Genetic, Social, and General Psychology Monographs, 113(4), 461–485

The Likelihood of Correlational Thinking in Adults: A Comparative Study and Methodological Critique

JUDITH A. McLAUGHLIN
Department of Psychology
Eastern Montana College

ROY D. PEA
Department of Psychology
Bank Street College

The Likelihood of Correlational Thinking in Adults: A Comparative Study and Methodological Critique

JUDITH A. McLAUGHLIN
Department of Psychology
Eastern Montana College

ROY D. PEA
Department of Psychology
Bank Street College

ABSTRACT. Previous investigators have argued that the Piagetian methodology for assessing correlational thinking is not statistically appropriate. Using alternative methodology, they claim to have demonstrated that most adolescents and adults do not think correlationally at the formal operational level. The purpose of the present study was to compare college students' correlational abilities under the two methodologies. Forty students were given the Piagetian task and the alternative task, the latter with contingency information presented under one of two conditions: (a) trial-by-trial, or (b) in summary form. Consistent with previous findings, on the alternative task more students used the statistically appropriate formula than trial-by-trial when presented information in summary form. However, all students demonstrated formal operational thought on both the Piagetian and the alternative tasks, and under both conditions of the latter. The adequacy of both methodologies for assessing adults' correlational abilities in everyday life is discussed.

IN EVERYDAY LIFE, we are frequently confronted with situations in which it is possible or even necessary to detect covariations between events. There are many events in everyday life that are highly correlated, but there are also many that occur together fortuitously. The ability to detect true covariations and to distinguish them from chance co-occurrences is of great importance in helping us to understand the relationships between the multitude of significant events in everyday life.

In research on the judgment of correlations between binary events, it is common to identify the event combinations in terms of the cells of a 2 × 2 frequency table:

	q	$-q$
p	a	b
$-p$	c	d

The letter a represents the frequency that p and q occur together. The letter b represents the frequency that p occurs without q, and c represents the frequency that q occurs without p. The letter d represents the frequency that neither p nor q occurs.

In studying the development of formal operational reasoning, Inhelder and Piaget (1958) suggested that adolescents and adults have rather impressive correlational abilities. To assess children's and adolescents' correlational reasoning, Inhelder and Piaget gave subjects decks of approximately 16 cards, each with the picture of a face with either blue or brown eyes and either blonde or brunette hair. When given one deck, the subjects were asked to determine whether there was a relationship between hair and eye color among the faces on the cards. When given two decks, they were asked to determine which deck had the stronger relationship between hair and eye color.

Inhelder and Piaget (1958) found that, before adolescence, children's ability to reason correlationally was poor. Children were unable to deal with the problem posed by the deck of cards and could only refer to the relationship between hair and eye color that they saw in real life. On the other hand, Inhelder and Piaget found that most adolescents were able to use formal operational reasoning for the task, but not all adolescents used all of the information available. Adolescents in Level 1 of the formal operational period, while able to reason hypothetically, considered only the frequency of the positive confirming cases (a). Adolescents in Level 2 considered the frequencies of both the positive and negative confirming cases, $(a + d)$, but ignored the nonconfirming cases. Only adolescents in Level 3 of the formal operational period considered all of the cases. They recognized both that the blue-eyed blondes and the brown-eyed brunettes $(a + d)$ confirmed the relationship and that the brown-eyed blondes and blue-eyed brunettes $(b + c)$ disconfirmed the relationship; Level 3 adolescents based their judgments on the difference in the proportion of these confirming and disconfirming cases.

An abridged version of this article was presented at the 53rd annual meeting of the Eastern Psychological Association, Baltimore, Maryland, in April 1982. The research was funded in part by a grant from Eastern Montana College.

Roy D. Pea is now at the Center for Children and Technology at Bank Street College.

Requests for reprints should be sent to Judith A. McLaughlin, Department of Psychology, Eastern Montana College, Billings, MT 59101.

Inhelder and Piaget used their study of the development of correlational thought in adolescence as a demonstration of formal operational logical reasoning. As evidence of formal reasoning, they took judgments that could be described by the logical formula:

$$r = \frac{(a + d) - (b + c)}{(a + b + c + d)}$$

This formula is a measure of the difference in the proportion of confirming cases and the proportion of disconfirming cases.

Several researchers have experimentally tested Inhelder and Piaget's theory of formal operational correlational reasoning and have provided some confirmation. Lovell (1961) found that among 26 secondary school students of high ability, over 75% demonstrated formal operational correlational reasoning, while nearly 40% of those were at Level 3. Adi, Karplus, Lawson, and Pulos (1978) and Green, Jurd and Seggie (1979) found similar results with secondary school students: The vast majority who were tested demonstrated some level of formal reasoning on correlation tasks, but only a minority demonstrated the highest level. Ross (1973) and Kuhn, Langer, Kohlberg and Haan (1977) found similar results when testing college students. In general, these results suggest that the majority of adolescents can solve Piagetian correlation tasks with formal reasoning, but most do not use the highest level of formal thought to do so. One study (Shaklee & Mims, 1981) had results that were somewhat discrepant from these. Of the 29 college students tested in this study, 72% were able to demonstrate the highest level of formal correlational reasoning.

In contrast to these studies based on Inhelder and Piaget's (1958) theory of correlational reasoning, other researchers have argued that the attainment of correlational thinking is not as universal as Inhelder and Piaget suggested. Shweder (1977, 1980) has argued that the concept of correlation is a relatively complex one that "is not spontaneously available to human thought" (1977, p. 638). He suggested that correlation is a nonintuitive concept that is not present in the thinking of most normal adults. Furthermore, Shweder argued that special circumstances must be contrived to elicit correlational thinking, such as having all of the data available for examination at one time. Shweder concluded that correlational reasoning is "generally absent from the thinking of most normal adults including social scientists" (1977, p. 641).

