In R.S. Nickerson (Ed.) Attention and Performance VIII. Hillsdale, New Jersey: Erlbaum, 1980.

38 Individual Differences in General Abilities Useful in Solving Problems

Jonathan Baron
Jennifer Freyd
Judith Stewart
University of Pennsylvania
Philadelphia, Pennsylvania
United States

ABSTRACT

We report some experiments designed to identify the components of general problem-solving ability, both capacities and strategies. The first experiment looked for group differences in a possible capacity, the ability to use weak retrieval cues of the sort provided by acrostic puzzles, an ability of general value in problem solving. Graduate students (good problem solvers, we assume) and a control group read words aloud and were then unexpectedly tested for memory of the words. The students were more able than the controls to use weak cues consisting of a few letters from each word but less able to recognize the words themselves. The group differences in recognition are ascribed to greater familiarity of all words to students. An attempt to equate effective frequency eliminated the group difference in recognition, a result suggesting that the groups do not differ in incidental-learning ability. The last two experiments look for group differences in strategies useful in discovery of rules or principles, such as spelling sound rules. Students are more likely to modify their own proposed rules on the basis of counterexamples and more likely to state principles spontaneously.

INTRODUCTION

Data from I.Q. tests (e.g., Thurstone, 1938) and other sources indicate that people who are good at solving one type of problem are generally good at solving other types of problems. These correlations have been ascribed to general intelligence. But to date, the concept *general intelligence* remains primarily a psychometric construct rather than a description of mental processes used more often or more quickly by good problem solvers than by less good problem solvers. The present chapter is a preliminary attempt to test some hypotheses about what these processes might be.

Following Baron (1978), we hypothesize that good problem solvers are characterized not only by less restrictive biological limits on information processing (capacities) but also by a greater probability of use of certain highly general strategies. Our strategy here is to seek direct measures of these strategies and capacities and then to look for differences between groups likely to differ in the abilities (strategies and capacities) in question. If the groups differ, we can conclude that there are individual differences in the abilities in question and that these abilities have something to do with the criteria by which the groups were selected (for whatever reason). (For example, motivational differences may be the cause of measured ability differences. If so, the differences are still of interest to the extent to which the same motivativations work outside the laboratory.) This research strategy, and its problems, are discussed further by Baron (1978) and Baron and Treiman (1980).

In the first experiment we report, we seek such group differences in the ability to retrieve information from memory using weak retrieval cues—quite possibly a capacity. We argue that this ability is crucial in problem solving of all sorts, from acrostic puzzles to philosophy. The next two experiments look for group differences in strategies useful in the discovery of general rules. The second experiment looks at the strategy of proposing a candidate rule, testing the rule by seeking examples and counterexamples, and modifying the rule when counterexamples are found. The third experiment examines a strategy somewhat more primitive than the second, the strategy of overtly stating (to one's self, at least) a principle, when solving a problem where a principle or rule needs to be found. (This strategy is part of the strategy examined by the second experiment.)

The subjects used in the present studies were drawn from two groups. One group consisted of graduate students in psychology, mathematics, and anthropology (one subject), at the University of Pennsylvania. These students were selected in part because they were good problem solvers, we assume, although many other criteria were surely involved. The control group consisted of people from a part-time employment service, matched roughly in age and sex to the students. All subjects were between 22 and 30 years of age. Three of the eight students and four of the nine controls were female. Some subjects were not used in some tasks, as we point out later. As a rough check on the validity of our selection with respect to differences in problem-solving abilities, each subject was given 10 problems from the Raven's Progressive Matrices. The students all solved at least 8 problems, and the controls at most 5 (except for one control subject who solved 8 of a different selection of problems used for this subject only).

EXPERIMENT 1: INCIDENTAL LEARNING AND WEAK-CUE RETRIEVAL

The ability to use weak cues for retrieval seems to be an essential part of problem solving (Nickerson, 1977). When we do a crossword puzzle, we try to retrieve a word, given a couple of letters and some associate of the word (often not previously perceived as an associate). When solving a problem such as Duncker's (1945) tumor problem, we must retrieve ideas, such as the idea of rotation (of the source of rays) when cued only with the problem of focusing the rays on the tumor. (Often such ideas must be modified before they are applied, a fact that makes their retrieval even more mysterious.) In such problems as the Progressive Matrices, the items retrieved are often rather general ideas or "frames" such as the idea of addition or of rotation. For a more realistic example, one of the authors solved the problem of how to remove 10-year-old stickers from some recently opened wedding presents by first trying to think of an organic solvent in the house and then retrieving the memory that there was some charcoal lighter fluid in the back hall; the interesting fact here is that lighter fluid was (apparently) never before encoded as an organic solvent.

