College Students

A discussion of differences in reflective thinking across the college years must also take into account the prevalence of adult

learners on college campuses, who are now approaching a majority (U.S. Department of Education, 1991). With this demographic change, a correspondence between age and class level can no longer be assumed. RJI mean scores have been reported for five studies that tested nontraditional-age students (Glatfelter, 1982; King, J. W., 1986; Schmidt, 1985; Shoff, 1979; Strange and King, 1981). These are listed in Resource B, Table B6.5. In these studies, a total of 137 adult learners were tested. Such studies are important in informing the question of whether educational level or age has a more potent influence on the development of reflective judgment. Among college students, this question is of particular interest since adult learners have been found to differ from traditional-age college students on such dimensions as learning style, motivation for attending college, and assertiveness (Apps, 1981; Brookfield, 1987; King and Bauer, 1988; Knowles, 1984; Mezirow and Associates 1990). The studies described here examined whether differences in reflective thinking exist as well.

The scores for the freshman adult learners ranged from 3.3, earned by a sample of entering continuing education students whose mean age was twenty-three years (Shoff, 1979), to 3.8, earned by a sample of twenty-two-year-old students (Strange and King, 1981). By contrast, the lowest mean for a sample of college senior adult learners (all over twenty-five years old) was 3.7 (Glatfelter, 1982), which is very close to the highest score for any freshman sample among the nontraditional samples. The highest RJI mean score for an adult college senior sample was 4.4 (Strange and King, 1981), earned by a group of twenty-six-year-old seniors. Adult college seniors scored significantly higher than adult freshmen ($p < .001$).

In contrast to the differences found on other educationally relevant dimensions, adult students do not appear to be dramatically different from their younger counterparts in terms of their reflective thinking. The mean scores by class are remarkably similar to those of traditional-age students at the same class level: the overall mean score for both adult and traditional-age freshmen was 3.6; the overall mean score for both adult and traditional-age seniors was 4.0. However, the range of scores

for the adult seniors was narrower (3.74 to 4.41) than that for the traditional-age seniors (3.56 to 5.02). The single small sample of sophomore adult learners scored higher than traditional-age sophomores (4.3 and 3.6, respectively).

Any interpretation of these findings about the reflective thinking of adult learners must take into account the variability in definitions of adult learner that have been used, reflecting a range of age-based criteria. Second, since many adult learners are enrolled part-time, readers are reminded that being a senior reflects a wide range of years of enrollment for many adult learners. Third, there is a much smaller body of research on the reflective thinking of adult learners ($n = 137$ students) compared with that of traditional-age students ($n = 966$ students). These caveats make for more cautious interpretations.

These studies suggest that adult students also hold assumptions about knowledge that limit their ability to make judgments on the basis of critical analysis (at least as measured by the RJI). As discussed at the beginning of this chapter, age alone is not predictive of the ability to think reflectively. But for students who have delayed their pursuit of a college degree, age may be strongly related to other relevant factors, such as readiness to learn or openness to being challenged or to entertaining divergent views, and these students may experience a developmental spurt during college as a result of this combination of factors. The data provide additional support for the argument that attending college plays an important role in the development of reflective thinking, and we encourage further research to help sort out and better understand the relationships between the attitudinal, readiness, and cognitive ability factors so that adult learners can be better served by college faculty and staff in pursuit of this goal.

Reflective Thinking Among Graduate Students

Graduate students have consistently earned the highest RJI scores of any students tested: their assumptions about knowledge and knowing are more consistent with the upper stages of the Reflective Judgment Model. Seven studies have used the

RJI to test a total of 196 graduate students (Brabeck, 1983a; King, Wood, and Mines, 1990; Kitchener and King, 1981; Kitchener, Lynch, Fischer, and Wood, 1993; Kitchener and Wood, 1987; Lawson, 1980; and Mines, King, Hood, and Wood, 1990). Their scores are listed in Resource B, Table B6.6. These studies included both graduate students who were working on a master's degree or were early in their doctoral studies (up to two years) and those who were at an advanced level of doctoral study (more than two years); they have been classified accordingly. (More precise classifications were not possible because of the general way in which several samples were described—for example, as “graduate students.” Where graduate samples were not identified by specific degrees, categorizations were based on other indicators of degree completion, such as the number of credits earned or other relevant descriptors.)