Shweder (1977) used a study by Ward and Jenkins (1965) to support his argument that adolescents and adults are disinclined to think correlationally. Ward and Jenkins gave college students 10 problems in which the students were asked to rate the degree of relationship between cloud seeding and the occurrence of rain on 50 days in certain states. One group of students was presented the information on a trial-by-trial basis, whereas a second group

was given the information in summary tables, and a third group received the information in both forms. Among the students given only trial-by-trial information, only 17% based judgments on both confirming and disconfirming cases, while 78% of those given only summary information did so. Ward and Jenkins concluded that "statistically naive subjects lack an abstract concept of contingency that is isomorphic with the statistical concept. Those who receive information on a trial by trial basis . . . generally fail to assess adequately the degree of relationship present" (p. 240).

On the basis of this research, Shweder (1977, 1980) argued that adolescents and adults are disinclined to think correlationally. He suggested that it is possible to elicit correlational thought under special circumstances, such as when summary information is given. However, when information is given trial-by-trial, which, Shweder argued, is how information is available in everyday life, a majority of adolescents and adults fail to use all of the information available, and thus fail to think correlationally. Shweder thus contended that the transition to formal operational thinking may not be as spontaneous as Piaget suggested and that such experimental research disconfirms Piaget's theory of formal operational development.

Other researchers have similarly argued that neither adolescents nor adults may have the correlational reasoning abilities proposed by Inhelder and Piaget. Nisbett and Ross (1980) suggested that "in the absence of theories, people's covariation detection capacities are extremely limited" in the perception of covariations in the social domain (p. 111). Shaklee (1979) concluded that research based on Piagetian theory may define the upper limits of human cognitive competence, but that in practice, people's correlational performance may be much more limited than suggested by Inhelder and Piaget.

Thus, while some of the research on correlational thinking has provided confirmation for Inhelder and Piaget's theory, other research has been taken as evidence against that theory. The researchers supporting the latter position have argued that the method of testing used by Inhelder and Piaget simplified the correlational task to enable subjects to demonstrate correlational competence, but that simplification made the task dissimilar to everyday situations in which people might use correlational thought. They argue that when the testing conditions are similar to everyday life, very few people are able to demonstrate correlational competence.

Furthermore, these researchers have argued that the formula by which Inhelder and Piaget defined the highest level of formal correlational reasoning is statistically inadequate. Ward and Jenkins (1965) pointed out that the Piagetian formula is appropriate only when the frequency of p equals the frequency of $-p$. If these two frequencies are not equal, the Piagetian formula may indicate a relationship between two variables when there is in fact no relationship. For example, in the following set,

	q	$-q$
p	8	2
$-p$	4	1

there is a positive relationship between p and q, according to the Piagetian formula, $r = (9-6)/15 = .20$. However, the conditional probability of q occurring, given p, $\text{pr}(q|p) = a/(a + b) = .80$, is equal to the conditional probability of q occurring, given $-p$, $\text{pr}(q|-p) = c/(c + d) = .80$. Thus, the occurrence of p is in fact independent of the occurrence of q. Ward and Jenkins suggested that this difference in conditional probabilities is a more appropriate formula statistically:

$$r = \frac{a/a + b - c}{c + d}$$

One assumption made by Shweder (1977, 1980) in arguing against the Piagetian position was that Inhelder and Piaget's correlation task provides an assessment of correlational competence comparable to that provided by Ward and Jenkins' (1965) summary condition task. In both tasks, Shweder argued, information is prepackaged for subjects, simplifying the correlational task. Shweder based his assumption that the two tasks measure the same abilities on the fact that subjects have been able to demonstrate formal reasoning on both tasks. However, there are important differences between the tasks. In the Piagetian task, subjects are asked to compare two sets of stimuli, to determine which has the greater relationship between two variables, while in Ward and Jenkins' summary condition task, subjects are asked to rate the relationship between two variables for a series of sets of stimuli. These two tasks have never been empirically compared to determine their equivalence. Therefore, one purpose of the present study was to compare subjects' performance on the Piagetian comparison task and on Ward and Jenkins' rating task, to determine if the two tasks do in fact provide comparable correlational reasoning assessments.

In examining the research on which Shweder's (1977, 1980) position is based, a second problem is apparent. Ward and Jenkins (1965) a priori specified seven formulae, representing seven types of correlational reasoning against which each subject's performance was evaluated. Rather than analyzing how the subjects actually solved the correlation problems, Ward and Jenkins instead assessed to which of the seven formulae each subject's judgments best conformed. A similar procedure was used by Shaklee and Mims (1981), with results that were discrepant from other studies of correlational reasoning. There are many possible ways of solving correlation problems, legitimate and not, and by a priori specifying only a limited set of possible formulae against which to test a subject's performance, that subject's correlational reasoning might well be misdiagnosed. Therefore, a second purpose of the present

study was to assess the accuracy of this method of determining correlational reasoning.

A final purpose of this study was to assess the accuracy of Shweder's (1977, 1980) contention that most adolescents and adults are disinclined to engage in correlational thought characterizable in terms of Piagetian formal operations. Shweder argued that very few people will use formal operations in solving correlation problems under conditions comparable to everyday life, even though those same people may be able to use formal operations on a Piagetian task. In this study, by testing subjects both on the Piagetian task and under trial-by-trial conditions, which according to Shweder are comparable to everyday life, it is possible to test the validity of that argument.

Method

Subjects

Forty undergraduate students at Elizabethtown College, Pennsylvania, served as subjects. Participation was solicited from students in psychology courses, and none of the volunteers had had formal statistical training in correlations. Each student was tested individually both on the Piagetian comparison task and on Ward and Jenkins' rating task. For the rating task, half the students received information in summary table (tables condition), while the remaining students were shown the information trial-by-trial (display condition). Within each condition, the order of presentation of the comparison and rating tasks was counter-balanced.