The present experiment is based on a type of problem in which weak-cue retrieval is introspectively prominent, the type of problem found in acrostic or crossword puzzles. In essence, we ask whether the students are better than the controls at solving problems of this sort in part because the students are better at retrieving words they know, given the weak cues provided by the puzzle. The main alternative hypothesis is that the students are better just because they know more words or because their memory traces for the words they know are stronger.

We first presented words to subjects in a way designed to equate (depth of) encoding as much as possible. The subjects were asked simply to read the words aloud five times, without being told that their memory for the words would be tested. We then tested recognition memory for the words by presenting (whole) words that either had been previously presented or had not. In addition to our presenting these strong retrieval cues (as Tulving, 1974, might call them), we also presented weak cues in the form of words with most of the letters replaced by dashes (e.g., C--------A as a cue for CALIFORNIA). Again, half of these cues were parts of previously presented words and half were not. Subjects were asked to decide whether a word fitting the frame had been presented and to guess what the word was, if possible, whether or not they thought it had been presented. Our main hypothesis is that the controls, relative to the students, will show a greater deficit in recognition of the weak cues than of the strong cues and a greater deficit in recall of the words than in strong-cue recognition.

One possible confounding variable in studies of group (including developmental) differences is effective familiarity with stimuli, which may affect performance (Baron & Treiman, 1980). A solution to this problem would be to use entirely novel material. But it is not clear what that would be, because all stimuli are somewhat similar to stimuli we have experienced, and prior experience might make the greatest difference when the similarity is low. In the present experiment, we measured the effect of familiarity on our tasks, by varying the frequency of the words presented, hoping that it would have equal effects on all major tasks. The words were separated into five lists, each list drawn from one-fifth of the frequency rank list of Carroll, Davies, and Richman (1971). As an additional measure of effective frequency for individual subjects, subjects were asked, at the end of the experiment, to indicate which presented words they did not know the meaning of. These words can be assumed to have the same effective frequency for all subjects.

In the recognition experiment, subjects were asked to give confidence ratings: one (very sure the word was presented), two (think it was), three (think it wasn't), or four (very sure it wasn't). Use of this measure with both strong and weak cues ensures that measures of performance will be based on the same response scale. (Tulving, 1974, and others who compare different kinds of retrieval cues by comparing recognition and recall, have not achieved this sort of comparability.)

The design also allows us to gather preliminary data relevant to other questions. First, we can ask whether our groups differ in learning ability. The groups did show large differences in an intentional learning task. However, such differences in intentional learning may be due to differences in the strategies used to encode the stimuli. It is thus of interest to examine incidental learning, where such encoding differences are minimized (as done by Cermak & Reale, 1978, for example, in a study of chronic alcoholics).

Second, we can ask whether the groups differed in reading speed, for we timed the list reading. Lists were read in order of decreasing frequency, five times through all lists. The words were all of three or more syllables, selected so as to be ambiguous in stress or pronunciation if read according to spelling-sound rules alone. Thus, initial reading time could be used as a rough measure of familiarity, whereas asymptotic reading time could be used to estimate a subject's reading speed.

In summary, each subject was asked to read five lists of 20 words each, in order of decreasing frequency, five times through the lists. Subjects were told that the task was a measure of reading speed. Then a few minutes were taken to explain the (unexpected) memory tests. In these tests, subjects were shown five lists of 40 words each, in order corresponding tot he original lists, with distractors from the same frequency range as the targets. Alternate words in each list had most letters replaced by dashes. The subject was to indicate confidence that each item (or a word containing the given letters, for weak-cue items) had been presented. The weak cues were chosen to be equally "good" across frequency classes (as is verified in the data). Finally, subjects indicated which words they did not know the meaning of. Three of nine control subjects could read so few of the words that the experiment was halted immediately. This left six controls and eight students.

