DOCUMENT RESUME

ED 374 156 TM 022 085

AUTHOR	Wood, Phillip
TITLE	A Secondary Analysis of Claims Regarding the Reflective Judgment Interview: Internal Consistency, Sequentiality and Intra-Individual Differences in Ill-Structured Problem Solving.
PUB DATE	Apr 94
NOTE	91p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Error of Measurement; Higher Education; *Individual Differences; Interviews; Longitudinal Studies; Meta Analysis; Models; Outcomes of Education; *Problem Solving; *Scoring
IDENTIFIERS	Confirmatory Factor Analysis; *Internal Consistency; Problem Structure; Rater Reliability; *Reflective Judgment Model; Sequential Analysis

ABSTRACT

The Reflective Judgment Model and associated interview (RJI) (Kitchener and King, 1981) measure the ability of individuals to reason about ill-structured problems. It has gained popularity as a measure of college outcomes associated with postsecondary education. This study examines general claims for the model made from existing data from 15 of 25 studies of the RJI, representing 1,334 subjects and 4 of 5 available longitudinal studies. About 0.3% of scores were found to have coding or clerical errors or to have been incorrectly read into statistical programs. Discrepant definitions of rater agreement were found for some studies, making it difficult to interpret reported agreement rates. An improved scoring system is proposed that results in improved internal consistency estimates for the RJI. Examination of data reveals a systematic pattern of increasing performance and variability of the RJI as a function of educational level, replicating past findings. Confirmatory factor analysis suggests that the dilemma topics of the RJI represent tau-equivalent measures with significant intra-individual variation for the individual topics of the RJI. Seven figures and 11 tables present analysis results. (Contains 67 references.) (SLD)

A Secondary Analysis of Claims Regarding the Reflective Judgment Interview: Internal Consistency, Sequentiality and Intra-Individual Differences in Ill-Structured Problem Solving

Phillip Wood

University of Missouri-Columbia

The author would like to thank Patricia King and Karen Kitchener for their assistance in securing the data. Reprint requests should be sent to:

Phillip Wood

210 McAlester Hall

University of Missouri-Columbia

Columbia, MO 65211

Abstract

The Reflective Judgment Model and associated interview (RJI, Kitchener & King, 1981) measure the ability of individuals to reason about ill-structured problems. Past research has reported that the Reflective Judgment Interview is an internally consistent measure of ability and supported the notion of Reflective Judgment as an invariant progression of increasingly more adequate solutions to problems which do not allow of a single correct answer. The model has gained particular popularity as a measure of college outcomes associated with post-secondary education. Recent work has suggested that the RJI can be reliably used to tailor classroom interventions to improve instruction.

The present study examines the general claims for the model made from the existing data. Twenty-five studies were identified which used the Reflective Judgment Interview. Data were secured from fifteen of these studies, constituting a 72% sample of the population of known studies, representing data from 1334 subjects and including four of the five available longitudinal studies.

On critical examination of the available data, approximately .3% of the scores were found to have coding errors, clerical errors, or to have been incorrectly read into statistical programs. Errors of statistical analysis or reportage were found in a minority of studies. Discrepant definitions of rater agreement were found for some studies, making it difficult to interpret reported agreement rates.

An improved scoring scheme is proposed which results in improved internal consistency estimates for the RJI. Internal consistency, agreement, and intra-class correlation coefficients between raters are reported by study and by educational level.

Examination of available data by educational level revealed a systematic pattern of increasing performance and variability on the RJI as a function of educational level, replicating past research findings. When individual studies are examined within educational levels, however, large differences exist between samples, suggesting large institutional, cohort, or experimental design differences. Results from Davison's (1979) test of sequentiality and spline regressions applied to these data reveal that although the RJI documents a sequential progression, higher levels of Reflective Judgment show considerably more variability across levels than lower levels of Reflective Judgment.

Confirmatory factor analyses of these data are presented which show that the dilemma topics of the RJI do not constitute parallel forms as previously thought, but represent tau-equivalent measures with significant intra-individual variation for the individual topics of the RJI.

A Secondary Analysis of Claims Regarding the Reflective Judgment Interview:
Internal Consistency, Sequentiality and Intra-Individual Differences
in Ill-Structured Problem Solving

In recent years, calls for educational reform at the post-secondary level have focused on the need for college students to reason complexly about issues which have no single correct answer. Specifically, such calls have concentrated on cultivating student awareness about difficult real-world problems and justifying these positions in a rational, defensible manner. One often-mentioned theme is the need to cultivate students' reasoning about problems which do not have a single correct answer:

> 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 are always are framed by context; for that reason, they always are contingent (Association of American Colleges, 1991, p. 13)

In addition, while the process of forming reasoned opinions about such complex issues is a hallmark of the educated individual, students are expected to differentiate between merely expressing a personal preference based on whim or habit and developing a defensible approach to the problem based on relevant evidence and argument. For example:

> Students need to learn... 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" (Association for American Colleges, 1991, p. 14)

Finally, educated students should be able not only to reason complexly about problems using the general rules of inquiry of a particular discipline or specialty, but should also be able to contrast, evaluate and choose between the reasoned arguments put forward by competing perspectives on the problem:

> 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).

The Reflective Judgment Model.

In recent years, the Reflective Judgment Model (Kitchener & King 1981) has gained increased acceptance as a measure of college students' ability to reason about such problems. The model systematically documents the emerging abilities of students to deal with complex problems according to the three themes outlined above. At early levels of Reflective Judgment, subjects fail to appreciate that a given complex problem may allow of more than a single correct answer which is "simply 'there.'" At intermediate levels of Reflective Judgment, subjects are aware that some problems may not have single correct answers, but have substantial difficulty differentiating arbitrary personal preferences from organized and developed interpretations and judgments. At advanced levels of Reflective Judgment, subjects move from the ability to formulate a valid approach to a problem to the ability to evaluate the general adequacy of an approach relative to other logical, internally consistent approaches. The Reflective Judgment model and the assessment instrument based on it, the Reflective Judgment Interview (RJI), describe qualitative changes in the abilities of students to understand complex problems and to justify their point of view. Pascarella & Terenzini (1991), in their review of twenty years of educational research describe the Reflective Judgment model as "... the best known and most extensively studied" model of post-formal operations (p. 123).[1]

Conceptually, the Reflective Judgment model describes increasingly adequate epistemic cognitions that a subject makes when faced with an "ill-structured problem." Wood (1983) described ill-structured problems as those for which there does not exist a single correct answer. The Reflective Judgment Model does not make a claim to being a comprehensive model for every conceivable type of problem which could possibly constitute a valuable college outcome or all

1. It should be noted that King & Kitchener (1994) do not describe Reflective Judgment as a measure of post-formal reasoning, since reasoning about the well-structured problems by means of formal operations and epistemic cognitions about ill-structured problems are distinct content areas.

types of ill-structured problems. It is rather a specific model of the epistemic cognition involved when the subject is faced with an "ill-structured problem" for which conflicting sources of information currently exist and about which even qualified experts can be expected to disagree. As an unique type of problem solving ability, the Reflective Judgment model has been differentiated from simple declarative cognitions about a problem or even meta-cognitive processes of self-evaluation of performance (Kitchener, 1983). Wood (1983) also conceptually differentiates the ill-structured problems of the Reflective Judgment Model from processes involving statistical inference, utility, and risk. While no study has directly compared statistical reasoning with Reflective Judgment, King et al. (1990) found supporting evidence for this claim in that mathematics and computer science graduate students performed lower on the RJI than graduate students in the social sciences (composed of psychology, sociology, and educational measurement). Ill-structured problems differ from "ill-defined" problem spaces typified by researchers who examine the role of "insight" in real-world problem solving (e.g., Duncker, 1945; Maier, 1933; Hoffman et al., 1963 also discussed in Sternberg, 1982) in that "ill-defined" problems, while not allowing of a clear deductive solution strategy, nevertheless allow of single correct answers about which qualified experts would not disagree.

Many studies and discussions of Reflective Judgment have had as their goal the differentiation of Reflective Judgment from competing or similar intellectual ability constructs which deal with the ability to solve "well-structured" problems (those having a single, correct answer) which are more commonly used in educational assessment. For example, Brabeck & Wood (1990), Wood (1990) have argued that traditional measures of critical thinking (such as the Watson-Glaser Critical Thinking Appraisal) are a necessary but not sufficient condition for Reflective Judgment. Specifically, Wood (1990) employed a statistical test to determine that low levels of critical thinking were accompanied by low levels of Reflective Judgment, while high levels of critical thinking are accompanied by moderately higher performance in Reflective Judgment as well as by a dramatic increase in variability in performance. Wood & Games (1991) found a similar necessary but not sufficient pattern of increasing heteroscedasticity using Terman's Concept Mastery Test, a general measure of verbal ability. Mines et al. (1990) used discriminant analyses to differentiate particular critical thinking skills related to particular levels of Reflective Judgment.

Additional studies have sought to differentiate the RJI from more developmentally based constructs. For example, Kitchener & Kitchener (1981) argued that, logically, development in Piagetian formal operations was insufficient to account for differences in Reflective Thinking ability. King (1977, also discussed in King, 1986) noted that, 91% subjects in her study scored as formal operational while modal Reflective Judgment scores ranged from Stage 2 through Stage 7. King et al. (1989) also examined the relationship of RJI to constructs such as moral development (as assessed by Rest's (1979) Defining Issues Test), ego development (as measured by Loevinger's Sentence Completion Test (Loevinger et al., 1970), and other measures of psychosocial development. (Glatfelter, 1982; Polkosnik & Winston, 1989). King & Kitchener (1994) summarize the results of these studies by noting that Reflective Judgment appears to develop independently of ego development and is only moderately or unrelated to measures of psychosocial development.

The Reflective Judgment Interview & Dilemmas.

The Reflective Judgment model of justification for beliefs and the RJI assessment procedure are most easily understood by an overview of the interview and Reflective Judgment Stage descriptions. The RJI consists of the presentation (in random order) of a brief complex issue (hereafter "dilemma"). These dilemmas represent an ill-structured problem dealing with an ongoing controversy. Most Reflective Judgment studies have used four traditional dilemmas given in Table 1. The subject is then asked to explain his/her position in response to a series of standardized probe questions given in Table 2. These questions are posed by a trained, certified interviewer who probes subject responses in order to secure an unambiguous summary of the subject's approach to the problem. Two facets of this epistemic cognition have been described at each level: 1.) The View of Knowledge:What subjects think can be known about a problem, and 2.) The Concept of Justification: How subjects can know when they know something. A brief summary of View of Knowledge and Concepts of Justification appropriate to Reflective Judgment Levels 1 through 7 is given in Table 3.

Insert Tables 1, 2 & 3 about here

Scoring of Reflective Judgment Protocols.

Details regarding the scoring of Reflective Judgment protocols are given in more detail elsewhere (Kitchener & King, 1981, 1985; King & Kitchener, 1994), but a brief description of the rating process will be given here. Each dilemma given to a subject is transcribed and all references to the educational level and gender of the subject are removed from the transcript. Transcripts of each dilemma are then given in random order to two trained raters who assign a Reflective Judgment rating by means of a three-digit scoring system. If only one style of reasoning is present in a given transcript, raters assign the same score to all three digits (e.g., if the transcript contained evidence for only Stage 3 reasoning, the rating would be 333). Occasionally, a transcript contains evidence of another level of reasoning in addition to the most predominant style. In these cases, the first digit of this rating represents the predominant style of reasoning in the protocol, with the less-evident stage occupying the second or third digits. The dominant stage is then repeated in the remaining stage. (For example, legitimate multiple stage ratings would be 334 or 343.) To obtain a Reflective Judgment score under this system, the three digits assigned by a rater are averaged. Scores, then follow the measurement scale in thirds (e.g., scores of 3, 3.3, 3.6, and 4 are possible). Transcript ratings are averaged across raters to form a final score for a dilemma. Finally, composite scores for the interview are formed by averaging across all four dilemmas.

Since Reflective Judgment scores are meant to convey evidence for a particular stage of reasoning in a transcript, safeguards have been built into the procedure to assure the final score assigned to a given dilemma actually reflects to a level of reasoning observed by the raters and not an average of ratings from other levels. For example, the rerating system for scoring the RJI ensures that it is not possible for a final rating of rating of 4 to occur for a dilemma by averaging a three-digit rating of 3-3-3 from one rater with a three-digit rating of 5-5-5 from the other. Under the rerating system, raters are asked to re-rate a transcript if the ratings assigned by both raters differ by one full stage or more. These discrepant transcripts are combined with transcripts which were initially in agreement and these transcripts are then rerated. If, on rerating the scores are still discrepant by at least one stage, the raters discuss the transcript in question and assign a final score to the transcript. For each transcript then, three possible scores are possible, Round 1 (or initial)

ratings, Round 2 (composed of rerated transcripts and transcripts which were initially in agreement) ratings, and Resolved (composed of Round 2 ratings and final scores as assigned after rater discussion) ratings.

Traditionally, measures of inter-rater agreement (the proportion of times that the two raters assigned scores within a stage of each other) is based on Round 1 and Round 2 scores. Internal consistency estimates based on the four dilemmas (such as coefficient alpha) are based on resolved ratings averaged across both raters, since these are thought to represent the most accurate estimates of a subject's score on the dilemma.[2]

Psychometric & Conceptual Claims for the Reflective Judgment and the RJI.

As noted below, 25 studies involving longitudinal and cross-sectional designs have been conducted to date which used the RJI. This represents a total of 1671 individuals. Based on these studies several general claims regarding the psychometric properties of the RJI and the general nature of Reflective Judgment ability have been made. A secondary analysis of the available RJI data allows for a closer examination of these claims. These claims concern: 1.) the adequacy of the existing rating rules and certification procedure to produce reliable, accurate data which are comparable across studies; 2.) the sequential nature of the Reflective Judgment model; 3.) claims concerning the gradual nature of change in Reflective Judgment over time; 4.) the view that the RJI dilemmas constitute essentially parallel forms; and 5.) the relationship of educational level to Reflective Judgment. These claims are discussed in turn below. For each of these claims, an effort will be made to note general consensus on some claims, discrepancies and their basis, as well as areas of inquiry which have not yet been addressed.

Before doing so, however, it is worthwhile to note that efforts to evaluate the data across studies to date also allows an investigation of whether the three-digit scoring procedure can be improved. As noted before, RJI scores to date have consisted of a simple average of the three digits of assigned by a rater and many or most ratings contain a mixture of two adjacent stages. In many cases the relative position of the minor stage conveys information about the relative salience of the minor level. For example, a rating of 343 is thought to contain more Stage 4

2. It should be noted, however, that some reports of internal consistency for the RJI have been based on Round 1 scores, evidently in the belief that such estimates are a more conservative estimate than use of Round 2 or Resolved scores. As shown below, this is not the case.

reasoning than a rating of 334. Below a new and simple scoring scheme is introduced which allows for more fine-grained scaling of Reflective Judgment ability. This improvement results in a slight gain in internal consistency of the instrument as well as permitting the computation of stage utilization scores which can be used to assess the sequentiality of the model.

The Rating and Certification Procedures for Reflective Judgment assure that RJI data is comparable across studies.

This claim is not explicitly addressed in RJI research, but is implicit in attempts to compare and interpret RJI performance across studies. Three ways of addressing the comparability of RJI scores have been attempted to date; comparison of general conclusions from several RJI studies; comparison of reported psychometric characteristics of the RJI across studies; and explorations of possible rater bias. Each of these approaches will be discussed and evaluated in turn.

Several studies have attempted to evaluate the overall performance of the RJI across studies and to summarize salient differences between subjects (such as educational level) (Kitchener, 1986; Kitchener & King, 1990; King & Kitchener, 1994). These studies have accepted the reported data and analyses "as is" without an examination of possible data entry errors or errors of analysis which could have occurred. In cases where scored transcripts or entered data were available, data integrity and the accuracy of reported statistical analyses can be examined. Specifically, in many cases, it is possible to determine whether errors occurred by proofreading coding sheets or scored transcripts. In other cases, errors can only be identified by examining the data for coded values outside the permissible range, and by writing computer programs to check whether discrepant scores of raters were overlooked in the rerating process.

After the data have been checked for accuracy, a closer examination of the psychometric properties of the RJI can be undertaken. Generally, as reported in the literature, the reliability of the for published RJI studies has been very good to exceptional (e.g., King, et al., 1990; King & Kitchener, 1994; Kitchener et al., 1989; Brabeck & Wood, 1990). King & Kitchener (1994), summarized the reported agreement and internal consistency estimates across all known RJI studies. For the thirty studies which reported agreements, the median agreement rates of 77%. Forty percent of the studies reported an agreement level of 87% and one quarter reported agreement levels of at least 90%. Interestingly, of the four studies which reported an agreement

levels less than 70%, King & Kitchener note that three were from samples that included adult learners and nonstudent adults. King & Kitchener also report a median coefficient alpha of .85 across studies which report it, with a range across studies from .50 to .96.

The findings regarding reported agreement and internal consistency, while informative, can be profitably supplemented by a reanalysis of the existing data. In examining the reported psychometric properties, four difficulties or ambiguities of interpretation present themselves: 1.) The reported internal consistency and agreement indices across studies are based on data composed of a variety of educational levels in some instances and relatively few in others; 2.) The internal consistency estimates reported in all studies, coefficient alpha or inter-rater correlations, assume that raters constitute a fixed effect, rather than a random effect, meaning that no effort has been made to assess the generalizability of the internal consistency to a larger pool of certified raters. 3.) The definition of inter-rater agreement seems slightly different across studies (discussed in more detail below); and 4.)The possibility of rater bias, agreement, and reliability information has not always been investigated across studies.

Heterogeneity of RJI ability across Studies. It is not, strictly speaking, possible to compare the agreement and internal consistency coefficients across all studies as reported in King et al. (1994) since some studies reported these statistics based on extremely small, homogeneous sample sizes used in a particular study, while other studies reported reliabilities taken across a variety of educational levels. If the RJI is to be used to promote educational interventions in particular educational settings it is necessary to understand how reliably individuals within a given educational level may be discriminated using the RJI. Also, while the reported reliabilities of the RJI in published articles are quite high, it is not known whether comparable psychometric properties are found for RJI studies which did not report them. Some notes from unpublished studies suggest this might be a problem. For example, Van Tine (1990) noted an initial 62% agreement rate for an early subset of her data which necessitated a recalibration of the raters before rating could continue.

Some research has investigated the use of the RJI within a given educational level, or as a result of educational interventions or environmental support (e.g., Kitchener et al., 1993; Lynch, 1990; Sakalys, 1984). In addition to investigating the range of observed RJI scores of given

educational levels (discussed below) it is also appropriate to investigate the internal consistency of the RJI within educational level, as opposed to internal consistency estimates which are based on a population of widely different levels of educational attainment. While it is expected that such internal consistency estimates would be lower, it is not known whether the decrease in internal consistency prohibits the use of the RJI for groups of students who are all at a given educational level.

Raters as Random Effects. Conceptually, however, there is also some question as to whether the internal consistency estimate, coefficient alpha, is appropriate for assessing internal consistency across raters. Coefficient alpha assumes a fixed effect model of reliability and the estimates of internal consistency may not be used to generalize to a larger universe of interest (in this case, the larger universe consists of the pool of certified raters). Also, simple comparisons of coefficient alpha estimates across studies do not take into account the issue of sampling variation in the reliability estimate. For example, heterogeneity of reliability estimates studies examining a particular educational level may identify important differences across samples in the internal consistency of the RJI, or may be attributable merely to sampling variation in the reliability estimate. To address these issues, internal consistency estimates using the random effects intraclass correlation coefficient (ICC, Shrout & Fleiss, 1979) will be computed. Random effects ICC's differ from fixed effects ICC's (which include coefficient alpha as a special case) in that they take into account the sampling variation between raters. Finally, it is also possible to calculate approximate confidence intervals in order to assess whether the RJI possesses differential internal consistency across samples.

Rater Agreement Criterion Differences. There is also some evidence that different criteria for rater agreement have been used across studies. For example, Welfel (1979) describes a three-point difference between raters as being in agreement, while other researchers report a three point difference as being discrepant (e.g., Brabeck, 1983; Glatfelter, 1982; King et al., 1983; King et al., 1990). In addition, some studies appear to assume that the ratings of a given transcript may only consist of at most two stages (such as 4-3-4, 4-4-3, and 3-4-3), while other studies mention that, on rare occasions, three stage ratings have been used (such as 4-5-6).[3]

Rater Bias. The issue of possible rater bias has been relatively under-explored in RJI research. Only two studies have investigated this possibility: Brabeck (1980) conducted an extensive comparison of the scores from both raters used in her study in an effort to identify possible rater bias and found a small but statistically significant difference between overall raters (difference=.1, p. 114). This bias appeared to be unrelated to educational level. Kelly (1993), using a Rasch measurement model, found no statistically significant differences between raters in King & Kitchener's (1994) 10 year longitudinal data.