Materials and Procedure

Comparison task. Each student was first given a practice problem on the Piagetian comparison task. In the practice problem, the student was given a single deck of cards, on each of which was printed a picture of a face with either blue or brown eyes and either blonde or brunette hair. Each student was asked to determine whether there was a relationship between hair and eye color among all the faces in the deck. Some students initially had difficulty understanding the task. For those subjects, alternate forms of questioning were tried (e.g., "Does a certain hair color go with a certain eye color?", "Can you find the eye color by looking at the hair color?", "Do you have a good chance of knowing the eye color if you know the hair color?"), until the subject's judgment was based in some fashion on the frequencies of the four combinations of hair and eye color among the faces in the deck.

Six test problems were then administered. For each problem, the student was asked to compare two decks of face cards similar to those used in the practice problem and was asked to determine which deck of cards contained the stronger relationship between hair and eye color. The student was also asked to explain the basis of that judgment. The number of faces with each combination of hair and eye color in each deck is shown in Table 1.

Following the procedure used by Inhelder and Piaget (1958), the students were urged to sort the cards into classes if they did not spontaneously do so. Also, if a student appeared to consider only one of the combinations of hair and eye color, an attempt was made to elicit consideration of all combinations by asking, "But for all the cards, is there a relationship?" The students were never told what combinations to consider, but only to consider the cards as a whole group. For each problem, the student's judgment on which deck had the stronger relationship was noted, and the student's rationale for that judgment was recorded.

Rating task. The procedure used for the rating task followed as closely as possible that used originally by Ward and Jenkins (1965). The problems were presented to the students as being the results of experiments on the effects of cloud-seeding on rainfall in various states. The students were shown the number of days that it did or did not rain, given that cloud-seeding had or had not occurred.

TABLE 1
Frequencies of Faces With Each Combination of Hair and Eye Color in Comparison Task Problems

		Blonde hair		Brown hair	
Problem	Deck	blue eyes [a]	brown eyes [b]	blue eyes [c]	brown eyes [d]
1	a	2	1	2	4
	b	5	3	1	3
2	a	4	3	1	4
	b	4	3	1	2
3	a	4	1	2	1
	b	4	4	2	2
4	a	6	3	1	2
	b	5	2	4	7
5	a	5	2	2	3
	b	3	2	1	3
6	a	3	3	5	1
	b	2	1	3	6

A total of 10 problems was given to each student. The first two were practice problems. The number of days on which each possible combination of events occurred for each problem is shown in Table 2. These frequencies are those used originally by Ward and Jenkins (1965). Following the procedure used by Ward and Jenkins, the students were given printed instructions in test booklets, which informed the students that they would be shown the results of the cloud-seeding experiment for each of 10 states and would then be asked to rate the amount of control that seeding exerted over rainfall in each state. A separate answer page was provided for each problem, on which was printed the question, "How much control does seeding the clouds have over the occurrence of rain in State *n*?", and a scale, marked in units of 10 from 0 to 100, with 0 labeled *no control* and 100 labeled *complete control*.

Students tested in the display condition were shown the information for each problem one day at a time. In Ward and Jenkins' original study, this information was presented by machine. In the present study, the results for each day were printed on 3 × 5 in. cards. For each problem, the student was given a deck of cards representing all the days in that problem and was instructed to go through the deck one card at a time. Ward and Jenkins' subjects were allowed to view the results for each experimental day for 2 s. In the present study, that degree of control of presentation time was not possible. However, the students tested in the display condition were allowed to go through each deck only once and were asked to spend between 1 and 2 min examining each deck. These deck examinations were timed, and subjects taking less than 1 min or more than 2 min were urged to examine subsequent decks more slowly or quickly, respectively. Students tested in the display condition in the present study spent a mean of 88.1 s examining each deck, which did not differ significantly from the amount of time allowed by Ward and Jenkins (1965), $t(19) = 1.87, p > .05$.

TABLE 2
Frequencies of Days With Each Combination of Seeding and Rain Used in Ratings Task Problems

Problem	Seed		No seed	
	rain [a]	no rain [b]	rain [c]	no rain [d]
1	23	2	7	18
2	5	15	5	25
3	13	12	12	13
4	41	4	4	1
5	10	5	5	30
6	43	2	2	3
7	22	3	3	22
8	25	15	0	10
9	10	7	8	25
10	25	0	15	10

Students who were tested in the tables condition of the rating task in the present study were given test booklets similar to those used in the display condition, except that on each answer page was printed a table of the following form:

Clouds seeded	_ days
rain	_ days
no rain	_ days
Clouds not seeded	_ days
rain	_ days
no rain	_ days

The 10 problems used in the display condition shown in Table 2 were also used in the tables condition. The tables condition subjects were allowed to work through the problems at their own pace. All subjects in both conditions were not allowed to examine previous judgments when rating each problem.

Scoring

Comparison task. Each student's judgment and the rationale given for that judgment to each of the six test problems in the Piagetian comparison task were analyzed in order to categorize those judgments into one of the following four levels of correlational reasoning.

At Level 0, judgment was not based on information presented in the deck of cards, but rather on the relationship existing in the real world (e.g., "Blue eyes are supposed to go with blonde hair"), or the judgment was not probabilistic (e.g., "There's some of each combination, so you can't tell what the relationship is"). At Level 1, judgment was based solely on the frequencies of events in one class (e.g., "There are more blue-eyed blondes in this deck than that, so this deck has the stronger relationship"). At Level 2, judgment was based on the frequencies of events in two classes (e.g., "There are more blue-eyed blondes and brown-eyed brunettes in this deck, so it has the stronger relationship"). At Level 3, judgment was based on the relative frequency of confirming to disconfirming cases (e.g., "Fifty percent of this deck is blue-eyed blondes and brown-eyed brunettes, but only 30% of that deck is, so this deck has the stronger relationship").