Retrieval Results

The measure of retrieval was the point biserial correlation between the subject's confidence rating and whether the item was presented. This measure resembles d' in that it attempts to remove influences of response bias. However, because it approaches an asymptote of 1.0, as performance improves, it resembles *hits minus false alarms* in its scaling properties.

Figure 38.1 shows retrieval as a function of cue type, group, and word frequency. Lower frequency yields higher performance (as found in all previous studies we know); for each cue type there is a significant negative correlation ($p < .05$, 3 *df*) between frequency and performance. This effect, as measured by the slope of the best fitting line, was if anything greater for weak cues than for strong cues.

As hypothesized, the students were more accurate than the controls with the weak cues ($p < .025$). This is not due to stronger memories of the stimuli, it appears, because they were less accurate with the strong cues ($p < .025$, 1 tailed). Lower performance with strong cues would be expected if the familiarity of the words was greater for the students than the controls, for familiarity impairs recognition. (Possibly the students read the words more quickly and thus encoded them less strongly, however.) But the students' superior performance with the weak cues is great enough to overcome this deficit, even though frequency itself has a negative effect on weak-cue retrieval as well as strong-cue retrieval.

We can ask whether students were more able to fill in a word that had been presented, on the basis of a weak cue. The students filled in a mean of 14.0 words; the controls, a mean of 6.8 [omitting one control who filled in no items at all, $t(11) = 2.74$, $p < .01$]. Because this difference could result if the students were simply more likely to fill in words, regardless of whether a word had been presented, we counted the number of filled-in words that had not been presented and that had the correct number of letters. The students filled in a mean of 18.5 of these; the controls, a mean 8.8. This difference was not significant [$t(11) = 1.2$]. To ask whether there was an interaction between

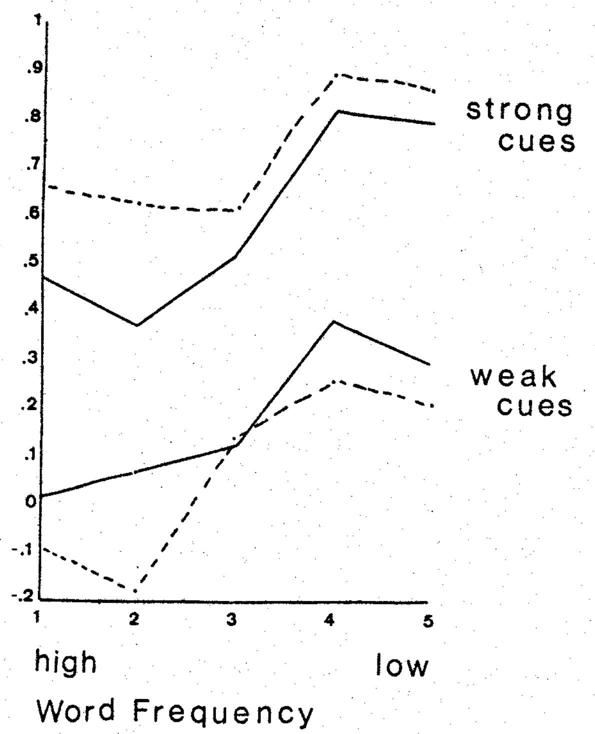

FIG. 38.1. Recognition performance (point biserial correlation between whether the stimulus had been presented and the subject's confidence rating) as a function of word frequency, for students (solid line) and controls (dashed line) for the two cue types.

group and presented versus nonpresented in these scores, we computed the correlation between group membership and each of the scores. (The scores cannot be compared directly, because there is a greater opportunity to fill in a nonpresented word—nonpresented words can be filled in for weak cues for presented words.) The correlation between group membership and the number of presented words filled in (0.64) was in fact significantly higher [$t(10) = 2.10$, $p < .05$] than the correlation for nonpresented words (0.31), given a correlation of 0.77 between these two scores across all subjects. (This analysis assumed that the nonpresented score is as reliable as the presented score, a safe assumption given the greater range of the former.)

