ADVANCES IN THE PSYCHOLOGY OF HUMAN INTELLIGENCE

Volume 3

Edited by

Robert J. Sternberg
Yale University

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1986 Hillsdale, New Jersey London

5 Cognitive Style and Its Improvement: A Normative Approach

Jonathan Baron
Peter C. Badgio
University of Pennsylvania

Irene W. Gaskins
The Benchmark School, Media, PA

THE DEFINITION AND MEASUREMENT OF COGNITIVE STYLES

Cognitive styles are supposed to be general dimensions of individual differences in thinking, attending, remembering, and perceiving other than those dimensions best described as abilities. For certain dimensions of style, it may be a good thing for a person to be located at a certain point on the dimension or even as close as possible to one end of it. Our concern with such styles is thus similar to our concern with ability: we would like to find ways of making everyone as close to the optimum point (or end) as possible. Styles may also help us to predict which students will benefit from which kinds of training (Cronbach & Snow, 1977), but this is not our concern here.

The first part of this chapter concentrates on the definition and measurement of cognitive styles, particularly the style of impulsiveness. First, we describe how style measures differ from, on the one hand, ability measures and, on the other, personality measures. Next, we outline a new approach to cognitive styles, in which a style is conceived in terms of the magnitude and direction of departure from the optimum on a certain dimension. The optimum may be defined by a normative model of performance in the task the subject is asked to perform, that is, a model of what a subject *should* do. Alternatively, departures from optimum may be inferred from the results of training studies: if change in style improves behavior unambiguously, the style must have been nonoptimal before training. We compare our new approach to impulsiveness to the extant literature on reflection-impulsivity, a related but not identical concept. We end this part by discussing possible relations between impulsiveness and other measures of style and ability.

The second part of this chapter reports a training study in which the cognitive style of a group of poor readers was modified with apparent success. The theoretical perspective of the first half of the chapter served as both motivation for the study and a framework for interpreting its results.

Cognitive Styles Defined

Measures of individual differences in aptitudes and abilities are almost universally based on speed, accuracy, or both. Ability tests differ in generality. For example, performance on the Graduate Record Examination Aptitude tests correlates fairly well with grades in a variety of academic subjects. However, the score on the Psychology Graduate Record Examination does not correlate with much except grades in psychology and perhaps the number of psychology courses taken.

In contrast to ability measures are personality measures, in which items do not have right and wrong answers and speed of answering is usually irrelevant. Personality measures generally concern characteristic tendencies, motives, beliefs, values, and attitudes. Like aptitude measures, they can be general or specific, e.g., one's attitude toward technology or one's attitude toward the TRS-80, respectively.

Cognitive style measures traditionally occupy a middle ground between aptitude measures and personality measures. As such, their status has been unclear, and this lack of clarity has occasioned several efforts to distinguish them from the other two kinds of measures (e.g., Kagan & Kogan, 1970; Kogan, 1980). By general agreement, cognitive style measures are "cognitive" in the sense that they concern attending, perceiving, remembering and thinking. Cognitive style measures resemble personality measures in that they do not concern the limits of performance. Rather, they may be defined so as to leave open the question of whether they measure abilities or characteristic tendencies. (Note that, though both style and personality measures can measure characteristic tendencies, the usual difference is that personality measures do this by questionnnaires whereas style measures use laboratory tasks designed to elicit the habitual behavior in question.)

Aptitude measures, in contrast to style measures, are defined so that it is never undesirable to be close to one end of the dimension, regardless of the situation. It is never harmful to have five extra IQ points; at worst, one can choose not to use them. On the other hand, style measures are designed "to account for the range of cognitive performances where the form and manner, rather than the sheer skill of performance," is at issue (Kogan, 1980). Style measures are thus defined so that they need not imply that more is always better. However, as Kogan (1976) pointed out, discussions of various styles frequently assume that one end of the scale is better than the other even though this is not implied by the definition of the style. Let us illustrate the various relations that exist between styles and values with a few examples from the literature.

Kagan, Moss, and Siegel (1963) asked children to classify pictures of common objects or put together those objects that went together best. In some cases, a large array of pictures was presented. In other cases, the subject was given three pictures that could be classified in two different ways. Classifications were scored as analytic (common attributes), inferential (common category membership), or relational (thematic relationship such as exists between a match and a pipe). For example, one triad of pictures showed a picture of a dog missing a mouth, a picture of a boy missing a mouth, and a picture of a dog house. The analytic classification would put together the first and second, the relational, the first and third. Although all three bases of classification were correct in satisfying the experimenter's instructions, relational classifications were considered least mature, analytic, most mature. However, we can imagine tasks, such as setting the table, in which relational classifications might be appropriate, so that a child with this style might actually do better on the first try. Nonetheless, Kagan et al. were quite explicit in arguing that a tendency to make analytic classifications was a superior style. Their argument was based on their findings that adults who made analytic classifications tended to share such traits as concern with intellectual mastery, independence, and high achievement motivation (as assessed by several personality measures). Likewise, among children, the tendency to give analytic responses was positively correlated with age and IQ and negatively correlated with behavioral impulsiveness and hyperactivity. With these results in hand, Kagan et al. went on to make further suggestions about the importance of the analytic style. For example, the task of learning to analyze words in reading (bat, cat, boy, etc.) is closely analogous to the classification task itself.

Another example of implicit values is the research inspired by the findings of Smith and Kemler (1977). In one experiment, Smith and Kemler gave children three squares, A, B, and C, cut out of paper varying in shades of gray. A and B were identical in one dimension, such as size, but very different in another, such as brightness. A and C were close in both dimensions but identical in neither. Children were asked to put together the squares that went together best. Older children tended to make the dimensional classification, A and B, and younger children tended to make the similarity-based classification, A and C. The only error was B and C, which were neither identical on a dimension nor similar overall; such errors were omitted from definition of the "style." In this case, dimensional classifications acquired positive value by virtue of their association with increasing age. Subsequent research has suggested that they are also correlated with intelligence in children (Kemler, 1982a, 1982b), but not in adults (Smith & Baron, 1981). (Although the workers in this field do not consider themselves to be students of cognitive style, it seems impossible to draw a principled distinction between their work and other work in the cognitive style tradition. It would seem that Kagan et al., 1963, might even accept this classification task as a measure of analytic style.) Kemler (1983) has gone on to argue that dimensional classifications are in fact more suitable for the kinds of tasks that older children are called upon to do, such as the learning of physical concepts. However, she also points out the value of similarity-based classifications for some of the tasks of early childhood, such as learning basic-level object categories.

For still other styles, such as "field dependence," the basis of measurement is accuracy or speed of performance, such as accuracy of setting a rod to the vertical, despite the tilt of one's chair and of a frame around the rod, or speed of locating embedded figures. Such measures differ from pure ability measures in that their purpose is to measure a property of perception that by itself is not a matter of accuracy alone, namely, the "effect of the perceptual field on selected elements within the field" (Linton, 1955). Thus, this is called a style measure because it might not always be desirable to ignore the perceptual field, although it is desirable to do so in the tasks that are used. (Widiger, Knudson, and Rorer [1980] have argued that field dependence does not exist as a style distinct from abilities such as those involved in measures of spatial intelligence. It turns out that measures of field dependence correlate as highly with these intelligence measures as the latter correlate with each other.)

Some measures are made up by looking at differences in speed or accuracy of two similar tasks. This method of measurement (discussed by Baron & Treiman, 1980a) allows for control of individual-difference variables (traits) that affect both tasks. We shall take these differential measures to be styles rather than abilities only when the trait that affects the tasks differentially is not one that is necessarily desirable in all situations. By this account, distractability, measured as the difference between doing a task with distraction and doing it without distraction, is an ability.

On the other hand, Baron (1979; Baron and Treiman, 1980b) defined a stylistic dimension of children's reading of isolated English words, called the Phonecian-Chinese dimension (following a suggestion of H. Gleitman). When reading such words, one may rely on word-specific associations between the word and its identity or on some kind of representation of spelling-to-sound rules. (These rules may be represented as more general rules for forming analogies.) At one end of the dimension, Chinese rely on word-specific associations; at the other end, Phonecians rely on rules. This style is measured on the basis of accuracy on two kinds of items: nonsense words, which can be read only by use of rules; and exception words, which require some use of word-specific associations. A child's position on the dimension (with Phonecian more positive, Chinese more negative) can be defined by subtracting accuracy on exception words from accuracy on nonsense words. (Because the two tests might not be equated for variance, it is desirable to normalize the scores first and then subtract the z scores instead of the raw scores; see Baron and Treiman, 1980a). Because any general abilities will, we assume, have the same effect on both tests, this subtraction will essentially remove effects of general ability on the style measure. Treiman (1984) has recently found parallel Chinese-Phonecian differences in spelling, and this continuum correlates with the reading continuum. A similar style measure has been developed by Freyd and Baron (1982) for reliance on knowledge of derivational morphology in giving word definitions. Some eighth-grade children seem unable to use derivational rules of the sort used to make up words such as "scenic," which these subjects might define as "a picture, or a nice view," ignoring the fact that "-ic" indicates an adjective.