After each student's six judgments and rationales were thus categorized, the pattern of levels to which those judgments were assigned was examined, and the student was categorized as being in the level of reasoning represented

by the majority of his or her judgments. All students were able to make probabilistic judgments based on the frequencies in the decks of cards, and thus no students were categorized as Level 0, indicative of concrete operational thinking.

Rating task. The procedure used by Ward and Jenkins (1965) for scoring the rating task was followed in the present study, the only difference being the number of formulae with which each subject's judgments on the eight test problems was compared. Ward and Jenkins generated seven formulae that they believed might be used in solving correlation problems. These formulae are identified in Table 3. In the present study, 39 additional formulae were identified. These, together with those used by Ward and Jenkins, can be divided into nine types, which are listed in Table 3. Note that, from a statistical viewpoint, some of these formulae are obviously more reasonable than others.

In scoring performance on the rating task, the eight test problems (problems 3 through 10) were first solved by each of the 46 formulae shown in Table 3. The ratings to the eight test problems given by each subject were then correlated, using Pearson's *r*, with the ratings generated by each of the formulae. Two formulae were identified for each subject from these correlations. As in Ward and Jenkins' (1965) study, the formula to which each subject's judgments were most highly correlated out of the seven formulae identified by Ward and Jenkins was designated as the formula used by that subject as a basis for judgment under the limited set of formulae criterion. Then, the formula to which each subject's judgments were most highly correlated out of all 46 formulae identified in Table 3 was designated as the formula used by that subject as a basis for judgment under the extended set of formulae criterion.

TABLE 3
Limited and Extended Sets of Formulae

Type	Formulae			
Frequency in one class	a*	b	c	d
Sum of two classes	$a+b$	$a+c$	$a+d$*	$b+c$
	$b+d$	$c+d$		
Sum of three classes	$a+b+c$	$a+b+d$	$a+c+d$	$b+c+d$
Difference in two classes	$a-b$*	$a-c$*	$a-d$	$b-a$
	$b-c$	$b-d$	$c-a$	$c-b$
Marginal values difference	$(a+b)-$	$(a+c)-$	$(b+d)-$	$(c+d)-$
	$(c+d)$	$(b+d)$	$(a+c)$	$(a+b)$
Ratio	a/b	a/c*	b/a	b/d
	c/a	c/d	d/b	d/c
Conditional probability	$a/(a+b)$*	$a/(a+c)$	$b/(b+d)$	$b/(b+d)$
	$c/(a+c)$	$c/(c+d)$	$d/(b+d)$	$d/(c+d)$
Ratio of confirming cases	$(a+d)/(a+b+c+d)$		$(b+c)/(a+b+c+d)$	
Difference in conditionals	$a/(a+b) - c/(c+d)$*		$a/(a+c) - b/(b+d)$	

Note. Formulae marked with an asterisk were used by Ward & Jenkins (1965) and constitute the limited set of formulae.

Under both criteria, a student was designated as using a formula only if the correlation between the student's ratings and the ratings generated by the formula equaled or exceeded .6664, the critical value of Pearson's *r*, with alpha = .05 and 7 degrees of freedom. Under the limited set criterion, 36 students gave ratings that correlated significantly at the .05 level with at least one formula. Of those, 30 were correlated significantly at the .01 level and 19 of those at the .001 level. Similarly, under the extended set criterion, the same 36 students gave ratings that correlated significantly at the .05 level with at least one formula. Of those, 32 were correlated significantly at the .01 level and 24 of those at the .001 level. Four students gave ratings that did not correlate significantly with any formula under either criterion.

Results

Comparison of Present Study to Ward and Jenkins'

Table 4 shows the percentage of students whose judgments in the rating task in the present study and in Ward and Jenkins' (1965) study showed the highest correlation to the three most frequently assigned formulae in the limited set (those formulae for which Ward and Jenkins reported percentages). In each condition of the rating task, the proportion of students assigned to each of these formulae in the present study was compared to the proportion reported by Ward and Jenkins, using the *z* test for a difference in proportions. The results of these analyses are also shown in Table 4. In both conditions, there were no significant differences between the two studies in the proportion of students assigned to each formula. Thus, the distribution of formulae assignments in the present study is not significantly different from that found by Ward and Jenkins.

Comparison of Limited and Extended Formulae Sets Criteria

Table 5 shows the number of students whose judgments in the rating task in the present study, under both conditions, showed the highest correlation to each formula in the limited set and extended set of formulas. Of the 40 stu-

TABLE 4
Percentage of Students Assigned to Most Frequent Formulae in Ward and Jenkins' and Present Studies

Condition	N	Formulae			
		$a+d$	$a/(a+c)$	$a/(a+c)-b/(b+d)$	others
Display condition:					
Present study	17	35.3%	35.3%	0.0%	29.4%
Ward & Jenkins'[a]	24	50.0%	21.8%	16.7%	12.5%
z*		−0.93	0.96	−1.76	1.35
Tables condition:					
Present study	19	15.8%	15.8%	57.9%	10.5%
Ward & Jenkins'[a]	23	13.1%	4.3%	78.3%	4.3%
z*		0.25	1.26	0.95	0.78

Note. [a]These data are from Ward & Jenkins, 1965, p. 237, and represent subjects tested in that study under the same conditions as used in the present study.

*$p > .05$ for all values of z shown.

dents tested, 30% were assigned to different formulae under the two criteria. Thus, for these students, there was a formula in the extended set to which their judgments correlated more highly than to the formula to which they would have been assigned if just the limited set of formulae were used. Using the highest correlation between judgments and formula ratings as the basis of assignment, the formulae to which these students would be assigned if just the limited set were used would, in effect, be a misdiagnosis.