The age/educational effect is again apparent for this sample: the early-level graduate students used epistemic assumptions consistent with Stages 4 and 5, and the advanced-level graduate students used assumptions consistent with Stages 5 and 6. Overall, advanced doctoral students scored almost three-quarters of a stage higher ($M = 5.3$) than did graduate students earlier in their programs ($M = 4.6$). It is important to note that students enrolled in graduate programs were the first to consistently demonstrate the use of Stage 5 assumptions on the RJI. Further, only at the advanced doctoral level was there evidence of the consistent use of Stage 6 assumptions on the RJI.

In general, the graduate students evidenced the ability to offer a balanced perspective on a controversial issue, such as understanding why different people might approach a given problem from different perspectives and how that would lead them to focus on different types of evidence. Moreover, at the advanced doctoral level, they evidenced the ability to explain how they arrived at a judgment about an issue, a judgment they defended as more reasonable or more plausible than other alternatives or as the best interpretation of the available evidence. In short, they more consistently demonstrated the ability to make truly reflective judgments.

The attainment of this goal may reflect a variety of fac-

tors. For example, graduate admissions committees may select students who have already demonstrated advanced reflective thinking skills. By contrast, the higher scores of the doctoral students could be attributed to the intellectual demands of graduate study, from course assignments to research projects. It may not be coincidental that students who scored at the advanced levels, predominantly doctoral students, were enrolled in educational programs that require the writing of a dissertation. This undertaking frequently demands a careful analysis of an issue that acknowledges competing perspectives, followed by the creative synthesis of ideas and the requirement of taking the responsibility for explaining one's own point of view and defending it in light of competing alternatives.

One study found disciplinary differences in reasoning about ill-structured problems: King, Wood, and Mines (1990) found that graduate students in social science disciplines scored significantly higher on the RJI than did their counterparts who were enrolled in mathematical sciences, a difference that remained significant even when GRE scores were partialled out. These authors suggest that this difference may be a function of the emphasis (or lack thereof) given to ill-structured problems within each discipline. Little is currently known about the specific demands of graduate study that contribute to (or hinder) students' abilities to make reflective judgments. Given the multitude of ill-structured problems that professionals face today and the need to make defensible, reasoned choices about these issues, we encourage more research into ways in which graduate study can be structured to better prepare students to make such judgments.

Reflective Thinking Among Nonstudent Adults

While the preponderance of the research on reflective judgment has focused on students at various educational levels, six studies have examined how adults who are not formally enrolled in educational programs reason about ill-structured problems (Glatfelter, 1982; Glenn and Eklund, 1991; Josephson, 1988; Kelton and Griffith, 1986; J. W. King, 1986; Lawson, 1980). In

most cases, the nonstudent adult samples were included as comparison groups with samples of college students, especially adult students; this design provides a control for age when studying educational effects. Eight samples of nonstudent adults have been tested to date ($n = 191$); means are listed in Resource B, Table B6.7. Comparing the RJI scores of those samples in which participants had and had not earned college degrees yielded a very distinct pattern in which those without college degrees consistently scored lower. Further, there was no overlap in the range of mean scores between the samples with and without college degrees: the mean scores for the samples without college degrees ranged from 3.5 to 3.7; the mean scores for the samples with college degrees ranged from 4.0 to 5.2. Even the sample of adults who had been enrolled in a collegiate program but then dropped out of college (Glatfelter, 1982, classified here under those without college degrees) fit this pattern. The two highest mean scores in this group were earned by samples with postbaccalaureate degrees (Glenn and Eklund, 1991; Josephson, 1988).

The strongest contrast between college-educated and non college-educated adults is provided in the Glenn and Eklund (1991) study. These researchers administered the RJI to two groups of participants who were at least sixty-five years old (thus controlling for age) but who differed in terms of their educational attainment. The first group consisted of adults with up to a high school education; their RJI mean score was 3.7, which is about half a stage higher than the overall mean score among high school seniors (3.3) and closer to the average for the college samples (3.8). The second group consisted of retired faculty members with doctorates; the RJI mean score for this group was 5.2, which is comparable to the scores earned by advanced graduate students.