Reflective Judgment documents a uniform series of sequential changes in the ability to reason about ill-structured problems.

The Reflective Judgment model is a complex stage theory, meaning that individuals are assumed to function at a variety of levels in addition to their predominant or preferred level of response (King & Kitchener, 1994; Rest, 1979). The sequential nature of the construct was first investigated based on Kitchener & King's (1981) initial cross-sectional study of advanced doctoral students, college juniors and high school juniors who were matched on verbal ability. Davison (1979) proposed a sequentiality test which revealed that subjects responded at levels adjacent to their predominant level (a property which this sample did not demonstrate for other developmental measures except for Rest's (1979) Defining Issues Test) (Davison et al., 1980). This form of "cross-sectional" sequentiality for the RJI was replicated on other samples of undergraduate and graduate students (Strange & King, 1981; Welfel, 1982). This type of sequentiality has been applied to longitudinal studies as well (Brabeck & Wood, 1990; King et al., 1983). In addition, longitudinal research involving two or three times of testing has reported increases or no change in performance between 84-100% of individuals (Brabeck & Wood, 1990; King et al. 1983; Kitchener et al. 1990; Kitchener et al. 1989; Sakalys, 1984; Schmidt, 1983, 1985; Welfel & Davison, 1986). To date, no evidence of stage skipping has been found. King & Kitchener (1994, Table B6-2) report none of the three educational levels in their longitudinal study increased one and a half stages or more between adjacent testings over the course of their 10 year longitudinal study. A reanalysis of all available longitudinal data described below under the

3. It should be noted that some examples of the scoring of the RJI indicate that three-stage ratings are acceptable, although rare (e.g., Kitchener, 1986). In rater training workshops, however, raters have been instructed not to assign three-stage ratings without additional confirmation. Some raters contacted personally by the author indicated that they didn't believe that three stage ratings were allowed.

proposed revised scoring scheme generally confirms this observation at the level of individuals, except that testings from three individuals in the Kitchener & King (1981) study and one individual from Kitchener et al. (1993) study showed evidence of stage increases between 2.1 and 2.45 stages.[4]

The reported data on sequentiality is in agreement and appears to hold for cross-sectional as well as longitudinal data. Such analyses have not been conducted in all applications of the model, raising the possibility that some populations might be identified for which the model may not hold. In a related vein, the relatively small sample size employed in most studies of Reflective Judgment (all studies employed samples less than 200 subjects), does not allow a more fine-grained examination of the stage and sequence of Reflective Judgment. It could well be, for example, that certain stages of Reflective Judgment are characterized by single approaches to ill-structured problems, while at other levels of Reflective Judgment, subjects exhibit much more variability in performance in addition to their preferred level of response. Specifically, no examination has been made looking at whether the complex stage model of the RJI represents a single level of response at some levels while showing more variability in response at other levels.

Growth in Reflective Judgment ability is not due to a test/retest effect and is gradual.

Three additional studies have investigated whether the obtained upward trend in scores in longitudinal studies could be due to a test-retest effect for the instrument. Kitchener & King (1990) report minimal differences between the traditional Reflective Judgment dilemmas and a new dilemma topic concerning nuclear energy in the 10 year follow-up testing of their subjects. Kitchener et al. (1993) found no differences in a two week follow-up assessment of the RJI, even when subjects received extensive information about the Reflective Judgment construct and studied examples of higher level responses. This pattern has been replicated in studies which employed relatively short assessment intervals. Sakalys (1984) reports small nonsignificant (growth=.1) gains in RJI score over a four month period. Polkosnik & Winston (1989) report a statistically significant change of .19 for a six month longitudinal study.[5] King & Kitchener

4. The three individuals from Kitchener & King's (1981) longitudinal study scored as follows:
time1=3.2 time2=3.6, time3=6.00 and time4=5.28
time1=2.7 time2=3.4, time3=5.7 and time4=5.5
time1=2.4, time2=3.4, time3=5.5 and time4=5.5
The Kitchener et al. (1993) subject scored time1=3.8 and time2=6.2

(1994) report that all longitudinal studies employing at least a year's duration found statistically significant increases, particularly among those involved in collegiate programs. A reanalysis of some of the longitudinal data with short testing intervals (such as Kitchener et al., 1993 and Brabeck and Wood's (1990) longitudinal data) could inform researchers whether the test/retest reliability is similar across different studies and samples.

Differences in Reflective Judgment as a function of educational level have documented a generally increasing pattern of Reflective Judgment.

Early undergraduate samples tend to score at about 3.5 on the RJI, indicating that at times subjects believe these issues are a merely a function of opinion and at other times they believe that opinions must be justified by facts, but are unsure of how to incorporate discrepant information. Undergraduate juniors and seniors, by contrast, score at about level 4, indicating that, on the average, they recognize the importance of facts in supporting an opinion, but do not systematically organize these facts into any logical internally consistent approach relative to a particular discipline or theory. Beginning graduate students generally score at about 4.5, indicating that they at times do not organize their views in terms of any internally consistent fashion, and at other times appear able to organize the information in an area in terms of a particular discipline or theory, but have no criteria for choosing between available alternative explanations/theories. Samples of advanced doctoral students score at about level 5.5, indicating that they at times appreciate that a synthesis across disciplines/theories is possible,[5] but they do not produce these syntheses themselves. At other times, these students appear able to organize the material only in terms of a particular theory or discipline.

King et al. (1994) summarized the reported patterns in all known RJI studies by educational level and found generally common patterns of RJI scores within educational level. A secondary analysis of available data could be used to more accurately assess the performance within and across educational levels. For example, the magnitude of institutional differences in Reflective Judgment ability and/or differences in RJI ability as a function of educational level can reveal important interindividual differences in students which would have direct implications for the structure and approach of teaching for Reflective Judgment. The relationship of educational

5. Which, as shown below, is not statistically significant.

level to Reflective Judgment could be explored to see whether it shows a necessary but not sufficient pattern, providing preliminary evidence for the presence of differential trajectories of growth in Reflective Judgment over time, the presence of distinct subgroups which have different overall levels of RJI scores, or the presence of some unmeasured variable(s) which interact(s) with educational level to produce a functional model of RJI ability (Wood, in press).

Differences in Performance Across RJI Dilemmas are either nonsignificant or negligible.

The overall RJI score for an individual is composed of a simple average across all four dilemmas, implying that RJI scores for each dilemma represent strictly parallel measures composed of equal amounts of error and true score variability. Some indirect support for this view has been advanced. For example, Kitchener et al. (1989) report that individual's modal score was consistent across problems 75% of the time. The remainder of the time, the mode was no more than one stage discrepant. Welfel (1979) examined the issue of dilemma differences from an analysis of variance framework and found that statistically significant differences did exist between dilemmas, but that the magnitude of these differences was quite small (.1 of a stage) and found that the creation evolution dilemma was lower than the other three. Kelly (1993), employing a generalized Rasch model, found no dilemma differences across the four dilemmas for longitudinal data based on Kitchener & King's (1981) study. Brabeck (1980, p. 114) reported that average scores on the chemicals dilemma were .2 higher than on the other three dilemmas (which were identical to one decimal place) but did not test to see whether these differences were statistically significant.

No study to date has attempted to directly explore the psychometric properties of the RJI from a classical test theory model. King et al. (1994) examined previously published inter-dilemma correlations for RJI studies and concluded that the magnitude of the correlations appeared to be the same for each of the four dilemmas, although no statistical test of this conclusion was made. The process of deriving an overall Reflective Judgment score for an individual, composed of the simple average of scores on all four dilemmas, assumes that each dilemma should have equal weight in computing a composite score. In psychometric terms, all research to date has assumed that the four dilemmas of the RJI are strictly parallel measures, and have not examined whether the dilemmas are parallel forms which are perhaps only tau-

equivalent (equivalent in terms of the true-score which is estimated containing different amounts of error variance, Lord & Novick, 1968) or congeneric (i.e., not equivalent in terms of the level of the true score estimated, but still indicators of the same underlying construct). Such explorations into the properties and utility of dilemma scores as parallel measures could result in improved scoring and interpretation of RJI data.

While all research to date converges on the view that overall differences in performance are, if present, quite small, the pattern of dilemma differences differs from study to study. No research to date has investigated whether these observed differences by dilemma could be due to other rater phenomena such rater bias. An examination of several data sets could be used to decide whether the observed differences between dilemmas are related to issues of rater bias or whether the obtained differences are the product of chance sampling variation in subjects. Further, all studies to date have examined the issue of dilemma differences based only on global samples spanning a wide range of educational levels. It seems reasonable to explore whether a differential pattern of differences across dilemmas may obtain for individuals as a function of educational experience. Finally, all tests of differences across dilemmas to date have sought to uncover systematic differences in performance on the dilemmas which obtain systematically across individuals. Although some studies have sought to explore whether interindividual differences in these dilemmas exist (by analyzing whether differences in level exist across curricular or education groups), no studies have sought to examine whether systematic intra-individual differences in performance exist which are unrelated to intact observable groups. Since it seems reasonable to believe that individuals may differ considerably in their interest and exposure to the dilemma topics, the issue of systematic intra-individual performance patterns across the RJI dilemmas will also be explored using an hierarchical factor model.

Data Collection.

Identification of Existing Studies of Reflective Judgment. An initial pool of existing studies employing the RJI was constructed by contacting raters and researchers known to the author and other Reflective Judgment researchers. In addition, computerized literature searches using Dissertation Abstracts International, PsychLit and ERIC were conducted. Twenty-three studies were identified through this process. The general design, sample sizes, and reported results

from these studies are reported in King et al. (1994). Two additional studies were also identified: Dove (1990) administered the other two RJI dilemmas to subjects who were administered two traditional dilemmas in the study reported by Kitchener et al. (1993). DeBord (1993), assessed twenty freshmen and twenty-two beginning graduate students. Six studies were identified in the computerized searches but were not used in the present study because studies did not involve the use of the traditional RJI and did not use trained raters. Of the 25 studies identified by this procedure, data sets from fifteen studies were gathered. Of the five longitudinal studies identified, four studies were ascertained. Six of the fifteen studies ascertained were unpublished reports or dissertations/theses. Of the studies not available, one was a presented paper, two were published in scholarly journals, and seven were unpublished dissertations/theses. In all, Reflective Judgment data from 1334 subjects were obtained, representing 5530[6] dilemma ratings by judges.

Insert Table 1 About Here

Identification of Subject Samples and Sample Sizes. The experimental designs of all studies were then examined in an effort to examine the discrete samples which were taken in each study and the number of subjects assessed. For example, the King et al. (1990) study of critical thinking examined the performance of twenty freshmen, and seniors and graduate students taken from either the social sciences or the natural sciences. This study, then, was counted as containing five samples of students. All samples were then totaled across studies to arrive at a number of individuals in the study (100). Studies, samples, and subjects who were ascertained and unavailable were grouped according to general educational level and are presented in Table 1. The ascertainment procedures were most successful for the High School and Graduate populations, where 100% of the available data were included. Sixty-eight percent of the subjects taking the RJI interview were undergraduates. Only two of the five known studies using the RJI on nonstudent populations were available, representing roughly 40% of the available population. Overall, 69% of the available subject data was collected, an acceptable rate, given the fact that the studies employed were conducted over the past fifteen years. In addition, as will be discussed, average

6. Number includes longitudinal testings and so is not a multiple of the number of subjects.

level and range of performance for individual grades are very similar to figures based on a meta-analysis of all reported results in King, et al. (1994). In two cases, separate studies investigated the performance of the same individuals. Dove (1990) administered the remaining two dilemmas to a sample of subjects in the Kitchener et al. (1993) study. Van Tine (1990) involved a retest of selected subjects in McKinney (1985). Data from these studies were merged in order to provide more complete records for each subject and to prevent a given individual from being included twice in the data set.

Although attempts to secure raw data over such a long time period are infrequently attempted in psychology, it appears that the proportion of data secured from available data is quite good. By comparison, the ascertainment rates of 72% overall and even the 36% of nonstudent populations and 63% of undergraduate samples compare favorably with Wolins (1962), where only 24% of the data sets from a sample of 37 published studies over a three year period were made available for secondary analysis.

Results

Accuracy Checks.

Data Verification. The first step in conducting any secondary analysis is to check the accuracy of the received data sets. For the present study, it was possible to secure original coding sheets and data sets only for ten studies.[7] Two studies were reentered from coding sheets or tables.[8] For the remaining studies, accuracy checking for these data sets was confined to: Checking to make sure that all RJI ratings were within reasonable bounds, proofreading the input formats for statistical programs which accompanied the data sets, and replicating the major analyses reported for each study. For all studies, it appears that random mis-keypunching did not occur, since all ratings for all dilemmas for all individuals fell within the range 1-7. Although no keypunching errors were found, in three instances a subject had ratings for one rater on a dilemma, and the second rater assigned only a single digit. This problem was encountered once

7. Studies with coding sheets were: Brabeck (1983), Brabeck & Wood (1990), DeBord (1993), Glatfelter (1982), King, Wood & Mines (1990) Kitchener & King (1981), Kitchener et al. (1993), Kitchener & Wood (1987), Polkosnik & Winston (1989), Strange & King (1981).
8. Dove (1990) and Polkosnik & Winston (1989)

each in Kitchener, et al. (1993), Van Tine, (1990) and Strange & King (1981). No corrections for these codings were found in the computer programs, making data from these ratings invalid. In these cases, it seemed most reasonable to code these ratings as three digits corresponding to the single digit rating. Finally, one rating from Kitchener & King (1981) Time 1 data was entered in the wrong column. The correct entry for this resolved score was ascertained by pulling the data from the original transcripts involved.

Errors of Analysis in Previous Research. Proofreading the input formats for the data sets revealed an analysis error for the results reported in Strange & King (1981). For these data, the input format statement for the food additives dilemma, resolved (consensus rating) was off by one column, resulting in an RJI dilemma score which was the sum of two digits divided by three, instead of all three digits. Rerunning the analysis of variance reported in Strange & King (1981) and Strange (1978) based on corrected data revealed the same pattern of statistical significance, however the significance levels of independent variables used to predict the RJI were uniformly higher than that reported. In addition, Polkosnik & Winston (1989) reported statistically significant growth in scores over a three month period for a sample of sixteen college students from the University of Georgia. Although it was possible to reproduce the Time 1 and Time 2 means for these two groups (3.42 and 3.44, respectively), this difference was nowhere near statistical significance using a repeated measures analysis of variance. This error appears to be due to an incorrect calculation of the Mean Square Error for the analysis. Three minor errors of reportage or analysis were also found.[9]

Rater Discrepancies. With three exceptions, raters in these studies were trained to a criterion agreement rate of 80% using a two point agreement criterion.[10] Some irregularities occurred across studies as to how discrepant raters could be before this divergence could be

9. 1.) Davison (1979) incorrectly states that the Reflective Judgment scale has 9 levels but that no individuals were identified at levels 1, 8, and 9. 2.) Welfel (1982, p. 495) incorrectly reports the degrees of freedom for an analysis of covariance as a one way design when it should have been a 2x2 design. The corrected test of the overall model analysis of covariance was $F(3,58)=2.88$, $p<.01$, retaining the significant educational level effect ($F(1,58)=10.75$, $p<.01$) 3.) Wood & Games (1991) incorrectly refer to the overall Concept Mastery Test scores as a vocabulary subtest of the measure.

10. Exceptions are the first, second, and third testings of the Kitchener & King (1981) study where the raters were Patricia King and Karen Kitchener, Strange & King (1981) where the raters were trained by Patricia King prior to the existence of the certification process, Kitchener and Wood, where the raters were Karen Kitchener and a trained rater and the fourth testing of Kitchener & King (1981) where the raters were a certified rater and either Patricia King or Karen Kitchener.

counted as a disagreement. Kitchener & King's sample at Time 1 contained seventeen scores which were discrepant by exactly three points. At Time 2 sixteen scores were discrepant by exactly three points. Welfel (1982) Time 1, and Kitchener & Wood (1987) defined disagreement as more than a three point discrepancy between ratings. All remaining available data sets[11] appear to have defined discrepancy as a difference of three points or more (including Welfel (1982) time 2, and Kitchener & King Times 3 and 4).[12] One rating from McKinney (1985) was discrepant by three points and was not resolved. Examination of the patterns of rating and rerating revealed that, on other occasions, raters had rerated transcripts discrepant by three points, so this discrepant rating appears to be an oversight.

In summary, the examination of data accuracy and rater discrepancies revealed that data entry, computer programming, or failure to resolve discrepant scores resulted in incorrect data for 18 dilemmas. 17.85% of the studies contained some coding error. Expressed as a function of the total number of dilemmas which were rated, the overall error rate for the studies was .33%. This also represents a 1.35% error rate of the data for any individual associated with a particular measurement occasion. Although the error rates noted for these studies must be interpreted as conservative estimates (since it was not possible to investigate all studies from original data), the error rates compare favorably with those found in previous psychological research. Rosenthal (1978), for example, notes a 1% error rate at the data entry level, which is much higher than the .33% error rate for individual dilemmas found here and is lower than all studies reviewed except one reviewed by Rosenthal (1978).[13]

Scoring Issues

Multi-Stage Ratings. A second minor source of diversity in the scoring of Reflective Judgment across studies occurs when some raters assigned three separate digits to indicate their

11. It was not possible to check the accuracy of Round 1 and Round 2 scores for some studies since these data sets only contained resolved ratings for both judges. Studies which did not separately code their round 1 and round 2 scores were the Lawson (1980), Welfel (1982), & Welfel & Davison (1986). Some studies did not record resolved ratings separately from rerated data. This occurred in Kitchener & King (1981), King et al. (1983), and Van Tine (1990). Strange & King (1981) did not use a rerating system, and Kelton & Griffith (1986) and Polkosnik & Winston (1989) used data from only one rater in analyses.
12. It is rather unusual that the Kitchener & King (1981) and Welfel & Davison (1986) studies would change their definitions of rater agreement from one time to another without documenting this fact. This finding appears to point to some early and undocumented variability in the operational definition of rater agreement.
13. Rosenthal (178) does not report error rates at the level of individual records.

rating of a protocol. For example, in these situations, raters assigned the ratings such as 4-5-6 to a protocol, indicating that the predominant reasoning style in the transcript was Stage 4 with evidence of reasoning belonging to Stages 5 and 6. These ratings occurred in the Kitchener & King (1981) data set, the McKinney (1985) and Van Tine (1990). Examination of available Round 2 ratings from all studies found that out of a total of 5530 available ratings, 1.61% (89) contained ratings from at least one judge which contained three different stages. Although (as discussed below) subjects may frequently show evidence of more than two levels across the four dilemmas of the RJI, an examination of the response patterns across judges reveals that such three-stage ratings are not advisable at the transcript level on two grounds: First, raters do not seem to be able to reliably detect more than two types of reasoning in a given protocol. Based on Round 2 ratings, judges never agreed that all three stages were present. In only one resolved rating session were judges able to agree that three stages were present in the protocol. Second, the practice of allowing more than three stages to be present in a protocol poses some difficulties for notions of rater agreement. Since discrepancy is defined as a three or more point difference between ratings, this means that a rating of 4-5-6 is discrepant with a rating of 4-4-4, even though both judges agree on the major level of reasoning in the transcript. Similarly, a 4-5-6 is counted as being in agreement with a rating of 6-6-5 even though the major level of reasoning in the protocol is different by two levels.

Proposed Overall and Stage Utilization Measures. As noted before, the score for a given protocol, dilemma, and interview consists of the average of the three-digit codes across raters. As noted above, this scheme, which measures ability in one third of a stage increments, does not take into account the relative position of the second stages assigned to a protocol. For example, the ratings 343 and 334 are possible ratings with the same average score (3.3). The first rating, however, is used by raters to indicate more Stage 4 reasoning in the protocol than the second. In addition, it is at times useful to examine protocol ratings in order to assess the percent of times in an interview that a given level of reasoning is present. Under the current system of averages, it is not possible to do this and a subject who receives an average score of 4.0 on the RJI based on four pure examples of Level 4 reasoning is not distinguished from a subject who reasons at Level 3 on two dilemmas and at Level 5 on the other two.

In order to enable a more fine-grained scale and in order to compute rough measures of subjects' stage utilization measures, compositing weight measures were assigned to each of the three digits of a rating. The first and second digits were assigned a weight of .4 and the third was assigned a weight of .2. These values were multiplied by the stage levels and then summed across the three digits to form a composite score. For example, a rating of 334 would receive a value of (3x.4)+(3x.4)+(4x.2)=3.2. Scale intervals are then measured in even increments of one fifth (Ratings of 333, 334, 343, 434, 443, and 444 would be 3, 3.2, 3.4, 3.6, 3.8, and 4.0, respectively). In addition, for those rare dilemmas which were assigned three different stages by a rater, the end rating reflects the major level assigned to the protocol (e.g., a rating of 456 receives an overall score of 4.8 instead of 5.0). This more fine grained scale provides modest improvements in the internal consistency of the RJI as will be described below.

