Form and Content in the Development of Deductive Reasoning

Willis F. Overton and Shawn L. Ward
Temple University
Jeffrey Black
Temple University

Ira A. Noveck
State University of New York at Binghamton
David P. O'Brien
New York University

A major issue in the field of deductive reasoning concerns the role of logic versus the role of semantic content in determining performance on reasoning problems. One theoretical approach (Overton, 1985; Staudenmayer, 1975) maintains that the development of a logical competence is a necessary prerequisite for solutions to deductive reasoning problems. According to this approach, specific content effects operate as moderators of logical competence to inhibit or enhance successful task performance. From this perspective, both formal reasoning processes and the real-time information-processing strategies related to content effects play essential roles in determining task performance. A second theoretical approach (Griggs, 1983; Mandler, 1983) asserts that formal reasoning processes play little or no role in task performance. According to this approach, task performance is virtually totally determined by context variables such as those provided by specific semantic contents.

The aim of the present research is to explore both the role of the development of formal logical reasoning and the role of semantic content in the solution of deductive reasoning problems. The development of formal logical reasoning is represented in Piaget's theory (Inhelder & Piaget, 1958) as a succession of increasingly powerful logical systems or competencies. Several recent studies (Byrnes & Overton, 1986; O'Brien & Overton, 1982; Overton, Byrnes, & O'Brien, 1985) have supported the prediction that a fully developed formal logical competence is not available before adolescence. These studies, which focused specifically on conditional inferences, were not, however, solely concerned with the acquisition of logical competence. In addition, the research explored several factors, including task interpretation, cognitive style, and semantic content, as moderators of the logical competence once that competence had been acquired. The present research is also framed by this competence-moderator-performance approach (Overton, 1985; Overton & Newman, 1982).

The competence-moderator-performance approach is a formulation, within a developmental perspective, of an integration of stage-structure (levels of competence) and information-processing (moderators) views of cognition. This approach maintains that performance on any cognitive task is a function of both level of logical competence and the individual's information-processing skills and capabilities. With respect to the present research, the main feature of this approach is that it distinguishes between logical knowledge necessary for task solution—which individuals may or may not have available—and informational processes (e.g., attention, encoding, representation, retrieval) that may facilitate or inhibit the accessibility of available logical knowledge.

Wason's (1966, 1983) selection task provides an appropriate domain for exploring issues concerning levels of logical reasoning competence and the effect of semantic content as the latter is mediated through informational processes. The standard, or abstract, form of this task presents subjects with an array of four cards showing, for example, respectively, the letter E, the letter K, the number 4, and the number 7. In addition, subjects are presented with a conditional rule: "If there is a vowel on one side of the card, then there is an even number on the other side." Subjects are told that each card has a letter on one side and a
number on the other and that the task is to select those cards and only those cards that would have to be turned over to decide with certainty whether the rule is true or false.

A fully adequate solution consists of the selection of the $E$ card and the $7$ card and only these cards. This solution recognizes that only a falsification strategy will provide logical certainty. The four showing cards may be described as the $E = p$, the $K = \neg p \ (not \ p)$, the $4 \ = q$, and the $7 \ = \neg q \ (not \ q)$. The conditional rule "If $p$ then $q$" is determined with certainty only under the condition in which it is falsified, that is, the $p$ and $\neg q$ condition. Only the $E$ card ($p$ and possible $\neg q$ on opposite side) and the $7$ card ($\neg q$ and possible $p$ on opposite side) could each satisfy this condition.

To date, there is little direct evidence concerning the developmental course of performance on the selection task (see, however, Bady, 1979; Moshman, 1979). The primary reason for this neglect has been the fact that, typically, when the standard abstract form is presented to adults, fewer than 10% give fully adequate solutions (Evans, 1982). Most of the research literature employing this task has been devoted to understanding the reasons for this poor performance. Several studies have demonstrated that the semantic content composing the conditional rule significantly affects level of solution.

An early phase of selection task research with adults suggested that the use of thematic or realistic content as opposed to abstract content led to substantial improvement. Later studies, however, failed to replicate the thematic effect or demonstrated only weak support for the effect (see Griggs, 1983, and Wason, 1983, for reviews). Recently, Griggs (1983) proposed that the assumed general thematic facilitation effect is, in fact, limited to the use of familiar content. Griggs investigated and supported this proposal in a series of studies (Cox & Griggs, 1982; Griggs, 1983).

Griggs (1983) interpreted the familiarity effect in accordance with a memory-cueing explanation holding that facilitation is a function of the subject’s being able to recall past experience with the problem content, the rule employed, and some counterexample of the rule. Further, Griggs (1983) was explicit in maintaining that this explanation "does not propose that 'logical' reasoning is facilitated. It only proposes that cued relevant experience leads the subject to make the correct selection" (p. 28). In a similar fashion Mandler (1983) proposed a schemata-cueing explanation and dealt with formal reasoning by stating that the "performance looks like a general understanding of deductive reasoning, but in fact it is only knowledge of what matters in a familiar setting" (p. 116).

From the perspective of the present research it is assumed that semantic facilitation operates on an underlying logical competence. It is therefore expected that both the nature of the logical competence and the nature of the semantic content affect task solution. Thus, the present research examines the relation between developmental and semantic effects in solutions to the Wason selection task.