How do various types of life experiences affect reflective thinking? Several provocative findings on the effects of given life experiences on moral development may inform our understanding of the life experiences that affect reflective thinking. In summarizing this research on moral development, Rest (1986, p. 57) notes: "The people who develop in moral judgment are those who love to learn, who seek new challenges, who enjoy

intellectually stimulating environments, who are reflective, who make plans and set goals, who take risks, who see themselves in the large social contexts of history and institutions and broad cultural trends, who take responsibility for themselves and their environs." It seems reasonable to assume that the same is true for people who develop in their ability to make reflective judgments.

Research on Gender Differences in Reflective Judgment

Do women and men use comparable sets of assumptions about knowledge when resolving ill-structured problems? This question has frequently been raised in light of the prevalence of models or theories of development that have been derived exclusively or predominantly from male samples (Chickering, 1969; Erikson, 1963; Kohlberg, 1969, 1984; Levinson, 1978; Perry, 1970). The underlying concern here is that such models may not offer accurate or adequate descriptions of the development of women and that research based on these models is inherently biased. Two recent developmental models that address gender-related differences in cognitive development are those of Baxter Magolda (1992) and Belenky, Clinchy, Goldberger, and Tarule (1986). (For analyses of persisting beliefs regarding gender differences in moral development, see Brabeck, 1983b, and Mednick, 1989.) Even though the Reflective Judgment Model was derived from interviews with equal numbers of women and men, questions of gender differences continue to be raised. Most of the studies reviewed in this chapter included both women and men and are used as the basis for our discussion of gender differences in reflective judgment. Studies that tested all-female samples (Brabeck, 1983a; Brabeck and Wood, 1990; Glatfelter, 1982; Hayes, 1981; J. W. King, 1986; Josephson, 1988; Nickerson, 1991; Sakalys, 1984) are not included in this discussion.

In addition to the longitudinal studies based on the original Kitchener and King (1981) study, seventeen other studies that used the RJI included both male and female participants (Bowen, 1989; Evans, 1988; Glenn and Eklund, 1991; Kelton and Griffith, 1986; King, Taylor, and Ottinger, 1993; King,

Wood, and Mines, 1990; Kitchener and Wood, 1987; Kitchener, Lynch, Fischer, and Wood, 1993; Lawson, 1980; McKinney, 1985; Polkosnik and Winston, 1989; Schmidt, 1985; Shoff, 1979; Strange and King, 1981; Van Tine, 1990; Welfel, 1982b; Welfel and Davison, 1986). Four of these were longitudinal retest samples. Of the seventeen studies, three (Evans, 1988; Kelton and Griffith, 1986; Polkosnik and Winston, 1989) did not report whether gender differences were obtained. Of the remaining fourteen studies, seven reported finding no differences by gender.

Of the seven studies that found gender differences, six reported that males scored higher than females (King, Wood, and Mines, 1990; Kitchener and Wood, 1987; Kitchener, Lynch, Fischer and Wood, 1993; Lawson, 1980; Shoff, 1979; Strange and King, 1981). The remaining study (Schmidt, 1985) reported no overall gender main effects but a class by gender interaction: female traditional-age juniors and nontraditional-age freshmen (but not traditional-age freshmen) scored higher than their male counterparts.

The class by gender interactions found by Schmidt (1985) and in our six-year and ten-year follow-up studies suggest that men and women may develop at different rates and that the rates may differ between cohorts of men and women. Wood (1993b) explored whether differential patterns of growth and change exist in the data reported by Kitchener, Lynch, Fischer, and Wood (1993). He found systematic differences in growth spurts between males and females, with males showing more evidence for spurts and sustained growth during the college and graduate student years, while females showed more evidence of dramatic growth in their late teens (eighteen to nineteen) and more moderate increases thereafter. These data suggest that there may be differences in the timing of developmental changes in men and in women. We encourage more research on this possibility.

Baxter Magolda (1992) offers a different perspective on the question of gender differences in intellectual development. This perspective is based on extensive interview data she collected from a cohort of college students over a five-year period. She claims that while there is little evidence for gender-based

structural differences in reasoning, gender-related differences do exist *within* epistemological structures. Since the RJI was not designed to elicit such information and since the data reported here have not been scored for this purpose, it cannot be said whether such differences might also exist in the context of the Reflective Judgment Model.