In addition to providing a finer scale of measurement, it is also possible to calculate approximate stage utilization scores for ratings. A rating of 334 indicates 80% Stage 3 utilization and 20% Stage 4 utilization. A rating of 343 indicates 60% stage 3 utilization, and 40% Stage 4 utilization. Such stage utilization scores are similar to the utilization scores developed for Rest's (1979) Defining Issues Test and allow an examination of subject's response repertoire which is not possible given only simple averages.

Psychometric Properties of the RJI.

Previous reviews of the RJI have emphasized the high reliability and internal consistency of the measure. Unfortunately, it difficult to judge whether the multiple pass rating system employed in the RJI results in improved reliability and internal consistency or to assess the degree to which the RJI possesses the same internal consistency and agreement rates across samples and educational levels; Some studies reported agreement only for Round 1 scores, some calculated these statistics only for the experiment as a whole and not for particular educational levels, and some studies sampled a wide variety of educational levels while others concentrated on students of relatively homogenous educational levels. Additionally, as noted above, some studies employed a two-point agreement criterion in ratings, while others employed a three-point agreement. In order to make a statement of the psychometric properties of the instrument, the available raw data from Round 1, Round 2 and Resolved ratings were re-examined.

Psychometric properties for the RJI may be divided into three general areas: Internal consistency, inter-rater agreement, and intraclass correlations between raters. Internal consistency as measured by coefficient alpha, which indicates the reliability of the RJI treating the four topic of the interview as separate items. Item scores are based on the average across the two raters for each topic. Agreement rates are estimates of the proportion of times that one rater assigns a score within one stage of the other rater. Intraclass correlations represent the proportion of true score variability between two raters. As such, two estimates of intraclass correlations can be considered, the intraclass correlation between raters for the overall score across all four dilemmas, and the intraclass correlation across raters across the individual dilemmas of the RJI.

Internal Consistency. The second column of Table 5 shows the internal consistency of the RJI for all available studies. In the top half of the table, alphas for studies using all four topics are presented. The bottom half of the table shows alphas for those studies which used some subset of the standard RJI topics. The internal consistency estimates are presented based on resolved data, since resolved scores represent the raters' best estimate of the score for a particular topic. Numbers in parentheses indicate overall internal consistencies based on Round 1 ratings, where available. As can be seen, many studies show little, if any improvement in internal consistency as a result of the rating process. Two studies (DeBord, 1993 and McKinney (1985) Time 2) showed a decrease in internal consistency as a result of the rating process, while only two studies note an increase in internal consistency (Kitchener & Wood, 1987; Kitchener & King, 1981 Time 4).[14].

Insert Table 5 About Here

The internal consistency estimates in Table 5 are, in most cases, slightly higher than those reported in the original studies, possibly due to the increased precision of measurement afforded by the finer-grained scoring procedure noted above. Van Tine's (1990) reported alpha of .87 is much higher than the .78 found for this study and may constitute a transposition error. Polkosnik

14. The increase in internal consistency for the Kitchener & Wood (1987) study is probably due to the poor agreement rates for these data (discussed below). The relatively slight change in coefficient alpha for the Kitchener & King (1981 Time 4) study may be due to the use of a split rating scheme involving three raters.

& Winston's (1989) reported alpha of .89 is higher than the .80 found for the Time 1 data, but may be due to the internal consistency estimate being based on all times of assessment.

It is not possible to directly compare the internal consistency estimates across all studies directly as a measure of the quality of RJI ratings, since some studies included a wide range of educational levels (e.g., Kitchener & King, 1981), while other studies examined only a narrow range of educational levels (e.g., McKinney's study of high school students). In order to facilitate comparisons across studies, separate internal consistency estimates were computed for each study by educational level and the resulting internal consistency estimates and confidence intervals were examined in an effort to detect possible patterns of differential internal consistency across studies. With one exception (noted below) the studies which employed all four standard dilemmas were roughly equivalent. Data were then grouped according to educational level and the resultant coefficient alpha estimates based on Time 1 data by educational level, accompanied by a 95% confidence interval are given in Table 6.

Insert Table 6 About Here

In order to identify possible differences across studies within educational level, coefficient alpha estimates were generated separately by study within each educational level. In general, samples taken from restricted ranges of high verbal ability (such as the undergraduate and high school samples in Kitchener & King (1977) which were matched on verbal ability to a graduate sample and Kitchener et al. (1993) which were also sampled on the basis of high verbal ability) were not different in terms of internal consistency than samples which were selected for high variability (such as Brabeck's samples of high and low critical thinkers, or McKinny's sample of high and low academic achievement high school students). Internal consistency estimates from one sample were, however, markedly different than remaining samples: The college freshman data from Welfel (1982) were markedly lower than that of other studies (coefficient alpha for these 32 subjects was .13). Accordingly, coefficient alpha was recomputed for the college freshmen excluding this sample.

Overall, the coefficient alpha within educational level ranges from .73 to .85 for undergraduate samples, and about the mid .80's for graduate samples. These internal consistencies are generally representative of the internal consistencies found when coefficient alpha was calculated for each educational level separately for each study. While these internal consistencies are helpful in informing an instructor or educational researcher as to the internal consistency which may be expected when the RJI is administered to individuals within a given educational level (as might happen in research using educational interventions to promote Reflective Judgment), it is also helpful to estimate the internal consistency of the instrument based on selected ranges of educational levels of interest (as might occur in studies which seek to document outcomes of higher education tied to the undergraduate or graduate experience, for example. Coefficient alpha across all undergraduate data was .81 (95% Confidence interval .78-.85). Coefficient alpha based on all graduate data was .86 (95% Confidence interval .82-.89). Based on only 8th-12th grade data coefficient alpha was .81 (95% Confidence interval .76-.86). When coefficient alpha is based on all data with ratings on all four dilemmas based on only first testing data, coefficient alpha is .92. When coefficient alpha is based on all data regardless of testing, this value is .94.

Of course, the internal consistency estimates presented here by educational level must be interpreted with some caution. On the one hand, these internal consistency estimates may represent overestimates relative to what a researcher may expect, in that they include institutional level differences and thus may reflect a larger range of variability than that found in any particular sample. On the other hand, these estimates may represent underestimates of the reliability that would be found in any given study, since unequal numbers of individuals were sampled at each educational level. If a researcher were to gather an equal number of data points at each educational level, the observed coefficient alpha would be higher than found in such unbalanced samples. Nevertheless, it seems reasonable to employ these figures as summary estimates of internal consistency since they so closely parallel the estimates arrived at based on individual samples. Additionally, those studies which did sample ranges such as freshman-senior or high school samples appear to have similar overall internal consistency estimates to the estimates based on such ranges of educational level.

Rater Agreement. Previously, two raw agreement measures have been calculated for the RJI: Raw agreement, the proportion of times that raters were less than three points discrepant in their ratings and agreement rates corrected for chance using Tinsley & Weiss' (1975) T-coefficient. Since Tinsley & Weiss' T-coefficient is a special case of the more general kappa coefficient, and since kappa does not take into account unequal marginal frequencies in its correction for chance agreement, it was decided to calculate three measures of agreement across studies: Raw agreement (defined as the percent of times that a composite score for a transcript from one rater was one level or more discrepant than the other rater's score), an unweighted kappa statistic, and a weighted kappa statistic which takes into account different marginal frequencies in the data (Bangdiwala, 1985). These agreement coefficients for each study are given in the right hand of Table 5. Overall, raw agreement rates for the individual studies fall between .7 and .8, with half of the studies containing a raw agreement rate 80% or higher. Only one study (Kitchener & Wood, 1987) contained an error rate lower than 70%. Two reasons exist for this low agreement rate: This study employed a three-point (as opposed to a two-point) agreement criterion between raters. Transcripts from this study were also translated from the original German, introducing a possible source of ambiguity in the transcripts. Kappa coefficients were uniformly lower than raw agreements, since these agreements contain a correction for chance agreement, with 55% of the samples demonstrating agreement rates of .70 or higher. Two studies, Kitchener & Wood (1987) and Van Tine's (1990) retest of a sample drawn from McKinney (1985) provided extremely low corrected agreement rates.

When these agreement rates are compared to those reported in the separate studies, the same raw agreement values were found with a few exceptions: King, Wood, & Mines (1990) reported an agreement rate of .90, which was based on the agreement rates of the composites of the two raters, and not on individual dilemma agreement. The agreement rates for Times 1 and 2 of Kitchener & King (1977) (.87 & .78) are higher than the reported .77 and .72, respectively. These differences could be due to a conservative approach to scoring agreement for transcripts which contained three stages of response.

Raw agreement rates were also examined separately for each study within educational level and then examined before being combined. Generally, no substantial differences by study

obtained for agreement rates and so agreements calculated separately by each educational level are presented in Table 7. As can be seen, raw agreement rates for each educational level appears to run from 80% to 90%, with Masters/Beginning doctorate students showing a 71% agreement rate, suggesting that these data may be more difficult for raters to agree on. Kappa for these data ranged from .60-.90. In most cases Bangdiwala's weighted kappa yielded estimates between unweighted kappa and the raw agreement.

Insert Tables 6 & 7 About Here

Internal Consistency between Raters. The internal consistency across raters for RJI, the degree to which the overall RJI score of one rater is consistent with that of another rater, has been traditionally assessed as the correlation between the summary scores of two raters (or, using raters as "items" the square root of this correlation, coefficient alpha, has been reported). This approach results in an overestimate of the degree of internal consistency between raters, because it assumes that raters constitute a "fixed effect." as would arise for a single study if the internal consistency of interest was composed of only a single pair of raters. In this study, however, it makes conceptual sense to estimate internal consistency taking into account the fact that raters were drawn from a larger pool of certified raters and it is to this pool that researchers wish to generalize their findings. The intra-class correlation coefficient (ICC, Shrout & Fleiss, 1979), a measure of internal consistency drawn from generalizability theory, is just such an internal consistency measure.[15] Under the assumption that raters are a fixed effect (and not randomly drawn from some larger population) the ICC is equal to coefficient alpha. Two forms of the ICC were calculated based on the present data. The ICC based on the composite rating for one rater with another was calculated. This estimate was also recalculated based on the individual dilemmas. Since no significant differences were found at the level of the individual dilemma, these dilemma-based ICC's were calculated across all available dilemmas. Since sample size varied substantially

15. The ICC formula used assumed that the same raters rated all transcripts within a given study. For studies which employed a three rater system, ICC's reported here are based on a rearrangement of the data records to create two ratings for each dilemma. No discrepancies in ICC between combinations of raters within such studies were found.

from one sample to the next, a Satterthwaite approximate 95% confidence interval was also calculated based on the formulae in Fleiss & Shrout (1978).

Based on the composite, internal consistency between raters appears quite high for studies which employed a wide range of educational levels (ranging from .73 to .97).The exceptions to this pattern were Kitchener & Wood (1987) (a study which employed a range of undergraduate and graduate subjects), with an ICC or .61, and King, Taylor & Ottinger (1989) a study of black undergraduate students (ICC=.32). Internal consistency estimates calculated separately by the four undergraduate levels for the King et al. (1989) study were particularly low, ranging from .13 to .64. While the low internal consistency of the Kitchener & Wood (1987) study appears to be due to the difficulties of translation (since one of the raters went on to successfully rate other studies) the ICC value (.32) for King et al. (1989) study are more difficult to interpret. Raters for the King et al. (1989) study, although certified, did not rate the data of any other study, making it difficult to determine if this pattern indicates a failure of the rating process for one study, or whether some rating problem exists for black students. Studies which employed a more restricted range of educational levels revealed much lower values (ranging from .12 to .57).

Insert Table 8 About Here

When these values are compared with reported coefficient alphas (or transformed, based on reported correlations), ICC's are generally .1 to .15 lower than their fixed effects counterparts. This discrepancy was less pronounced for studies with wide ranges in educational level.

When ICC's are based on the individual dilemmas of the RJI, these ICC's are much lower, due to the smaller amount of information present for raters to judge. Interestingly, when one compares the observed ICC based on the composite with a predicted ICC based on a Spearman-Brown estimate drawn from the individual dilemma-level, we find that the observed reliability of the composite is consistently lower by anywhere from .05 to .2 than the observed composite ICC. The exceptions to this pattern are King & Kitchener Time 1 and Glatfelter (1982), which yielded the same ICC and King & Kitchener, Time 4, which yielded a value .18 lower than that found for

the composite. This pattern will be interpreted in light of the investigation of the RJI's dilemmas as parallel forms described below.

Internal consistency estimates computed separately by educational level ranged from .56-.79, and do not appear to vary systematically as a function of educational level, when the confidence intervals for these ICC's are examined (Table 9). ICC's associated with individual dilemmas appear to generally run from .31 to .70.[16]

Insert Table 9 About Here

Taken together, the agreement rates and internal consistency estimates from these data indicate that, with a few notable exceptions, the Reflective Judgment Interview can be reliably and accurately scored by trained raters. Although the process of blindly rerating transcripts which are initially discrepant improves the agreement levels of the interview, the improvements in internal consistency based on these reratings is slight, suggesting that rater differences on a particular transcript "average out" over the four dilemmas yielding an acceptable indication of an individual's overall ability. In light of these patterns, the finding that some studies employed a three-point criterion for rater agreement versus the two-point criterion is probably not a significant threat to the quality of the data. For example, the lower agreement rate found for Kitchener & Wood (1987) does not appear to result in a dramatically different estimate of internal consistency based on coefficient alpha. Some notable exceptions were found, however. In two studies, the internal consistency and agreement rates of the data appear to be quite low compared with other studies. Welfel's (1982) freshman data appear to have a much lower internal consistency than other freshman samples. This could be due to raters scoring these data first, causing some calibration errors or rater bias (discussed below) to come to light. In addition, King et al. (1989) found exceptionally low internal consistency in a study of undergraduate black students at Bowling Green State University. This divergence from the general pattern could pose a threat to the generalizability of the RJI and its scoring to these populations, or could mean that the

16. Although Welfel's freshman sample demonstrated a significantly lower pattern of internal consistency across dilemmas, this differential pattern was not found for the ICC's. Recalculating the ICC values for freshmen without this sample resulted in no noticeable improvement.

certification process is not stringent enough; that rater "drift" occurs, with raters becoming less reliable after a period of time after certification.

Rater Bias. As noted above, few studies have explored whether some raters assign statistically significantly higher or lower values in their transcript ratings. Rater bias in this study was operationally defined within an analysis of variance framework as the presence of two effects: An overall main effect for rater (indicating that one rater awarded systematically higher scores than the other) and a Rater by Dilemma interaction (indicating that one rater assigned higher or lower scores than the other for particular dilemmas, but not necessarily over all dilemmas). In order to assess the presence of bias and its possible differential effect across studies and or educational levels a general linear model was specified with dependent variable of final assigned Reflective Judgment score and independent variables of Dilemma (4) x Study (12)[17]. Rater effects were specified as nested within Study. This analysis revealed a significant effect for Dilemma, a significant effect for study, and a significant study x dilemma interaction. (Type III Sums of squares F's: $F(3,6410)=4.67$, $p<.01$; $F(11,6410)=139.91$, $p<.01$; and $F(31,6509)=3.03$, $p<.01$ respectively.[18]

While these models may be taken to indicate little rater bias in the RJI, separate analyses of variance were conducted for each study, using transcript score as a dependent variable and Rater and Dilemma as independent variables. These models allow a more fine grained exploration of rater bias and also allow exploration of the dilemma by study interaction mentioned above. None of the twelve studies found a main effect for Rater. Eight of the twelve studies contained a statistically significant Dilemma effect. However, as indicated by the interaction of Dilemma with Study, the pattern of dilemma differences was not consistent across studies. Across most studies, regardless of statistical significance, the magnitude of overall dilemma differences never exceeded .15 of a stage. Two studies, however, (Kitchener & Wood, 1987 and Welfel, 1982) found a statistically significant interaction between Dilemma and Rater indicating differential

17. Comparable analyses were also conducted based on Round 1 and Round 2 data. These results yielded the same pattern of statistical significance reported here.
18. An additional general linear model included the variable of educational level (a categorical variable having 11 levels indicating educational level from 8th grade through advanced graduate study described in the next section). This model contained an additional significant effect for educational level, but no significant combinations of interactions between rater, study, educational level, or dilemma were found.

rater bias across dilemma. For the Welfel (1982) data, one judge consistently rated the creation/ evolution dilemma lower than the remaining three dilemmas (lower by .33 than the average across the remaining three dilemmas) while the other rater's scores were relatively the same.[19] For the Kitchener & Wood (1987) data, one rater awarded scores on the Pyramids dilemma which were on the average .47 higher than the other rater, while dilemma scores were comparable across the remaining dilemmas. Since gains of .3 to .4 of a stage represent the average two year gain for some populations (see Table 11 below) the magnitude of the interactions found for these studies is practically significant.

Some evidence also exists that the rater bias effects may be reduced by the rerating procedure. When separate analyses of variance similar to the ones described above were conducted on Round 1 data, a significant ($p<.05$) Rater effect was found for two studies, Brabeck (1983) and Van Tine (1990). In the Brabeck (1983) data, one rater awarded slightly higher scores than the other (.09). For the Van Tine (1990) data, one rater awarded overall scores .13 higher than the other.

To summarize, little evidence was found for systematic rater bias in these data as a whole. Some evidence exists, however, that some raters assign differentially high or low scores to a given dilemma within the interview, suggesting a greater need to extend the certification procedures to incorporate more examples of each dilemma at each level of Reflective Judgment. The statistically significant dilemma differences were not consistent across studies, suggesting either that small differences in expertise exist across the populations researched with the RJI, or that rater bias, if it exists, consists of a small and systematic bias at the level of individual dilemmas.

Level and Variability of RJI Scores as a Function of Educational Level.

In order to assess the level and variability in performance in RJI scores as a function of educational level, a box plot of final (resolved) composite RJI scores was formed For convenience, these distributions of scores are grouped into High School grades 8, 9, and 10, High

19. Given the markedly poorer internal consistency of the Welfel (1982) data for the freshman sample, additional follow-up analyses were run including educational level as a separate effect and also investigating whether the bias patterns were different for the freshman and senior samples. These merely confirmed the general pattern of bias in the study as a whole, but did not uncover a significant educational level by rater by dilemma interaction. No rater by dilemma interactions significant for either educational level group, presumably due to the lower statistical power of these analyses.

School grades 11 & 12, early undergraduate (Freshmen and Sophomores), advanced undergraduate (Juniors and Seniors), and Beginning and advanced doctoral students.

Since box plots are infrequently reported, a word of explanation regarding their design is appropriate. Box plots are designed to summarize the characteristics of level and shape of distributions. They convey five important features of a set of data: typical or central value, variability, shape (symmetry or skewness), outlying data points, and behavior in the tails of the distribution. The central box for each educational level extends from the lower quartile (Q1) to the upper quartile (Q3). Thus, the length of the box is the interquartile range showing the middle 50% of the observations. Behavior in the tails of the distribution (often termed "adjacent values") is indicated by the single lines above and below the central box. If the inter-quartile range (Q3-Q1) is denoted by IQR, these adjacent values are computed as: Q3+1.5IQR and Q1-1.5IQR, respectively. If the data are normally distributed, this range corresponds roughly to the 99%ile range for the data. More extreme values which are located inside 3IQR are denoted by right-leaning bars, and values outside this range are denoted by left leaning bars. Median values (Q2) are denoted by a middle line in the box, and mean values are denoted by a"+".

These univariate statistics conveyed via the box plots reveal a pattern of systematically increasing trend in RJI score as a function of educational level. In addition, as one compares higher educational level students to lower educational level students, it is apparent that the distribution of scores becomes simultaneously more variable and more positively skewed. Mean scores by educational level are quite similar to those based on reported means based on a survey of all available studies reported in King et al. (1994, Tables B6-3 through B6-6) and reported on the bottom of the figure.

Insert Figure 1 About Here

Part of the reason that these distributions increase in variability could be due to the increased variability of samples across the studies examined. In order to examine the relative magnitude of between sample differences, separate analyses of variance were conducted on the data for each educational level, treating the study of interest as an independent variable (High

School samples in the 8th grade and the 10th grade were only examined in one study and were therefore excluded from these analyses). (A single Manova of these data was not possible, given the fact that not all studies explored all educational levels.) For these analyses, each sample within a given educational level was compared against all other available samples of the same educational level. The results of these analyses must be interpreted with caution, since the design of the analysis is extremely unbalanced with sample sizes as small as eight individuals being included with studies as large as 67. Nevertheless, some common patterns emerged across these analyses. First, differences across samples within educational levels were statistically significant in all cases ($p<.01$) except in the high school 8th and 10th grades (where only data from Kitchener et al. (1993) was available for experimental and control conditions[20]) and the advanced graduate group ($F(3,66)=2.55$; $p=.06$).