The Phonecian end of the dimension (and the end with more reliance on morphological rules) turns out to be better, in that it correlates with overall performance or speed of learning, measured in other ways (Baron & Treiman, 1980b; Freyd & Baron, 1982). This is an empirical finding, not part of the definition of the style. This sort of finding can turn a style dimension that is at first evaluatively neutral into a dimension with more value placed on one end than the other. Such findings have characterized many style dimensions (Kogan, 1976, 1980). Even when this occurs, the conclusion that one end is better is not implied by the definition of the dimension or trait.

In sum, ability measures define a dimension so that value increases as one moves toward one end of the dimension. There is no situation in which it is better to be less able to do something, rather than more able. Style measures are not defined this way. For style measures, there may be situations in which it is better to be at one end, situations in which it is better to be at the other, and still other situations in which it is best to be at some intermediate point. For some style measures, it may be true, or assumed to be true, that it is usually better to be at one end than the other, or farther in that direction than most people are, but this assumption is not implied by the definition of the style.

In this chapter, we suggest a new approach to the definition of cognitive styles, which is formulated explicitly on the basis of normative considerations. The basic idea is to make explicit the normative assumptions behind style measures, be defining a style only relative to the optimum point on a given dimension for a given person in a given situation. We illustrate our approach primarily with reference to a style dimension that has been called reflection–impulsivity (Kagan, 1965), but that we redefine on the basis of our normative approach. (Baron[in press] develops a framework that specifies some other style dimensions that can be treated in this way.) We conclude this part with some suggestions about the relation between impulsiveness and other styles, particularly the Phonecian-Chinese dimension, and dimensional versus similarity classifications.

Reflection-Impulsivity

Reflection-impulsivity is usually defined as a dimension of individual differences in style of performance on tasks involving response uncertainty. Specifically, this dimension refers to differences among subjects in the trade-off of speed and accuracy. In most tasks involving thinking, errors may be avoided, up to a point, by taking more time and being more careful. According to traditional definitions (e.g., Kagan, Rosman, Day, Albert, Phillips, 1964; Messer, 1976), subjects who respond quickly and make many errors relative to other subjects are labeled impulsive. Those who respond more slowly and make fewer errors are labeled reflective. Two other types of subjects, those with short response latencies and low error rates, and those with long response latencies and high error rates (labelled "fast-accurates" and "slow-inaccurates", respectively), are in the middle of the reflection-impulsivity dimension, along with children who are simply average in all respects.

Reflection-impulsivity has been studied most extensively in children. The most widely used test for the assessment of impulsivity is the Matching Familiar Figures Test (MFFT) developed by Kagan and his colleagues (see Kagan et al., 1964). In this task, subjects are asked to decide which one of four, six or eight test stimuli is identical to a target stimulus. Stimuli consist of drawings of familiar objects and animals, with all but one of the test stimuli differing in one or more details from the target. The common finding is that error rates on this test are negatively correlated with response latencies; typically, r is about -.48 (Messer, 1976). Subjects may be classified as impulsive, reflective, fast-accurate, or slow-inaccurate according to their performance relative to a two-way median split of these negatively correlated measures. Other investigators use a composite of standardized response latencies and error scores (Salkind and Wright, 1977) or treat latencies and error scores as separate variables in correlational analyses (Kagan, Lapidus, & Moore, 1978; Smith, Smith, & Baron, 1984).

Reflection-impulsivity thus defined appears to be a general trait that is stable across a variety of problem-solving and reasoning tasks. Impulsive children make more nonsensical errors when reading aloud (Kagan, 1965). Response latencies and error rates of the MFFT are correlated with response latencies and error rates, respectively, on Raven's Colored Progressive Matrices (Hall & Russell, 1974). Kagan, Pearson, and Welch (1966) reported that children's response latencies on the MFFT were highly correlated with response latencies on the Haptic Visual Matching Task. These investigators also found that impulsive children had faster response times and higher error rates on verbal reasoning tasks ("guess the item") and on nonverbal reasoning tasks (story completion and series completion). Many similar results were reviewed by Messer (1976) and Kogan (1983), and more are reported in the second section of this chapter.

There is some limit to this consistency, to be sure. The most recent attempt to characterize the relevant domain is that of Kagan et al. (1978)

> The reflection-impulsivity dimension is only applicable to those problem situations where (a) the child believes an aspect of intellectual competence is being evaluated, (b) the child has a standard for the quality of his performance in the task, (c) the child understands the problem and believes he or she knows how to achieve its solution, (d) several equally attractive response alternatives are available, and (e) the correct answer is not immediately obvious and therefore the child must evaluate the differential validity of each potential solution hypothesis. (p. 1004)

This statement is not quite clear about what "available" means. In most research, but not all, alternatives are provided in multiple-choice format. However, it seems reasonable to take "available" to mean something like, "within the child's ability to produce as an answer, given sufficient time." (By this definition, a question about the Swahili equivalent of some English word would have no available response alternatives for most of us.) Assuming this interpretation, (d) and (e) amount to a statement that thinking is involved, that is, (in the sense of Dewey, 1933, and Baron, 1985b) an intentional effort is made to resolve doubt. The essence of (b), we think, is that the child have the same standard as the experimenter. This requirement seems a matter of convenience. Essentially all the research on reflection-impulsivity has used tasks that the experimenter can easily score as correct or incorrect (perhaps in degrees). For all we know, the dimension might apply quite well, say, to artistic creation; however, the problem of scoring the answers would be difficult, because the experimenter and the subject might not agree on the standards to be used. Property (c) essentially excludes cases in which the expectation of the value of thinking is so low that the child gives up immediately. This implicitly acknowledges part of the reinterpretation of impulsivity that we are making here. Property (a) suggests a motivational explanation. One determinant of the value of success is the function of success in promoting self-esteem. It may well be that individual differences in this sort of motivation are major determinants of individual differences in reflection-impulsivity.

Reflectives tend to do better on IQ tests, especially when the test involves response uncertainty and when long latency is not heavily penalized (Messer, 1976). In general, MFFT errors show strong negative correlations with IQ, and depending on the test, latencies show zero or weak positive correlations. Findings of positive correlations between IQ and some measure of latency are of interest, especially since long latency is typically treated as a negative factor in the scoring of the IQ test itself. Yet such findings are common (see Smith et al., 1984; Sternberg, 1984; and results reported later in this chapter). These findings suggest that many people do not perform as well as they could if they were to take more time in intellectual tasks.

Reflection-impulsivity may be modified in a way that transfers to other tasks (e.g., Egeland, 1974; Meichenbaum, 1977). Thus, reflection-impulsivity is at least to some extent dispositional. If people stop too soon, it is not because they have no choice in the matter. Impulsivity also appears to be stable over time (Kagan, 1965; Messer, 1970).

Impulsivity has been thought to decline developmentally. Children tend to become more reflective as they grow older, that is, they increase their response times and decrease error rates on the MFFT (Ault, 1973; Salkind, Kojima & Zelnicker, 1978). Given these results, it might be argued that the negative correlations between impulsivity (or other style dimensions) and the various measures of good performance and adaptation are mediated by a third variable, rate of development. Children who develop more quickly would thus be both better adapted to school and less impulsive. These findings suggest that the former correlation may be mediated by individual differences in rate of development. In the case of impulsiveness, several studies now show individual differences among adolescents and young adults, populations that ought to show minimal effects of differences in rate of development. Cegalis and Ursino (1979), Drake (1970), Lösel (1980), Smith et al. (1984) and Ward (1983), found either a negative correlation between latency and errors (suggesting that differences in these measures are not a result of differences in ability alone), a correlation between impulsivity and some other tasks, or both results.

In some studies (e.g., Block, Block & Harrington, 1974; Wolf, Egelston, & Powers, 1972) latency measures by themselves appeared to be unreliable or invalid as measures of reflection-impulsivity; they failed to correlate with latency in other tasks or with external variables that one might think would be related to reflection-impulsivity. A possible reason for this is that mean latency measures are sensitive to both ability and impulsivity. If we are looking for effects of style on latency, uncontaminated by effects of task ability on latency, it seems reasonable to exclude correct responses from the analysis, because the latency on these responses will be more heavily affected by ability than will the latency of the remaining responses. The latter are drawn from trials on which the subject either gives up or makes an error of commission. In either case, short latencies are likely to result from premature cessation of work rather than high ability at the task. Smith et al. (1984) compared different latency measures in a study of reflection-impulsivity in young adults. They employed two matching tasks that were more difficult than the MFFT and also tested subjects on a number of problem-solving tests: Raven's Matrices, anagrams, and number series completion. Individual differences in reflection-impulsivity were more clearly apparent when correct responses were excluded from the latency measures than when mean latencies over all trials were used. For example, the median correlation between latency in one task and error rate in another was .45 when correct responses were excluded, and .35 when all responses were included. The median correlation of latency in one task with latency in another was .45 for the measure excluding errors and .34 for the measure including all responses. (In the matching tasks, the subject had no way of deciding that two stimuli were the same except by failing to find a difference; thus, all same responses were counted as giving up and were thus not considered correct. This was also true of giving-up responses in the anagram task, where the subject knew that some anagrams were insoluble.)