The result for filled-in words shows that the probability of filling in a word is affected both by group membership and by whether the word had been presented. This result rules out a simple explanation in terms of group differences in motivation. It also rules out a model in which subjects generate possible words, consistent with each weak cue, and then check each generated

word as if it were a strong-cue item. By this model, the students might perform better simply because they were better at generating possible words. But the group difference in filling in presented words is larger than that in filling in nonpresented words, so differences in the latter ability cannot account entirely for the former. (We assume that both groups followed the instructions to fill in even nonpresented words. In fact, this part of the task was emphasized to all subjects.) Note that in the model we just rejected the generation of words is unaffected (or equally affected for both groups) by whether the generated word had been presented. If we assume that the students were particularly prone to generate words that had been presented, the results are consistent with a less restrictive generate-and-test model.

As an additional check on the possibility that the group differences are due to differences in word knowledge, which would allow the students to generate more posssible words, a separate analysis was done of the recognition data for only the words each subject did not know the meaning of. Again, the students were more accurate with weak cues and less accurate with strong cues. Neither result was significant by itself, but the interaction between group and cue type was significant at the 0.025 level. (The measure of accuracy was the mean difference in ratings between each unknown presented word and the mean rating of all the unpresented words of the same cue type in the same list. The interaction was determined from the difference between dependent correlations between group and this measure for the two cue types. The relevant measures were: students—strong, 2.0; controls—strong, 2.5; students—weak, 0.8; controls—weak, 0.2.)

There are two possible artifactual explanations of our major results, one having to do with the quality of the weak cues as a function of frequency and the other with possible group differences in stimulus encoding. By the first account, the weak cues were more useful for low-frequency words than for high-frequency words (despite our efforts to prevent this); if the cues were equally useful, weak-cue recognition would actually be positively correlated with word frequency, and the group differences could then be explained in terms of differential effects of frequency on the two measures. To test this account, we defined cue effectiveness as the conditional probability of filling in the word from which each weak cue had been derived, given that a guess was made. For words that had not been presented, the cues were if anything more effective for the two higher-frequency lists (.52 and .42) than for the two lower-frequency lists (.36 and .31). This suggests that our efforts to make cues equally effective were successful. (Note that the words presented were chosen at random from the list of 40 test items for each list. Note also that although cues were apparently more effective for low-frequency presented words than for high-frequency presented words—.63 and .67 versus .84 and .79, respectively—this result simply reflects the negative effect of frequency on weak-cue recognition, as shown in Fig. 38.1.)

The second possible artifact is that the groups might have encoded words differently. For example, the students might have relied more heavily on phonetic representations, and this might have made the weak cues more effective for them. We cannot rule all hypotheses of this sort, and further tests of the generality of our findings are in order. However, a post hoc analysis suggests that the groups did not differ in semantic versus phonetic encoding. We selected five distractors that were semantically related to some presented item (e.g., government—Washington) and five that were phonetically (and possibly graphemically) related (e.g., experience—ebullience). Items in each of these sets were more likely to lead to false positive responses than were unrelated items ($p < 0.005$ for semantic, 0.05 for phonetic). The students falsely recognized 13 semantic distractors and 9 phonetic ones, and the controls falsely recognized 4 of each type. Thus, there is no evidence that the students were more prone to encode the words phonetically. Further evidence against the involvement of encoding differences is provided by the analysis of words that subjects did not know the meaning of, which could not, we assume, be encoded semantically very well. Recall that the major group differences were found for these items alone.

In sum, our results indicate a group difference in ability to retrieve with weak cues. This result is not due to differences in knowledge of words nor to differences in strength of encoding of words presented nor to other kinds of knowledge that might affect ability to generate words. The group difference is most easily ascribed to some more general ability. It is possible that the students are more prone to use a generate-and-test strategy, possibly because they are more motivated. It is also possible that the difference represents a less modifable capacity difference, possibly a capacity to retrieve with weak cues and possibly a more general capacity such as the amount of mental energy available. Further research must be directed at the generality of this sort of group difference, as well as its nature. (Note that even motivational differences are of interest if they are sufficiently general.)

Reading Times

An analysis of the reading times might serve two purposes. First, we might obtain a measure of asymptotic reading speed. Second, we might use times for first reading to estimate the familiarity of each list to each subject. By plotting strong-cue recognition accuracy against this measure, we can ask whether the groups differ in incidental learning, with word knowledge and encoding roughly equated.