Within each educational level, differences across samples revealed practically significant differences (R^2 ranged from .31-.56 for all groups except college freshman ($R^2=.11$), college seniors ($R^2=.12$), and advanced graduate groups ($R^2=.10$)). For each of the educational levels average scores for each sample and the standard error of the mean are expressed as a dot plot in Figure 2. For this figure, a dot represents the mean associated with the group, and the horizontal line expresses the standard error associated with the mean. For some samples (e.g., Polkosnik & Winston's (1989) study of University of Georgia students, Kitchener et al.'s (1993) study of Colorado students) the standard errors are quite large, due to the small sample sizes associated with these studies within educational level. For the 9th grade high school samples, data taken from Kitchener et al.'s (1993) study were higher than the McKinney (1985) students who were identified as high academic achievement students. These students were, in turn, higher than those identified as low academic achievement from McKinney's study. This pattern is expected, given that the Kitchener et al. (1993) data were taken from students of high verbal ability. For high school juniors, Kitchener & King's (1981) sample (mean=2.8) were lower than data from Kitchener et al. (1993) (mean=3.7). This is unusual, given the fact that both of these samples were selected for high verbal ability. This pattern may demonstrate either large sample to sample

20. Preliminary analyses of variance examining only the experimental and control pre-test conditions for the Kitchener et al. (1993) study revealed no Time 1 differences between experimental and control conditions for any educational level.

differences in Reflective Judgment for this educational level, or a substantial cohort effect. For the high school senior data, Kitchener et al.'s (1993) seniors were higher than all other samples. McKinney's (1985) low and high achieving seniors were lower than the other samples. Brabeck's (1983) high critical thinking sample scored higher than her low critical thinking sample, with Glatfelter's (1982) sample falling between these two groups.

Insert Figure 2 About Here

For the college freshman data, less clear-cut differences were found. Data from Kitchener et al.'s (1993) and Kelton & Griffith's (1986) studies performed higher than all other samples other than Strange & King's (1981) study of University of Iowa freshmen. In addition, data from Strange & King's (1981) study were significantly higher than that from Mines et al.'s (1990) study of freshmen from the same institution. This pattern suggests that freshmen from selective private institutions score roughly .4 to .5 of a stage higher on the RJI than their public university counterparts. The finding that two samples taken from the University of Iowa differ by .4 of a stage may be due to the fact that neither the Mines et al. (199) study nor the Strange & King (1981) study were random samples from the freshman population. The freshman sample for the Mines et al. (1990) study were drawn from an Introductory Rhetoric class, a course which many college freshmen are able to test out of by demonstrating successful writing skills. Data from the Strange & King (1981) study were closely matched on the basis of ACT composite score to a sample of freshmen. To the extent that attrition may be related to academic aptitude as defined by the ACT, Strange & King's (1981) sample of freshmen may be composed of higher academic aptitude students than the freshman sample as a whole.

For the sophomore data, three distinct groups emerged. Data from Polkosnik & Winston's (1989) data were comparable with Brabeck's (1983) sample of low critical thinking students taken from institutions described as small, private Catholic institutions in New England. Brabeck's (1983) high critical thinking students from these same institutions were comparable to King et al.'s (1989) study of black students from Bowling Green State University. Finally, data

from Kitchener et al.'s (1993) study of Denver University students were higher on the RJI than all other groups.

For the college junior data, data from Kitchener & Wood's (1987) study of German university students and Kitchener et al.'s (1993) study of Denver University students were comparable. Data from Brabeck's (1983) high critical thinking sample was higher than data taken from King et al.'s (1989) study of black students at Bowling Green State University, Brabeck's (1983) low critical thinking sample, and Polkosnik & Winston's (1989) study of University of Georgia students. Data from Kitchener & King's (1981) study of University of Minnesota students fell between these two groups.

For the college senior data, samples from Kelton & Griffith (1986), Strange & King (1981), and Kitchener et al. (1993) were higher than samples from Polkosnik & Winston (1989), King et al. (1989), and Glatfelter (1982). Although the data taken from two private selective institutions again scored at the high end of available samples, data from only three public university samples were lower than these two samples.

For the beginning graduate student data (defined as either entering, master's level, or beginning doctoral level) data from King et al.'s study of social science doctoral students and Kitchener et al.'s (1993) students from Denver University were higher than samples from King et al.'s (1990) University of Iowa mathematics and computer science students, University of Missouri counseling psychology students, and Brabeck's (1983) sample of high critical thinking graduate students. Brabeck's (1983) low critical thinking sample and DeBord's (1993) sample of clinical psychology graduate students were lower than all other groups. While this pattern replicates King et al.'s observed differences between samples of social science and natural science students, the finding that DeBord's counseling psychology sample was not different than the natural science sample and the finding that the clinical psychology student sample was lower than the natural science sample indicate that the area of social science study per se is not necessarily associated with higher levels of Reflective Judgment. Some of this difference may be due to the differential amount of educational experience these samples may have had. Samples from King et al.'s (1990) study were composed of some individuals who had completed master's degrees, and all had completed at least two years of graduate study. DeBord's (1993) clinical psychology

sample were all beginning graduate students who were tested in the fall of their first year. None of these students had master's degrees. DeBord's (1993) sample of counseling students were also beginning doctoral students, but 10 of the 15 reported they had already completed master's degrees. As mentioned above, no significant differences were found in the advanced graduate samples.

The pattern of sample differences for the Beginning Graduate samples indicates that students who elect advanced study in the social sciences are not composed of individuals who score at higher levels of Reflective Judgment than their natural science counterparts. Although this may point to a need to replicate the general findings of King et al. (1990) study, there is also reason to believe that the observed differences between samples may also be due to the differential amount of educational experience across samples. As such, this pattern of samples suggests that early levels of graduate study, particularly in the social sciences, may be accompanied by relatively rapid growth in the ability to reason about ill-structured problems.

In summary, the patterns of means for the individual samples are striking. Although the general pattern of performance across all samples reveals a distinct upward trend as a function of educational level as shown in Figure 1, the range of samples for many undergraduate levels is large and roughly a stage, ranging from stage 3 to stage 4. The magnitude of differences between these samples is substantial: Recall that differences of .4 of a stage are about the same size as the two year longitudinal effect for the instrument, and differences of a full stage constitute roughly changes associated with six to ten years of longitudinal growth in Reflective Judgment. As such, when applying Reflective Judgment theory to a particular institution or classroom, it would be misleading to classify a particular classroom or sample of students as scoring at a giver level in the absence of RJI data. This points to the need for assessment of Reflective Judgment level before proceeding with educational interventions geared to a particular level of Reflective Judgment reasoning. These is also preliminary evidence to suggest that students from selective private institutions score higher than samples taken from public universities. The absence of longitudinal data, more detailed information on institutional admissions and attrition patterns precludes any firm conclusion in this regard. It seems reasonable to conclude, though, that educational institutions who wish to use to RJI as an instrument to document institutional

effectiveness must also carefully attend to issues of student characteristics (such as major and critical thinking level), attrition, and institutional admissions criteria before claiming a superior performance of their students as a result of their educational experiences at a particular institution.

General Sequentiality of the RJI

The sequentiality of the Reflective Judgment as a complex stage theory model of development was also investigated. A complex stage theory model assumes that individuals progress according to the ordered stages proposed by the theory, but does not assume that subjects reason at the same stage in all situations due to differential task demands or random subject performance variation. Under a complex stage model, response patterns from developmental data conform to the sequentiality hypothesis if the second most frequently used stage is adjacent to the first, if the third most frequently used stage is adjacent to either the first or the second stage, etc.

Davison (1979) proposed a test of the sequentiality for such complex developmental sequences by means of a probabilistic unfolding model. Davison's test involves the modeling of a contingency table of the major or predominant stage which an individual demonstrates by their minor, or second most frequently used stage. By definition, the contingency table which results has zeros on the diagonal. For this test, the contingency table was calculated based on the percent stage utilization scores described above for a data set composed of all resolved RJI data across all measurement occasions and represents data from 1579 records.[21] This contingency table of major and minor stage is given in Table 10. The data reported in Table 10 contains 1566 records, since twelve records contained evidence for only one stage of Reflective Judgment and were excluded from this analysis. For seven individuals, ties resulted which could be broken in either a favorable or unfavorable manner to the sequentiality hypothesis. Since the analyses presented here found that the Reflective Judgment model was differentially sequential, the results reported here are based on breaking these ties randomly. Results based on breaking the ties favorably or excluding tied data from the analysis were identical.[22] As can be seen from Table 10, the data from 46

21. Four comparable analyses were also conducted based on only time one data across, all available Round 1 and Round 2 data and based on only studies for which all four standard dilemmas were used. In each case, the pattern of statistical significance was also identical to that reported here.

22. It should be noted, however, that when the analyses are rerun based on data which break ties unfavorably, that the analyses are the same except that the χ^2 for the modified sequentiality hypothesis remains significant. Analyses which either exclude such ties as ambiguous data or break such ties randomly seem a more appropriate test of sequentiaility.

individuals not consonant with the sequentiality hypothesis and are located in cells 2/4, 2/5, 3/5, 4/2, 4/6, 4/7, 5/3, 5/7, 6/3, 6/4, 7/4 and 7/5. These 46 observations represent only 2.94% of the data on which the sequentiality test was based.

Insert Table 10 About Here

Davison (1979) proposed a two-step process for determining whether the response frequencies correspond with those predicted by a given developmental sequence. The first step involves testing whether the data conform to a quasi-independence model, which assumes no sequentiality in the data. The predicted frequencies from this model form the basis of a χ^2 test. If this χ^2 is statistically significant, the data are tested in a second step against a sequential model, which includes a single sequentiality parameter for adjacent stages and indicates an added probability of occurrence predicted by the developmental sequence. If the χ^2 from this model is not statistically significant, this is taken as confirmation of the sequential nature of the data. For these data, the independence χ^2 is highly significant (χ_{19}^2=2475.87; p<.001) but Davison's sequential model was also statistically significant (χ_{18}^2=52.05; p<.05). An examination of the actual and predicted frequencies under the sequential model (shown in the first and second columns of Table 10) revealed that the largest discrepancies from the data occurred in the sequential cells of the table and in situations where Davison's model predicted more nonsequential responses on the basis of chance than in fact occurred. Conceptually, lower Reflective Judgment stages appear more clear-cut than more advanced stages. Individuals with dominant scores of 3 or 4 appear to have fewer upper-stage responses than one would expect under the sequentiaility model, while individuals with dominant stages of 5 and 6 had fewer adjacent stage frequencies than expected under the sequential model. A small modification of Davison's test was made to test this possibility: separate response frequencies for adjacent stages were modeled for each dominant stage. This fit of this model was also statistically significant, indicating the probabilities of dominant and subdominant stages were different, even within major stage (χ_{13}^2=29.03, p<.01). An additional sequential model was also specified which allowed the adjacent stages to be of unequal frequencies for each level. By definition, this model results in a

perfect fit of the adjacent stages, while fitting an independence model to the remaining cells. Predicted frequencies for the modified sequentiality model are given as the fourth entry in the table cells of Table 10. The Reflective Judgment data conformed well to this modified sequential model (χ_9^2=8.30; p>.5).

Even though no nonsequential cells of the contingency table differed from rates predicted under chance by more than 3, it is helpful to examine which studies contained stage discrepant ratings in an effort to understand if certain populations or studies contained higher rates of nonsequential ratings. Nonsequential ratings, though rare, seemed to occur most frequently in three studies: testings based on King & Kitchener's (1981) longitudinal study, King, Wood, & Mines' (1990) study of mathematics and social science graduate students, and Kitchener & Wood's (1987) study of German university students.[23].

Taken together, the results of the sequentiality analyses show that the Reflective Judgment interview and scoring system document a complex developmental sequence. Nonsequential responses are quite rare and, given that some of the nonsequential responses occurred predominantly in three of the studies, suggests that these deviations may be the result of minor scoring variability between raters. Given the translation required for Kitchener & Wood's study of German university students, it is also possible that these few deviations represent difficult or ambiguous translations. To date, these results stand in contrast to previous applications of Davison's test to other developmental theories, which have failed to reject the quasi-independence model. The finding differential sequentiality by dominant levels of Reflective Judgment is discussed in the context of spline regressions discussed below.

Stage Utilization as a Function of Reflective Judgment Level.

One of the reasons for the patterns of increased variability as a function of Reflective Judgment level may be that subjects are more stage-homogeneous in their responses at earlier levels of the model and show a greater variability in performance on the RJI at higher levels. To explore this, a series of spline regressions was conducted on the data predicted stage utilization scores for Levels 2-7 as a function of overall rescaled RJI score[24]. (Since Reflective Judgment

23. Five of the 11 4/6 ratings were taken from longitudinal tests from Kitchener & King's (1981) study; Two of the four 7/5 ratings, two of the five 5/7 ratings were taken from King et al. (1990); and the two 7/4 ratings, three of the five 4/7 ratings, and one of the four 6/4 ratings were from Kitchener & Wood (1987)

Level 1 responses occur only rarely, they were excluded from the present analysis). The results of these regressions are given in Figure 2.

Generally, Figure 2 shows that stage utilization patterns are more diverse for more advanced levels (stages 5 and 6) of Reflective Judgment than for earlier levels. For example, the average percent stage utilization for individuals with an overall score of 5.0 is only 50%, with the final score reflecting a composite of Level 4 and Level 6 reasoning. It is worthwhile to note that no individuals evidenced nonadjacent stage utilization patterns which ran counter to the Reflective Judgment model (e.g., no protocols contained examples of only Level 4 and Level 6 reasoning). High percent stage utilization scores for levels 2, 3, 6, and 7 may be due to the fact their location at the ends of the measure and not reflective of the degree of high stage utilization for these levels (e.g., it is mathematically possible to have an overall score of 6.9 on the RJI only if a substantial proportion of the dilemma scores are at level 7).

Growth and Stability of the RJI over Time. In addition to assessing the psychometric properties of the RJI within a given testing, an examination of the longitudinal testings of the RJI permits an examination of Reflective Judgment over time, particularly as an examination of the test/retest correlation of the instrument and an examination of the size of changes in RJI over time as a function of initial educational level. The test/retest correlations, magnitude of observed differences, and initial mean scores of available longitudinal testings is given in Table 11. Three general patterns emerge from an examination of these test/retest correlations: First, as expected, the test/retest correlations go down as a function of the length of time between testings. In order to assess the relative comparability of these correlations, it is possible to generate estimates of the test/retest correlation for a year's duration (Rindskopf, 1984). These estimates appear for all testings of length six months through 10 years in parentheses after the test/retest correlations in Table 11. When these annualized correlations are examined, it appears that longitudinal data based Kitchener & King's (1981) study show estimated test/retest correlations between .85-.98. Data taken from other studies show a lower pattern of test/retest correlation, such as Brabeck's (1983) high school sample, Polkosnik & Winston's (1989) study, and Kitchener et al.'s (1993) skill theory study of short term change in RJI. Kitchener et al.'s (1993) short term skill theory

24. As Darlington (1990) notes, response curves which asymptote at particular values are not well-suited for the more familiar polynomial regression techniques.

study also presents some insight into the test/retest effects for the RJI. Within educational level, Kitchener et al.'s (1993) data appear to show lower test/retest correlations than one would expect from the test/retest correlations from other studies. Although Kitchener et al.'s short term study used only two dilemmas, it does not seem reasonable to attribute the lower test/retest correlations to the lower reliability of the shorter RJI (see Table 5). Generally, subjects in the control group showed a slightly higher test/retest correlation than subjects in the experimental group, as would be expected. Over all educational levels, the test/retest correlation was .88 for the control group and .79 for the experimental group. It could be that the observed differences in test/retest correlations are a function of the characteristics of the samples involved. Random samples of students show a lower pattern of test/rest correlation (Brabeck, 1983; Polkosnik & Winston, 1989, and Welfel & Davison, 1986) while studies of high verbal ability students (Kitchener & King, 1981; Kitchener et al., 1993) show a higher pattern of test/retest correlations.

Insert Table 11 About Here

An examination of patterns of average change between testings also provides some insight as to whether differential trajectories of growth obtained on the RJI as a function of initial educational level. Kitchener & King's (1981) sample showed an average different of .2 to .4 for the high school samples, while Brabeck's (1983) longitudinal data showed almost no growth. Since Brabeck's study involved samples of high and lower critical thinkers, perhaps this difference in growth rates could be due to a differential growth rate for high ability students, whereby they gain more on Reflective Judgment than their lower verbal ability counterparts. Interpolated different scores for a single year for the early undergraduate samples range from .14 (Welfel & Davison, 1986), .22 (Polkosnik & Winston (1989) Times 1 & 3 and .32 (Polkosnik & Winston, Times 1 & 2). For the late undergraduate samples, annual differences ranged from .1 (Kitchener & King 1981, Times 1 &4) to .6 (Polkosnik & Winston, 1989, Times 1 and 3).

For the graduate samples, test/retest differences were much higher for data from Kitchener et al. (1993) than would be expected based on an examination of earlier change patterns. The 24 early graduate students in the study showed a difference of .59 over the period of two weeks, an

indication of a possible test/retest effect of the RJI for this population, or evidence that the educational intervention presented in Kitchener et al. (1993) was particularly effective for this group. Even though the difference score for the 13 control individuals in this group did not show much gain, replication of a test/retest study for these individuals seems warranted. Similarly, for advanced graduate students, the gains in overall RJI score for the Kitchener et al. (1993) data are much higher than would be expected from the Kitchener & King (1981) data. Some of this difference could, however, be also due to the fact that these subjects tested much lower at time 1 than the subjects in Kitchener & King (1981).

When separate analyses of variance of the Kitchener et al. (1993) data was conducted separately for the experimental and control groups, the main effect for time for the control group (.013) was not significant, while the main effect for time in the experimental group (.23) was statistically significant. These results must be interpreted with caution, since a general linear model of the combined data failed to find a condition by time interaction. It is interesting to note however that a comparable analysis of variance which employed chronological age as opposed to educational level failed to find these patterns of significance. As such, these analyses may be taken to support the previous general finding in Reflective Judgment research that educational experience, as opposed to chronological age, is a more important determinant of level of Reflective Judgment.

Reflective Judgment: Inter and Intra-Individual Differences by Dilemma.

A structural equations approach was designed to test whether the RJI represents a single psychological construct (as argued from the internal consistency estimates and previous research discussed above). The structural equation approach proposed here also allows an assessment of the relative magnitude of rater bias and systematic differences in Reflective Judgment as a function of dilemma topic. Before discussing a classical test theory approach to Reflective Judgment, it helps to examine some properties of the instrument which have not been examined or explained thus far in the paper. As noted above in Table 6, the items of the RJI demonstrate a rough general item equivalence by educational level. If any trend can be extracted from these data, it would appear that the instrument is more internally consistent for advanced undergraduates (college juniors and seniors) and graduate students than for early undergraduates.

If the RJI measures a single ability, one would expect to find that dilemma scores for the RJI would be closer to each other for more advanced educational levels than for earlier educational levels. An examination of the patterns of scores assigned to individual dilemmas, however, finds that the opposite is the case. In an examination of the distribution of dilemma scores, it was found that 21.75% of individuals demonstrated dilemma scores which were at more than one major level and that multiple stage ratings were more common among advanced educational levels. For the junior, senior, beginning graduate and advanced graduate samples, the percentages of individuals with dilemma scores at more than one level of reasoning were 17%, 20%, 14%, and 16% respectively. Although 14% of the college freshman sample contained more than one stage rating across dilemmas, for no other group did more than 7% of the individuals show evidence of more than one stage. It is paradoxical, then, that groups which appear to show the largest variability in performance across dilemmas (which would normally indicate the presence of increased measurement error) would demonstrate the greatest amount of internal consistency. The structural equation approach outlined below is designed to assess whether the variability in performance in ratings across dilemmas constitutes measurement error, systematic differences in Reflective Judgment ability as a function of dilemma topic, or some pattern of unmeasured systematic bias in the rating of the RJI.

Structural equation modeling has enjoyed increased popularity with psychologists as a way of testing form equivalence (Lord & Novick, 1968; Loehlin, 1992). In parallel forms models, performance on each observed variable (in this case the rater's evaluation of a dilemma transcript) is thought to be a product of an unobservable true score (or scores) and a unique component of the variable due to measurement error. For models in which all variables are measures of the same underlying construct, these relationships may be represented by means of a path diagram such as that found in Figure 3. As can be seen, such models demonstrate the close relationship between classical test theory and confirmatory factor analysis. For the present study, a set of initial models was tested using all available data where all four RJI dilemmas were available from two raters (N=668). Eight observed variables were measured for this model, each rater's scores on each of the four dilemmas. All structural models presented below are maximum likelihood estimates.