Using the extended set of formulae resulted in an additional improvement. The extended set accounted for a larger percentage of the variance in students' judgments than did the limited set. Under the limited set of formulae criterion, the formulae to which students were assigned accounted for a mean 81.7% of the variance in the students' judgments. Under the extended set of formulae criterion, a mean of 84.1% of the variance in the students' judgments was accounted for. The difference in the mean percentage of variance accounted for under the two criteria was significant, $t(35) = 3.20, p < .01$.

In summary, it does appear that using only a limited set of formulae with which to evaluate correlational judgments carries the possibility of misdiagnosing the basis of those judgments. Furthermore, use of only the limited set to evaluate students' correlational judgments accounts for less of the variance in those judgments than use of the more extended set makes possible.

Comparison of Conditions in Rating Task

The percentage of students assigned to each type of formulae in the extended set under each condition of the rating task was compared. The number of students tested in each condition assigned to each type of formula is shown in Table 5.

Among the students tested in the display condition, who were given information trial-by-trial, 41.2% appear to have based judgments on either a single cell (a or b) or on the difference in two cells ($a - b$ or $a - c$). Among students tested in the tables condition, who were given summary information, only 10.5% were assigned to those formulae. The difference in these proportions is significant, $z = 2.12, p < .05$.

There was no significant difference in the proportion of students in the two conditions who appear to have based their judgments on a conditional probability: ($a/(a + b)$ or $a/(a + c)$), $z = -0.28, p > .05$. Among display condition students, 23.5% were assigned to formulae based on a single conditional probability, while 26.3% of the tables condition students were so assigned.

TABLE 5
Number of Students Assigned to Each Formula, Under the Limited and Extended Formula Sets Criteria

Formula	Limited set		Extended set	
	Display	Table	Display	Table
Frequency in one class:				
a	3	1	2	1
$-b$*	–	–	2	0
Difference in two classes:				
$a-c$	2	1	2	0
$d-c$*	–	–	1	1
Sum of confirming cases:				
$a+d$	6	3	6	2
Conditional probability:				
$a/(a+b)$	6	3	6	2
$a/(a+c)$*	–	–	0	2
Difference in conditional probabilities:				
$a/(a+b)-c/(c+d)$	0	11	0	4
$a/(a+c)-b/(b+d)$	–	–	0	6
Unknown:	3	1	3	1

*Not included in limited set of formulae.

There was a greater proportion of display condition students who were assigned to the sum of confirming cases formula, 35.3%, than the proportion of tables condition students assigned to that formula, 10.6%. This difference, however, did not reach significance, $z = 1.78, p > .05$.

The final difference between the two conditions was that there were significantly more tables condition students than display condition students assigned to formulae based on the difference in two conditional probabilities, $z = 3.52, p < .05$. In the tables condition, 52.6% of the students were assigned to formulae based on a difference in conditional probabilities. However, in the display condition, no students were assigned to those formulae. In general, then, students given trial-by-trial information were more likely to use the less statistically appropriate formulae, while students given summary information were more likely to base judgments on a difference in conditional probabilities.

Correspondence Between Students' Judgments and Formulae Ratings

In the rating task, formulae were assigned on the basis of the correlation between the students' judgments and the ratings generated from the formulae. While students may have given judgments whose rank orders corresponded to that of the formula ratings, so that the correlation between the two was high, there may have been considerable discrepancy between the actual values of the students' judgments and formula ratings. In order to assess the accuracy of the formulae as descriptors of the students' judgments, it was necessary to evaluate the differences, if any, between those judgments and the formulae ratings.

Table 6 shows the mean deviation of students' judgments on each test problem from the ratings generated by the formulae to which the students were assigned for the three most frequently assigned types of formulae. The number of students assigned to the other formulae were too few to analyze statistically.

For most of the problems, the students assigned to the sum of confirming cases formula gave judgments that were lower than the ratings generated by that formula. The mean deviation from the formula rating was not significant, however, for any of the problems. Of the 8 students assigned to the sum of confirming cases formula, 5 students gave judgments whose mean deviation from the formula ratings was less than 10 points. Considering that the students made their judgments on a scale marked in units of 10 points, their judgments showed a high degree of correspondence to the formula ratings. The remaining three students assigned to the sum of confirming cases formula had mean deviations that were within 20 points of the formula ratings.

The students assigned to formulae based on a single conditional probability gave judgments for all of the problems that were lower than the ratings

TABLE 6
Mean Deviations of Students' Judgments From Ratings of Assigned Formula in Ratings Task

Problem	3	4	5	6	7	8	9	10
				Sum of confirming cases[1]				
M	−15	2	−3	−1	−4	−13	−8	−8
SD	20	5	9	2	10	18	19	15
t	−2.08	1.06	−0.90	−1.32	−1.11	−2.05	−1.17	−1.50
				Conditional probability[2]				
M	−2	−6	−2	−8	−5	−7	−1	−6
SD	5	7	9	8	7	11	10	9
t	−1.36	−3.03*	−0.76	−3.20*	−2.27*	−2.13	−0.12	−2.25*
				Difference in conditionals[3]				
M	18	15	5	12	9	18	17	12
SD	24	24	24	13	5	17	23	19
t	2.42*	2.04	0.62	2.99*	5.30*	3.23*	2.31*	1.92

Note. [1]n = 8. The ratings for the sum of confirming cases formula were calculated as [100 × (a + d)/50], to convert those ratings to the same scale as the students' judgments. [2]n = 11. [3]n = 10.

*$p < .05$.

generated by the formula to which they were assigned. The mean deviations from those formulae ratings were significant for problems 4, 6, 7, and 10. Under the $a/(a + b)$ formula, those four problems were given the highest ratings, while under the $a/(a + c)$ formula, three of the four problems received the highest ratings. Thus, the students' judgments were significantly lower than the formula ratings for those problems on which the formula ratings were the highest. However, even though the students' judgments were lower than the formula ratings, the degree of deviation was not great. For 9 of the 11 students assigned to single conditional probability formulae, the mean deviation of judgments from the formula ratings was less than 10 points. The mean deviation of judgments from ratings for the remaining 2 students was less than 20 points.