One of the nice things about reflection-impulsivity, as compared to other styles, is that it is defined clearly enough for an investigator to design new measures of it and be sure that he is studying the same dimension as have others (provided that the new measure is reliable). If the new measure fails to correlate with other measures, we would say that the style is limited in its generality, not that the new measure is poor. In the absence of such a definition, the study of a style becomes the study of the correlates of a test (or set of tests) used traditionally to measure it, and the test can never be improved. The style then becomes no more than a reification of the test scores themselves, hence of no possible theoretical interest. The advantages of clear definition are not widely appreciated. Many investigators assume that impulsivity is defined in terms of the MFFT, despite the fact the definition of the style can be used to decide whether any test, including the MFFT itself, is a good measure of that style. For example, Zelnicker and Jeffrey (1976) have suggested that reflection-impulsivity is better described as a preference for an analytic strategy. Their results indicate that indeed this is what the MFFT tends to measure, but this fact should be taken as a criticism of the MFFT as a sufficient measure of the dimension, not (necessarily) as an empirical finding about the dimension itself. (Similar considerations apply to the results of Block et al., 1974.)

Use of Normative Models

We feel that the definition and measurement of cognitive styles can be improved by more direct consideration of normative questions, that is, questions about what is optimal or best for each subject in each task. For example, in the case of reflection-impulsivity, a simple interpretation of the literature is that many subjects are *too* impulsive, they stop thinking too soon. This sort of consideration leads to two types of inquiry. One is the examination of training. If training a child to think longer leads to a decrease in error rate, and if we can assume that the decreased error rate is worth the cost in extra time, we can conclude that the child was truly impulsive, in the sense of thinking too little, before training. If we can devise measures that distinguish children who can benefit from training in this way from those who cannot, we can say that these measures correlate with impulsiveness in this sense. This is the approach we illustrate later.

The second type of inquiry involves testing our interpretation directly by comparing a subject's performance to a normative model that specifies just when the subject ought to stop thinking in a given task. One way to do this is to define the optimum stopping point, for a given subject performing a given task, in terms of an expected utility model (see Baron, 1985b, for discussion of the concept of utility as used here). Optimal performance involves spending the amount of time working on a problem that maximizes the expected utility of a response, that is, roughly, how much good it can be expected to produce. A subject should continue thinking just as long as he expects thinking to do some good; when the expected utility of continuing becomes negative, thinking should stop. To respond "too soon," means to spend less than the optimal amount of time as defined by a normative model. Impulsive subjects are those whose cognitive style is suboptimal in the sense that their expected utility of responding would be increased by spending more time. A subject who responds quickly and makes many errors will not necessarily be impulsive by this criterion; it simply may be that further thinking would not do enough good for it to be worthwhile. Of course, a subject could have a suboptimal style if he spends too much time on each problem. The normative model can tell us when a person persists too much or too little. It's spirit is very much as W. C. Fields put it (quoted by Janoff-Bulman and Brickman, 1982, p. 218): "If at first you don't succeed, try, try, again. If you still don't succeed, quit. No use being a damn fool about it."

More generally, the use of normative models in the study of cognitive style can specify an optimal point on a style continuum, and subjects may be classified for each task according to which side of the optimum they are on. A major question in this kind of research is whether most people are on one side; for example, whether they are impulsive rather than reflective, in a given type of situation. If so, we should ask whether this is true for all tasks or whether deviations from the optimum are the result of unresponsiveness to relevant properties of tasks. From our perspective, the study of individual differences in cognitive style should be concerned with differences in the degree of deviation from the optimum point rather than differences in position on the continuum.

Normative models have an additional, descriptive value in specifying a set of parameters, and a set of formulas for combining them, that describe optimal performance for a given task. For example, relevant parameters for the study of impulsiveness are the cost of thinking per unit time, the probability that the subject's best response will be correct (a function of time spent thinking and the subject's ability at the task), and the utility of being correct as opposed to incorrect. In appropriate mathematical models, parameters such as these can be part of a formula that specifies the optimum stopping point. We can attempt to describe individual differences in style of performance on the task in terms of these parameters. Even when subjects perform suboptimally, their behavior may be better understood if it can be described as a systematic deviation from some aspect of the normative model. For example, subjects may be insensitive to certain parameters, or they may behave as if they have a systematic bias in estimating them. Alternatively, subjects might behave as if they had used the wrong formula for deciding when to stop thinking about a problem. Of course, subjects do not actually use a formula; rather, the formula describes how the subjects' behavior depends on variations in the parameters. However, for simplicity of language, we shall speak as if subjects did actually use a formula. Any such deviation from the normative model would lead to suboptimal styles. Further, individual differences in the size and direction of such deviations will lead to individual differences along a dimension of cognitive style. Thus, as a descriptive device, a normative model can provide guidance in the search for possible determinants of individual differences in cognitive styles.

Let us elaborate this kind of metholodolgy by considering further the expected utility approach to tasks involving response uncertainty outlined earlier. To simplify discussion, let us suppose that the subject can stop at any one of a number of discrete points in time, for example, the end of every second. Further, we suppose that the subject has an intended response (including "don't know") that can be made at any time he stops. At issue is the choice of stopping at a given point or continuing to the next point. In general, one should continue if the expected utility is positive. The expected utility of continuing is a function of: the cost of continuing, the utility of being correct (as opposed to being incorrect), and the increase in the probability of being correct at the next point, as opposed to stopping and making the intended response. The last parameter, the expected increase in the probability of being correct, usually depends on the subject's confidence in the intended response, the one that would be made at the current point. A subject who is overconfident in his intended response, that is, if his belief in its correctness is irrationally strong, he could underestimate the gain in certainty that could come from further thinking and could therefore respond too quickly. To examine whether impulsiveness is the result of overconfidence in this way, the investigator can either (a) attempt to model subjects' actual behavior by adjusting the relevant parameter in the normative model or (b) elicit subjects' estimates of the parameter directly and test the relationship between misestimation of the parameter and suboptimal performance.

A normative approach to practically any style is possible, as long as the style is well defined. However, it may be necessary to think of styles somewhat differently in order to use this perspective. For example, consider the analytic and dimensional styles discussed earlier, which involve choices among different classifications of stimulus triads. In order to define an optimum point on the continuum of dimensional vs. similarity, for example, we need to have some purpose other than simply doing the task to please the experimenter, because the experimenter would be pleased by either type of classification. For example, we could have the subject learn rules of various sorts, and position on the style continuum would determine the kinds of generalizations the subject made. If the rule were based on similarity, as may be the case for "basic level" concepts such as "cup" (Kemler, 1983), generalization on the basis of similarity would be effective; but, if the rule were based on abstraction along a single dimension, such as "three" (the dimension being numerosity), generalization on the basis of dimensional identity would be necessary. Just as we can move the optimum stopping point by manipulating payoffs, we can also move the optimum position on the style dimension of dimensional-vs.-similarity by varying the type of rule to be learned. Of interest is the failure to adapt optimally to the contingencies we have imposed (as examined by Kemler, 1982a, 1982b).

In sum, normative models are used to determine how subjects ought to perform in a given cognitive task. Subjects' cognitive styles are then defined by comparing how they do perform with how they ought to perform. Normative models are used descriptively to articulate a set of parameters and formulas that are assumed to be operative in cognitive tasks involving uncertainty. Individual differences in suboptimal styles of performance on such tasks can, perhaps, be explained in terms of biased estimates of these parameters and/or use of incorrect formulas.

This approach contrasts with one traditional approach to measurement of style in which the experimenter carefully avoids setting up the task in a way that allows an optimum style to be defined. In this regard, the present approach is similar to the use of ability tests to assess style, for example, the use of the embedded figures test to measure field dependence. However, we do not measure ability directly, in terms of time or errors; rather we measure deviation from optimum on the style dimension of interest. In this way, we can learn about the direction of deviation from the optimum point, even when a deviation in either direction would increase errors. The approach in which no optimum can be defined is based on the assumption that the subject's spontaneous tendency is elicited only under such conditions. Our reply is this: If the "spontaneous tendency" gives way to optimal performance as soon as explicit contingencies are imposed, the style cannot have much consequence outside of very special laboratory situations. If, on the other hand, stylistic factors still affect performance even when contingencies allow an optimum to be defined, the subject's spontaneous tendency can still be measured. Further, measurement under such contingencies might more accurate because the experimenter controls more precisely just which aspects of the situation are ambiguous and which are not.

Models of Optimal Stopping

The measurement of impulsiveness by this method involves comparing behavior to a model of optimal stopping. One such model is Edwards (1965, following Wald), which specifies when one should stop sampling evidence in certain kinds of tasks. For example, suppose the subjects are in a position to observe data from a well-defined data generating machine. They know that the machine is one of two types. One machine can be an urn filled with 70% red balls and 30% blue ones, the other machine, an urn with 30% red and 70% blue, and each datum could be drawing a ball from an urn (and replacing the ball). It is their task to decide which type of machine they are observing. They are paid off in some amount for being correct but must pay a fixed cost for each ball drawn. Thus at each point in time, the subjects must choose between asserting that it is a type 1 machine (hypothesis H1), asserting that it is a type 2 machine (hypothesis H2), or observing another datum from the machine.