We also offer a word of caution about interpreting these findings because of differences in sampling and in the treatment of related variables. First, the criteria for sample selection varied widely across studies (including high school students matched on academic aptitude to a sample of graduate students, as well as samples of college students selected at random). Second, some reported results accounted for the effects of such factors as academic aptitude while others did not. (For a discussion of the necessary but not sufficient relationship between aptitude and reflective judgment, see Chapter Seven.) It is important to remember that these samples vary on a wide range of variables beyond gender and are not always comparable. Therefore, whether the reported gender differences described above result from differences in factors such as ability, rates of maturation, the timing of developmental spurts, or differential opportunities for involvement in intellectually stimulating environments remains to be seen.

Research on Cross-Cultural Differences in Reflective Judgment

Are the epistemic assumptions and concepts of justification found in samples from the United States unique to this culture, or are they also found in samples from other countries? Only one study has used the RJI to investigate this question with a certified interviewer and certified raters. Kitchener and Wood (1987) tested a sample of forty-eight college students divided evenly between men and women from a traditional German university. (The university is located in what was West Germany.) One-third of the participants were beginning students; one-third were advanced students completing work for the master-diplom or diplom ingeneur degree; and the remainder were doctoral students. The

participants' mean ages were twenty, twenty-five, and twenty-nine and a half, respectively. All participants took the German Reflective Judgment Interview (GRJI) as well as the analogies portion of the Intelligenz Struktur Test (IST), which is a measure of verbal ability (Amthauer, 1955). The GRJI was identical to the RJI with two exceptions: it was translated into German, and the creation/evolution problem was omitted since in a pilot study using German participants, this problem did not have salience for the sample. A new problem on the safety of nuclear energy was written for use in this study.

Interrater reliability was .83, and interrater agreement ranged from .79 to 1.00 when calculated separately for each problem; these are comparable to figures for samples from the United States. Similarly, the response patterns of German students conformed to those predicted by the sequential ordering of the Reflective Judgment Model when evaluated by Davison's (1979) test for sequentiality. An educational level by gender ANOVA revealed significant main effects for education: the doctoral students scored significantly higher on the GRJI (5.3) than the two younger groups (4.7) and (4.6). A main effect for gender was also obtained, perhaps reflecting the fact that the men were older than the women, having completed their compulsory military service prior to enrollment. The educational level effect, but not the gender effect, remained significant after covarying out IST scores. While German university students as a group are different in many respects from U.S. students (notably selectivity), the GRJI scores of the doctoral students are comparable to RJI scores of U.S. samples. The scores of the entering and diplom students were higher than those typical of undergraduate samples in the United States but similar to those earned by U.S. postbaccalaureate students.

Although Germany and the United States are both highly industrialized societies, academic socialization differs in the two countries. This was particularly true in the mid-1980s, when the data for this study were collected. If the stages of the Reflective Judgment Model were found to reflect only the influence of academic socialization in the United States (which was not the case here), other concepts of justification and assertions about

knowledge or perhaps a different developmental sequence would have been identified in the German sample. This study provides preliminary support for the cross-national applicability of the Reflective Judgment Model. We encourage more research on this topic.

Overview of Research Findings

This chapter has summarized the findings of thirty-three studies of more than seventeen hundred individuals who have taken the RJI. These findings can be used as the basis for answering the six major questions posed at the beginning of the chapter: (1) Does reflective judgment develop between late adolescence and middle adulthood? (2) How do high school, college, and graduate students reason about ill-structured problems? (3) Does students' reasoning improve with additional exposure to and involvement in higher education? (4) Do adult learners differ from traditional-age students in their reflective thinking? (5) How does the reasoning of adults who have not participated in higher education compare to the reasoning of those who have earned a college degree? (6) Are there gender or cross-cultural differences in the development of reflective thinking?

The first of these questions is particularly important because data on changes in reflective thinking over time test the developmental assumptions of the Reflective Judgment Model and are necessary for its validation. Documenting change over time requires longitudinal data on individuals who are tested and retested over time. Although longitudinal data are scarce to nonexistent for most models of adult cognitive development (Brabeck, 1984), for the Reflective Judgment Model, longitudinal data have been collected from 241 individuals in nine studies.

Ten-Year Longitudinal Study of the Development of Reflective Judgment

The ten-year longitudinal study of the development of reflective thinking is not only the study of longest duration but in-

cluded the most testings (four) and the largest sample size (eighty at the first testing and fifty-three at the fourth testing ten years later). In 1977 (the year of the first testing), the participants in this study were from three educational levels: high school juniors, college juniors, and advanced doctoral students. The students in the younger two groups were matched by gender and academic aptitude to the graduate sample. (This procedure yielded a sample with rather distinctive characteristics.) All were invited to continue their participation in the study again in 1979, 1983, and 1987.