In the preadolescent years children lack the type of formal logical competence that would permit complete and consistent task solutions. We assumed that although partial solutions on the Wason task may occur with some frequency at this level (e.g., correctly choosing the $p$ alternative), complete and consistent falsification solutions across problems will not be prevalent. During adolescence, formal propositional and predicate logical competence emerges, and this permits systematic deductive reasoning with logical relationships within a complete propositional system. Thus, given the conditional, reasoning at this level involves recognition that "$p$ and not $q$" is the inverse or negation of the conditional; "if $q$, then $p$" is the reciprocal; and "not $p$ and $q$" is the inverse of the reciprocal. This recognition is necessary both to distinguish the conditional from other logical forms (e.g., the biconditional, "if and only if $q$") and to systematically understand the conditions under which the conditional can be tested for its truth or falsity. We assume that the availability of this formal competence in adolescence will result in a prevalence of complete falsification solutions and consistent falsification solutions across problems.

In addition to the primary focus on developmental effects, the present research explores the generality of semantic facilitation by examining the effects of employing several familiar-content problems. Prior investigations by Griggs (1983) employed only a single familiar problem and thus the generality of his conclusions is necessarily limited.

Three experiments are presented that examined the performance of 4th, 6th, 8th, 10th, and 12th graders on the selection task using several semantic and abstract contents. We expected a general increase in reasoning performance with age and that familiar content would enhance performance. Further, we predicted that, consistent with the competence-moderator-performance position and earlier developmental research (O'Brien & Overton, 1982; Overton et al., 1985), performance with familiar-content problems would be poor prior to adolescence.

In Experiment 1, subjects in the 8th, 10th, and 12th grades were tested on three familiar-content and two abstract-content problems. In Experiment 2, subjects in Grades 4, 8, and 12 were tested on the same familiar problems of Experiment 1. In Experiment 3, subjects in the 4th, 6th, and 8th grades were tested on three familiar-content problems, three familiar-content problems with the conditional clauses reversed, one general thematic problem (meaningful but not familiar), and one abstract-content problem.

### Experiment 1

#### Method

**Subjects**

Forty-four students in each of three grades ($N = 132$) participated: 8th grades ($M = 13$ years, 3 months, $SD = 5.9$ months; 22 male and 22 female), 10th graders ($M = 15$ years, 3 months, $SD = 5.4$ months; 22 male and 22 female), and 12th graders ($M = 17$ years, 2 months, $SD = 5.5$ months; 24 male and 20 female). All of the students were enrolled in middle-class schools in suburban Philadelphia, Pennsylvania.

#### Tasks and Procedure

Five variations of Wason's (1966) selection task were used. Three problems used familiar content: (a) The drinking age problem was taken from Cox and Griggs's (1982) research and presented the conditional rule "If a person is drinking beer ($p$), then the person must be over 21 ($q$)"; (b) the motor vehicle problem was novel to the present research and stated the rule "If a person is driving a motor vehicle, then the person must be over 16"; and (c) the school problem was also novel to
this research and presented the rule “If a student has more than 10 absences in a school year, then the student must repeat the school year.” The remaining two problems used abstract content and were similar to those used by Wason (1966) and Johnson-Laird, Legrenzi, and Legrenzi (1972): (d) The vowel problem presented the rule “If a card has a vowel on one side, then it has an odd number on the other side”; and (e) the abstract problem presented the rule “If a card has a $D$ on one side, then it has a $J$ on the other side.”

For each problem there were four response alternatives. These response alternatives correspond to the affirmation and denial of $p$ and $q$. The correct solution to each problem consisted of the selection of only those two alternatives that could falsify the rule, that is, the $p$ alternative (yielding a $p$ and possible $-q$) and the $-q$ alternative (yielding a $-q$ and possible $p$).

Test booklets were constructed, each containing an instruction page and five problems (experimental group) or two problems (control group) presented on separate pages. The instruction page informed the subject that the booklet contained several problems and that each presented a rule that was supposed to be followed. The experimenter read aloud the instructions while the subjects read them. The subjects were told they would find with each problem four cards with information pertinent to four people or things described on each side of each card, along with a rule that applied to the four people or things. The subjects were instructed to select the card or cards that determined whether or not the rule was being broken.

Using the drinking age problem as an example, the format of each page was as follows: At the top of the page the four response alternatives—drinking beer ($p$), 22 years of age ($q$), drinking Coke ($-p$), and 16 years of age ($-q$)—were presented in rectangular boxes. Below these alternatives were the following instructions:

Each of the above cards has information about a person sitting at a table. On one side of a card is a person’s age and on the other side of the card is what the person is drinking. Here is a rule: IF A PERSON IS DRINKING BEER, THEN THE PERSON MUST BE OVER 21. Pick the card or cards that you would definitely need to turn over to decide whether or not the rule is being broken.

Space was provided below these instructions for the subject’s choices. A totally correct falsification solution required the selection of the $p$ card and the $-q$ card, not the selection of the $-p$ or $q$ cards.

The subjects were assigned randomly to either an experimental or control group. The experimental group was administered a booklet containing all five selection problems. The problem order was randomized across subjects within each content type, and the familiar set always preceded the abstract set. The order of response alternatives was randomized across problems. The control group received only the two abstract-content problems, and the order of presentation was randomized. Prior to receiving these problems subjects in the control group engaged in a paper-and-pencil spatial ability task to balance time spent with the experimenter. Testing was conducted in groups of four or fewer.