We assumed, as a first approximation, that reading time approaches an asymptote by a constant proportion on each trial. This assumption yields an exponential practice curve, $R(i) = B[\exp(-Ci)] + A$, where $R(i)$ is the time for the ith trial, A is the asymptote, B represents the point at which practice

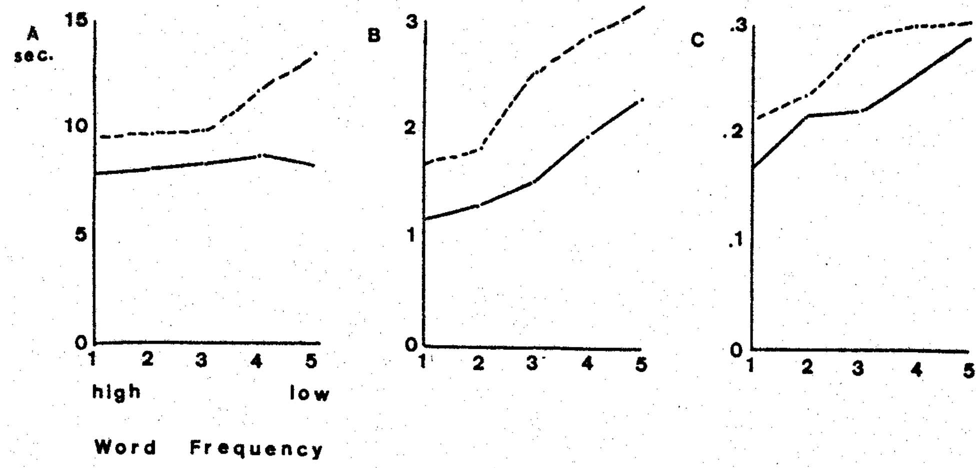

FIG. 38.2. *A*, *B*, and *C* parameters of the best-fitting exponential practice curve for students (solid line) and controls (dashed line).

started, and C represents the rate of improvement with practice. We use B as our estimate of list familiarity. A, B, and C were estimated for each list for each subject, using a program that minimized squared deviations of $\log[R(i)]$ from predicted values.

The mean parameter values for the two groups are shown in Fig. 38.2. The most encouraging result is that A is independent of frequency, as it should be, for the students and for all but the least frequent words (containing many unknown words) for the controls.

The group difference in A was significant ($p < 0.01$). [However, for the three highest-frequency classes, where A was unaffected by frequency, the difference was not quite significant: $t(12) = 1.75$.] Although there are many possible explanations of this result, it does suggest that it is possible to measure group differences in mental abilities in a way that removes confounding effects of familiarity with stimulus materials.

Figure 38.3 shows group means in strong-cue accuracy for each list as a function of $\ln(B)$. Arguably, the points lie along the same curve. If they really do, it would appear that there are no group differences in incidental learning when effective frequency (B) is held constant. We also found the best fitting line, for each subject, for strong-cue accuracy as a function of $\log(B)$. We then found the intercept of each line for $\log(B) = 2$, a point within the range of values for six subjects in each group. For these subjects, the strong-cue accuracy of the students was slightly lower than that of the controls. For all subjects, the accuracy at the intercept was nonsignificantly higher for the students [0.72 versus 0.69, $t(12) = 0.38$]. These results suggest that there are no differences in incidental learning between two rather extreme groups.

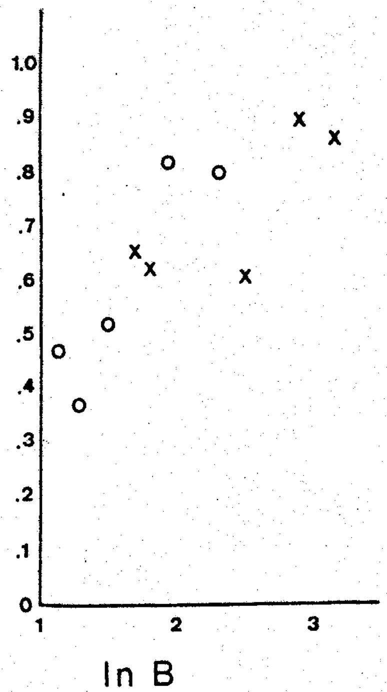

FIG. 38.3. Performance in recognition with strong cues (same measure as Fig. 1) as a function of the *B* parameter for the practice curves (used as a measure of familiarity). Each point represents a subject. The O's are students; the X's are controls.