RJI scores differed by group in 1977, with the more advanced students earning successively higher scores. The rate and amount of change over time were found to differ by group: while the scores of the former doctoral students remained relatively stable (at Stages 6 and 7), the scores of the other two groups increased steadily over time. The scores of the former high school students showed the most growth over the decade, about two and one-half stages, from 2.9 to 5.5; the scores of their former college counterparts increased by about a stage, from 3.8 to 5.1. By using a new interview problem in 1987, we determined that these changes were not attributable to familiarity with the standard RJI problems.

Because grouped data mask changes in individual scores, individual mean and modal scores are also reported. The RJI scores of fully 92 percent of the longitudinal sample increased between 1977 and 1987; for over two-thirds of the sample, scores increased or remained stable between each testing interval. The sequence of stage usage generally was found to follow the order hypothesized by the model, with higher-stage usage following lower-stage usage. Few regressions were observed. Individual scores also indicated that most of the participants used reasoning that is characteristic of more than one stage in the Reflective Judgment Model at any given time (see Figure 6.2), with Stages 5 and 6 being the most variable. Individual changes over time are also depicted in verbatim examples from the ten-year longitudinal study. Each transcript illustrates development along at least one dimension of reasoning described by the Reflective Judgment Model.

In the 1987 testing (but not in prior testings), men scored higher than women. In examining the educational attainment of these individuals (excluding those who were already enrolled in doctoral programs in 1977), we found that almost 50 percent of the men but only 15 percent of the women had earned postbaccalaureate degrees. Since graduate programs often require students to make independent judgments grounded in the knowledge base of a discipline, this gender difference may be attributable to the men's higher level of participation in educational programs in which higher-stage reasoning is an explicit goal, in which higher-stage reasoning is taught, or in which students are given feedback on their progress. The larger educational issues raised by this finding, including the effectiveness of graduate programs in teaching students to think reflectively, clearly merit additional research.

Other Longitudinal Studies of Reflective Judgment

The patterns of scores from seven other longitudinal studies in which 241 individuals were tested over time have also been examined. In every sample tested, the scores either stayed the same or increased over time. Further, the amount of change over time was strongly related to the duration of time between testings: the longer the duration, the more growth between testings. These studies provide additional evidence that development in reflective thinking occurs steadily and slowly over time, following the sequence described in the Reflective Judgment Model.

Cross-Sectional Studies of Reflective Judgment

The results from twenty-five cross-sectional studies in which more than fifteen hundred students took the RJI have also been summarized. The students ranged in age from the teen years to middle adulthood and were from diverse geographical regions of the United States. The subsamples within studies have been grouped into ten educational levels that range from early high school to advanced graduate school plus nonstudent adults. Our purpose here is to provide a broad view of how students' ability

to make reflective judgments differs by educational level, not to provide a specific prediction of students' reasoning based solely on their educational level. Because of the variability in sampling procedures across studies (for example, some controlled for academic ability and some did not), differences between educational groups should be interpreted cautiously.

The findings from these studies suggest that RJI scores increase slowly but steadily from high school ($M = 3.2$) to college ($M = 3.8$) to graduate school ($M = 4.8$). Among nonstudent adults, the average RJI score for those without a college degree was 3.6; for those with a college degree, the mean RJI score was 4.3. The high school students consistently evidenced the assumptions of the pre-reflective stages of the Reflective Judgment Model—for example, making decisions on the basis of what they “wanted to believe” or “what felt right” or defending beliefs as personal opinions that are not subject to evaluation. Since these assumptions may still be in place at the time of a student's enrollment in college, faculty and staff may also hear such statements. Although such thinking is antithetical to true reflective thinking, we encourage all educators to recognize that these assumptions ring true to some students as a framework for understanding and resolving ill-structured problems and that educators should not assume that these students are aware of the limitations of this framework.