In opposition, the students assigned to formulae based on the difference in two conditional probabilities gave judgments that were greater than the ratings generated by those formulae and showed more deviation from those formula ratings. Half of the 10 students assigned to these formulae did give judgments that deviated on average less than 10 points from the formula ratings. However, 4 students gave judgments that deviated more than 20 points

from the formula ratings. Thus, the correspondence between student judgments and the formula ratings was lowest for those students assigned to formulae based on the difference between conditional probabilities.

In summary, the correspondence between the students' judgments and the ratings generated by the formulae differed to some degree between the types of formulae. Students assigned to single conditional probability formulae or to the sum of confirming cases formula appeared to give judgments that were lower than the formula ratings, but generally by less than 10 points (or one scale unit). Students assigned to formulae based on the difference between conditional probabilities appeared to give judgments that were higher than the formula ratings and some were more deviant from those ratings.

Performance on Rating and Comparison Task

In order to compare performance on the rating task and the comparison task, each student was assigned to one of the Piagetian levels described in the Methods section on the basis of performance on the rating task, in addition to the assignment made on the basis of performance on the comparison task. Students using formulae based on the difference between conditional probabilities were categorized as Level 3, as those formulae involve the utilization of all confirming cases and the total number of cases. For the same reason, the use of a confirming cases/total cases[1] formula was also categorized as Level 3. Students using the sum of confirming cases formula were categorized as Level 2, as that formula is based on the frequency of positive and

[1]On the basis of the criteria originally used by Ward and Jenkins (1965), it is impossible to distinguish between students who based judgments on the sum of confirming cases formula (Level 2), and students who based judgments on a confirming cases/total cases formula (Level 3). All of the problems constructed by Ward and Jenkins had a total number of 50 cases. Therefore, the rank orders generated by the sum of confirming cases formula and the confirming cases/total cases formula are the same. In fact, the correlations that Ward and Jenkins report for the sum of confirming cases formula (Ward & Jenkins, 1965, p. 237) are actually the correlations that result from the confirming cases/total cases formula (i.e., they are ratios, rather than integers). In order to replicate Ward and Jenkins' study, we used the test problems exactly as constructed by them and subsequently established a further criterion to distinguish Level 2 and Level 3 performance on that task. If students' ratings consistently corresponded to the ratings generated by the confirming cases/total cases formula (i.e., on 7 or 8 of the 8 problems, the students' ratings were within .04 of the calculated ratings), they were assigned to Level 3; otherwise, to Level 2. Only 4 subjects (2 display and 2 tables condition) of the 9 students whose patterns of judgment correlated most highly to the sum of confirming cases formula were assigned to Level 3 on this basis.

negative confirming cases. Students using any of the remaining formulae were categorized as Level 1. Each of the remaining formulae is based on a comparison of positive confirming cases to one of the other classes, but disregards the negative confirming cases, and thus all are consonant with Level 1 correlational thought.

The number of students categorized in each of the levels on the basis of the comparison task and on the basis of the ratings task are shown in Table 7. The Wilcoxon t test was used to assess whether the students scored at a higher level on one of the two tasks. For students given summary information in the ratings task, there was no significant difference overall in assignment to levels between the ratings and comparison tasks (Wilcoxon $t(14) = 20, p > .05$). Thus, overall, students scored at a comparable stage on the comparison task and the tables condition of the ratings task. By contrast, students were significantly more likely to score at a higher level on the comparison task than in the display condition of the ratings task (Wilcoxon $t(14) = 20.5$, $p < .05$).

In analyzing response patterns of individuals, several further interesting cross-task differences emerged. For the display condition, there was a signif-

TABLE 7
Number of Students in Each Level of Formal Operations as Assigned on Basis of Ratings and Comparison Tasks

Ratings task	Comparison task			
	Level 1	Level 2	Level 3	Total
Display condition				
Level 1	3	6	4	13
Level 2	2	1	1	4
Level 3	1	1	1	3
Total	6	7	7	20
Tables condition				
Level 1	0	6	0	6
Level 2	0	1	0	1
Level 3	1	7	5	13
Total	1	14	5	20
Both conditions				
Level 1	3	12	4	19
Level 2	2	2	1	5
Level 3	2	8	6	16
Total	7	22	11	40

icant tendency for shifts from Level 1 up to Levels 2 or 3 to favor the comparison task (sign test, $n = 12, X = 3, p = .019$). Thus, students who characteristically considered just positive confirming cases (Level 1) in the trial-by-trial condition of the ratings task took into account more of the relevant information (Levels 2 and 3) for their judgments in the comparison task. Also, there was a significant trend for students revealing Level 3 thought on the ratings task tables condition to take into account less relevant information (Levels 1 and 2) for the comparison task (sign test, $n = 8$, $X = 0$, $p = .004$). These more revealing intra-individual analyses reveal a hierarchy of contexts likely to elicit more statistically appropriate thinking. The tables condition of the ratings task elicits the most advanced correlational thought, followed by the comparison task, and finally, the difficult display condition of the ratings task.

Discussion

One of the principal purposes of this study was to assess whether students are able to demonstrate formal operational reasoning on correlational tasks. This comparative study of two methods of assessing correlational thinking revealed that *all* college students tested performed at some level of Piagetian formal operations, both on the Piagetian comparison task and on Ward and Jenkins' ratings task, whether given summary information or when required to collect information trial-by-trial. Shweder's (1977, 1980) claim that young adults do not generally achieve correlational thinking characteristic of formal operations is thus refuted.