Insert Figure3 About Here

RJI Dilemmas a Tau-Equivalent versus Congeneric Parallel Forms. As an initial test, two models were compared which are drawn from classical test theory. In the first model, RJI ratings across dilemmas and raters were through to represent tau-equivalent measures, meaning that each rater's score on each dilemma is roughly the same in terms of its efficacy as a measure of underlying Reflective Judgment ability. Each dilemma rating may, however, contain a different amount of error variance. This model of Reflective Judgment ability may then be compared with a model of congeneric parallel forms, in which each variable is thought to be related to the underlying construct of Reflective Judgment, but variables differ in their relative strength as indicators of Reflective Judgment. If a congeneric model of Reflective Judgment were true, compositing weights for the different dilemmas could then be used so that more reliable Reflective Judgment dilemmas would be more heavily weighted in the estimation of overall Reflective Judgment ability. Congeneric models failed to improve the fit of the model by reference to Type II fit measures (TLI=-.10, SBC_2=-.01, AIC_2=.00[25]). In contrast to the factor loadings shown in Figure 4 which show the congeneric factor model, unstandardized factor loadings for the tau-equivalent model were .85 across all dilemmas.

Insert Figure 4 About Here

Identification of Topic-Specific & Rater-Specific Sources of Covariation. Given the eight indicator variables of this study, it is also possible to estimate the relative magnitude of rater bias and systematic effects due to dilemma topic by means of a Schmid-Leiman transformation (Schmid & Leiman, 1957). Since the issue of dilemma differences and rater bias has been explored by means of analysis of variance models (e.g., Brabeck 1980; Wood, 1981; Welfel, 1979) some explanation and justification for the present approach is necessary. The means for the

25. It should be noted that the fit measures here are Type II fit measures relative to the null model of tau-equivalence and not to the independence model which is the usual basis for fit measures. See Marsh et al. (1988) for a discussion of the superior performance on Type II fit measures and their computational definitions.

four traditional dilemmas for the 668 subject data pool were almost identical (means were 4.06, 4.08, 4.00 and 4.05 for the Pyramids, News, Evolution, and Food Additives dilemmas respectively[26]). This means that, across all individuals, no dilemma appears to be systematically easier or more difficult than any other. Such statistics, though, may mask important sources of interindividual differences across dilemmas. For example, suppose that, unknown to the researcher, exactly half of the individuals in this study were very health conscious and as a result scored quite highly on the Food Additives dilemma. Further suppose that the other half of these subjects were quite interested in news reporting and thought quite a bit about whether it was biased. This state of affairs is shown graphically in Figure 5. The average scores by dilemma across these two groups might appear nearly identical. Clearly, if the researcher did not have access to information about these sources of expertise or interest, she might conclude that, on the average no differences existed between these two dilemmas when, in fact, quite dramatic unmeasured interindividual differences in expertise were evident. In analogous fashion, it is also possible to identify previously unmeasured effects at the rater level by examining common patterns of high or low assignment across all dilemmas of a particular rater by examining whether any significant covariation between the four dilemmas rated by a particular rater exists above and beyond that explained by reference to the overall Reflective Judgment Level (or dilemma-specific factors) found in the data.[27]

Insert Figures 5 & 6

Although it is not possible in this present study to detect level of interest/involvement or differential expertise across these dilemmas, some estimate of the presence and extent of such context effects can be gleaned from consideration of the data from two raters of these dilemmas. Figure 6, for example, shows the bar chart patterns of the scores of two raters for these

---

26. These means are based on all available data with four dilemmas and is a larger sample than that discussed in the structural models below. Overall means taken from studies with two raters for all dilemmas yielded identical mean values to two decimal places.
27. Since the analyses presented below constitute data from several raters and studies, the statistical power of detecting differential rater patterns at this level is probably quite low. Nevertheless, the model of rater effects provides a test for the existence of systematic rater effects across all studies.

unmeasured groups, indicating that, even though the unmeasured variable of subject interest/ experience is not included in the study, some evidence for such effects can be gleaned correlationally. Specifically, if context effects exist, we should expect to find that one rater's score for a dilemma transcript should correlate with another rater's score for the same transcript above and beyond the score one could expect based on overall Reflective Judgment ability.

In terms of a structural equation model, then, a Schmid-Leiman (1957) hierarchical factor analysis was used to determine if specific dilemma effects were present in the data in addition to an overall Reflective Judgment ability. Patterns of covariation between rater's scores for a dilemma over and above global Reflective Judgment ability were investigated by testing whether a significant improvement in model fit resulted from the addition of latent variables which had, as their indicators, only a particular dilemma. Two types of dilemma-specific Reflective Judgment models were explored: One which forced dilemma effects to be equal for all dilemmas, and another which allowed such effects to vary from one dilemma to another.[28]

In like fashion, the issue of unmeasured rater bias for certain types of individuals can be explored by testing to see whether a given rater's scores across the four dilemmas share common patterns of covariation beyond that accounted for by overall Reflective Judgment ability. These patterns of bias could, conceivably be different for the four dilemmas of the RJI.

Insert Figure 7

Since models which include topic specific and rater specific effects above and beyond overall Reflective Judgment ability are nested with the single factor Reflective Judgment model, it is possible to compare the incremental fit for nested model by reference of Type II incremental fits. Using the tau-equivalent model of Reflective Judgment as a null model, and an incremental fit criterion of .9 (Marsh et al., 1988), the structural model indicating that Reflective Judgment is a tau equivalent model with context effects unique to each dilemma (set equal across each dilemma) yielded a superior fit (TLI=.97, SBC_2=.95; AIC_2=.97). The standardized solution for these data

28. Since each of the context effect factors contained only two indicators, it was necessary to constrain the context effects to be the same for each dilemma and factor variances for these factors were set to 1 in order to obtain mathematically identified solutions.

are given in Figure 7. As can be seen, overall, each RJI rating is a relatively strong indicator of overall Reflective Judgment ability. (The standardized factor loading for overall Reflective Judgment level may be interpreted as a measure of the internal consistency of a given judge's rating of a particular dilemma. When adjusted for the effects of the length of the instrument, this value (.95) closely approximates the values of coefficient alpha for the overall group reported above. In addition to this overall ability, however, a substantial context effect also exists for the data (.46) but such context effects are not as reliably indicated as overall Reflective Judgment ability.[29]

Consonant with the overall estimates of rater bias using analysis of variance described above, little evidence was found for rater bias under this intra-individual differences model (defined as two unmeasured variables which had as manifest variables, all of the dilemma ratings associated with a particular rater. TLI increments for the effects of rater bias above the proposed model were all <=.6) The magnitude of these effects when estimated accounted for less that .03% of the variability in Reflective Judgment scores, a practically nonsignificant amount.

Of course, given the substantial literature and theory assuming the Reflective Judgment is a single ability, supplementary analyses were conducted to determine if the observed dilemma effects could be due to other characteristics of the data, such as differential gender effects, sampling variation, or differential reliability induced by the inclusion of data with lower internal consistencies.[30] Analyses which excluded suspiciously low reliability or unusually high- and low-scoring samples relative to other studies yielded the same choice of reliability model as the general analysis as well as similar factor loadings. Separate follow-up analyses based on general educational level (collapsed into high school, freshman and sophomore undergraduate, junior and senior undergraduate, and graduate samples to provide a minimally satisfactory sample size), revealed that the RJI is more reliable measure of overall ability for graduate samples (evidenced

---

29. Additional analyses were also conducted which employed the Schmid-Leiman transformation outlined here, except that dilemma effects were allowed to vary as well as factor loadings for the general RJI factor. These analysis generated fit measures similar to the TLI indices for the model described here, and incremental fit measures of this target model using the model adopted here did not reveal any incremental improvement in fit (TLI=-.06, SBC_2=-.28, and AIC_2=.09.

30. Specifically, the data were rerun excluding the data of Welfel's (1982) freshman sample, Van Tine's (1990) data and retest of McKinney (1985). Computer runs excluding low- or high-scoring samples excluded Kitchener et al. (1993), Strange & King (1981), DeBord (1993), and Brabeck (1983).

by a factor loading of .90 as op osed to .82 for the general model) and that dilemma effects were more pronounced for this group as well (.68 as opposed to .46 for the general model). As such, these analyses can be seen as a corresponding to the results from the spline regressions of stage utilization, which suggested that higher stages of Reflective Judgment are accompanied by more variability in response than lower levels of Reflective Judgment. In addition, the Type II fit measures were less clear-cut for the High School and Beginning Undergraduate analyses (TLI=.81, SBC_2=.56, AIC_2=.81 and TLI=.88, SBC_2 .72 and AIC_2=.88 for the Beginning Undergraduate and High School samples, respectively). These fit patterns also indicate that, for lower levels of RJI ability, the salience of unique dilemma effects and the assessment of Reflective Judgment as a single ability are less pronounced (loadings associated with overall Reflective Judgment were .67 and dilemma-specific effects were .55). Some caution, however, is appropriate in making this interpretation, since these patterns could also arise due to the smaller range of ability in the high school and beginning undergraduate samples relative to the graduate samples.

Discussion

At this point, a summary evaluation of the major claims for the Reflective Judgment model based on RJI data is appropriate, highlighting implications and directions for future research using the Reflective Judgment interview.

Summary of Findings

Data Accuracy and Replication of Analyses. Although effort was expended to identify and correct errors in the available data, it appears that, on the whole, Reflective Judgment data has been carefully and accurately entered. Clearly, the most serious errors in the data set occurred not from faulty data entry, but from mistakes in computer programming of data once it was entered. While, in one sense, any incorrect data is unacceptable in scientific research, the error rates found in these data appear to be lower than those found in previous psychological research. This increase in accuracy may be due to the increased sophistication of data entry and error correction which have occurred in psychological research since the 1970's. The few remaining errors appeared oversights which could be detected or avoided by suitable computer programming rather than relying on hand calculation. In no cases did correction of these errors result in different

patterns or different decisions regarding statistical significance. In contrast to earlier research (Rosenthal, 1978), data errors in Reflective Judgment do not appear to be in the direction of the researchers' hypotheses.

Some studies deviated from the usual reported procedure of defining a discrepancy as a difference of one stage between raters. While this resulted in a lower agreement rate, the psychometric differences between these studies (based on internal consistency estimates) appeared slight. The process of rerating discrepant scores between raters, while having little impact on the general internal consistency of scores, appeared to serve to identify and to some extent control for idiosyncratic rater biases at the dilemma level for some studies. While most studies indicated excellent internal consistency and agreement rates, some studies show some evidence of "rater drift" suggesting that a periodic review or recertification of raters may be advisable, prior to rating for a new research project.

In addition, a few additional errors and analysis decisions made it inappropriate to combine information from selected studies. In two cases incorrect statistical analyses were reported (Polkosnik & Winston's (1989) repeated measures anova and Welfel's (1982) follow-up analysis of covariance). King, Wood & Mines' (1990) reported agreement level was based on rater mean scores across all dilemmas and not the dilemma-level agreements found in other studies. Minor errors also occurred as a result of reporting raw agreements versus agreements which were corrected for chance.

General Psychometric Claims for the RJI. On the whole, the RJI appears to be an internally consistent instrument, even when these internal consistencies are computed within educational level. Coefficient alpha estimates ranged from .72 to .90 within educational level and indicate that the instrument demonstrates acceptable reliability for use as a research instrument within a given educational level setting in addition to its previous use as a general measure of the outcomes of college education across a variety of educational levels. Two samples showed a lower than expected degree of internal consistency. No ready explanation other than rater drift seems available for scores from Welfel's (1982) freshman sample. King et al.'s (1989) study of black undergraduates showed a clear pattern of low internal consistency across all educational levels. Clearly, additional work using the RJI on minority populations is warranted in order to

ascertain if these findings are due to faulty rater training or due a differential reliability of the RJI for minority populations. Two of the three high school samples also demonstrated a low internal consistency. Further work establishing the internal consistency of the RJI for high school students also seems warranted.

Test/Retest effects for the RJI show negligible gains in improvement as a result of taking the instrument, and a test/retest correlation are in general agreement with the internal consistency estimates produced earlier. The one exception to this pattern are the gains in performance for the graduate samples from Kitchener et al. (1993). The dramatic gains for early graduates in the experimental condition should be replicated before one can conclude that such environmental supports are particularly efficacious for this group. The gains of roughly a third of a stage for the advanced graduate samples may indicate that this group has a significant test/retest effect and may explain difficulties that Kitchener et al. (1993) encountered in attempting to document a developmental growth spurt for subjects in the older age groups of their study. Clearly research which examines the role of chronological age which controls for the effects of educational level are warranted in order to establish the presence of growth spurts for Reflective Judgment under high support conditions.

Little evidence was found for overall rater bias in these studies (meaning that one rater assigned systematically higher or lower scores to individuals across all dilemmas). Some evidence, however, does exist that some raters differentially score given dilemmas higher or lower than the remaining dilemmas. Although this problem is corrected to some extent by the rerating process, it seems advisable for future researchers to conduct these analyses to determine if such differential bias is present. Perhaps this finding points to the need for a more extensive certification process.

The Relationship of Reflective Judgment to Educational Level. On the average, samples of Reflective Judgment scores increase as a function of level of educational attainment from 8th grade high school samples to advanced doctoral samples. In addition, greater variability in performance is found at higher levels of educational attainment, even after taking sample differences into account. However, substantial differences were found across samples with the same level of educational attainment, suggesting that unmeasured subject characteristics (such as

verbal ability) or differential collegiate admissions policies or attrition rates may affect the average level of Reflective Judgment level. Some preliminary evidence exists that student populations from highly selective private institutions are higher in Reflective Judgment than their public university counterparts. Examination of samples taken from other institutions indicates that, contrary to the conclusions of King et al. (1990), graduate students in the social sciences do not necessarily demonstrate high levels of Reflective Judgment relative to samples of natural science students. This discrepancy may be due to differential amounts of educational attainment for these samples, and suggests that further research is necessary in early graduate samples (discussed below).

Claims Regarding Sequentiality. The analyses of these data support the view that the Reflective Judgment model describes an internally consistent series of increasingly more adequate solutions to ill-structured problems. No nonsequential response patterns were identified using loglinear techniques, however the sequentiality analyses and spline regressions show that lower levels of Reflective Judgment show less variability in performance than higher levels of Reflective Judgment, suggesting that the ability to reason complexly about some ill-structured problems may be different from topic to topic at more advanced levels of Reflective Judgment. Perhaps this is a function of differential interest or expertise across the Reflective Judgment topics (e.g., some individuals may be aware of conflicting general contexts for organizing information in the area of food additives, but are unaware of the conflicting interpretations of data in the Evolution dilemma).

Claims Regarding Reflective Judgment as a Single Intellectual Ability. Currently, researchers and educators view Reflective Judgment as a single ability. Kitchener et al. (1993), based on subjects' ability to paraphrase prototypic Reflective Judgment statements, argue that individuals at a given level of Reflective Judgment are unable to understand and interpret statements much higher than their own. As such, they argue that educational interventions should provide appropriate support conditions to the individual based on his/her level of Reflective Judgment. The results of the confirmatory factor analysis suggest that this approach may be most appropriate for use with earlier undergraduate and high school populations, but that advanced undergraduate and graduate students do not have a "single score" on Reflective Judgment, but

demonstrate differences in their ability to reason across the individual dilemmas of the RJI. As such, this analysis suggests a different type of educational intervention for such advanced students: When students are presented with a given ill-structured problem, instructors may attempt to identify other types of ill-structured problems which the student thinks more complexly about. Substantial improvements in Reflective Judgment about a particular dilemma could be realized by getting students to appreciate the ways in which different ill-structured problems share common epistemological characteristics.

The examination of test/retest effects for the graduate population also provides some insight to the nature of the existence and timing of growth spurts for older age groups. Kitchener et al. (1993) noted that significant improvements in performance existed as a function of chronological age, but were not able to determine a distinct chronological age period which improved performance in older students. Since the Kitchener et al. (1993) study did not control for level of educational level, the failure to control for educational level may explain the failure to document critical ages associated with the higher levels of Reflective Judgment.

Conceptually, the identification of problem-specific intra-individual differences in Reflective Judgment is an example of the utility of the developmental perspective for models of adult cognition such as the Reflective Judgment Interview. Baltes, Reese, & Nesselroade (1977, p. 4) describe the focus of life-span developmental psychology as "the description, explanation, and modification (optimization) of intraindividual change in behavior across the lifespan, and on interindividual differences (and similarities) in intraindividual change." Lamiell (1981) has termed such a combination of idiographic and nomothetic approaches "idiographic." Further research which investigates the presence and extent of such intraindividual differences in performance may attempt to identify systematic interindividual differences in performance (by identifying target groups which differ in Reflective Judgment topics) or may further document the systematic intraindividual variation as a function of topic at the dilemma level. To some extent, this work has already begun: DeBord (1993) assessed two new dilemmas in psychology in addition to the traditional Reflective Judgment dilemmas. He found that systematic differences in Reflective Judgment were not found for the traditional versus psychological dilemmas in beginning undergraduates, but found that significant differences were found for doctoral students

in psychology, with the psychology graduate students showing higher levels of Reflective Judgment for the psychological dilemmas than for the traditional dilemmas.

General Implications.

In general, the Reflective Judgment model and interview is a theory of adult cognition which documents systematic advances made by individuals in reasoning about a class of problems which have no single correct answer. The present study represents a body of information which was drawn from several educational settings with a variety of populations. As such, some caution regarding the generalization of this body of information to any comparable individual sample of individuals is appropriate. To the extent possible, however, every effort has been made to investigate the comparability of individual samples of data prior to combining these samples with other information. With the exception of one study of minority populations and one undergraduate sample, the interview appears to document these differences in an internally consistent fashion. In contrast to many theories or adult development, the RJI documents sequential patterns of response at both the level of individual's response (as evidenced by Davison's (1979) test, and longitudinal patterns of change (as evidenced by test/retest patterns). This attention to the development and scoring of the RJI together with the rater certification program documents such changes in reasoning to a degree not found in other models of adult development (except for Rest's (1979) instrument measuring moral development) (Davison et al., 1980). In addition to these requirements, the present manuscript has set a higher standard for internal consistency than previously employed in other studies by calculating internal consistency as a random effect and estimating internal consistency of the RJI dilemmas separately by educational level as opposed to the general reliability estimates for individual studies which are based on a variety of educational levels.

While the RJI possesses highly desirable characteristics relative to other models of development, this present manuscript enables some useful observations not readily apparent from the individual studies themselves. First, if the RJI is to be used as a measure of individual differences for college and university instructors at the classroom level (as suggested by King & Kitchener's (1994) discussion of implications of the model), such differences between students may not be as reliably unidentifiable as previously thought, even when based on information

equivalent to a full RJI interview scored by two certified raters. This finding does not devalue the utility of Reflective Judgment as a general overall goal of higher education, the value of the Reflective Judgment model for promoting ill-structured problem-solving, or the use of an overall average across individuals as an indicator of average problem-solving ability. It does, however, suggest that the interview may not have sufficient reliability for use in aptitude-treatment interaction studies. As such, further assessment of individuals may be necessary or the use of latent variable techniques may be necessary to estimate aptitude treatment interactions for such studies (such as proposed by Kenny & Judd, 1984).

Implications for Future Research

It is highly unlikely, given the labor-intensive nature of the administration and scoring of the RJI, that a complete replication can be conducted of the educational level and context effects found here for the RJI. The present findings do, however, point the way to the design of a number of studies which could replicate the findings presented here. In addition to the replication of the use of the RJI with minority populations mentioned above, it also seems advisable to test samples of graduate students (particularly masters and early Ph.D. levels) longitudinally, in an effort to see whether a significant test/retest effect exists for this group or whether the early graduate experience constitutes a time of significant, rapid growth in the ability to reason about ill-structured problems.

The other major findings of the present study concern the nature and interpretation of variability across the dilemmas of the RJI. On the one hand, a large proportion of variability in scores due to a single underlying factor is due to individuals' overall level of Reflective Judgment ability. Seen at the level of intraindividual differences, however, quite a different pattern emerges, with more educated individuals (and individuals with higher overall Reflective Judgment scores) showing a greater pattern of variability of responses across dilemmas. In the terminology of developmental psychology, the Reflective Judgment interview shows a greater pattern of "horizontal decalage" for higher levels than for lower levels. Dilemma-specific patterns of response in the RJI were also found which were, in general, more pronounced for individuals from late undergraduate and graduate samples. Since DeBord (1993) has documented systematic differences in reasoning as a function of topic for graduate students in the area of psychology,

dilemma-specific effects found in this study are interpreted as differential unmeasured intraindividual interest or expertise differences. The possibility also exists that the dilemma effects documented here represent an unmeasured rater by dilemma bias pattern. Clearly, additional research is indicated which involves the use of separate dilemmas dealing with the same general subject area in an effort to more accurately assess the role and extent of such content-specific patterns of covariation.