**Scoring**

A general solution score that gave partial credit for partial solutions was used as the basic dependent measure. For each problem a subject received one point for each of the following: (a) choosing $p$, (b) choosing $-q$, (c) not choosing $-p$, and (d) not choosing $q$. A second method of scoring consisted of the specific combinations of response alternatives selected by subjects on each problem. In this case a $p$, $-q$ selection pattern indicated a complete falsification solution.

**Results and Discussion**

A $3^{(grade)} \times 5^{(problem)} \times 4^{(response \ alternative)}$ analysis of variance (ANOVA) with repeated measures for problem and response alternative was computed on the general solution scores for the experimental group. There were significant main effects for grade, $F(2, 63) = 3.68, p < .05$; problems, $F(4, 252) = 66.40, p < .01$; and response alternative, $F(3, 189) = 12.87, p < .01$. There was also a significant Problem $\times$ Response Alternative interaction, $F(12, 756) = 8.03, p < .01$. No other interactions were significant.

The developmental effects will be examined first. The mean scores were .66, .73, and .74 for the 8th, 10th, and 12th grades, respectively. Newman-Keuls comparisons for the between-grade effects ($p < .05$) demonstrated that the 10th and 12th graders performed better than the 8th graders but did not differ from each other.

To assess the general developmental contribution more closely, we examined complete falsification solutions at each grade level for the experimental group. The complete falsification solution (select $p$, select $-q$, not select $-p$, not select $q$) is one of 16 possible selection combinations that can be given for each problem. Considering all five problems (i.e., familiar and abstract) 8th graders gave complete falsification solutions 29% of the time, 10th graders, 42%, and 12th graders, 46%. When only the three familiar problems are considered, 8th graders gave complete falsification solutions 48% of the time, 10th graders, 62%, and 12th graders, 68%. A test for the difference between two proportions revealed that only the 12th grade selected this solution more frequently than the 8th grade, $z = 2.33, p < .05$. The percentages for the familiar problems are presented graphically in Figure 1.

Confidence that an individual possesses a formal logical competence is increased to the extent that the individual is consistent in presenting logical solutions. As a consequence, a final developmental comparison was made of subjects who consistently gave complete falsification solutions across familiar problems. Because the subjects performed significantly more poorly on one of the problems (the school problem, discussed later) it was not included, and the consistency score for each subject was the complete falsification solution on each of the two remaining problems. Expressed as percentages of total number of subjects at each grade level these scores are 8th grade, 45%; 10th grade, 68%; and 12th grade, 86%. A test for the difference between two proportions indicated that the 12th graders were more consis-
tent than the 8th graders, \( z = 2.86, p < .05 \). The consistency scores are presented graphically in Figure 2.

The developmental findings are consistent with the perspective that, regardless of the specific nature of the propositional content, there is a systematic age-related increase in reasoning performance. It is also clear that familiar semantic content enhances performance. Given the ages of the subjects in this experiment, however, the question of whether a formal logical competence is available before adolescence awaits exploration in Experiments 2 and 3.

The second issue to be examined is the specific role of particular types of propositional content. Newman-Keuls comparisons (\( p < .05 \)) for the main effect of problem (see Table 1) revealed better performance for the drinking age and motor vehicle problems than for the school, vowel, and abstract problems. Performance on the school problem was significantly better than performance on the vowel and abstract problems. There was no difference in performance on the two abstract-content problems. The mean for each familiar problem is significantly greater than chance (50%, test for significance of a proportion). Correct choices were also significantly different from chance on the vowel problem but not on the abstract problem.

These findings support the prediction that familiar semantic content facilitates reasoning performance. With respect to the drinking age problem, the results directly support Griggs’s earlier findings. With respect to the motor vehicle problem, the results extend the generality of the familiar-content facilitation effect. However, the relatively poor performance on the familiar school problem introduces a note of caution. It appears that although familiarity may be a necessary condition, it is not a sufficient condition for facilitating performance (see also Wason, 1983).

Also explored were the patterns of selection combinations on each problem. For the familiar-content problems the complete falsification solution combination \( \neg p, \neg q \) pattern occurred significantly more frequently than chance (6.25%, test for significance of a proportion) for the drinking age problem (79%), for the motor vehicle problem (73%), and for the school problem (27%). A test for the difference between two proportions indicated that this combination occurred less frequently for the school problem than for the other two problems. The selection of all four alternatives was significantly greater than chance only for the school problem (19%). This again suggests the anomalous character of the school problem despite its familiar semantic content.

For abstract-content problems, the \( p, q \) combination and selection of all four alternatives occurred more frequently than expected by chance (6.25%) for both experimental and control groups. This finding is highly consistent with earlier research and has been interpreted as a failure to apply reasoning to the problem (Wason, 1983).

The next area for consideration concerns the main effect of response alternative and the Problem \( \times \) Response Alternative interaction found in the main ANOVA. Newman-Keuls comparisons of the four response alternatives by the experimental group showed better selection performance for the antecedent than consequent choices (see Table 1). The selection of the \( p \) alternative was correct more often than the selection of \( q \) and \( \neg q \) alternatives. The \( \neg p \) alternative differed only from the \( q \) alternative.