Shweder (1977) assumed that subjects' failure to use a formula based on the difference in conditional probabilities was evidence of a disinclination to use formal operational thought. That assumption, however, is not consonant with the Piagetian definition of formal operational thought. The formulae based on the difference in conditional probabilities are consistent with the logical strategies of Level 3, the most advanced level of formal reasoning. Subjects' failure to use those formulae can certainly be taken as evidence of a disinclination to use the most advanced level of formal thought, but cannot be taken as evidence of a disinclination to use *any* level of formal thought. All subjects in the present study, for all tasks, exhibited formal reasoning at one of the three levels, even when tested under the most difficult trial-by-trial condition. Therefore, Shweder was incorrect in contending that most adolescents and adults are disinclined to use formal operational thought.

A second purpose of the present study was to assess the accuracy of evaluating subjects' correlational reasoning by a priori stipulating a limited set of formulae to which that reasoning might correspond. Ward and Jenkins (1965) assumed that the formula to which a subject's judgments correlated most highly was the formula that best described the reasoning underlying that

subject's judgments. However, they stipulated only seven formulae against which to compare subjects' judgments. They thus assumed that one of those seven formulae would adequately describe each of their subjects' correlational reasoning.

In the present study, students' correlational judgments were compared both to Ward and Jenkins' limited set of formulae and to a more extended set of 46 formulae. Thirty percent of the students were assigned to different formulae under these two criteria, which places the validity of those formula assignments in question. The fact that, when the extended set of formulae were used, 30% of the subjects in the present study were assigned to formulae not in Ward and Jenkins' limited set demonstrates that the limited set of formulae was not adequate to describe the correlational reasoning of all subjects tested. Thus, there well may be more than a limited number of ways that people judge covariational relationships, and the use of only a limited set of formulae to evaluate those judgments may result in misdiagnoses.

The same problem, of course, may occur for any set of formulae specified a priori, no matter how large the set is. Even for the extended set of 46 formulae used in this study, there is no way to ascertain that the formula to which students were assigned actually described the students' correlational reasoning. It always remains possible that a subject may use some form of reasoning not represented by a formula in the specified set, but which correlates with the ratings of a formula in the set. In this case, the subject would be assigned to that formula, although it does not actually represent the subject's correlational reasoning. This problem remains inherent whenever the types of correlational reasoning possible are specified a priori.

The third purpose of the present study was to compare subjects' performance under the two methods of assessing correlational reasoning. In the present study, students used the least statistically adequate reasoning when information was presented trial-by-trial. Under those conditions, no students appear to have based judgments on the difference in conditional probabilities, which Shweder (1977, 1980) suggested was the most adequate form of correlational reasoning. Additionally, most students who used statistically inadequate reasoning when presented information trial-by-trial were able to use more adequate reasoning on the Piagetian comparison task. This demonstrates that higher levels of correlational reasoning were available to those subjects who did not access that ability when presented information trial-by-trial. Similarly, some students based judgments on the difference in conditional probabilities when given summary information on the ratings task, but utilized less information on the Piagetian comparison task. Thus, it is apparent that there are differences in the type of reasoning subjects use to solve the three tasks.

One obvious respect in which these three tasks differ is in working memory demands. Trial-by-trial information requires incremental frequency

counts of the four types of cases possible. In this situation, the patterns of response revealed a bias in favor of confirming cases as if they are remembered selectively. Although the reasons for such bias are poorly understood (Cohen, 1981; Wason, 1960), this study suggests that demands on working memory play a critical role as one factor inhibiting the accessibility of more consummate correlational thought. It is possible that subjects were unable to use the more statistically adequate formulae because they did not collect sufficient data in the trial-by-trial condition. A similar point applies to the Piagetian comparison task. The smaller number of events to process in that task may account, in part, for the better performances on it than on the trial-by-trial condition of the ratings task.

Shweder (1977, 1980) assumed that the Piagetian comparison task and the tables condition of the ratings task, in which subjects are given summary information, provide comparable assessments of correlational reasoning. He also assumed that the display condition of the ratings task, in which subjects are given information trial-by-trial, is more difficult than the other two. The evidence of the present study supports the latter assumption: Subjects were more likely to use higher-level reasoning on the Piagetian comparison task and on the tables condition of the rating task than on the display condition of the ratings task. However, the results of the present study do not support the first assumption: The majority of subjects did not use the same level of reasoning on the comparison task and on the tables condition of the ratings task. Thus, it is doubtful that these two methods provide equivalent assessments of subjects' correlational reasoning.

Shweder (1977, 1980) also argued that the trial-by-trial condition of the ratings task is more representative than the Piagetian comparison task of situations in everyday life requiring correlational thought. As few statistically naive subjects used statistically adequate reasoning when given information trial-by-trial, Shweder argued that people are disinclined to use correlational thinking in circumstances comparable to everyday life. While we do not intend here to argue that the Piagetian task is more representative of everyday life, we do suggest that the laboratory trial-by-trial task is as dissimilar to everyday life as other laboratory correlational tasks, including the Piagetian comparison task. Usually, in everyday life, event pairings do not impinge upon the subject at the rate of one every 2 s. Furthermore, events for which people seek correlational information in everyday life probably have more motivational significance than the relationship between hair and eye color or between seeding and rain on a deck of cards.

A critical assumption made by Shweder is that judgments that are not based on a statistically appropriate formula are inadequate. That assumption is certainly accurate from a statistical point of view. However, statistical theory is not the only standard for determining adequacy. We must ask—inadequate in terms of what behavioral goals? The relative adequacy of any

cognitive performance must be evaluated with respect to the ends or goals that specific performance was intended to achieve. The use of a statistically adequate formula, requiring the conscientious tabulation of four separate frequencies and adroit mental calculations, may be unnecessarily complicated for everyday use, despite its statistical adequacy.

Statistical theory was created by scientists and mathematicians to meet the demands of science for precision and objectivity. Galton first invented correlation coefficients only a century ago (Pearson, 1930), before which no one would have been surprised at the lack of correlational thinking in adults! The need for that degree of precision and objectivity in everyday life may not generally be so great that it justifies to the everyday mind the expenditure of time and effort that would be required to base correlational judgments on a statistically appropriate formula.