References

Association of American Colleges (1991). The challenge of connecting learning. Project on Liberal Learning, Study-in-Depth, and the Arts and sciences Major. Washington, D.C: Association of American Colleges.

Bangdiwala, S.I. (1985). A graphical test for observer agreement. Proceedings of the 45th Session of the International Statistical Institute, Amsterdam, Holland, 307-308.

Baltes, P.B., Reese, H.W., & Nesselroade, J.R. (1977). Life-span developmental psychology: Introduction to research methods. Monterey, CA: Brooks/Cole.

Brabeck, M.M. (1980). The relationship between critical thinking skills and development of reflective judgment among adolescent and adult women. Doctoral dissertation, University of Minnesota, Dissertation Abstracts International, 41/11A, p. 4647.

Brabeck, M.M. (1983). Critical thinking skills and reflective judgment development: Redefining the aims of higher education. Journal of Applied Developmental Psychology, 4, 23-34.

Brabeck, M.M. & Wood, P.K. (1990). Cross-sectional and longitudinal evidence for differences between well-structured and ill-structured problem solving abilities. in M.L. Commons, C. Armon, L. Kohlberg, F.A. Richards, T.A. Grotzer, and J.D. Sinnott (Eds.) Adult Development 2: Models and Methods in the Study of Adolescent and Adult Thought. (pp. 133-146) New York: Praeger.

Darlington, R.B. (1990). Regression and linear models. New York: McGraw-Hill.

Davison, M.L. (1979). Testing a metric unidimensional qualitative unfolding model for attitudinal or developmental data. Psychometrika, 44, 179-194.

Davison, M.L., King, P.M., Kitchener, K.S. & Parker, C.A. (1980). The stage sequence concept in cognitive social development. Developmental Psychology, 16, 121-131.

DeBord (1993). Promoting Reflective Judgment in counseling psychology graduate education. Unpublished masters thesis. University of Missouri-Columbia.

Dove, W.R. (1990).The identification of ill-structured problems by young adults. (Doctoral dissertation, University of Denver, 1990). Dissertation Abstracts International, 51-06B, 3156-3358.

Duncker, K. (1945). On problem-solving. Psychological Monographs, 58(5, Whole No. 270).

Fleiss, J.L. & Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika, 43, 259-262.

Glatfelter, M. (1982). Identity development, intellectual development, and their relationship in reentry women students. Doctoral dissertation, University of Minnesota. Dissertation Abstracts International, 43, 3543A.

Hoffman, L.R., Burke, R.J. & Maier, N.R.F. (1963). Does training with differential reinforcement on similar problems help in solving a new problem? Psychological Reports, 13, 147-154.

Kelly, C. (1993). The fundamental measurement of Reflective Judgment. Unpublished doctoral dissertation, University of Denver.

Kelton, J. & Griffith, J.V. (1986). The learning context questionnaire for assessing intellectual development. unpublished manuscript.

Kenny, D.A. & Judd, C.M. (1984). Estimating the nonlinear and interactive effects of latent variables. Psychological Bulletin, 96, 201-210.

King, P.M. (1977). The development of reflective judgment and formal operational thinking in adolescents and young adults. doctoral dissertation, University of Minnesota. Dissertation Abstracts International, 38, 7233A.

King, P.M. (1986). Formal reasoning in adults: A review and critique. In R. Mines and K. Kitchener (Eds.), Adult Cognitive Development. Praeger: New York.

King, P.M. & Kitchener, K.S. (1994). The development of Reflective Judgment in adolescence and adulthood. Jossey-Bass: San Francisco.

King, P.M., Kitchener, K.S., Davison, M.L., Parker, C.A., & Wood, P.K. (1983). The justification of beliefs in young adults: A longitudinal study. Human Development, 26, 106-116.

King, P.M., Kitchener, K.S. & Wood, P.K. (1994). Research on the Reflective Judgment model. in King, P.M. & Kitchener, K.S. The development of Reflective Judgment in adolescence and adulthood. Jossey-Bass: San Francisco.

King, P.M., Kitchener, K.S., Wood, P.K., & Davison, M.L. (1989). Relationships across developmental domains: A longitudinal study of intellectual, moral, and ego development. In M.L. Commons, J.D. Sinnott, F.A. Richards & C. Armon (Eds.), Adult Development:

Vol. 1 Comparisons and applications of developmental models (pp. 57-72). New York: Praeger, 1989.

King, P.M., Taylor, J.A. & Ottinger, D.C. (1989, November). Intellectual development of black college students on a predominantly white campus. Paper presented at the meeting of the Association for the Study of Higher Education, Atlanta, Georgia.

King, P.M., Wood, P.K. & Mines, R.A. (1990). Critical thinking among college and graduate students. The Review of Higher Education, 13(2), 167-186.

Kitchener, K.S. (1983). Cognition, metacognition and epistemic cognition: A three-level model of cognitive processing. Human Development, 4, 222-232.

Kitchener, K S. (1986). The reflective judgment model: Characteristics, evidence, and measurement. In R.A. Mines & K.S. Kitchener (Eds.), Adult Cognitive development: Methods and models (pp. 76-91). New York: Praeger, 1986.

Kitchener, K.S. & King, P.M. (1981). Reflective judgment: Concepts of justification and their relationship to age and education. Journal of Applied Developmental Psychology, 2, 89-116.

Kitchener, K.S. & King, P.M. (1985). Reflective judgment scoring manual. (Available from K.S. Kitchener, School of Education, University of Denver, Denver, CO 80208 or P.M. King, Department of Higher Education and Student Affairs, Bowling Green State University, Bowling Green, OH 43403.)

Kitchener, K.S. & King, P.M. (1990). The reflective judgment model: Ten years of research. In M.L. Commons, C. Armon, L. Kohlberg, F.A. Richards, T.A. Grotzer & J.D. Sinnott. Adult Development: Vol. 2. Models and methods in the study of adolescent and adult thought. (pp. 63-78) New York: Praeger.

Kitchener, K.S., King, P.M., Wood, P.K. & Davison, M.L. (1989). Consistency and sequentiality in the development of reflective judgment: A six year longitudinal study. Journal of Applied Developmental Psychology, 10, 73-95.

Kitchener, K.S. & Kitchener, R.F. (1981). The development of natural rationality: Can formal operations account for it? Contributions to Human Development: Social Development in Youth: Structure & Content, 5. Karger: Basel, Switzerland.

Kitchener, K.S., Lynch, C.L., Fischer, K.W. & Wood, P.K. (1993). Developmental range of reflective judgment: The effect of contextual support and practice on developmental stage. Developmental Psychology, 29, 893-906.

Kitchener, K.S. & Wood, P.K. (1987). Development of concepts of justification in German university students. International Journal of Behavioral Development, 10, 171-185.

Lamiell, J.T. (1981). Toward an idiothetic psychology of personality. American Psychologist, 36, 276-289.

Lawson, J.M. (1980). The relationship between graduate education and the development of Reflective Judgment: A function of Age or Educational Experience. Doctoral dissertation, University of Minnesota. Dissertation Abstracts International, 47, 402B.

Loehlin, J.C. (1992). Latent Variables Models (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.

Loevinger, J., Wessler, R. & Redmore, C. (1970). Measuring ego development. Jossey-Bass: San Francisco.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley

Lynch, C.L. (1990). The impact of a high support condition on the exhibition of reflective judgment. (Doctoral dissertation, University of Denver). Dissertation Abstracts International, 50/09, p. 4246.

McKinney, M. (1985). Reflective judgment: An aspect of adolescent cognitive development (Intelligence, Wechsler, Verbal Achievement, Ability). Intellectual development of younger adolescents. (Doctoral dissertation, University of Denver). Dissertation Abstracts International, 47, 402B.

Maier, N.R.F. (1933). An aspect of human reasoning. British Journal of Psychology, 24, 144-155.

Marsh, H.W., Balla, J.R. & McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Mines, R.A., King, P.M., Hood, A.B. & Wood, P.K. (1990) Stages of intellectual development and associated critical thinking skills in college students. Journal of College Student Development, 31, 537-547.

Pascarella, E. & Terenzini, P. (1991). How college affects students: Findings and insights from twenty years of research. San Francisco: Jossey-Bass.

Polkosnik, M.C. & Winston, R.B. (1989). Relationships between students' intellectual and psychological development: An exploratory investigation. Journal of College Student Development, 30, 10-19.

Rest, J. (1979). Development in judging moral issues. University of Minnesota: Minneapolis.

Rindskopf, D. (1984). Using phantom and imaginary latent variables to parameterize constraints in linear structural models. Pychometrika, 49, 37-47.

Rosenthal, R. (1978). How often are our numbers wrong? American Psychologist, 33, 1005-1008.

Sakalys, J.A. (1984). Effects of an undergraduate research course on cognitive development. Nursing Research, 33, 290-295.

Schmid, J. & Leiman, J.M. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53-61.

Schmidt, J.A. (1983). The intellectual development of traditionally and nontraditionally aged college students: A cross-sectional study with longitudinal follow-up. (doctoral dissertation, University of Minnesota). Dissertation Abstracts International, 44/09A, p. 2681.

Schmidt, J.A. (1985). Older and wiser? A longitudinal study of the impact of college on intellectual development. Journal of College Student Personnel, 26, 388-394.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Sternberg, R.J. (1982). Handbook of human intelligence. New York: Cambridge University Press.

Strange, C.C. (1978). Intellectual development, motive for education and learning styles during the college years: A comparison of adult and traditional-age college students. (Doctoral dissertation, University of Iowa). Dissertation Abstracts International, 39/08A, p. 4768.

Strange, C.C. & King, P.M. (1981). Intellectual development and its relationship to maturation during the college years. Journal of Applied Developmental Psychology, 2, 281-295.

Tinsley, H.E.A. & Weiss, D.J. (1975). Inter-rater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22, 358-376.

Van Tine, N.B. (1990). The development of reflective judgment in adolescents. (Doctoral dissertation, University of Denver), Dissertation Abstracts International, 51/08A, 2659.

Welfel, E.R. (1979). The development of Reflective Judgment: Its relationship to year in college, academic major, and satisfaction with major among college students (Doctoral dissertation, University of Minnesota). Dissertation Abstracts International, 40/09A, p. 4949.

Welfel, E.R. (1982). How students make judgments: Do educational level and academic major make a difference? Journal of College Student Personnel, 23, 490-497.

Welfel, E.R. & Davison, M.L. (1986). The development of reflective judgment in the college years: A four year longitudinal study, Journal of College Student Personnel, 27, 209-216.

Wolins L. (1962). Responsibility for raw data. American Psychologist, 22, 657-658.

Wood, P.K. (1983). Inquiring systems and problem structure: Implications for cognitive development. Human Development, 26, 249-265.

Wood, P.K. (1990). Construct validity and theories of adult development: Testing for Necessary but not sufficient relationships. In M.L. Commons, C. Armon, L. Kohlberg, F.A. Richards, T.A. Grotzer, & J.D. Sinnott (Eds.), Adult development: Vol. 2. Models and methods in the study of adolescent and adult thought. (pp. 113-132). New York: Praeger.

Wood, P.K. (in press). The effect of unmeasured variables and their interactions on structural models. in A. von Eye and C. Clogg (Eds.) Analysis of latent variables in developmental research. Beverly Hills, CA; Sage.

Wood, P.K. & Games, P. (1991). Rationale, detection and implications of interactions between independent variables and unmeasured variables in linear models. Multivariate Behavioral Research, 25, 295-311.

Table 1
Traditional Reflective Judgment Dilemmas

Topic	Text
Pyramids.	Most historians claim that the pyramids were built as tombs for kinds by the ancient Egyptians, using human labor, and aided by ropes, pulleys and rollers. Others have suggested that the Egyptians could not by themselves have build such huge structures, for they had neither the mathematical knowledge, the necessary tools, nor an adequate source of power. They claim that the Egyptians were aided by visitors from other planets.
News.	Some people believe that news stories represent unbiased, objective reporting of news events. Others say that there is no such thing as unbiased, objective reporting and that even in reporting the facts, news reporters project their own interpretations into what they write.
Evolution.	Many religions of the world have creation stories, These stories suggest that a divine being created the earth and its people. Scientists claim, however, that people evolved from lower animals forms (some of which were similar to apes) into the human forms known today.
Food Additives.	There have been frequent reports about the relationship between chemicals that are added to foods and the safety of these foods. Some studies indicate that such chemicals can cause cancer, making these foods unsafe to eat. Other studies, however, show that chemical additives are not harmful, and actually make the foods containing them more safe to eat.

Table 2
Reflective Judgment Interview - Probe Questions[1]

Probe Questions	Rationale/Purpose
What do you think about these statements?	To allow the participant to share an initial reaction to the problem presented. Most respondents state which point of view is closer to their own (e.g., that the Egyptians built the pyramids, that news reported is biased).
How did you come to hold that point of view?	To find out how the interviewee arrived at the point of view, and whether and how it has evolved from other positions on the issue.
On what do you base that point of view?	To elicit the basis of a participant's point of view, for example, a personal evaluation of the data, consistency with an expert's point of view, a specific experience, etc.
Can you ever know for sure that your position on this is correct? How or why not?	To better understand the participant's assumptions about whether issues like this can be known absolutely, what s/he would do in order to increase the certainty, or why that wouldn't be possible.
When two people differ about matters such as this, is it the case that one opinion is right and the other one is wrong? If yes, what do you mean by "right"? If no, can you say that one opinion is in some way better than the other? What do you mean by "better"?	The question is not designed to assess moral rightness, but factual accuracy. The first purpose is to see if the respondent hold a dichotomous, either/or view of the issue (characteristic of the early stages). The second purpose is to allow the participant to give criteria by which s/he evaluates the adequacy of arguments (information that helps differentiate high from middle-level stage responses).
How is it possible that people have such different points of view about this subject?	This question is designed to elicit comments about the participant's understanding of differences in perspectives and opinions (what they are based on and why there exists such diversity of opinion about the issue.

[1] Adapted from King & Kitchener (1994)

Table 2 (cont.)
Reflective Judgment Interview - Probe Questions

Probe Questions	Rationale/Purpose
How is it possible that experts in the field could disagree about this subject?	This question is not designed to tap their view of experts and authorities in terms of their decision-making about controversial issues, such as what role experts might play (if any), whether their viewpoints are weighted differently, and if so, why this would be done.

If the person does not take a stand on the issue (does not endorse a particular point of view) on the first question, the following questions are asked:

2. Could you ever say which was the better position? How/Why not?
3. How would you go about making a decision about this issue?
4. Will we ever know for sure which is the better position? How/Why not?
5. When people differ about matters such as this, is it the case that one opinion is right and one is wrong?
   (If yes) What do you mean by "right"?
   (If no) Can you say one opinion is in some way better than the other? What do you mean by better?
6. How is it possible that people can have such different points of view about this subject?
7. What does it mean when experts in the field disagree about this subject?

Table 3: The Reflective Judgment Model: Stage Related Assumptions about Knowing[1]

Stage 1- A person knows what she or he has observed. Facts and judgments are not differentiated.

View of Knowledge: Knowledge is assumed to exist absolutely and concretely. It can be obtained with absolute certainty by direct observation.

Concept of Justification: Beliefs need no justification since there is assumed to be an absolute correspondence between what is believed and what is true. Alternatives to one's view are not perceived.

Typical Judgment: "I know what I have seen."

Stage 2- Authorities and facts are related: authority figures are sources of facts and, therefore, truth.

View of Knowledge: Knowledge is assumed to be absolutely certain, or certain but not immediately available. Knowledge can be obtained directly through the senses (such as direct observation) or via authority figures.

Concept of Justification: Beliefs are unexamined and unjustified, or justified by their correspondence with the beliefs of an authority figure (such as a teacher or parent). Most issues are assumed to have a right answer, so there is little or no conflict in making decisions about disputed issues.

Typical Judgment: "If it is on the news, it has to be true."

Stage 3- Absolute answers are assumed to exist, but to be temporarily inaccessible. In the absence of absolute truth, facts, and personal beliefs are seen as equally valid.

View of Knowledge: Knowledge is assumed to be absolutely certain or temporarily uncertain. In area of temporary uncertainty, only personal beliefs can be known until absolute knowledge is obtained. In areas of absolute certainty, knowledge is obtained from authorities.

Concept of Justification: In cases in which certain answers exist, beliefs are justified by reference to the authorities views. In areas in which answer do not exist, beliefs are defended as personal opinion since the link between evidence and beliefs is unclear.

Typical Judgment: "When there is evidence that people can give to convince everybody one way or another, then it will be knowledge; until then, it's just a guess."

Stage 4- Evidence is now seen as important to the construction of knowledge claims, along with the acknowledgment that a belief cannot be known with absolute certainty for pragmatic reasons. Thus, knowledge claims are idiosyncratic to the individual.

View of Knowledge: Knowledge is uncertain and knowledge claims are idiosyncratic to the individual since situational variables (e.g., incorrect reporting of data, data lost over time or disparities in access to information) dictate that knowing always involves an element of ambiguity.

Concept of Justification: Beliefs are justified by giving reasons and using evidence, but the arguments and choice of evidence are idiosyncratic, for example, choosing evidence that fits an established belief.

Typical Judgment: "I'd be more inclined to believe evolution if they had proof. It's just like the pyramids: I don't think we'll ever know. Who are you going to ask? No one was there."

[1] **Adapted from King & Kitchener (1994)**

Stage 5- Types of evidence are differentiated within perspectives (e.g., historical or scientific evidence). Further, different rules of inquiry across perspectives or disciplines are recognized. Quality of evidence is also evaluated as strong/weak, relevant/irrelevant, etc. Evidence is not an end in itself, but is used to construct interpretations.

View of Knowledge: Knowledge is contextual and subjective since it is filtered through a person's perceptions and criteria for judgment. Only interpretations of evidence, events or issues may be known.

Concept of Justification: Beliefs are justified within a particular context using the rules of inquiry for that context and by context specific interpretations of evidence. Specific beliefs are assumed to be context-specific or are balanced against other interpretations, which complicates (and sometimes delays) conclusions.

Typical Judgment: "People think differently and so they attack the problem differently. Other theories could be as true as my own, but based on different evidence."

Stage 6- Generalized rules of inquiry may be applied across perspectives (e.g., the weight of the argument, likelihood of the conclusion being correct, acknowledgment that judgments are tentative). Interpretations are subject to critique and judgment for coherency, consistency with the evidence, explanatory power, etc.

View of Knowledge: Knowledge is constructed into individual conclusions about ill-structured problems based on information from a variety of sources. Interpretations that are based on evaluations of evidence across contexts and on the evaluated opinions of reputable others can be known.

Concept of Justification: Beliefs are justified by comparing evidence and opinion from different perspectives on an issue or across contexts, and by constructing solutions that are evaluated by criteria, such as the weight of the evidence, the utility of the solution or the pragmatic need for action.

Typical Judgment: "It's very difficult in this life to be sure. There are degrees of sureness. You come to a point at which you are sure enough for a personal stance on an issue."

Stage 7- Judgments are seen as the outcome of a process of rational inquiry; they are based on a variety of interpretive considerations (e.g., the explanatory value of the interpretations, the risks of an erroneous conclusion, consequences of alternative judgments) and the interrelationships of these factors.

View of Knowledge: Knowledge is the outcome of a process of reasoned inquiry in which solutions to ill-structured problems are constructed. The adequacy of those solutions is evaluated in terms of what is most reasonable or probable based on the current evidence, and is reevaluated when relevant new evidence, perspectives, or tools of inquiry become available.

Concept of Justification: Beliefs are justified probabilistically based on a variety of interpretive considerations, such as the weight of the evidence, the explanatory value of the interpretations, the risk of erroneous conclusions, consequences of alternative judgments, and the interrelationships of these factors. Conclusions are defended as representing the most complete, plausible, or compelling understanding of an issue, based on the available evidence.

Typical Judgment: "One can judge arguments by how well thought out the positions are, what kinds of reasoning and evidence are used to support it, and how consistent the way one argues on this topic is as compared with other topics."