Table 1: Mean Correct Responses for Each Problem by Response Alternative

<table>
<thead>
<tr>
<th>Problem</th>
<th>( p )</th>
<th>( q )</th>
<th>( \neg p )</th>
<th>( \neg q )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1: Experimental group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drinking age</td>
<td>.88</td>
<td>.92</td>
<td>.95</td>
<td>.88</td>
<td>.91</td>
</tr>
<tr>
<td>Motor vehicle</td>
<td>.88</td>
<td>.91</td>
<td>.95</td>
<td>.88</td>
<td>.91</td>
</tr>
<tr>
<td>School</td>
<td>.70</td>
<td>.45</td>
<td>.74</td>
<td>.67</td>
<td>.64</td>
</tr>
<tr>
<td>Vowel</td>
<td>.82</td>
<td>.36</td>
<td>.61</td>
<td>.50</td>
<td>.57</td>
</tr>
<tr>
<td>Abstract</td>
<td>.86</td>
<td>.27</td>
<td>.65</td>
<td>.41</td>
<td>.55</td>
</tr>
<tr>
<td>Total</td>
<td>.83</td>
<td>.58</td>
<td>.78</td>
<td>.67</td>
<td>.72</td>
</tr>
<tr>
<td><strong>Experiment 1: Control group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vowel</td>
<td>.80</td>
<td>.36</td>
<td>.71</td>
<td>.26</td>
<td>.53</td>
</tr>
<tr>
<td>Abstract</td>
<td>.83</td>
<td>.30</td>
<td>.76</td>
<td>.27</td>
<td>.54</td>
</tr>
<tr>
<td>Total</td>
<td>.82</td>
<td>.33</td>
<td>.74</td>
<td>.26</td>
<td>.53</td>
</tr>
<tr>
<td><strong>Experiment 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halls</td>
<td>.92</td>
<td>.82</td>
<td>.88</td>
<td>.92</td>
<td>.88</td>
</tr>
<tr>
<td>Absent</td>
<td>.98</td>
<td>.62</td>
<td>.98</td>
<td>.88</td>
<td>.87</td>
</tr>
<tr>
<td>Radio</td>
<td>.88</td>
<td>.37</td>
<td>.87</td>
<td>.50</td>
<td>.65</td>
</tr>
<tr>
<td>Envelope</td>
<td>.82</td>
<td>.45</td>
<td>.63</td>
<td>.45</td>
<td>.59</td>
</tr>
<tr>
<td>Abstract</td>
<td>.60</td>
<td>.37</td>
<td>.50</td>
<td>.58</td>
<td>.51</td>
</tr>
<tr>
<td>Halls-reverse</td>
<td>.33</td>
<td>.25</td>
<td>.25</td>
<td>.23</td>
<td>.27</td>
</tr>
<tr>
<td>Absent-reverse</td>
<td>.40</td>
<td>.12</td>
<td>.28</td>
<td>.03</td>
<td>.21</td>
</tr>
<tr>
<td>Radio-reverse</td>
<td>.25</td>
<td>.28</td>
<td>.30</td>
<td>.22</td>
<td>.26</td>
</tr>
<tr>
<td>Total</td>
<td>.65</td>
<td>.41</td>
<td>.59</td>
<td>.48</td>
<td>.53</td>
</tr>
</tbody>
</table>

Note. Maximum score = 1.0.
the familiar semantic content has a transfer effect to abstract content. To answer this question, a 3(grade) \times 2(experimental/control group) \times 2(problem) \times 4(response alternative) ANOVA with repeated measures for problem and response alternative was computed on the general solution scores for abstract-content problems. There were main effects for grade, $F(2, 126) = 5.19, p < .01$, and response alternative, $F(3, 378) = 40.63, p < .01$, and a significant Problem \times Response Alternative interaction, $F(3, 378) = 4.92, p < .01$. The pattern of these effects did not differ from that given by the primary ANOVA. It should be noted in passing, however, that the significant grade effect adds support to the view that some process other than memory-cueing based on semantic familiarity is operating on performance on the task.

With respect to the issue of transfer there were no significant group effects, and hence the experimental experience with familiar content did not transfer to the abstract content. This absence of a transfer effect is quite consistent with the literature in this field (see Griggs, 1983; Wason, 1983). It light of this absence, Griggs (1983) suggested the presence of positive transfer as the criterion for the presence of logical reasoning. This suggestion, however, has major weaknesses. From a theoretical perspective, it can as readily be argued that transfer, or the lack thereof, between familiar and abstract content is irrelevant to the logical-reasoning/semantic-content issue. Consider, for example, the interpretation (Wason, 1983, p. 61) that the basic difficulty with the standard abstract task is that it presents a cognitive overload of eight items of information that cannot be readily integrated into a coherent whole. If this interpretation is correct, then familiar content may serve an integrative function allowing the subject to represent the problem as a coherent whole in working memory while proceeding to apply formal reasoning to the individual selection items. Given this interpretation, there is no reason to expect positive transfer from familiar to abstract content because the abstract material, regardless of prior experience, lacks coherence.