A normative subject—one who maximizes the expected value of his decisions—will choose on the basis of the values (payoffs) associated with choosing correct or incorrect hypotheses, the cost of sampling data, and his current probability for the moxt likely, or favored, hypothesis. This probability is revised after each ball is drawn (see Phillips and Edwards, 1966, for a simple exposition of how this should be done). The amount of revision depends on the *diagnosticity* of the evidence. This is a measure of how well each datum distinguishes the two hypotheses: in this particular case; it can be expressed by the ratio p(D/H1)/p(D/H2), where D is the datum (color of the ball drawn), p(D/H1) is the probability of that color if H1 is correct, that is, the probability of drawing that ball from urn 1, and similarly for p(D/H2). In our example, the diagnosticity of a red ball is .7/.3.

One essential insight behind Edwards' model is that the optimal stopping point can be defined in terms of the probability of the favored hypothesis alone, for any given task. Before this cutoff value of the probability of the favored hypothesis is reached, one should continue sampling data. The number of balls drawn so far is irrelevant, except insofar as it affects the current probability of the favored hypothesis. This is because the diagnosticity of the data is fixed, so the change in probability of the favored hypothesis to be expected from drawing another ball is determined only by the probability of that hypothesis before the ball is drawn. This aspect of the model is counterintuitive, and its neglect is similar to a common mistake of naive poker players who, when deciding whether or not to fold, consider how much money they have already put in the pot rather than the amount that might be gained or lost as a result of continued play (Fried and Peterson, 1969).

Edwards (1965) also derives a formula by which we can compute the expected amount of data one will need to sample before the optimum stopping point is reached. It is not necessary to go through the mathematics here. However, one consequence of the formula is of interest: for a given set of payoffs and costs, the optimum amount of data is an inverted U function of the diagnosticity of the data. To get an intuitive sense of why this is so, consider the two extreme cases. When the data are worthless, that is, when the proportion of balls of each color is the same in urn 1 as in urn 2, there is no point in drawing any balls at all. At the other extreme, when diagnosticity is perfect, when one urn has all red balls and the other all blue, it will suffice to purchase only one datum. At intermediate levels of diagnosticity, each datum is valuable but not definitive and one will have to purchase many data.

A number of studies have asked about optimal stopping, as defined by Edwards' model, in bookbag and pokerchip experiments. In such experiments, subjects sample pokerchips from bookbags in order to decide which of two bags of red and blue chips they have been presented with. The distribution of red and blue chips in each bag is known to the subjects, but the particular bags sampled is a matter of chance. Snapper (1971) examined the amount of information required as a function of the diagnosticity of the data. He found that subjects bought more information than they should have when diagnosticity was low, but less than they should have when it was high. In general, it appeared that subjects did show an inverted U curve for amount of evidence sampled as function of diagnosticity, but the peak was shifted to the left (and was also lower than it should have been, in one experiment). In another study, Edwards and Slovic (1965) presented subjects with a complex task in which both payoffs and diagnosticity varied from trial to trial. Some subjects were overcautious, buying more information than they should have, and others were impulsive, buying less than they should have; these individual differences were consistent across different tasks. These experiments must be interpreted with caution because of the nature of the bookbag and pokerchip task, which calls attention to the numerosities of different kinds of events as evidence. For example, Kahneman and Tversky (1972) found that subjects based probability of the favored hypothesis on what amounted to the ratio of red to blue balls they had drawn so far, whereas what is actually relevant is the difference. The ratio might be simply a salient property of numerical evidence. In this case, the subjects seemed to be using incorrect formulas more consciously than they might have in more naturalistic situations.

What is needed is the application of this model, or one like it, to tasks that are more realistic, more like the situations in which impulsiveness is usually found. Badgio and Baron (in research to be reported elsewhere) have used two such tasks. In one, each trial consists of a series of 4 sec. looks at pairs of complex visual patterns. Each series is generated by a hypothetical machine, which makes either all "same" pairs (with identical members) or all "different" pairs. The subjects can look at as many pairs as they like in order to decide whether the machine is a "same" or a "different". They pay (for example) 5 points per look and get 100 points if they correctly identify a machine as "different." In this task, the subject's ability to detect differences corresponds to diagnosticity. (The subjects rarely say "different" falsely; the pairs all appear to be "same" until a difference is found.) If we assume constant diagnosticity throughout each series (an assumption that may be checked), we can calculate an optimal stopping point and compare this to the point at which the subject actually gives up.

In the second task, the subjects solve an anagram problem on each trial. The subjects know that some proportion (e.g., 50%) of the problems have no answer. Subjects pay, say, 2 points per second of work on each problem and get, say, 100 points for answering a soluble problem. Here, the optimum point for giving up on the problem is more difficult to calculate, for diagnosticity, the subjects' probability of solving the problem at time T, given that it has not been solved yet, depends on T. We must thus estimate "diagnosticity" as a function of T before calculating the optimum. (In this task, the subject's thinking probably involves searching for possible answers rather than searching for evidence, in the sense of Baron, 1985b, so the term "diagnosticity" is not really appropriate.)

Possible Determinants of Impulsiveness

Tasks such as these can be used to analyse the determinants of impulsiveness. We consider the determinants of thinking too little, rather than too much, only because we suspect that the former is more prevalent, for reasons we indicate later.

According to our normative model, individuals are impulsive when they (behave as if they) underestimate the value of further thinking relative to its cost. There are several possible sources of this error:

Utilities and Costs. The optimal amount of thinking for a given task depends in part on the utilities of correct and incorrect responses and the cost of thinking. Individuals may behave as if they differ in relevant utilities, so that these differences in utilities determine differences in the time spent thinking. There is a sense (discussed later) in which a person's utilities might be irrational, and a person can be impulsive if his utilities are irrationally different from what they should be. It seems likely that this source of impulsiveness might be found in academic settings. The utility a student places on being correct in a homework assignment, for example, might depend on whether the homework will be graded, but not on the effect of understanding his homework might have on his final grade or its utility for life beyond the course in question. Similar failures could occur in nonacademic settings, for example, in consumer purchases, personal decisions, professional decisions, or decisions concerning the proper stance to take toward issues of public concern. In any of these cases, a person may place too low a utility on making the best decision, relative to the cost of the thinking involved. In sum, one likely source of irrationally premature stopping is reliance on the utilities of the moment, the pain of thinking, the fear of failure, without adequate consideration of the long-term utilities that one would ideally want to apply. This source will more likely lead to impulsiveness than to its opposite because, in general, the costs of thinking are immediate, and the benefits are in the future and hence less effective (Baron, 1985b).

Subjects will also stop thinking too soon if, other things equal, they overestimate the actual cost of the thinking to be done. The real costs associated with thinking include such things as the cost of cognitive effort (Russo & Dosher, 1983; Shugan, 1980) and the cost of actual time spent (e.g., one might prefer to be doing something else). These costs might be poorly estimated one way or another. One source of systematic overestimation of costs might be the association of thinking with experiences of failure, which could make a person anticipate that thinking will be more costly than it actually will be, given a fixed expectation of success.

We must consider the question of what utilities a person should have, for we need to know this in order to say that a person is impulsive because his or her utilities are irrational. To do this question justice would take us considerably beyond the scope of this chapter (see Baron, 1985b), but we suggest the following: expected utilities and costs are defined in terms of the thinker's utilities as they should be on a rational amount of reflection, that is, what they would be if the thinker were acting in accord with a rationally chosen life plan (Rawls, 1971). It is likely that people often act consistently with utilities other than these. Further, (as Rawls points out) rational self-interest alone might not fully take into account the utility of a person's thinking (or other activities) for others. Thus, in general, a teacher may have the right to encourage more thinking than students would want for themselves even after purely self-interested reflection.

Probabilities. In addition to the utilities and cost parameters, estimates of various probabilities are necessary to determine the optimal amount of time one should spend thinking in a given situation. There are two biases in the estimation of probabilities that can lead to impulsiveness: overconfidence in one's favored hypothesis and underconfidence in the effectiveness of one's thinking.

If one overestimates the probability that his favored hypothesis is correct (overconfidence), then he is likely to overestimate the utility of responding without further thought. In the extreme, if one is certain that his favored hypothesis is correct, then he will make his decision without any further thought. (Note that we are not assuming that subjects have numerical probabilities in their heads. Rather, we assume that it is possible for the subject to behave as if affected by some internal parameter analogous to such numerical probabilities.) In fact, a number of studies indicate that people's confidence judgments are generally inaccurate. For example, if we consider cases in which subjects claim to be 100% certain of their answer to a factual question, they may actually be correct 70% of the time or less (see Lichtenstein, Fischhoff, and Phillips, 1982, for a review).