Twenty studies tested a total of 966 traditional-age college students. In general, the freshmen in these samples evidenced the use of Stage 3 and Stage 4 assumptions of the Reflective Judgment Model ($M = 3.6$). The mean RJI score for the seniors was higher ($M = 4.0$), and their scores were more variable. Although the numerical difference between the averaged scores of freshmen and seniors is not large (about half a stage), the qualitative difference in reasoning it reflects is noteworthy. The Stage 4 reasoning prevalent among the senior samples was clearly more adequate and defensible than the Stage 3 reasoning it replaced: with Stage 4 reasoning comes the acknowledgment that uncertainty is not just a temporary condition germane only to a select set of problems but that it is an ongoing characteristic of the knowing process. It is also at this stage that indi-

viduals begin to understand the role of evidence in making judgments. While some educators may see this as so basic that it need not even be mentioned, the role of evidence (or what constitutes evidence) is not well understood or self-evident to many college students. While students are beginning to understand that evidence can be used as a tool in building arguments, they are also beginning to understand that the relationship between evidence and judgment is sometimes a complicated one, such as when the same evidence can be used to build conflicting arguments; this often leaves them feeling confused. Barry Kroll (1992b) has described this type of development as the abandonment of "ignorant certainty" (the dogmatic approach that is characteristic of earlier stage reasoning) in favor of "intelligent confusion." Kroll's observation should remind educators that the intelligent confusion students express should not be dismissed out of hand as merely confused thinking. The questions that the confusion generates show that students are acknowledging the complexity of an issue; this acknowledgment paves the way for more thoughtful, reasoned judgments. In this sense, the development from Stage 3 to Stage 4 reasoning across the college years should be understood as an important step in the process of learning to make reflective judgments. Nevertheless, although the scores of the undergraduate students showed progress toward the goal of learning to think reflectively, they did not show the achievement of this goal, as indicated by the prevalence of quasi-reflective reasoning among the college seniors.

Adult learners in five studies (a total of 137 students) have taken the RJI; these studies help us understand the relative effects of educational level and age on the development of reflective judgment. Contrary to studies that have alerted educators to differences between traditional- and nontraditional-age students, these studies revealed that the adult students in the samples were not dramatically different from their younger counterparts in terms of their reflective thinking. In fact, despite the variability in the operational definitions of the adult samples, the mean scores by class were remarkably similar to those of traditional-age students at the same class level. Again, among the adult learners, the seniors scored higher than the freshmen, providing

additional support for the positive influence of higher education on reflective thinking. While age alone does not predict reflective thinking ability, it may be strongly related to other relevant factors (readiness to learn, openness to being challenged, prior experience entertaining divergent views); these factors may help explain the perceived developmental spurt many adult learners report during college. However, these studies also show that adult students, too, hold assumptions about knowledge that limit their ability to make judgments on the basis of critical analysis; their style of reasoning is best characterized as quasi-reflective. Only samples of advanced graduate students have consistently yielded evidence of reasoning at Stage 5 or above.

A total of 196 graduate students have taken the RJI; their average RJI score ($M = 4.8$), was a full stage higher than that of the undergraduate students. Further, early-level graduate students (master's degree students or those who were in the first two years of doctoral study) used epistemic assumptions consistent with Stages 4 and 5, while advanced-level graduate students (those with more than two years of doctoral study) used assumptions consistent with Stages 5 and 6. In their interviews, graduate students tended to offer a balanced perspective on a controversial issue, explaining why different people might approach a given problem from different perspectives and how that leads them to focus on different types of evidence. At the advanced doctoral level, students tended to defend their judgments as more reasonable or more plausible than other alternatives or as the best interpretation of the available evidence. They also typically explained how they arrived at their judgment, with explicit reference to how they sorted and weighed the evidence in support of this judgment. Such reasoning demonstrates the ability to make truly reflective judgments. Whether these findings are a reflection of graduate school selection criteria, the intellectual demands of graduate study, more individualized mentoring, or unknown maturational factors is unclear. What is clear is that professionals in all fields face a multitude of ill-structured problems about which they are expected to make defensible, reasoned choices and that graduate programs have the responsibility to help students learn to make reflective judgments. The

data suggest that there is still room for improvement in meeting this goal.