**Experiment 2**

The findings of Experiment 1 demonstrate that 8th graders perform relatively poorly but 10th and 12th graders generally perform at a high level on familiar-content problems. To further explore this developmental effect Experiment 2 tested subjects at Grades 4, 8, and 12 on the same familiar problems used in Experiment 1. In addition, Experiment 2 represented an attempt to explore the reason for the poor school-problem performance in Experiment 1. We noted that the response alternatives for this problem contained a $q$ choice involving a negative ("not promoted") and a $\neg q$ choice involving an affirmative ("promoted"). We thought that this inversion of negative and affirmative statements might have increased the cognitive load on this problem and led to poor performance. Consequently, in Experiment 2 we changed the $q$ alternative to an affirmative statement and the $\neg q$ alternative to a negative statement.

**Method**

**Subjects**

Forty students in each of three grades ($N = 120$) participated: 4th graders ($M = 9$ years, 9 months, $SD = 4.8$ months; 20 male and 20 female), 8th graders ($M = 14$ years, 2 months, $SD = 5.3$ months; 20 male and 20 female), and 12th graders ($M = 18$ years, 1 month, $SD = 5.4$ months; 20 male and female). All of the students were enrolled in middle-class schools in suburban Philadelphia, Pennsylvania.

**Tasks and Procedures**

The same three familiar-content problems used in Experiment 1 were used here. The modification made on the school problem consisted of employing the $q$ response alternative the statement "repeat the school year" and employing as the $\neg q$ response the statement "not repeat the school year." The tasks were administered and scored in the same manner described in Experiment 1 except that subjects were tested individually.

**Results and Discussion**

A 3(grade) \times 3(problem) \times 4(response alternative) ANOVA with repeated measures for problem and response alternative was computed on the general solution scores. There were significant main effects for grade, $F(2, 117) = 14.35, p < .001$; problems, $F(2, 234) = 37.50, p < .001$; and response alternative, $F(3, 351) = 18.85, p < .001$. There were significant Grade \times Problem, $F(4, 234) = 8.59, p < .001$, Problem \times Response Alternative, $F(6, 702) = 3.68, p < .01$, and Grade \times Response Alternative, $F(6, 351) = 3.23, p < .01$, interactions.

For the developmental effects, the Grade \times Problem interaction demonstrated significant improvement between the grades on both the drinking age, $F(2, 234) = 10.08, p < .01$, and motor vehicle, $F(2, 234) = 14.85, p < .01$, problems ($Ms = 4th, .66$; 8th, .82; 12th, .94) but no improvement on the school problem ($Ms = 4th, .64$; 8th, .60; 12th, .68). The Grade \times Response Alternative interaction revealed that both 4th graders and 8th graders made more errors with the $\neg q$ alternative than with other alternatives.

As in Experiment 1, complete falsification solutions, ($p, \neg q$) were assessed at each grade level. This solution was given by the 4th graders 14% of the time; 8th graders, 42%, and 12th graders, 61%. A test for the difference between two proportions indicated that the 8th and 12th graders selected this solution more frequently than the 4th graders ($z = 4.83$ and $z = 7.52, p < .05$, respectively). The 12th graders made this selection more frequently than the 8th graders, $z = 2.94, p < .05$. As can be seen in Figure 1, these results are developmentally consistent with the results obtained in Experiment 1.

Also, as in Experiment 1, a formal logical consistency score was computed for each subject (i.e., the complete falsification solution for both the drinking age and motor vehicle problems). Expressed as percentages of the total number of subjects at each grade level, these scores are as follows: 4th, 7%; 8th, 40%; 12th, 75%. A test for the difference between two proportions ($p < .05$) indicated that the 4th graders performed more poorly than both other grades and that 8th graders performed more poorly than 12th graders. Inspection of Figure 2 shows that these results are developmentally consistent with the results obtained in Experiment 1.

The developmental findings support and extend the conclusion of Experiment 1. It is clear that regardless of the specific nature of familiar propositional content, reasoning performance increases systematically with age. Further, there is no
evidence of adequate formal reasoning at the 4th-grade level, whereas the 8th grade appears to mark something of a transition to more adequate formal reasoning.

With respect to the specific content, the main effect of problem and the Grade × Problem and Problem × Response Alternative interactions all indicate that the school problem remained anomalous despite the modification introduced in this experiment. The means for the three problems were as follows: drinking age, 81%; motor vehicle, 79%; and school problem, 64%. Newman-Keuls comparisons (p < .05) for the main effect of problem indicated significantly poorer performance on the school problem than on either of the other two problems.

Experiment 3

The preceding experiments show that 10th- and 12th-grade subjects give adequate logical solutions to familiar-content problems but that there is a marked drop in this performance at Grade 8 and again at Grade 4. In Experiment 3 new content more specifically familiar to the younger subjects was introduced in an effort to examine whether this content would significantly promote adequate performance. To this end subjects in Grades 4, 6, and 8 were tested, and three new, age-appropriate, familiar-content problems were constructed and assessed in relation to a general thematic problem (i.e., meaningful but not familiar) and one abstract problem.

A final feature of Experiment 3 represents an effort to more fully investigate the relative contribution of reason and content. Here the three familiar-content problems were assessed with the content of their clauses reversed. This manipulation was also studied by Cox and Griggs (1982, Experiment 3), who found it produced a response pattern (−p, q) not previously reported in the literature. They argued that this pattern reflects a failure to reason analogically from the recall of familiar content. It would seem, however, that this pattern is more adequately interpreted as an attempt to transform the counterintuitive content into familiar content, along with the application of rules of formal logic for solution. To more fully consider this explanation a replication and larger data base are necessary.