The second way in which subjects' biased estimates of probabilities can lead to impulsiveness is underconfidence in the efficacy of further thought. In terms of Edwards' (1965) model, a subject might act as if the diagnosticity of further evidence were low, even if the same person assigned a high probability to his favored hypothesis. By underestimating the diagnosticity of future evidence, one underestimates the expected utility of further thinking and, therefore, stops thinking too soon.

It may seem paradoxical that one possible determinant of impulsiveness is overconfidence and another is underconfidence. This apparent paradox is resolved when we notice that the confidence applies differently in the two cases. The overconfidence is in the thinking that one has done so far; the underconfidence is in the thinking one would do if one were to continue. Both biases lead to low expectations concerning the utility of further thought.

Failure experiences may affect subjective confidence, subjective diagnosticity, or both. Individuals may differ in what is affected. For some, failure may lower their confidence without lowering their diagnosticity, which would make them work harder the next time. For others, the effect of failure would be the reverse. The same could be said, mutatis mutandis, for success. In fact, Diener and Dweck (1978, 1980) have found individual differences among children in subjective responses to success and failure. Of interest is how these subjective responses affect the parameters of our model.

In general, we would expect these probability biases to operate in the direction of producing impulsiveness rather than overcautiousness. One reason for this (Baron, 1985b) is that biases that produce overcautiousness often correct themselves through experience. A person who spends more time than necessary on a type of problem or decision can learn that the extra time is not doing any good, that his final answer is the same as he would have given much earlier. However, a person who spends too little time cannot know what would have happened if more time had been spent.

Misweighing. A final source of impulsiveness is the failure to weigh correctly the expected utility of thinking against its cost. A person may be fully affected by the utilities of his best plan (or by proxy utilities such as the desire to please Daddy), but may still stop too soon, simply because he has not learned to weigh feelings of expectation against feelings of cost. For example, if a person overweighs cost relative to expectation, that is, if his stopping point is affected by costs more than by expected utilities, then (other things equal) he will stop too soon. However, there is no reason to expect this sort of bias to be on one side of the optimum rather than the other. A person may just as well overweigh feelings of expectation relative to costs and thus spend too much time thinking. Note that misweighing is indistinguishable from certain other biases, such as misestimating expected payoffs by a constant proportion of their actual utility.

Generality. Impulsiveness as defined here might or might not be general across tasks. It might be general across tasks within a certain domain. The extent to which impulsiveness is a general trait depends on the source of the impulsiveness. For example, if impulsiveness results from low subjective diagnosticity, and if this in turn results from certain failure experiences, a person's interpretation of the failure experiences will affect the generality of the impulsiveness. Given the same grade of D in a math course, one student might conclude he is bad at calculus, another that she is bad at school, others that they are bad at everything. The resulting impulsiveness would generalize accordingly.

There are many idiosyncratic kinds of personal histories that would lead to impulsiveness in different situations for different reasons. Hence, we see no reason to expect any orderly principle to govern the domain over which impulsiveness, or any other style, is general. However, because some reasons for impulsiveness will have some generality, there will, on the average, be some correlation between impulsiveness in one situation and impulsiveness in another. This is consistent with the evidence available. The more similar the situations from the perspective of the determinants of individual differences, the higher the correlation. The question of generality of traits is less important than the question of how people might be taught to be unbiased or optimal in a way that will apply across situations. This question is answered most directly by studies of training. General training may be possible even when there is, at present, little generality of individual differences. (We do not mean to suggest that training in cognitive styles is useless if it turns out to be specific to certain classes of situations, such as school work.)

Styles, Intelligence, and Development

We may think of intelligence as a set of traits that can be defined without regard to the situation in which they are measured and that promote the rational formation of plans and their successful execution (Baron, 1985b). Many such traits seem likely to increase with development; thus, the study of intellectual development is in part the study of the development of intelligence. Given this (admittedly broad) view of the nature of intelligence, cognitive styles may qualify as components of intelligence (Baron, 1985a; Sternberg, 1984). A style makes its greatest contribution to intelligent behavior when a person behaves in accord with the normative model for that style. In this section, we consider the relation between impulsiveness, (as an example of a style) intellectual success, and development. In the next section, we consider the relation between impulsiveness and other styles.

Earlier, we discussed the evidence for a relation between impulsivity in the traditional sense—that is, position on the speed-accuracy tradeoff—and other measures. These findings might not hold up when impulsiveness is defined as a deviation from the optimum. For example, it might turn out that correlations between speed of responding and other measures result from the presence of subjects whose optimum latency is short, for example, subjects who would not gain much from further thinking. On the other hand, if individual differences in latency are not largely the result of individual differences in optimum latency, the positive findings of that literature would still hold. (As noted, negative results may be due to the use of poor measures, e.g., including correct responses when computing the latency measure.)

The finding that impulsivity correlates negatively with IQ and school performance suggests that, if we were to make everyone less impulsive, schoolwork would, on the average, improve. This suggests that, by our new criteria, people are generally impulsive. One reason for accepting this suggestion consists of the arguments given earlier to the effect that people ought to be more often impulsive than overcautious (in the absence of compensatory education).

Now let us consider the fact that impulsivity correlates with school performance even when IQ is held constant (Messer, 1976). To begin, note that accuracy and speed are both involved in IQ tests and in schoolwork. In the case of IQ tests, speed is often measured and taken into account according to some sort of formula. In the case of schoolwork, speed is relevant for several reasons: tests are often time limited; a "slow thinker" will have trouble keeping up with what a teacher says; and the time available for homework may in fact be limited both by motivational and practical considerations. In essence, any measure of IQ or school performance reflects a combination of accuracy and speed. At issue is the relative weight of these two in each measure. The simplest interpretation of the evidence is that the relative weight of accuracy and speed is different in schoolwork and in IQ tests. It ought to be possible to improve the power of IQ tests to predict school performance by weighing speed less heavily in their scoring. (And, other things equal, current IQ tests that do not stress speed should be better predictors than those that do.)

One reason that speed may be less important than test designers think is that speed may not be as stable a property of performance of a given task (e.g., reading) as impulsiveness. Most of the tasks done in school, or in work, will ultimately be practiced quite a bit. Practice, as we know, increases speed. Because people may differ in their sensitivity to practice and in the amount of practice they have, a person's ultimate speed on a task may be very hard to predict from speed at an early stage of practice. However, impulsiveness at an early stage of practice may be much more predictive of later performance. First, impulsiveness may be largely unaffected by practice, so that an impulsive person will continue to make many errors even as speed increases. Second, low accuracy in the early stages of learning may be particularly harmful when the learning is a component of some more advanced skill, as the decoding of single words is a component of reading (Baron, 1977). What holds for skills may hold even more for understanding. Errors in understanding the basis of addition and subtraction (e.g., the place system) will become more and more harmful as more advanced subjects are learned. Thus, a student with initially low accuracy may fall further and further behind, whereas a student who is initially careful will acquire a firm basis for later learning, and that student's speed will improve with practice. In sum, then, it seems that test designers (including the teachers and professors who design their own tests) underestimate the benefits of thinking and thus design tests that tend to penalize those students who appreciate how great those benefits can be relative to costs in the long run.

Other results (cited previously) suggest that impulsivity declines with age. We would expect latency to decrease, given the evidence that mental speed itself increases with age (Chi & Gallagher, 1982). Thus, the fact that latency increases at all, even if only a little, is impressive evidence of a decrease in impulsivity, and this decrease may correspond to a decrease in impulsiveness normatively defined (perhaps resulting from educaton). (However, an increase in latency may result from an increase in the true effectiveness of extra thinking, so that age changes in impulsivity, traditionally measured, need not reflect age changes in impulsiveness.) Whatever the reason for this developmental change in impulsivity, its effects on performance of various tasks may be substantial. We simply do not know how much of the literature on intellectual development can be accounted for in terms of impulsivity alone.

For example, many of the tasks studied by Piaget and his followers (see Flavell, 1977) may be affected by impulsiveness. Klahr and Wallace (1970), provided this sort of account of a number of Piagetian classification tasks. They developed a computer model of performance of these tasks. Some of the tasks required many more steps than others. After each step, the model used a parameter called MOTIVE to decide whether to continue. (The name of the parameter seems to reflect a particular account of impulsiveness; however, Klahr and Wallace make no explicit commitment to this account.) Younger and older children are assumed to differ primarily in the strength of MOTIVE. The model accounts for the developmental sequence of the tasks, that is, the order in which the various tasks are first "passed." Of course, as in the case of most models, there are alternative accounts. For example, the more complex tasks may place more burden on working mamory.

More generally, it is at least conceivable that a great many of the tasks mastered in childhood are affected by impulsiveness. Barstis and Ford (1977) found that impulsive kindergarteners performed less well than reflectives in Piaget's tests of conservation of number and quantity. For adults, many conservation tasks seem to be "tricks." We are tempted to give the wrong answer, but we stop ourselves. A child, however, may give the first answer that comes to mind.