Further evidence is needed, however, both on the generality of this result and on the validity of *B* as a measure of effective frequency.

EXPERIMENT 2: THE DISCOVERY OF RULES

In the next experiment, subjects were asked to discover spelling-sound correspondence rules. For example, they were asked to give a rule for when the letter *a* is pronounced as in *mate*. This task is similar to other members of a large class of problems that require the strategies of proposing principles or rules, testing the rules by trying to think of examples or counterexamples, and modifying the rules when the examples do not fit. Our task involves the sort of reasoning done in linguistics, which can serve as a model for scientific and philosophical reasoning in general. (For experimental purposes, linguistic tasks have the advantage that most speakers are familiar with the data to be accounted for by the rules sought.) The same kinds of strategies, involving formulation and tests of hypotheses, are undoubtedly useful in solving problems of other sorts, from debugging a computer program to buying a car.

The subjects used in Experiment 1 were used here as well. Three of the controls were eliminated because they did not know that spelling-sound rules

existed. (They believed that they read by words or syllables only.) This left 8 students and 5 controls. The idea of rules was first explained to each subject, using the example of *r* (where the rule is very simple) and *g* (where the rule is complex). The rule for *g* as a function of the next letter and of position in the word was explained with examples. Subjects were asked to discover the conditions under which each letter was pronounced as in the following words: o—go, got; *i*—bit, bite; *a*—rat, rate; *t*—with, nation, native; *c*—cent, can't. Subjects were asked to talk aloud while working, and the experimenter asked questions freely, to elicit information about the strategies used. Subjects were told that the experimenter would help them think of examples if they asked. No subject gave a fully adequate rule for any item. But the purpose of the experiment was to assess strategies, and this assessment was easily made.

Written transcripts (and tapes when needed) of the interviews were scored for presence of certain strategies by Baron and Stewart. Seven of 13 were scored independently to check reliability, which was 84%. All disagreements were resolved by discussion, and the rest of the interviews were scored by both judges together.

The scoring categories used were the following:

Rule Statement. A stated rule had to specify that a letter was pronounced in a certain way (in a context more general than one word) or that no rule was possible for choice between two pronunciations. Veracity was irrelevant.

No Hypothesis. This represented absence of rule statement, for example, in cases in which a subject responded only by giving examples.

Rule Transfer. A parallel between two rules (e.g., the three vowels, or *C* and *G*) was mentioned.

Example. The subject gave an example before stating a rule (as if trying to use the example to arrive at a rule) or asked the experimenter for an example.

Confirming Example. An example was given after a rule statement, the example obeying the rule stated. Confirming examples were scored + (plus) if the example met a condition specified by the rule. For example, "*C* is pronounced hard at the end of a word, like *TIC.*" Confirming examples were scored – (minus) if the condition was not met and the pronunciation was not given (e.g., *lace* for the rule just given). Although + was used 27 times, – was used only once (by a student).

Disconfirming Example (Counterexample). A counterexample can also be +; when the condition is met and the pronunciation is not given, or –, when the condition is not met and the pronunciation is given. For the final-*C* rule stated, *tick* is – and *bosc* is +. [Wason & Johnson-Laird, 1972, seem to use "confirming" to mean + (meeting the condition) and "disconfirming" to mean –. We find that our terms clarify their findings if not ours.] Positive counterexamples may be real exceptions, of course.

Unused Counterexample. A counterexample was given, either by the subject or the experimenter, but was not accounted for. If the subject said that something was an exception, we used this category only when the subject was blatantly wrong (to avoid the need to set a criterion for true exceptions).

Changed Rule. This category was used for changes in rules already stated (often to account for counterexamples) but not for new statements of disjunctive conditions for a certain pronunciation of a letter; a changed rule had to alter the pronunciation of some examples dealt with by the old rule.

Implicit Rule. This category was used when a stated rule contained a subcondition apparently designed to account for a small set of counterexamples. For example, a few subjects said that *th* is pronounced as in *with* except when the two letters are parts of different words, as in *pothole.* This kind of statement suggests that the subject had already thought of a counterexample and changed the rule to account for it.