Method

Subjects

Twenty students in each of three grades (N = 60) participated: 4th graders (M = 10 years, 4 months, SD = 6.0 months; 8 male and 12 female), 6th graders (M = 12 years, 0 months, SD = 5.8 months; 10 male and 10 female) and 8th graders (M = 14 years, 3 months, SD = 7.1 months; 11 male and 9 female). All of the subjects were enrolled in middle-class schools in upstate New York.

Tasks and Procedure

Eight variations of Wason's (1966) selection task were used. Three of the problems were novel to the present research and presented rules taken from the school discipline code that were familiar to the subjects: (a) the halls problem presented the rule "If a student is in the school halls (p), then the student is not running (q)"; (b) the absent problem, "If a student is absent, then the student brings a note from his or her parent"; (c) the radio problem. "If a student is in school, then the student is not playing a radio." Two problems similar to those used by Johnson-Laird et al. (1972) and Wason (1966) presented rules that were not familiar to the subjects. The first employed meaningful but not necessarily familiar content: (d) the envelope problem, "If an envelope is sealed, then it has at least a 20 cent stamp on it." The second was an abstract-content problem: (e) the abstract problem, "If a card has a letter B on one side of it, then it has the number 4 on the other side of it." The final three problems were identical to the three familiar problems (a, b, and c), but the antecedent and consequent clauses were reversed from their usual semantic positions (Cox & Griggs, 1982). These reversed-clause rules were (f) the halls-reverse problem, "If a student is not running, then the student is in the school halls"; (g) the absent-reverse problem, "If a student brings a note from a parent, then the student was absent"; and (h) the radio-reverse problem, "If a student is not playing a radio, then the student is in school."

Test booklets were constructed, each containing an instruction page and eight problems presented on separate pages. The instruction page informed the subject that the booklet contained several problems and that each problem presented a rule that was supposed to be obeyed and several examples to which the rule applied. For each example the subject was required to decide whether or not more information was needed in order to know whether or not the rule was being obeyed. Using the halls problem as an example, the format of each page was as follows. On the page was printed the following:

A school has a rule for its students. It is your job to make sure that each student is obeying the rule. The rule is: IF A STUDENT IS IN THE HALLS, THEN THE STUDENT IS NOT RUNNING. To find out whether or not the students listed below obey the rule or not, decide which students you have to find out more about.

Following this were printed four situations (response alternatives):

(a) The first student is in the school halls (p). Do you need to find out whether or not the student is running to know whether or not the student is obeying the rule? (b) The second student is not in the school halls (¬p). Do you need to find out whether or not the student is running to know whether or not the student is obeying the rule? (c) The third student is not running (¬q). Do you need to find out whether or not the student is in the school halls to know whether or not the student is obeying the rule? (d) The fourth student is running (q). Do you need to find out whether or not the student is in the school halls to know whether or not the student is obeying the rule?

As in Experiments 1 and 2 a totally correct falsification solution required selection of the p and the ¬q alternatives only.

The order of presentation of the eight problems was randomized across subjects, and the order of the four alternatives was randomized across problems. Subjects were tested individually. The experimenter read aloud the instructions while the subject read the instruction page. Each rule and its response alternatives were read aloud by the experimenter and by the subjects. Subjects were required to respond yes or no for each alternative.

Scoring

Responses were scored in the same manner described in Experiments 1 and 2.

Results and Discussion

A 3 (grade) × 8 (problem) × 4 (response alternative) ANOVA with repeated measures for problem and response alternative was computed on the general solution scores. There was a significant main effect for problem, F(7, 399) = 91.78, p < .001, and for response alternative, F(3, 171) = 20.99, p < .001, and a
significant Problem × Response Alternative interaction, $F(21, 1197) = 4.35, p < .001$. There was no main effect or interaction involving grade.

Although grade effects were not significant for the general solution scores, it is important for purposes of comparisons with Experiments 1 and 2 to first examine the developmental data. Complete falsification solutions were examined at each grade level. Considering first the five nonreversal problems, which is the appropriate comparison with the five problems of Experiment 1, 4th graders gave complete falsification solutions 24% of the time; 6th graders, 29%; and 8th graders 39%. To compare this performance with that in Experiment 1, it should first be noted that the 8th-grade subjects in the present experiment were 1 year older than the 8th graders of Experiment 1 (14 years 3 months vs. 13 years 3 months). Thus, if subjects are ordered according to age across the two experiments, the following monotonic increase in performance is observed: 4th, 24%, 6th, 29%; 8th (young), 29%; 8th (old), 39%; 10th, 42%; 12th, 46%.

When only the three familiar problems (halls, absent, and radio) are considered, the 4th graders gave complete falsification solutions 40% of the time; 6th graders, 47%; and 8th graders, 59% (see Figure 1). A test for the difference between two proportions revealed that the 8th graders selected this selection more frequently than the 4th graders, $z = 2.08, p < .05$. No other comparison was significant. Despite the differences on specific content across the three experiments, there is a consistent age-associated increase in performance. However, it is also clear that semantic content exerts an influence at all ages.