Galotti, Baron, and Sabini (in press) presented evidence that individual differences among college students in performance on categorical syllogisms can be accounted for by impulsiveness. In one experiment, good and poor reasoners were selected by a test of performance on such syllogisms as, "All French books are large. Some large books are blue. What can you conclude about the relation between subject and color?" Subjects were asked once again to solve such syllogisms, first giving an initial answer within 20 sec. and then giving a final, cosidered answer. Although the good reasoners were no slower than the poor reasoners at giving the initial answer, they were slower in giving the final one, and the interaction was significant. The good reasoners thus spent more time evaluating their initial answers; they were also more likely to correct errors.

In sum, there is reason to investigate the possibility that developmental trends and individual differences in intellectual tasks can be partially accounted for by impulsivity, in the old sense, and perhaps also by changes in impulsiveness, in the normative sense. This holds for tasks that are taught in school as well as for those mastered naturally. For school tasks, impulsiveness may play a particularly important role, given the fact that impulsiveness in the early stages of learning can have cumulative negative effects, whereas the effects of overcautiousness on latency in the early stages may be largely overcome through practice.

Impulsiveness and Other Styles

Individual differences in impulsiveness may account for the individual differences observed in other style dimensions and their correlations with intellectual success. One example is the analytic style of Kagan et al. (1963). Kagan et al. (1963, 1964) found that analytic responses took longer than nonanalytic ones and that instructing subjects to respond quickly rather than slowly reduced the number of analytic classifications. (Denney, 1972, found no correlation between nonanalytic responding and MFFT impulsivity. However, here and elsewhere, this result could be interpreted as showing that the style is not sufficiently general to affect both tasks. Causal hypotheses about impulsivity are best tested in the tasks of primary interest, in this case, the classification task itself.)

Similar results have been found for the style of dimensional classification (Smith and Kemler, 1977), described earlier. The tendency to classify stimuli dimensionally rather than by overall similarity seems to be related to the time one takes in the classification task. Smith and Kemler Nelson (1984) found that younger children took less time than older children or adults in a free classification task. Further, adults who made predominantly similarity-based classifications took less time than did those who made more dimensional classifications. If adults were forced to speed up in the task, their classifications shift from mostly dimensional to mostly similarity based (also found by Ward, 1983). When adults are told what kind of classification to make, similarity or dimensional, accuracy is higher for dimensional than for similarity classifications at slow speeds, but higher for similarity classifications at high speeds. Making dimensional classifications thus seems to require time (for unknown reasons).

Impulsiveness may also underlie the developmental trend from similarity-based classifications to dimensional ones in free classification tasks. The adult data on free classification under speed instructions may be compared directly to the data from children with no instructions about speed. When this is done, it appears that the proportion of similarity responses the adults make is approximately the same as the proportion made by children who take the same amount of time. Ward (1983) found similar results. Dimensional responding in college students was correlated with long latencies in the task itself and with "reflective" responding in the MFFT. In children, the MFFT did not predict classification performance (poor generality again?), but, once again, dimensional responses were slower than similarity responses, and subjects who made dimensional responses were slower in general. Finally, instructions to respond quickly reduced the proportion of dimensional classifications. In sum, to a first approximation, it appears that the developmental trend in free classification performance may be accounted for entirely by the developmental trend in the amount of time allotted to the task. If so, the empirical facts adduced to argue that dimensional classifications are normatively superior become irrelevant. The same may hold for the analytic style of Kagan et al. (1963). However, we do not know whether children could make dimensional classifications if they were forced to take longer. Conceivably, children forced to try to make dimensional classifications would take much longer than adults. Although children may be impulsive in the old sense, they might not be impulsive in the new sense, for there may be little to be gained from taking extra time.

Finally, let us consider the Phonecian-Chinese dimension described earlier. There are a couple of reasons to think that impulsiveness might predispose a child to read by word specific associations rather than rules. First, as just noted, impulsiveness may be related to use of dimensional (vs. similarity) classifications. And Phonecians appear to be more inclined to classify spoken syllables dimensionally. Treiman and Baron (1981) asked children to say which of the following syllables went together best: BIH, VEH, BO. BIH and VEH are (on the basis of independent ratings) similar overall, and BIH and BO are identical in one "attribute" (namely, the sound of the letter B) but dissimilar overall. Children who were better readers, and, in particular, children who were Phonecians, tended to put together BIH and BO, as opposed to BIH and VEH. The finding that use of rules is correlated with the tendency to analyze speech into letter sounds is consistent with much other evidence (Baron et al., 1980; Baron and Treiman, 1980b). It could be argued that the classification of syllables is unlike other free classification tasks. However, Smith and Baron (1981) found that individual differences in similarity versus letter-sound classification of syllables correlated as highly with individual differences in similarity versus dimensional classification in two other tasks as these individual differences correlated with each other. The two other tasks invoked classification of squares differing in size and brightness and angles differing in angle and side length. Thus, the tendency to classify syllables according to common letter-sounds seems to be an instance of a more general continuum of individual differences in use of similarity versus common attributes. This continuum, in turn, is correlated with individual differences in impulsivity (traditionally measured) and also with the Phonecian-Chinese continuum. There is thus an empirical basis for thinking that Phonecians will be less impulsive.

Phonecians may be less impulsive because the use of rules in early reading is a task in which extra thinking pays off. To see how this could be, consider an oversimplified account of what might go on when a child reads a word either by rules or by use of a word-specific association. For the latter method, the child must do nothing but recall the identity of the word while looking at its letters. In the early stages of learning, memory retrieval does take time, of course, but not that much time. Thus, this method will be relatively insensitive to early cessation of the attempt. When rules are used, the reader must go through several episodes of memory retrieval, perhaps each letter sound or perhaps some relevant analogies (Baron, 1977), as well as some additional thinking about how to use what has been retrieved. The advantage of this extra time is the ability to read words one could not read on the basis of word-specific associations alone and thus the ability to learn specific associations for these words by oneself. The situation is in some ways analogous to that studied by Corbett (1977), who found that memory retrieval based on visual imagery mnemonics was initially slower than rote retrieval, although more effective if sufficient time were allowed.

Summary

Cognitive style measures occupy a middle ground between personality measures and ability measures. Although it is not always a good thing to be at one end of a style dimension, researchers have generally behaved as if they thought one end to be better than the other. For example, there have been no studies of how to make people more impulsive or less analytic.

We have proposed a new approach to the definition and measurement of cognitive styles based explicitly on normative considerations. Rather than asking how a person falls on a style dimension relative to other people, we ask where he falls relative to where he ought to fall. One way to discover where he falls is to move him, by training, and ask whether he is better off as a result. Another way is to specify where he ought to fall in terms of a precise normative model. The advantages of this approach are as follows:

1. The normative approach captures the evaluative nature of cognitive style measures. The evaluative assumptions are made explicit. A subject's style is described in terms of how he deviates from what he ought to do if he wishes to behave optimally. For example, impulsive subjects are those whose fast response times are suboptimal in the sense that the expected utility of responding would be increased by further thought.

2. With traditional measures, subjects can be positioned along style dimensions only relative to other subjects. For a given subject in a given task, it may be adaptive to be at one end of the dimension relative to other subjects; hence, the negative implication of being at this end of the dimension would be inappropriate. For example, subjects with low efficacy in a task but who respond quickly may not benefit from further thought. To call these subjects impulsive would not distinguish their style from that of subjects with comparable response latencies and who *would* benefit from further thought. The normative approach avoids this problem since a subject's behavior in a task is compared to the optimum for that subject and not to the behavior of other subjects.

3. The use of normative models in the study of cognitive styles is helpful in isolating possible determinants of suboptimal styles. The models specify a set of parameters and a set of formulas for combining them that are assumed to be operative in subjects' behavior in a given task. Any error or bias on the part of the subject in estimating these parameters or applying the formulas will, other things equal, lead to suboptimal styles.

4. The discovery of such biases underlying sub-optimal styles could be useful in developing programs for the training of cognitive style. For instance, if a major bias underlying impulsiveness turns out to be overconfidence, then one way to get people to be less impulsive might be to teach them to be less overconfident (Koriat, Fischhoff, and Lichtenstein, 1980).

5. The normative approach may help us to understand the relationships between

styles and other measures, such as school performance, IQ, and indices of cognitive development.

6. Cognitive styles, normatively defined, may constitute components of intelligence in a broad sense of the term. If so, these components may be malleable and hence good targets for educational efforts to increase intelligence.

A TRAINING STUDY

In the study described here, we attempt to modify the cognitive styles of a group of poor readers. Our work is in the tradition of previous studies concerned with the reduction of impulsiveness (e.g., Egeland, 1974; Meichenbaum, 1977; for reviews, see Blackman and Goldstein, 1982; Kogan, 1983 and Messer, 1976). We deal with other styles as well.

Training studies are important within our framework for a few reasons. First, if style training can improve the situation of individuals by standards they would accept on reflection (and thus increase expected utility), we have evidence that their styles were biased before the training, from the normative point of view. Second, if a test of style can pick out those people who will benefit in this way, we have evidence that this test is a valid measure of biased style. In the present study, we show that a teacher questionnaire about impulsiveness predicts who will benefit from training; it thus appears that teachers are sensitive to departures from optimality, normatively defined. Finally, successful training studies argue for the practical value of attempts to improve cognitive style by whatever means are available.