Each category was coded at most once per letter (to avoid decisions about what was the same instance and to minimize effects of subject verbosity).

If students are better problem solvers, they ought to use optimal strategies more consistently. In particular, they ought to be more likely to use *changed rule* and *implicit rule* and less likely to use *unused counterexample.* Students were in fact more likely to change rules (1.9 out of 5 letters versus 1.0) and to give implicit rules (0.5 versus 0.0) and less likely to give unused counterexamples (0.5 versus 1.0). Other means are, for students and controls, respectively: rule statement, 4.0, 3.6; no hypothesis, 0.1, 0.6; rule transfer, 1.8, 0.8; example, 0.5, 0.2; confirming example, 1.9, 2.6; disconfirming example, 1.5, 1.6. Although significance tests are inappropriate for a study of this scope in which so many comparisons may be made, it does appear that the students were more likely to use strategies that a linguist should use. Whereas we would like to argue that the controls have a remediable deficiency in tendency to use optimal strategies, it is also possible that the controls were unable to think of rules to account for counterexamples even when they tried.

The following quotations illustrate the different styles of the two groups. The first is from a student; the second from a control:

1. *A* is generally long when it's by itself or when it comprises the initial syllable of a word. Is this true? Are there examples? Abate. No, forget that one. (Experimenter offers *appraise.*) Appraise, abate. No, it's short. Alphabet. No, *A*'s at the beginning of words generally seem to be short. (Coded as: rule statement, disconfirming example, changed rule.)

2. In *with* you don't have a *t* sound, you have a *t, h* sound. So really the *t* is not heard. . . . In *nation* you don't hear the *t* also. You hear an *s, h* sound. . . . In *native* you do hear the *t*. . . . Don't know why you do hear the *t* . . . (Experimenter: Why don't you hear it in *nation*?) Well, *tion* is a sound by itself. (Coded as: rule statement.)

Another finding of possible interest is that five of the eight students and none of the controls dealt with the three vowels as examples of a single general rule. This might indicate a greater tendency to try to generalize rules, a useful strategy not included in our scoring.

In sum, these preliminary results show that strategies for rule abstraction can be studied in a naturalistic (for academics) task. They suggest that groups differ in the tendency to use optimal strategies. Whether these tendencies are easily modifed is best determined by studies of training and transfer.

EXPERIMENT 3: STATEMENTS OF PRINCIPLES IN EVERYDAY PROBLEMS

The discovery of spelling-sound rules *requires* that the subjects state principles, so the data did not allow us to discover group differences in the tendency to state principles spontaneously. Spontaneous statements of principle (if only to one's self) may aid in the recognition of problems, in the discovery of general rules and, more generally, in the discovery of order in the world (Baron, 1978). In the present experiment, we look for group differences in the tendency to state principles in an open-ended interview. Subjects were asked questions about familiar topics, picked so that appropriate answers might or might not include principles: (1) what factors determine how fast a hot liquid will cool? For example, if you make some hot soup, what kind of things could you do to make it cool faster? (2) How could you compare foods to find out if you are getting the most nutritional value for the money? What do you have to consider? (3) When it is right or wrong to say something false about yourself? Suppose an adolescent asked you this. What would you say?

Subjects were encouraged to introspect and to say all they had to say about each problem. The experimenter frequently asked what the subjects were thinking, during silences, or what the reasoning had been behind an unclear statement.

After the subject answered the questions, the experimenter asked the subject to go through the problems again and to state a rule or principle for each problem. The questions were restated one by one for this purpose (e.g., can you state a general principle that accounts for the rate of cooling?). Subjects were told that they could give a number of rules instead of a single rule and that they could say they had already stated a principle in the first part instead of repeating it. Subjects seemed to understand these instructions, as every subject succeeded in giving at least one rule in this part, and most recognized when they had already stated a rule in the first part. The purpose of this second part was to find out whether any failure to give a rule in the first part was due to inability to arrive at a rule or to failure to make the attempt to do so. The latter possibility is implied by the claim that principle statement is a general strategy used more frequently by good problem solvers. Thus, we hypothesize that students will be less likely than controls to state principles only when asked to do so.