As further support for the need to consider the *adequacy* of forms of correlational thinking as relative to the goals of that thought, we must refer to the statistical *inadequacy* of the formula advocated by Shweder (1977) and Ward and Jenkins (1965). They argued that the formula used by Inhelder and Piaget (1958) was statistically inadequate as an indication of the highest level of correlational reasoning. When the marginal frequencies of p and $-p$ are unequal or when the marginal frequencies of q and $-q$ are unequal, the Piagetian formula may yield a positive correlation when there is in fact no relationship between the variables.

Ward and Jenkins (1965) suggested that a formula based on the difference between the conditional probabilities, pr($q|p$) and pr($q|-p$), yields statistically adequate judgments. However, ratings based on this formula may also be inaccurate, when the marginal frequencies are not equal. For example, Problems 8 and 10, used by Ward and Jenkins in their ratings task, yield different correlations, as calculated by the difference in conditional probabilities formula:

Problem 8			Problem 10		
	q	$-q$		q	$-q$
p	25	15	p	25	0
$-p$	0	10	$-p$	15	10

According to Ward and Jenkins' formula, the data in Problem 8 yields a correlation of .625, while the data in Problem 10 yields a correlation of .40. According to Pearson's formula for the correlation of dichotomous data, both of these sets of data yield a correlation of .50. Ward and Jenkins' formula yields the same value as Pearson's only when both sets of marginal values are equal or when the correlation is zero.

Pearson's formula generates the same correlation for these two sets of data because it corrects for marginal imbalance in the denominator:

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Note that for subjects to give statistically adequate correlational judgments using Pearson's formula, they would be required to compute the square root of the product of the four marginal frequencies.

Furthermore, both Ward and Jenkins' formula and Pearson's formula are statistically inadequate taken by themselves because they are insensitive to the size of the total number of cases considered. In using these formulae, the reasoner is left not knowing how to interpret their outcome without recourse to statistical tables whose interpretation depends upon the further understanding of distributions, hypothesis testing, and other theoretical constructs of statistical theory. Additionally, selecting the events upon which the calculations will be made will depend on such complex concepts as adequate sampling. Our point here is not that we expect people to demonstrate all of this knowledge in laboratory tasks, but rather that the use of either Ward and Jenkins' or Pearson's formula does not in itself make judgments statistically adequate.

A fundamental difference between Piaget's and Shweder's approaches to the study of correlational thinking lies in the distinction between competence and performance. Inhelder and Piaget (1958) purposefully constructed tasks that did not place extraordinary demands upon the subject's memory or computational skills. Their efforts were directed towards attempting to best manifest the subject's correlational competence. Piaget did not argue that, on the basis of such performance, all subjects demonstrating correlational thought on his tasks would use that ability in all circumstances in everyday life (Broughton, 1981). In fact, he suggested that whether one uses formal thought depends to a large degree upon the types of materials and problems with which one is confronted (Piaget, 1972).

Shweder (1977, 1980), on the other hand, was primarily concerned with how the subject's performance reveals gaps in competence. If a subject failed to use the statistically appropriate formula on a task, Shweder then assumed that the subject was disinclined to use correlational thought in everyday life. Shweder's arguments point to the difficulties and limitations of our inclination to engage in correlational thought. Piaget searched for optimal correlational competence; Shweder searched for limitations in correlational performance.

Piaget and Shweder are at opposite poles: Piaget described what we optimally *can do;* Shweder, what we *do not do*. Considerable research still needs to be conducted to fill in the gap between the two poles, to describe the extent and potential range of our correlational abilities. Rather than focus on limitations of correlational thinking performances, we would urge, as do Stone and Day (1980), the developmental study of the positive situational

characteristics that elicit higher levels of correlational thinking, with the aim of facilitating statistically appropriate correlational thought when it best achieves the goals of the thinker's activities.

REFERENCES

Adi, H., Karplus, R., Lawson, A., & Pulos, S. (1978). Intellectual development beyond elementary school, VI: Correlational reasoning. *School Science and Mathematics, 78,* 675–683.

Broughton, J. (1981). Piaget's structural developmental psychology II: Logic and psychology. *Human Development, 24,* 195–224.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences, 4,* 317–370.

Green, S., Jurd, M., & Seggie, I. (1979). Formal thinking about correlation. *Scandinavian Journal of Psychology, 20,* 119–125.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Kuhn, D., Langer, J., Kohlberg, L., & Haan, N. S. (1977). The development of formal operations in logical and moral judgment. *Genetic Psychology Monographs, 95,* 97–188.

Lovell, K. (1961). A follow-up study of Inhelder and Piaget's *The growth of logical thinking. British Journal of Psychology, 52,* 143–156.

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

Pearson, K. (1930). *The life, letters, and labours of Francis Galton.* (Vols. 1–3). Cambridge: Cambridge University Press. (Original work published 1914)

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15,* 1–12.

Ross, R. J. (1973). Some empirical parameters of formal thinking. *Journal of Youth and Adolescence, 2,* 167–177.

Shaklee, H. (1979). Bounded rationality and cognitive development: Upper limits on growth? *Cognitive Psychology, 11,* 327–345.

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development, 52,* 317–325.

Shweder, R. A. (1977). Likeness and likelihood in everyday thought: Magical thinking and everyday judgments about personality. *Current Anthropology, 18,* 637–658.

Shweder, R. A. (1980). Rethinking culture and personality theory, Part III: From genesis and typology to hermeneutics and dynamics. *Ethos, 8,* 60–94.

Stone, C. A., & Day, M. C. (1980). Competence and performance models and the characterization of formal operational skills. *Human Development, 23,* 323–353.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19,* 231–241.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140.

Received March 13, 1984