In Experiments 1 and 2 confidence that subjects possessed a formal logical competence was expressed in terms of the consistency of presenting logical solutions across familiar-content problems. In each of the earlier experiments the one anomalous familiar problem was dropped and the consistency score was based on complete falsification solutions to each of the two remaining familiar problems. This criterion was again employed in the present experiment when the radio problem (discussed later) was dropped. These logical consistency scores were as follows: 4th grade, 20%; 6th grade, 35%; 8th grade, 60% (see Figure 2). A test for the difference between two proportions indicated that only the 8th graders were more consistent than the 4th graders, $z = 2.58, p < .05$.

When the developmental comparisons are considered as a whole across the three experiments, it is evident that regardless of specific propositional content, reasoning performance increases systematically between the 4th and 12th grades. It is also evident that familiar semantic content enhances the developmental performance. However, the performances of the 4th and 6th graders suggest that formal logical reasoning is not generally present during these years.

The next issue concerns the role of various types of propositional content. Newman-Keuls comparisons ($p < .05$) for the main effect of problem (see Table 1) revealed better performance for the familiar halls and absent problems than for all other problems. The familiar radio problem, however, resulted in significantly poorer performance than either of the other familiar problems. The radio and envelope problems did not differ, and both yielded better performance than the abstract problem. The reverse-clause problems all yielded significantly poorer performance than any of the other problems.

Again, as in Experiments 1 and 2, these findings support the view that formal semantic content facilitates reasoning performance. They also show that general thematic material is somewhat facilitative. With respect to the halls and absent problems the results extend the generality of the familiarity effect. However, as in Experiments 1 and 2, one of the new familiar-content problems (radio) yielded relatively poor performance, and this again introduces a cautionary note with respect to the sufficiency of familiar content in accounting for adequate reasoning performance on the task. On the other hand, the theoretical importance of familiarity is enhanced by the poorer performance on the thematic (meaningful but not familiar) envelope problem.

As in Experiment 1, selection patterns were also examined to explore differences among the eight problems. For the familiar-content problems the complete falsification solution combination $p, \neg q$ was selected 67% of the time on the halls problem, 52% on the absent problem, but only 27% on the radio problem. The fact that the $p, q$ pattern was most frequently selected for both the radio and abstract problems again suggests the anomalous nature of the radio problem as a familiar-content problem.

The results of the reversed-clause problems present valuable information on the role of formal reasoning and semantic content. The poor performance on these problems was described with respect to the general solution scores, and here it should first be pointed out that the $p, \neg q$ pattern falsification solutions were virtually absent for these problems. Most important, however, is the fact that the $\neg p, q$ pattern, and only this pattern, significantly exceeded chance (6.25%, test for significance of a proportion) on all three reverse-clause problems (halls-reverse, 42%; absent-reverse, 42%; radio-reverse, 37%). This pattern would in fact be the logically correct complete falsification solution if subjects mentally transposed the clauses to make semantic sense of the sentences but continued to apply formal logical reasoning to the surface structure of the sentence. That is, if subjects recognized that the logical solution requires selections that can possibly negate the consequences of a conditional sentence, but at the same time mentally transformed the clauses of the sentence to their familiar interpretation, actual selections would conform to the $\neg p, q$ pattern.

The finding of a high proportion of $\neg p, q$ combinations for reverse-clause sentences was also present in Cox and Griggs's (1982, Experiment 3) results, but they interpreted it as a failure to reason analogically. The systematic nature of the solutions and their logical character clearly and directly implicate the interpretation that both formal reasoning (competence) and semantic processing (moderators) are operating here.

As a final check on this effect, the developmental consistency scoring described earlier (i.e., complete correct solutions on both the halls-reverse and absent-reverse problems) was applied to these data but with the $\neg p, q$ pattern as the criterion. This demonstrated that only the 8th graders (30%) performed above chance (6.25%, test for significance of a proportion). Although 30% is too small a portion to generate confidence, this finding gives some further tentative support to the view that formal reasoning competence becomes available during adolescence.

A final area to be briefly examined concerns the main effect of response alternative and the Problem × Response Alternative interaction found in the ANOVA on the general solution scores.
(see Table 1). Newman-Keuls comparisons assessing the role of the four response alternatives revealed that, as in Experiment 1, judgments concerning antecedents ($p$ and $\neg q$ response alternatives) are correct more often than judgments concerning consequents.

Inspection of paired comparisons for the Problem × Response Alternative interaction revealed that the difficulty of the radio problem relative to the halls and absent problems results from differences with the $q$ and $\neg q$ response alternatives, although performance does not differ across the three problems for $p$ and $\neg p$ alternatives. This interaction also reveals that subjects make more errors with the $q$ alternative on the absent problem than with the halls problem; they have more difficulty with the $q$ alternative relative to the $\neg q$ alternative on the abstract problem, and they find the consequent alternatives particularly difficult on the absent-reverse problem.