In the setting of this study, the Benchmark School (grades 1–8), students at least one year behind in reading (as assessed by a reading specialist) and with IQs of at least 90 (WISC-R) are taught to read at their level in about four years. However, teachers find that these students often exhibit characteristics other than poor reading that interfere with school achievement. Some of these characteristics are of a sort that could be called deficiencies in cognitive style in the sense we have been discussing. Our discussions with teachers, and several attempts to develop the questionnaire described later, revealed three particularly troublesome styles, which we called impulsiveness, nonpersistence, and rigidity. It seems likely to us that these styles are typical of poor readers or students whose overall school performance is lower than otherwise expected. (See Butkowsky & Willows, 1980; Denney, 1974; Hood & Kendall, 1975; Kagan, 1965, for suggestive evidence on this point.)

The main question we ask is whether a program of general training can modify these styles. The training is general in three senses: (1) it seeks to change styles so that the changes will transfer to novel situations; (2) it seeks to train all deficient styles at once; (3) it combines several different methods that have been effective individually in previous studies. Previous attempts to modify impulsivity (cited previously) have led children to take more time on the tasks used in training. In some studies, training reduces error rates as well. Typically (e.g., Egeland, 1974), training transfers to tasks other than those used in the training itself. We know of no attempts to evaluate training directed at several styles at once or combining several methods. When teachers perceive several deficient styles, often together in individual students, it seems reasonable to set up a program to deal with all these styles, rather than just one. Likewise, when several methods have been effective, it seems reasonable to combine them. Our question is, therefore, whether such a realistic program can have a beneficial effect on styles. The question of which components of training are effective is best answered by a different type of study (e.g., Chapin & Dyck, 1976).

We define impulsiveness, as we did earlier, in terms of the costs and benefits of thinking. The value of success in thinking depends on the value of what is thought about. We assume here that the value of thinking about schoolwork is higher than most students believe it to be, both for society and for the students themselves in the long run. This is consistent with the general approach of the school, which is to prepare students for serious academic work, despite their difficulties in reading.

Rigidity (which plays a smaller role in this study) is, in essence, insensitivity to evidence against a favored possibility, e.g., an answer to a teacher's question. Rigidity is essentially what Baron (1985b) calls belief perseverance, the result of overweighing evidence in favor of a likely possibility and underweighing evidence against it. Nonpersistence, as we define it, is actually a form of impulsiveness, but in the completion of projects rather than single problems.

Experiment 1

The purpose of our first study was to determine whether these three styles, impulsiveness, rigidity, and nonpersistence, could be modified for the better. Our measures of the styles consisted of laboratory tasks and teacher questionnaires. There were three laboratory tasks: a set of arithmetic problems of increasing difficulty, a set of logic problems, and a visual matching task, similar in conception to the MFFT. These tests were designed primarily to measure impulsiveness. A measure of rigidity was included, but it turned out to be of questionable validity. We thus had to rely entirely on the questionnaires for measures of nonpersistence and rigidity. For a measure of academic performance, we were forced to rely entirely on a follow-up measure. (Most work at Benchmark is not graded; instead, students are required to complete every assignment correctly. Further, homework and seatwork varies considerably from class to class. Finally, students perceive achievement tests as severely time limited, so that even those who finish them are likely to have rushed through.)

Our training methods were drawn from recent literature on the training of reflective thinking and the modification of behavior, in particular:

1. Meichenbaum's (1977) methods of cognitive behavior modification, including teaching self-control through self-instruction in the new style.
2. More traditional behavior modification theory, involving the setting of goals and the provision of feedback.
3. The literature on metacognition, particularly its claims that stable and general modification of styles and strategies must be based on understanding of the reasons for the changes (see Brown, Bransford, Ferrara, & Campione, 1983, for a review).
4. Whimbey and Lockhead's (1980) use of talking aloud about one's thoughts while solving problems individually and in pairs.
5. Dweck's (1975) work on the importance of teaching students to attribute failure to lack of effort rather than to stupidity or external factors.

Although our primary purpose was the evaluation of the training program, we also report new data on the validity of our measures. In particular, we believe we have demonstrated for the first time that laboratory measures of impulsiveness correlated with teacher ratings of impulsiveness, and, as noted, the same ratings predict which students will benefit from training. The laboratory measures themselves are novel in that they rely on latencies only for errors, and our results thus validate this new method of measurement.

Method

Overview. The study consisted of the following parts: (a) construction of the teacher questionnaire; (b) completion of the questionnaire by the teacher who knew each student best, at the end of the school year, spring, 1981; (c) selection of experimental and control groups on the basis of the questionnaire; (d) administration of laboratory pretest measures, fall, 1981; (e) eight months of training of the experimental group; (f) completion of posttest questionnaires and administration of posttest laboratory measures, at the conclusion of training, spring, 1982; (g) follow up of graduates of Benchmark. We describe each of these parts in turn.

Questionnaire and Subject Selection. Our purpose was to design a questionnaire that measured styles that could be described in terms of the theoretical framework sketched above, that were perceived by teachers as common problems, and that could be defined in terms of observable classroom behaviors. Each item of the questionnaire described a behavior that is unambiguously desirable (about half the items in each scale) or undesirable, in keeping with our normative approach.

An initial pool of items was first constructed on the basis of a preliminary theoretical framework. This questionnaire was shown to a few teachers, who completed it for a few children and then suggested changes. This process was repeated several times. In the course of revision, we eliminated items concerning styles that were less problematic or more peripheral (taking initiative, caring more about getting the right answer than about understanding it), and items not meeting the criteria just listed. A second-to-last draft of the questionnaire was drawn up, using the format "When (situation), he (behavior)." Answers were on a six-point scale from almost never (0% to 10%) to almost always (90% to 100%). The scales are available from the first author. Each scale was defined as the sum of scores of relevant items, with signs reversed for positive items, so that positive scores indicated less desirable behavior. Items that did not correlate more highly with their own scales than with the other two scale scores were omitted, revised, or reassigned to another scale when it made theoretical sense to do so. Items with low correlations with any scale were also omitted or revised. The best three items from (the male form of) each scale are (with correlations with full scale score in parentheses):

Impulsiveness: when doing seatwork, he proceeds slowly enough to avoid careless errors (.82); when proofreading, he goes so fast that errors are missed (−.81); when answering a question, he takes the time to think through the answer to the question before responding (.83).

Nonpersistence: when given homework, he completes it (.75); when he does not enjoy a task he should complete, he gives up (−.76); when he announces a plan to do something, he carries out the plan to completion (.76).

Rigidity: when presented with convincing evidence or arguments against his answer or opinion, he is willing to change the answer or opinion (.84); when the teacher explains why an answer is wrong, the student continues to maintain that it is correct (−.84); when on the losing side of a dispute that others feel has been settled, he continues to argue (-.88).

The questionnaire was then completed for all 145 students in the school (110 males, 35 females). There were separate male and female forms. The reliabilities of the three scales (alpha) were: impulsiveness, .91; nonpersistence, .89; rigidity, .94. The scales also correlated with each other for the entire sample of subjects: .58 for impulsiveness and nonpersistence, .52 for impulsiveness and rigidity, and .38 for nonpersistence and rigidity. Every item except one correlated more highly with its own scale than with any other scale. This result, the fact that correlations among scales were lower than reliabilities, and other data we present later suggest that the scales measure different, but correlated, traits.

We considered setting up three different training groups, one for each deficient style. We rejected this idea because there were many children who would need training on all three styles according to any criterion of need we could set. In addition, in view of our goal of making the training realistic, we felt that the three styles would most appropriately be trained together, so that their interrelationships could be discussed. Thus, we used a single training condition. Of the 145 students, we eliminated 43 who graduated in 1981 and two others who were taking stimulant medication, leaving 100, from whom we selected the 60 with the lowest (least desirable) total scores. (The reliability of the total score was .94.) Eight of the 60 were girls. The mean age of these 60 subjects was 11.6 years (s.d. = 1.5; range, 8.3 to 14.7), the mean IQ was 109 (s.d. = 11.4; range, 86 to 139), the mean Metropolitan Achievement Test grade-equivalent score for reading (the previous spring) was 5.3 (s.d = 1.8), and the mean score for math was 5.6 (s.d. = 1.5). Experimental and control groups were chosen randomly, except that the questionnaire scores of the two groups were matched as closely as possible. The groups did not differ significantly in age, IQ, or achievement-test scores. Most students came from middle- or upper-middle class backgrounds. All subjects but two (one experimental, one control) were Caucasian.

Laboratory Pretest Measures. Each of the 60 selected subjects was given three laboratory tests, Arithmetic, Logic, and Visual Matching. The Arithmetic and Logic tests each consisted of a series of problems presented on the screen of a PET computer. The Matching test, also presented on the PET, consisted of a series of trials in which the subject was to decide whether two complex visual forms were the same or different. All three tests allowed us to measure latency and accuracy and hence to measure impulsiveness in the sense of high speed and low accuracy, relative to other subjects. In the Arithmetic and Logic tests, the difficulty of problems was adjusted so that each child missed 1/3 of them on the average. The latency measure was latency on errors, given this level of difficulty. For the Matching task, we used the latencies on correct "same" responses. The Matching task was such that "same" responses represented an admission of failure to find a difference; the two stimuli in each pair appeared to be identical from the outset, and differences were found only after search. Thus, again, the "same" latency represented the time before giving up. Measures of accuracy in the three tasks are described later.