Table 4

Study, Sample and Subject Ascertainment Rates by Educational Level

Education Level	Number of Studies	Number of Samples	Number of Subjects
High School			
(8th Grade - 12th Grade)			
Ascertained	5	12	196
	(100)[1]	(100)	(100)
Not Ascertained	0	0	0
Undergraduate[2]			
(Freshman - Senior)			
Ascertained	16	37	728
	(70)	(64)	(63)
Not Ascertained	7	20	428
Graduate			
(Masters & Doctoral)			
Ascertained	8	13	196
	(100)	(100)	(100)
Not Ascertained	0	0	0
Non-Student Populations			
Ascertained	2	2	55
	(40)	(40)	(36)
Not Ascertained	3	3	96
Total			
Ascertained	15[3]	63	1334
	(60)	(73)	(72)
Not Ascertained	10	23	524

1. Numbers in parentheses indicate percentages
2. Includes data from Traditional & Non-Traditionally aged student
3. Study numbers do not total since many studies investigated students from many education levels

Table 5

Internal Consistency Estimates Based on Topics of the RJI for All Available Studies[1]

Study	N	Coefficient Alpha	95% Confidence Interval Lower	Higher	Raw Agreement	Kappa	Bangdiwala
Studies Using All Four Dilemmas							
Brabeck (1983)	119	.76(.75)	.68	.82	.79	.67	.77
Time 2	25	.77(.76)	.58	.89	.82	.69	.80
Time 3[2]	22	.84(.86)	.70	.93	.84	.72	.85
Dove (1990) Time 1[3]	44	.91(.89)	.86	.95	.80	.70	.71
Kelton & Griffith (1986)	16	.95	.89	.98	n.a.[4]		
King, Wood & Mines (1990)	91	.89(.90)	.84	.92	.72	.62	.66
Kitchener & King (1981)	80	.95(.95)	.93	.97	.87	.84	.82
Time 2[5]	58	.96(.95)	.93	.97	.78	.71	.75
Time 3[6]	54	.92(.92)	.87	.95	.75	.67	.60
Time 4[7]	50	.88(.83)	.81	.93	.76	.69	.60
Lawson (1980)	80	.81	.73	.87	n.a.		
McKinney (1985)	56	.54(.55)	.30	.71	.98	.96	.97
Time 2[8]	21	.77(.86)	.55	.90	.70	.42	.59
Van Tine (1990)	42	.79(.82)	.67	.88	.90	.81	.83
Welfel (1982)	64	.63	.45	.76	n.a.		
Time 2[9]	25	.76	.55	.88	n.a.		
Studies Not Using All Four Dilemmas							
DeBord (1993)[10]	42	.83(.94)	.68	.91	.73	.57	.66
Glatfelter (1982)[10]	80	.87(.85)	.79	.91	.97	.94	.97
Kelton & Griffith (1986)[11]	125	.95	.94	.97	n.a.		
Kitchener et al. (1993)[12]	112	.86(.88)	.80	.90	.79	.71	.75
Time 2 Control[12]	53	.90(.91)	.83	.94	.85	.79	.83
Time 2 Experimental[12]	104	.93(.94)	.89	.95	.83	.77	.74
Kitchener & Wood (1987)[13]	48	.88(.79)	.80	.93	.61	.49	.48
King, Taylor & Ottinger (1989)[14]	146	.53(.53)	.38	.65	1.00	.99	1.00
Polkosnik & Winston (1989)[14]	19	.80	.58	.92	n.a.		
Time 2[14]	15	.89	.73	.96	n.a.		
Time 3[14]	15	.94	.86	.98	n.a.		
Strange & King (1981)[15]	64	.72(.73)	.58	.82	.75	.64	.67

1. Coefficient alpha based on Resolved Scores (numbers in parentheses indicate Round 1 alphas), Agreement measures based on Round 1 data
2. Published in Brabeck & Wood (1990).
3. Data from Kitchener et al. (1993) were merged with Dove (1990) to form complete records.
4. Data were not available because Round 1 ratings were not coded in the data set. For two studies, agreement information is unavailable because only one rater was used (Kelton & Griffith, 1986; Polkosnik & Winston, 1989).
5. Published in King, Kitchener, Davison, Parker & Wood (1983)
6. Published in Kitchener, King, Wood & Davison (1989).
7. Published in Kitchener & King (1990).
8. Conducted by and reported in Van Tine (1990)
9. Published in Welfel & Davison (1986)
10. Based on two randomly selected dilemmas
11. Based on three randomly selected dilemmas individuals with all four dilemmas not included
12. Based on subjects not included in Dove (1990) having only Additives and Pyramids dilemma. Experimental and control groups combined since Time 1 assessment is a pretest.
13. Based on Pyramids, News, and Additives Dilemmas
14. Based on News, Evolution, and Additives dilemmas
15. Based on Pyramids, News, and Evolution dilemmas

Table 6
Internal Consistency Estimates Based on Topics as Items[1]

Education level	N	Coefficient Alpha	95% Confidence Interval Lower	Higher
High School[2]				
9th grade	53	.84	.76	.90
11th Grade	44	.90	.84	.94
12th Grade	62	.72	.58	.82
Undergraduate				
Freshman	64	.56	.35	.71
Freshmen w/o Welfel (1982)	32	.79	.64	.89
Sophomore	32	.73	.53	.85
Junior	77	.85	.78	.90
Senior	119	.79	.72	.85
Graduate Students				
Masters/Beginning Doctorate	92	.82	.75	.87
Advanced Doctoral	42	.85	.76	.91

1. Studies not using all four dilemmas not included in Education Level alphas. Based on Resolved scores.
2. Data from 8th and 10th grades excluded since some dilemma covariances contained negative values, yielding an invalid estimate for coefficient alpha.

Table 7

Agreement Estimates for Individual Dilemmas by Educational Level[1]

Education level	N Dilemmas	Raw Agreement	Kappa	Bangdiwala
High School				
8th grade	20	.85	.71	.82
9th grade	220	.93	.87	.88
10th Grade	40	.95	.91	.92
11th Grade	112	.94	.90	.90
12th Grade	306	.89	.82	.88
Undergraduate				
Freshman	826	.94	.89	.91
Sophomore	252	.92	.86	.92
Junior	476	.83	.74	.80
Senior	817	.90	.86	.89
Graduate Students				
Masters/Beginning Doctorate	529	.71	.60	.63
Advanced Doctoral	238	.84	.80	.75

1. Based on all studies with available Round 1 data and calculated on a per dilemma basis. Numbers slightly higher reported in Table 5 since individuals were included who had one or more missing dilemmas for some items.

Table 8

Internal Consistency Estimates Based on Raters of the RJI for All Available Studies[1]

Study	N	Based on Composite			Based on Dilemma		
		Intra-Class Correlation	95% Conf. Interval Lower	Higher	Intra-Class Correlation	95% Conf. Interval Lower	Higher
Studies Using All Four Dilemmas							
Brabeck (1983)	119	.64	.57	.73	.50	.43	.56
Time 2	25	.48	.19	.71	.31	.14	.46
Time 3[2]	22	.57	.20	.80	.38	.18	.54
Dove (1990)[3]	44	.81	.62	.90	.60	.50	.69
King, Wood & Mines (1990)	91	.83	.75	.89	.65	.58	.70
Kitchener & King (1981)	80	.97	.81	.99	.90	.87	.92
Time 2[4]	58	.93	.79	.97	.81	.76	.85
Time 3[5]	54	.80	.64	.88	.62	.53	.70
Time 4[6]	53	.89	.71	.95	.38	.18	.54
McKinney (1985)	56	.75	.60	.85	.65	.57	.72
Time 2[7]	21	.12	.05	.38	.11	-.01	.24
Van Tine (1990)	42	.22	.04	.43	.24	.13	.35
Studies Not Using All Four Dilemmas							
DeBord (1993)[8]	42	.60	.36	.77	.56	.39	.69
Glatfelter (1982)[8]	80	.61	.45	.73	.27	.15	.39
Kitchener et al. (1993)[9]	112	.73	.62	.81	.69	.71	.75
Time 2 Control[9]	53	.82	.66	.90	.78	.68	.85
Time 2 Experimental[9]	104	.84	.76	.89	.79	.73	.84
Kitchener & Wood (1987)[10]	47	.61	.38	.76	.59	.41	.73
King, Taylor & Ottinger (1989)[11]	146	.32	.16	.45	.31	.22	.39
Strange & King (1981)[12]	21	.65	.29	.84	.59	.41	.73

1. Coefficient alpha based on Resolved Scores (numbers in parentheses indicate Round 1 alphas), Agreement measures based on Round 1 data
2. Published in Brabeck & Wood (1990).
3. Data from Kitchener et al. (1993) were merged with Dove (1990) for form complete records. Experimental and Control conditions combined since Time 1 assessment is a pretest.
4. Published in King, Kitchener, Davison, Parker & Wood (1983).
5. Published in Kitchener, King, Wood & Davison (1989).
6. Published in Kitchener & King (1990).
7. Conducted by and reported in Van Tine (1990)
8. Based on two randomly selected dilemmas
9. Based on subjects not included in Dove (1990) having only Additives and Pyramids dilemma. Experimental and control groups combined since Time 1 assessment is a pretest.
10. Based on Pyramids, News, and Additives Dilemmas
11. Based on News, Evolution, and Additives dilemmas
12. Based on Pyramids, News, and Evolution dilemmas

Table 9
Internal Consistency Estimates Based on Raters of the RJI by Education Level[1]

Educational Level	N	Based on Composite			Based on Dilemma		
		Intra-Class Correlation	95% Conf. Interval Lower	Higher	Intra-Class Correlation	95% Conf. Interval Lower	Higher
High School							
9th Grade	57	.61	.42	.75	.59	.50	.67
10th Grade	15	.71	.35	.89	.31	.01	.57
11th Grade	54	.60	.39	.74	.49	.38	.59
12th Grade	83	.70	.57	.79	.61	.54	.68
Undergraduate							
Freshman	181	.56	.45	.65	.47	.39	.53
Sophomore	72	.75	.63	.83	.62	.54	.70
Junior	132	.79	.72	.85	.58	.52	.64
Senior	137	.70	.60	.78	.52	.44	.59
Graduate							
Masters/Early Ph.D.	144	.67	.57	.75	.52	.45	.59
Advanced Graduate/Ph.D.	49	.74	.59	.85	.70	.62	.77

1. Agreement measures based on Round 1 data Sample sizes do not total to all complete studies since McKinney Time 2 data were included as 11th graders, and eight 8th graders are not included in this table.

Table 10

Observed and Expected Response Pattern Frequencies for Combined Reflective Judgment Interview Data

Predominant Stage	Minor Stage 2	3	4	5	6	7
2	-[1]	21[2]	3	1	0	0
		23.21[3]	.73	.33	.43	.31
		21[4]	1.69	2.11	.13	.08
		7.09[5]	11.74	4.01	1.44	.71
3	102	-	360	4	0	0
	99.01		355.04	3.67	4.85	3.43
	102		360	3.65	.22	.13
	28.24		286.99	98.07	35.32	17.39
4	3	454	-	269	11	5
	3.95	450.39		273.38	8.36	5.92
	1.91	454		269	10.66	6.43
	59.42	365.29		206.37	74.32	36.59
5	0	3	106	-	72	5
	.74	1.95	114.54		66.67	1.11
	1.00	3.65	106		72	3.36
	9.70	59.62	98.58		12.13	5.97
6	0	1	4	57	-	54
	.82	2.17	2.94	56.84		53.23
	.06	.23	4.71	57		54
	5.42	33.32	55.09	18.83		3.34
7	0	0	2	4	43	-
	.49	1.29	1.75	.78	44.69	
	.03	.13	2.60	3.24	43	
	2.22	13.67	22.60	7.72	2.78	

1. Dashes indicate response patterns (cells) which cannot occur.
2. The First number in each cell represents the observed frequency.
3. Second numbers indicate predicted cell frequencies under Davison's sequentiality model.
4. Third numbers indicate predicted cell frequencies under differential sequentiality model.
5. Fourth numbers indicate predicted cell frequencies under quasi-independence model.

Table 11

Test/Retest Correlations and Growth in Reflective Judgment[1]

Education Level	N	Time Interval	Test/Retest Correlation	Difference	Mean Initial Score
Junior High/Early High School (8th-10th Grades)					
Kitchener et al. (1993)					
Control	11	2 wks.	.59†	-.19*	3.45
Experimental.	20	2 wks.	.63**	.11	3.38
McKinney (1985)[2]	21	2 yrs.	.43†(.66)[3]	.50**	2.76
Late High School (Grades 11&12)					
Kitchener et al. (1993)					
Control	7	2 wks.	.37	-.00	3.67
Experimental	15	2 wks.	.87**	.09	3.77
Brabeck (1983)					
Times 1 & 2	25	1 yr.	.52**	.15†	3.40
Times 1 & 3	22	2 yrs.	.44*(.66)	.02	3.44
Kitchener & King (1981)					
Times 1 & 2	17	2 yrs.	.85**(.92)	.82**	2.78
Times 1 & 3	15	6 yrs.	.54*(.90)	2.09**	2.84
Times 1 & 4	13	10 yrs.	.59*(.95)	2.35**	2.82
Early Undergraduate (Freshman & Sophomore)					
Kitchener et al. (1993)					
Control	6	2 wks.	.61	.25	3.85
Experimental	14	2 wks.	.26	.06	4.23
Polkosnik & Winston (1989)					
Times 1 & 2	8	3 mo.	.88**(.60)	.08	3.29
Times 1 & 3	7	6 mo.	.46(.21)	.11	3.39
Welfel & Davison (1986)					
Times 1 & 2	25	4 yrs.	.20(.67)	.54	3.64
Late Undergraduate (Junior & Senior)					
Kitchener et al. (1993)					
Control	13	2 wks.	.94**	-.06	4.45
Experimental	19	2 wks.	.45†	.10	4.38
Polkosnik & Winston (1989)					
Times 1 & 2	19	3 mo.	.86*(.55)	.07	3.37
Times 1 & 3	8	6 mo.	.92**(.85)	.30**	3.45
Kitchener & King (1981)					
Times 1 & 2	27	2 yrs.	.77**(.88)	.42**	3.74
Times 1 & 3	27	6 yrs.	.51**(.89)	1.14**	3.76
Times 1 & 4	26	10 yrs.	.54**(.94)	1.03**	3.87

1. Based on Resolved Scores. Time intervals involving other than Time 1 as an initial score not included since many subjects changed educational level from one time to another.
   †=p<.1, *=p<.05, **=p<.01
2. Time 2 Conducted by and reported in Van Tine (1990).
3. Numbers in parentheses indicate annualized test/retest correlations

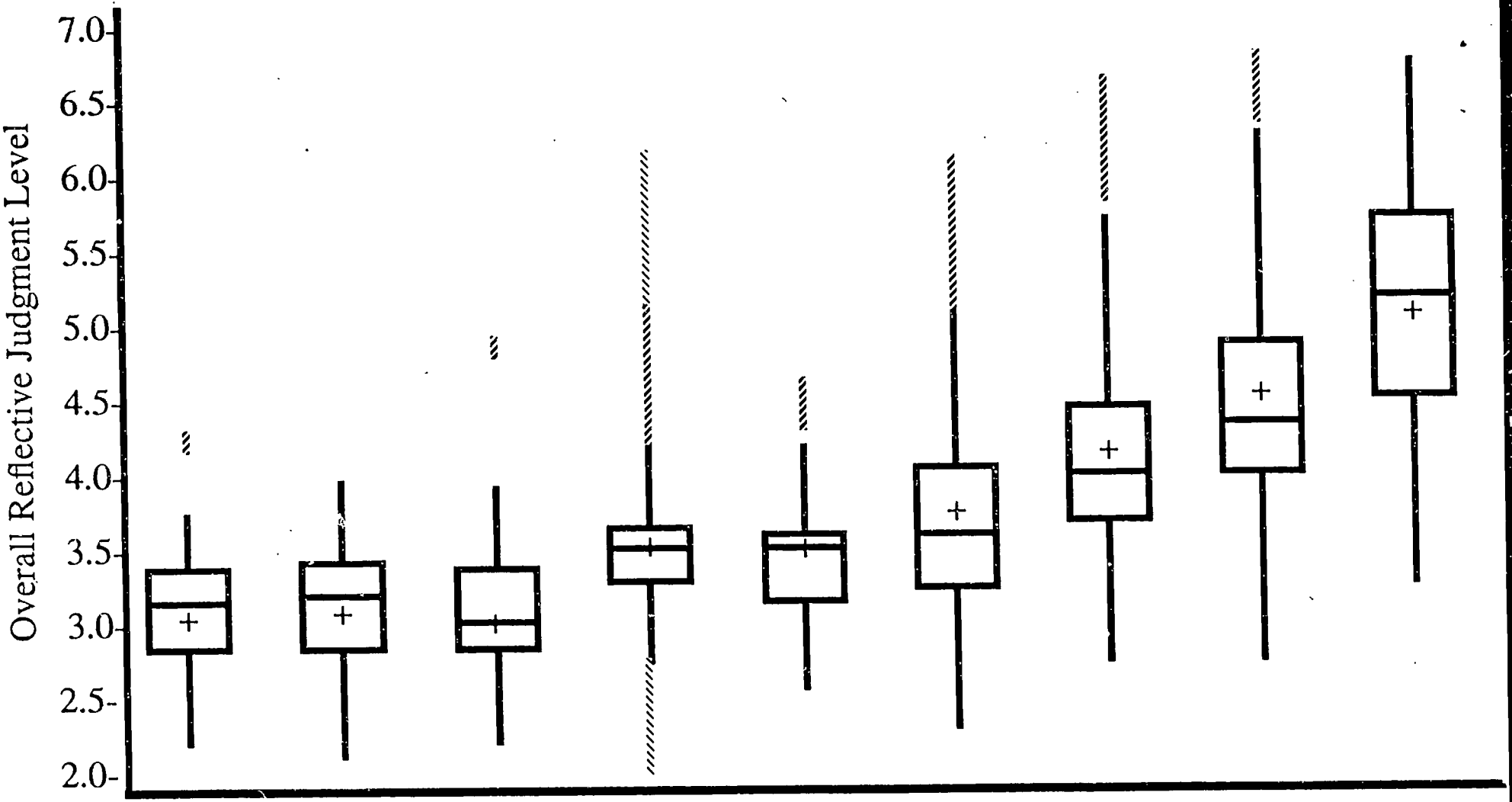
(Table Continues)

Table 11 (cont.)