General Discussion

The findings of the three experiments provide evidence that performance on logical tasks is a function of both the development of formal reasoning processes and the influence of semantic content. The relatively poor performance of the 4th and 6th graders, the variable adequacy of the 8th graders across conditions, and the consistently adequate performance of the 10th and 12th graders support the position that the formal competence necessary for successful propositional deductive inference becomes available in adolescence. The developmental increase between Grades 8 and 12 on abstract problems (Experiment 1) and the systematic responding of the 8th graders on the reversal problems (Experiment 3) give further tentative evidence for the position that, regardless of the specific nature of the content, individuals demonstrate performance that accords with the rules of formal inference once the competence for such performance becomes available in adolescence. On the other hand, the findings also clearly demonstrate that familiar semantic content significantly enhances performance and that this effect is more general than shown in earlier research.

The general form–content issue has been framed by the claim made by Griggs (1983) and others (e.g., Mandler, 1983) that familiar semantic content constitutes a sufficient condition for adequate performance on tasks that have traditionally been interpreted as entailing deductive reasoning processes. To establish this strong claim it would be necessary to demonstrate that the age effects and the differential familiarity effects found in the present research were themselves a function of familiarity. Consider first the developmental effects. If all problems in the present research were familiar to adults but not to children, the strong claim might be substantiated. This argument could be advanced for the drinking age and motor vehicle problems of Experiments 1 and 2. Here it might be claimed that the legal ages of drinking and driving are sufficiently removed from the concerns of 4th- and 8th-grade groups that they constitute non-familiar content and thus account for the resulting poor performance. The argument, however, cannot be supported by Experiment 3, in which all three familiar-content problems were equally familiar for all age groups and yet poor performance was apparent in both the 4th and 6th grades.

When the familiarity effects are considered, the sufficiency claim runs into difficulty with respect to the poor performance found in each experiment for one familiar-content problem (Experiments 1 and 2: school; Experiment 3: radio). To maintain the sufficiency claim it would have to be demonstrated that each of these problems was in some way less familiar than the two companion problems in each experiment. There is no obvious way in which this could be accomplished, and in fact it appears that the school problem with its school-relevant content would be more familiar to each of the age groups than the drinking age problem, which presents content that will not become directly relevant until several years following school completion.

Although the strong claim for the sufficiency of familiar contents cannot be supported, a weaker claim can be made by asserting that both formal reasoning competence and familiar contents are necessary conditions for adequate performance on propositional reasoning tasks. This claim arises from the theoretical context of the competence-moderator-performance model (Overton, 1985; Overton & Newman, 1982). According to this approach, and as demonstrated by the age effects of both the present research and other contemporary findings (Moshman, 1979; O'Brien & Overton, 1982; Overton et al., 1985), the development of propositional and predicate logical knowledge is a necessary prerequisite for adequate and consistent performance on formal reasoning tasks. Further, as predicted in and supported by the present research, formal logical competence is not fully available until adolescence. The claim we are making about competence is not that there is a lack of logical knowledge prior to adolescence, but rather that preadolescent reasoning is based on a qualitatively different level of competence than that which becomes available in adolescence.

A related issue concerning the development and availability of a formal reasoning competence in adolescence involves the fact that competence does not operate in a vacuum. That is, logical knowledge must operate on and in the context of world knowledge, and it is at this point that the necessity of both form and content, competence and moderators, logical processes and informational processes, becomes evident. If it is accepted that the world presents the individual not with a *ganze Feld* but with a patterned environment, then it would seem reasonable that (a) the varying nature of the patterns must be attended to, encoded, and represented by informational processing systems, and (b) the specific type of pattern will either facilitate or inhibit this processing. It is, then, this encoded and represented world knowledge that engages logical knowledge to arrive at a solution. From this perspective faults may appear in either the logical or the world-knowledge systems. For example, from the perspective of logical competence, the necessary level of competence may not be available. On the other hand, tasks may be designed in such a way that they are difficult to encode, and hence they inadequately access the proper logical competence.

Within the context of this analysis, the results of the present research, as they specifically concern semantic content, are adequately explained in terms of difficulties of problem representation. If specific allowances are made for the fact that Wason does not believe that logical competence plays a role, Wason’s (1983) own explanation of semantic effects is compatible with the present results. As discussed earlier, Wason suggested that the basic difficulty with the standard form of the selection task is that it
presents subjects with an overload of information that cannot be adequately represented as a coherent whole. Thus, within the parameters that engage an available propositional competence, semantic material promotes the facilitation of performance to the extent that it enhances a coherent representation of the problem in working memory.

Arguing from the fact that nonfamiliar, meaningful material (general thematic) does produce some facilitation, Wason (1983) proposed schemata (see Mandler, 1983) as a primary mechanism for the necessary representational integration. According to this explanation, familiar content produces the greatest facilitation because it most readily evokes schemata that operate as frames or integrating devices. Meaningful but nonfamiliar material is indirectly related to world-knowledge schemata and hence is less adequately integrated and less facilitative, whereas abstract content fails to evoke any integrating mechanism. This explanation accounts for the findings of the present research that familiar content and, to a lesser extent, general thematic content (envelope problem) facilitate the performance of those individuals who have a formal logical competence available. The explanation is also consistent with the lack of transfer between familiar and abstract content. It should be made clear, however, that this explanation, in contrast to Wason’s proposal (see also, Griggs, 1983; Mandler, 1983), does not include the inferential process within the schemata themselves. We propose that logical competence requires and interacts with information-processing systems, including schemata, but that competence is distinguishable from these systems.

References


Received September 19, 1985
Revision received June 12, 1986