As judged by the experimenter (on the basis of notes and, when needed, tapes), five students states rules for all three problems in both parts and three stated rules for two problems, again in both parts. Thus, no student stated a rule only in the second part. Five of the seven controls, however, stated rules in the second part but not the first for at least one problem. The difference between groups (five out of seven versus zero out of three, three being those students not at ceiling on the first part) is significant ($p < 0.05$) by Fisher's exact test. Thus, it does seem that students are more prone to state rules (at least overtly) in situations in which rules are not explicitly required.

Other findings of possible interest are: (1) Six of the eight students as opposed to two of the controls changed the statement of a rule at least once (and two students did this twice). Although this difference is not significant, it is consistent with the findings of the last experiment, where students were more prone to try to improve their own rules. (2) Five of the eight students and one of the controls mentioned the impossibility of solving the second (food) problem (a problem formulated poorly with the intention of eliciting this strategy). Although not significant, this result suggests that students are more prone to use the strategy of converting an apparently impossible problem into a manageable one.

SUMMARY AND CONCLUSIONS

In the first experiment, we found that students were less able than controls to recognize words when given strong cues consisting of the words themselves but more able to recognize words with weak cues. The controls seem to have a differential deficit in retrieval with at least the sort of weak cue we used, whatever the ultimate source of this deficit. This experiment also suggested that the groups differ in reading speed, with the effects of word familiarity removed. Finally, the groups do not seem to differ in incidental learning (tested with strong-cue recognition), where group differences in encoding strategies are minimized. In looking for this difference, we attempted to remove the effects of effective frequency of the stimuli for each subject. Because our groups do differe dramatically in an (unreported) paired-associate learning task, this result suggests that individual differences in learning ability may be due largely to encoding strategies (and retrieval abilities).

The second and third experiments sought group differences in strategies useful at least in problems involving the discovery of rules or principles. Such strategies might be sufficiently general in their usefulness to constitute part of general intelligence in the sense discussed by Baron (1978). In the second experiment, subjects were asked to discover spelling-sound rules. Students were more prone to try to improve their initial statements of rules by thinking of counterexamples and then changing their rules to account for the counterexamples. In the third experiment, students were more likely to state principles spontaneously in response to open-ended questions that did not call for such statements. Controls were, however, able to state principles when asked—a result suggesting that principle statement is a teachable strategy (Baron, 1978).

The last two experiments are properly conceived as exploratory studies, because we have not tried to rule out the sorts of familiarity artifacts addressed in the first experiment. Further, these apparent strategy differences may be entirely the result of less modifiable capacity dfferences, such as available mental energy or weak-cue retrieval ability (if that is not itself the result of a strategy). The most optimistic interpretation of these results is that they represent differences in the probability of use of general strategies and that these differences result from differences in education (broadly or narrowly conceived). If so, it ought to be possible to improve people's problem-solving abilities through education.

ACKNOWLEDGMENTS

The work was supported by N.I.H. Grant MH 29453 (Baron, principal investigator). We thank D. G. Kemler for helpful discussion.

REFERENCES

Baron, J. Intelligence and general strategies. In G. Underwood (Ed.), *Strategies in information processing*. London: Academic Press, 1978.

Baron, J., & Treiman, R. Some problems in the study of differences in cognitive processes. *Memory and Cognition*, 1980, in press.

Carroll, J. B., Davies, P., & Richman, B. *American heritage word frequency book*. New York: American Heritage, 1971.

Cermak, L. S., & Reale, L. Depth of processing and retention of words by alcoholic Korsakoff patients. *Journal of Experimental Psychology: Human Learning and Memory*, 1978, *4*, 165–174.

Duncker, K. On problem solving. *Psychological Monographs,* 1945, *58,* Whole No. 270.
Nickerson, R. Crossword puzzles and lexical memory. In S. Dornic (Ed.), *Attention and performance VI.* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.
Thurstone, L. L. *Primary mental abilities.* Chicago: University of Chicago Press, 1938.
Tulving, E. Cue-dependent forgetting. *American Scientist,* 1974, *62,* 74–82.
Wason, P. C., & Johnson-Laird, P. N. *Psychology of reasoning: Structure and content.* Cambridge: Harvard University Press, 1972.