Test/Retest Correlations and Growth in Reflective Judgment

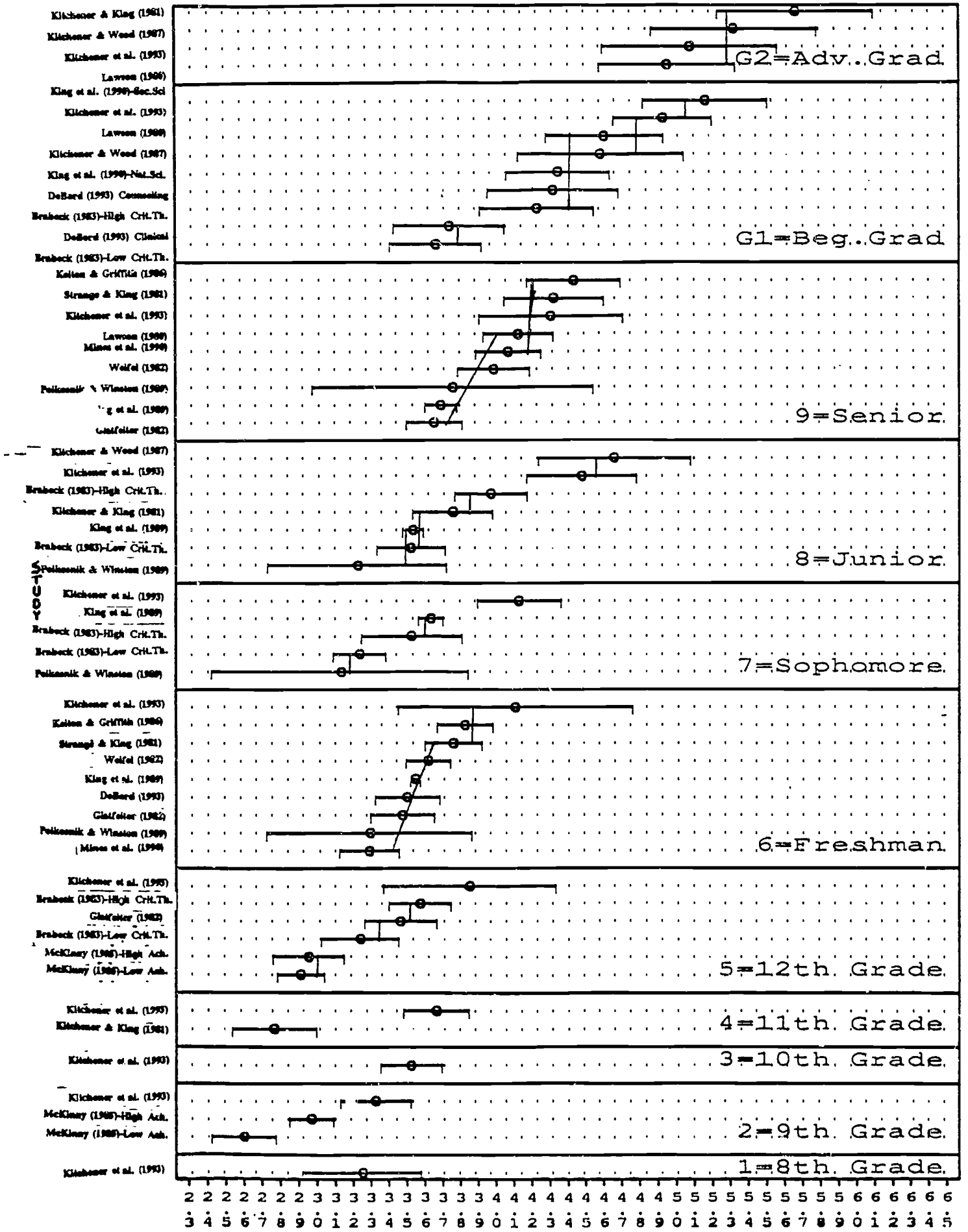
Education Level	N	Time Interval	Test/Retest Correlation	Difference	Mean Initial Score
Early Graduate					
Kitchener et al. (1993)					
Control	13	2 wks.	.68**	-.02	4.95
Experimental	24	2 wks.	.65**	.59**	4.91
Advanced Graduate					
Kitchener et al. (1993)					
Control	3	2 wks.	1.00**	.33*	4.42
Experimental	11	2 wks.	.69*	.32	5.26
Kitchener & King (1981)					
Times 1 & 2	15	2 yrs.	.81**(.90)	.19†	6.03
Times 1 & 3	13	6 yrs.	.87**(.98)	.22†	6.00
Times 1 & 4	14	10 yrs.	.83**(.98)	.09	5.93

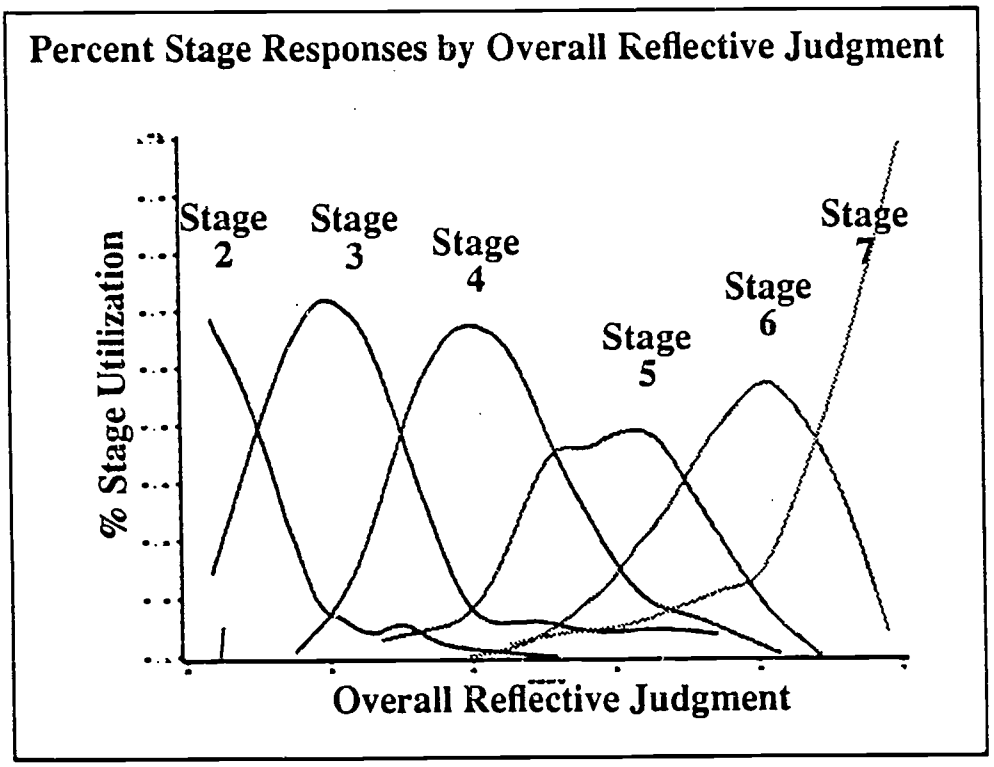
Figure Captions

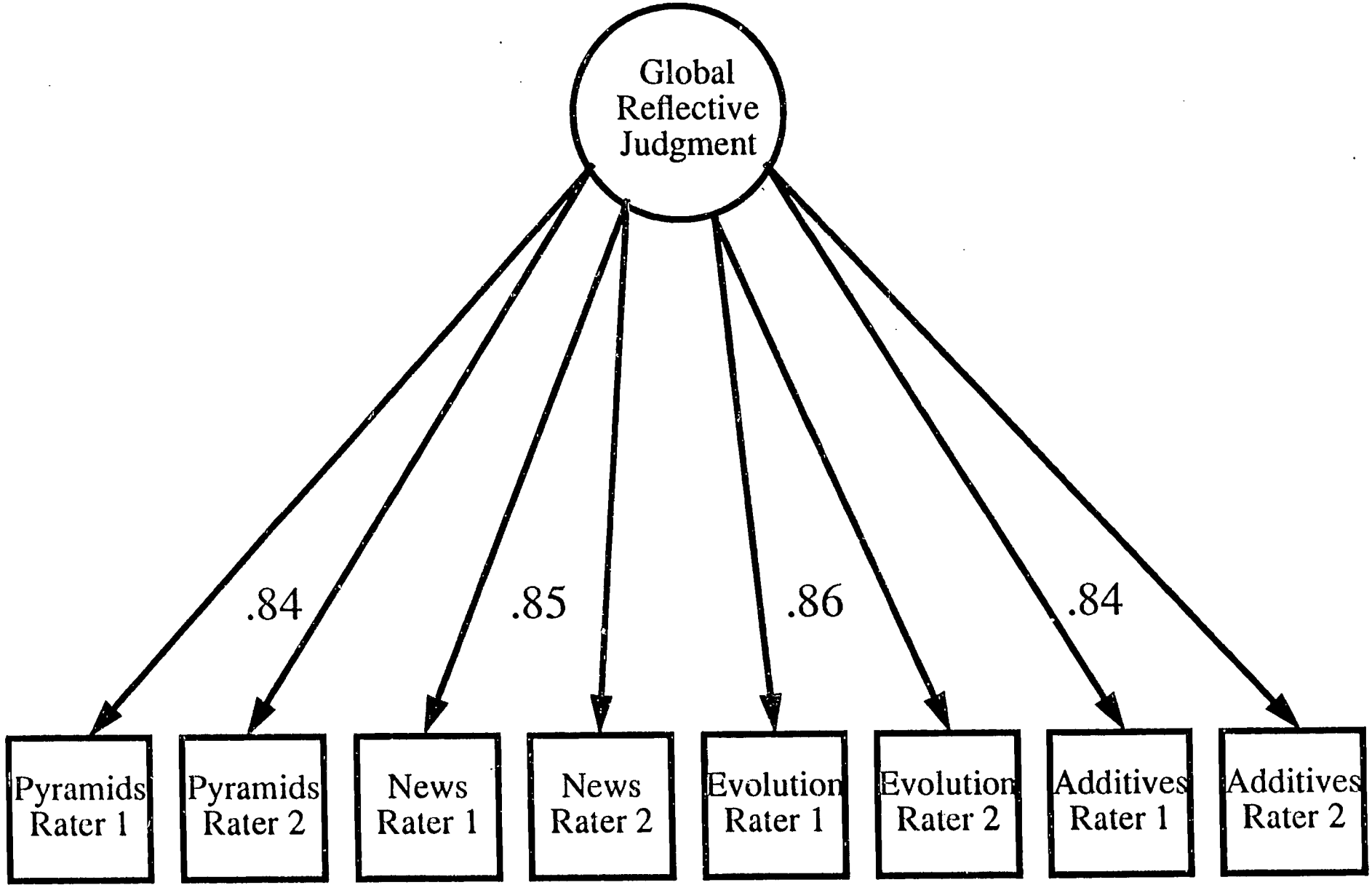


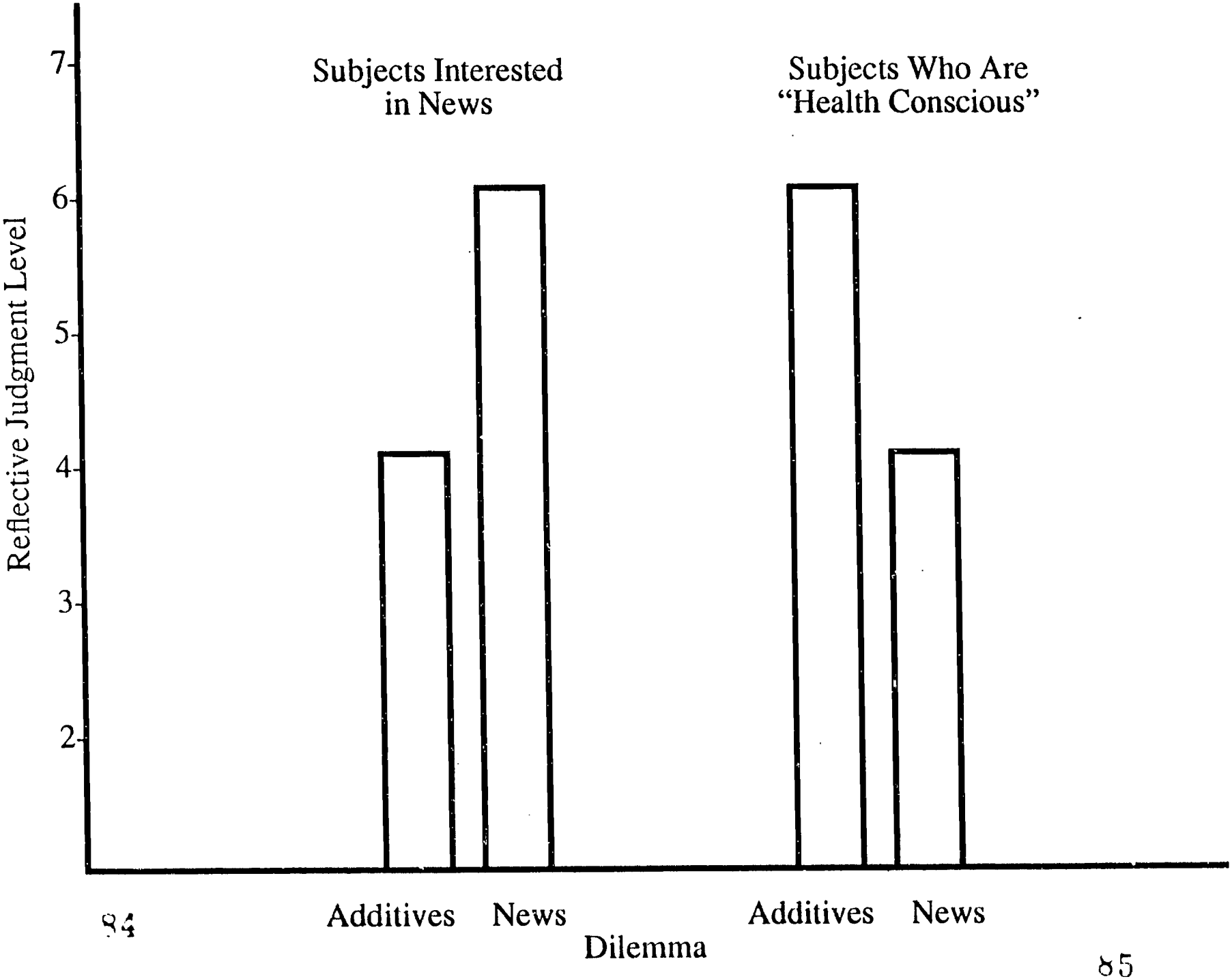
High School			Undergraduate				Graduate	
3.08-3.46[1]	3.12	3.27	3.63	3.57	3.74	4.00	4.62	5.27
3.25-3.52[2]	3.17	3.30	3.65	3.57	3.90	4.12	4.53	5.26

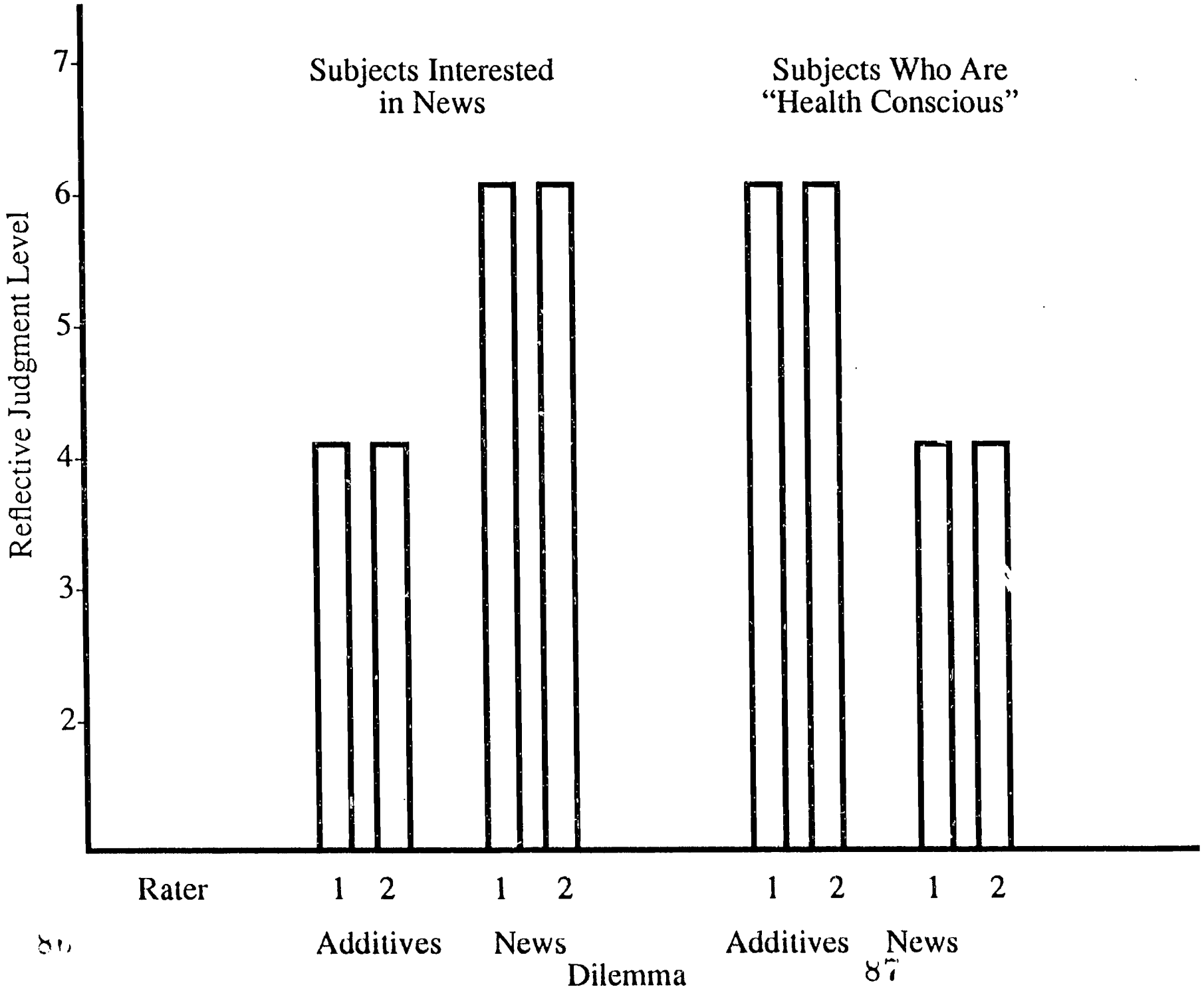
[1] Weighted means reported in King et al. (1994) from traditionally aged college students. 8th Grade means not reported.
[2] Means computed from secondary analysis based on traditionally and non-traditionally-aged students. Revised Scoring Scheme Scores.

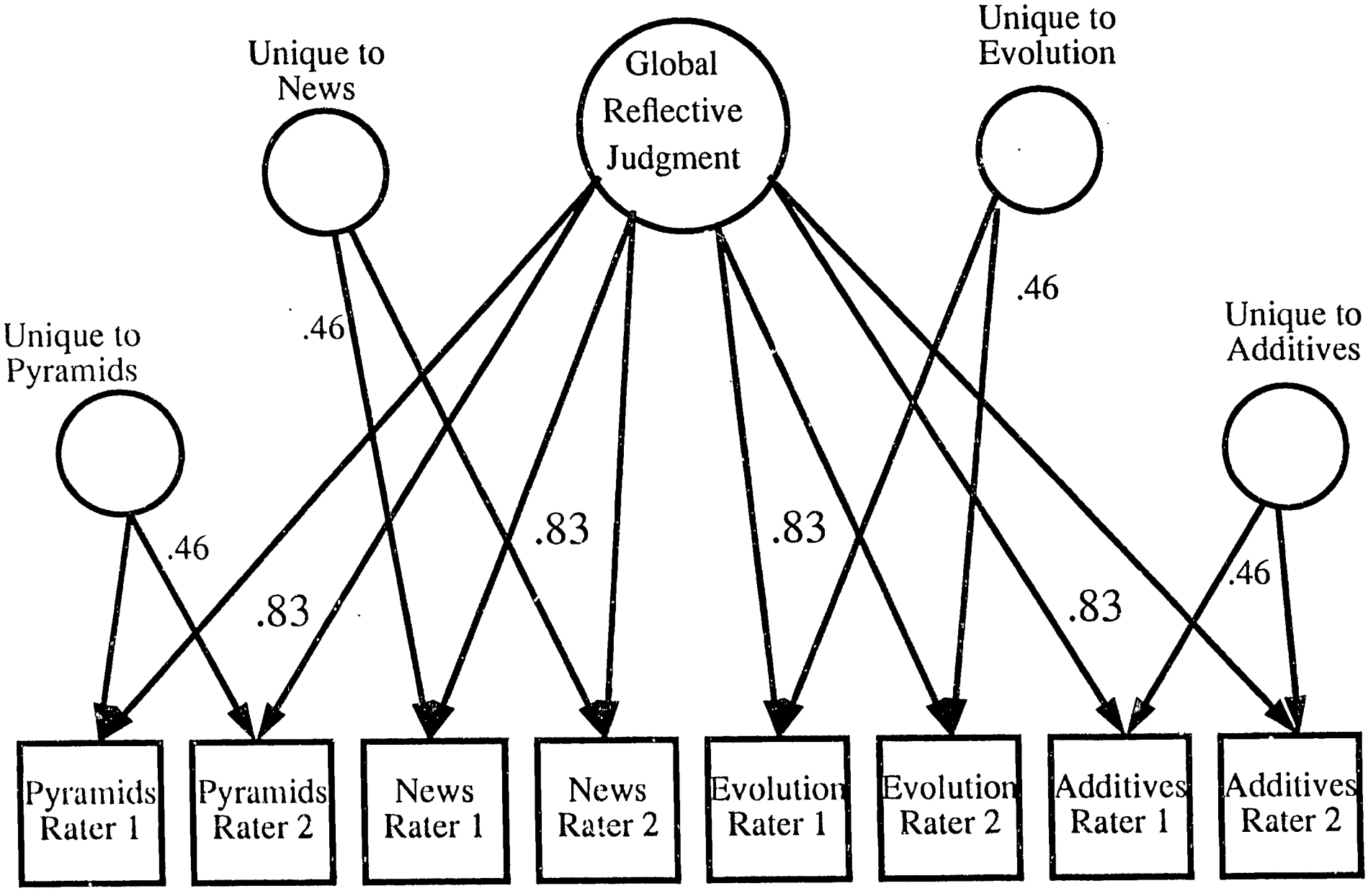



Percent Stage Responses by Overall Reflective Judgment
% Stage Utilization
Stage 2
Stage 3
Stage 4
Stage 5
Stage 6
Stage 7
Overall Reflective Judgment


Global
Reflective
Judgment
.84
.85
.86
.84
Pyramids
Rater 1
Pyramids
Rater 2
News
Rater 1
News
Rater 2
Evolution
Rater 1
Evolution
Rater 2
Additives
Rater 1
Additives
Rater 2


Subjects Interested in News
Subjects Who Are "Health Conscious"
Reflective Judgment Level
7
6
5
4
3
2
Additives
News
Additives
News
Dilemma




Unique to News
Global Reflective Judgment
Unique to Evolution
Unique to Pyramids
Unique to Additives
.46
.46
.46
.83
.83
.83
.46
Pyramids Rater 1
Pyramids Rater 2
News Rater 1
News Rater 2
Evolution Rater 1
Evolution Rater 2
Additives Rater 1
Additives Rater 2