The six studies of 191 adults who were not enrolled in educational programs provide a point of comparison with several of the student samples (especially adult learners) since the inclusion of nonstudent adults provides a control for age when studying educational effects. Adults who had earned a college degree scored consistently higher ($M = 4.3$) than those who had not done so ($M = 3.6$). In one study of adults over age sixty-five (Glenn and Eklund, 1991), the RJI scores of retired college faculty members were found to be slightly higher than the average score of graduate students in other reflective judgment research (see Table 6.6). The retired faculty members were matched on age with adults who had earned a high school diploma but who had not gone to college or graduate school. The scores of this comparison group were higher than those of high school students (their educational-level counterparts, albeit of a much younger age) and close to the average RJI score earned by college students in other studies. One provocative interpretation of these findings is that life experiences outside of formal education are also associated with the development of reflective thinking. While this may seem commonsensical, there is remarkably little evidence pointing to specific types of experiences outside of formal educational settings that foster cognitive development, a topic that merits additional research.

In general, the increases in reflective judgment across educational levels in the cross-sectional studies probably reflect a variety of factors, such as personal maturity, students' level of involvement with academic and cocurricular offerings, students' success in negotiating academic and social systems, or other concurrent life experiences, such as changing family responsibilities and participation in the work force. These increases may also reflect the effectiveness of educational programs designed to teach students to think reflectively and to make reasoned, interpretive arguments. While the cross-sectional studies did not evaluate the impact of specific educational programs on reflective thinking, the strong relationship between educational level and RJI scores does suggest that postsecondary education

has been influential in promoting reflective thinking. We encourage research that helps identify which aspects enable or hinder the achievement of this goal.

Research on Gender Differences in Reflective Judgment

Many developmental models have been criticized as gender biased because they were derived exclusively or predominantly from male samples (Chickering, 1969; Erikson, 1963; Kohlberg, 1969; Levinson, 1978; Perry, 1970). In light of the continuing interest in gender-related issues within developmental models, the data on gender differences from studies using the RJI have been reviewed even though the Reflective Judgment Model was derived from interviews with equal numbers of women and men and there was no a priori reason to suspect that such differences exist. Of the seventeen studies that included both males and females (excluding the ten-year study, discussed separately), seven reported differences by gender and seven reported finding no gender differences; three did not report on the matter. Of the seven reporting differences, males scored higher RJI than females in six of the studies, and the seventh study reported a class by gender interaction. While these findings may suggest a trend of higher scores for men, we think such a conclusion is premature in light of the wide variations in sampling used across these studies (for example, some were matched to characteristics such as academic aptitude while others were not). In particular, the reported differences (and *lack* of differences) may be attributable to a variety of factors in addition to gender, such as academic aptitude, leadership opportunities, and different rates of maturation, or differences in the timing of growth spurts, such as those found by Wood (1993b).

Research on Cross-Cultural Differences in Reflective Judgment

In the one study to date that investigated cross-cultural differences (Kitchener and Wood, 1987), the RJI was adapted for use with German samples. The resulting reflective judgment

scores were comparable to those obtained with samples of doctoral students from the United States; the scores of two other groups of young adults were also comparable, though somewhat higher. The significant differences between educational levels remained even after controlling for the effects of verbal aptitude.

Summary

The studies described in this chapter have collectively contributed to the validation of the Reflective Judgment Model and provided a picture of reflective thinking of individuals at different educational levels and in different settings. While the longitudinal studies provide a stricter test of development by following individuals over time to chart their development, the cross-sectional studies provide more diverse samples across different types of students at different types of institutions across the United States.

Overall, the findings from the longitudinal and cross-sectional studies of reflective judgment suggest a slow but steady pattern of development. The general findings from these studies can be summarized as follows: first, people who are engaged in educational activities tend to improve in their reasoning about ill-structured problems. Second, development typically follows the stage-related patterns described by the Reflective Judgment Model. The consistently higher mean scores among older, more highly educated individuals in cross-sectional studies, the consistent increase in mean scores over time in the longitudinal studies, and the more fine-grained analyses of the sequence of changes within individuals support this claim. Third, being in an educational setting seems to facilitate development. Although we cannot identify the specific components of an educational environment that make the difference, cross-sectional comparisons of student and nonstudent adults, comparisons of older adults with and without advanced degrees, as well as longitudinal data on those who completed a college degree and those who did not, all suggest that those who enter college and earn a college degree have higher reflective judgment scores than those who do not.

Understanding intellectual development and its many influences clearly requires careful examination of characteristics of people (such as students' epistemic assumptions) and characteristics of the multiple environments that stimulate, nurture, and challenge them (such as college classrooms and residence halls). Finer-grained analyses of both types of characteristics offer more detailed information on which to base instructional and other educational decisions. Chapter Nine of this book explores such strategies in more detail.