In the Arithmetic and Logic tests, there was also a measure of rigidity. Occasionally, the computer would instruct the subject, "Please rethink your answer," as a teacher might to encourage consideration of alternatives. A less rigid subject would take the suggestion seriously and possible change the answer. However, failure to change could also be interpreted as impulsiveness. There was no measure of persistence in the laboratory tasks.

The Arithmetic test consisted of a list of 75 problems, in order of difficulty as judged by Baron and Gaskins. Here is a sample, in order, chosen to illustrate the variety of problems used:

- What number comes after 5?
- If I have 3 apples and 2 peaches, do I have more apples or more fruits?
- Terry has 6 marbles. He gives 2 to Dave. How many does Terry have now?

- John has 3 pencils, and gave 1 to Bill. Bill and John then had the same number of pencils. How many did Bill have to start?
- If I have a lot of blocks that are 3 inches high and a lot that are 5 inches high, can I make a tower that is exactly 11 inches high?
- When it's 3 p.m. in New York, it's 9 p.m. in London. What time is it in London when it's 11 p.m. in New York?
- If the sum of two positive numbers is 10, what is their greatest possible product?

(More difficult problems are underrepresented here, as they were reached by few subjects.) The problems were chosen to be real problems rather than exercises in applying what had been learned. Ideas for some were suggested by Whimbey and Lockhead (1980).

Here are some examples from the 69 Logic problems:

- Paul is taller than Peter. Who is taller?
- None of the blue jars have marbles. Does any blue jar have marbles?
- John is shorter than Bill. Bill is shorter than Susan. Who is tallest?
- If Judy drives to work, it snows. Judy took the train. Did it snow?
- All the large jars and all the brown jars have bolts. Is every jar without bolts brown?
- If my grandfather's granddaughter's uncle is not my uncle, what is he to me?

The tests were given during August and September of 1981, just before the training began. Subjects tested in August (random with respect to experimental vs. control) were brought to school by their parents; other subjects were taken out of classrooms at teacher's convenience. The three tests were usually done in a single session, in the order Arithmetic, Logic, Matching. (For a couple of subjects, tests were postponed because of computer malfunction; one subject was never able to complete the Logic test, and his Arithmetic test was used twice, as explained later.) Baron conducted all the tests, and he was blind to group assignment throughout the study. The subjects and their parents were told that the tests were to find out how thinking changed as a result of being at Benchmark and that the results would not affect placement or other evaluation. Subjects were told to take the tests seriously, and all did so in the sense of completing them without complaint and with occasional expressions of interest.

Before the Arithmetic and Logic tests, the subjects were warned that sometimes the answer might be "can't tell"; this was emphasized before the logic problems. The subjects were told that the problems would start out easy but eventually become "so difficult that even Dr. Gaskins couldn't solve them." It was explained that the purpose of this was to discover how the subjects dealt with difficulty. Subjects were told that the session would take a fixed amount of time.

For the Arithmetic and Logic tests, each trial proceeded as follows: Before each problem was presented, the screen prompted, "Press space for next problem." The experimenter pressed the space when he felt the subject was ready. When the space was pressed, the screen was erased, and, after a 1-second delay, the problem was printed. A timer started when the printing was complete. The experimenter read the problem aloud at a measured pace. When the subject answered, the experimenter pressed the space, which stopped the timer, and then typed in the subject's answer. On alternative error trials, starting with the second, and on every fifth correct trial, starting with the fifth, the computer would print, "Please rethink your answer." (Errors were defined here as mismatches between the answer stored in memory and what was typed in. The rethink suggestion was given on some correct trials as well as on errors, so that the subject could not rely on it as a sign of an error.) After this rethink signal, the original problem remained on the screen. The timer started again, and the experimenter read the rethink suggestion and then typed in the subject's answer as before. (The time to rethink was not included as part of the basic latency measure.) Finally, the computer gave the experimenter a chance to correct typing errors (so that these would not be confused with subject errors) by printing on the screen any responses counted as errors. Only one subject seemed to discover what this meant. Otherwise, the subjects received no feedback except for general encouragement, and they were explicitly told, if they asked or looked puzzled, that the rethink suggestion did not necessarily indicate an error.

Each test began with the easiest problem. After each problem, the overall error rate was computed. If the error rate was less than 33%, and if the last response was correct, the computer went ahead in the list by five to choose the next problem; if the last problem was an error, the computer went back by one. If the error rate was greater than 33%, the computer went ahead by one after a correct response and back by five after an error. If the problem chosen in this way had already been presented, the computer chose the next available (not previously presented) problem after the chosen one, following a correct response, or the next available problem before the chosen one, following an error. If either end of the list of problems was reached, the computer chose the closest available problem to that end.

For the Matching task, on each trial the subject saw two figures on the screen of the computer. The subject was told to press "=" if the figures were the same and "-" if they were not. (These keys were adjacent and in the lower right corner of the keyboard; no subject had difficulty finding them.) The stimuli, used previously by Smith et al. (1984), were designed so that the small differences between figures would be immediately apparent once they were found. Each figure was bounded by a rectanglar gray frame 40 mm high and 32 mm wide (internally). The sides of the frames were 8 mm wide, and the two frames were separated horizontally by 16 mm. Each figure was a four-by-four matrix of 16 rectangles, each 8 mm wide and 10 mm high. Each of these rectangles contained one of 16 possible elements, which could be thought of as arranged in a series. At the beginning of the series was a blank rectangle, next a white vertical stripe covering the leftmost 1/8 of the rectangle, then a stripe covering the leftmost 1/4, then 3/8, and so on up to a filled rectangle. The next element removed the bottom 1/8, then the bottom 1/4, and so on up to a horizontal stripe covering the top 1/8. Thus, elements could differ by as little as a stripe of 1/8 of the height or width of the small rectangle. The left figure of the two was constructed by randomly permuting these 16 elements among the 16 positions. If the right figure was supposed to be different on a given trial, R elements were replaced by elements P places further up in the series (starting over at the end). Usually, R and P were 1, so that there was only one different element out of 16, and the difference was as small as possible. To start the session, however, R was set at 16 and P at 8, the greatest possible change. Only different pairs were presented, and each time the subject responded correctly, R was cut in half until it reached 1 (a difference in 1 element only). Then each time the subject was correct, P was cut in half until it reached 1. As soon as two errors were made, this adjustment process was terminated, and the experiment proper began, with the subject seeing no discontinuity. In the experiment, the probability that the figures were the same was .5. The purpose of the adjustment procedure was to set the difficulty of the task for each subject. As it turned out, all but a couple of subjects reached R = P = 1 before 2 errors were made, and the others seemed to find the task particularly easy. Thus, this adjustment was probably unnecessary. In the posttest, each subject was run under the settings used for him or her in the pretest.

Each of the three tests ended when the total time working on the problems (the sum of all latency measures) exceeded 7 minutes. Thus, more impulsive subjects did more trials. In general, the whole session took about 50 minutes.

Training

The 38 training sessions took place over an 8-month period. The 30 experimental subjects were divided into two groups, one group of 16 to be trained by I.G. and the other, a group of 14 to be trained by B. B. The two trainers took turns choosing students, with a view to maximizing their chances of working effectively with the students in their respective groups.

The first session lasted 40 min. and included all 30 subjects and both trainers. Of the remaining sessions, 22 were 20-30 min., and consisted of groups of 2-4 subjects plus a trainer; 15 sessions were 15 min. meetings of a trainer with a single subject. Following the 21st session, the 30 subjects were observed in their classrooms by two observers. The observers focused on one or two subjects at a time and recorded all apparent instances of impulsiveness, nonpersistence, or rigidity. On the basis of these observations, specific goals were set for each subject, corresponding to the subject's observed weaknesses in style. Goals were subsequently monitored on "goal sheets" by classroom teachers or aides, and subsequent training sessions consisted largely of reviewing each student's progress as evidenced by the goal sheets.

The training emphasized the following concepts:

1. The control of thinking by the thinker, including the roles played by attribution, expectation, perception, attention, and active involvement, and the need to deal appropriately with frustration, ambiguity, and anxiety.
2. The regulation and monitoring of one's basic abilities, knowledge and strategies by: (a) taking time to think; (b) sticking with a problem or task; and (c) considering all the alternatives. These three points served as key phrases emphasized throughout the training.
3. The use of specific strategies such as talking to oneself and using mnemonics.
4. The transferability of strategies and styles to new situations.

The following techniques were used in addition to the monitoring already described:

1. The above concepts were discussed, with a view to indoctrinating the subjects in their importance, and were periodically reviewed.
2. Homework forms on which the three phrases of point 2 were printed were given to the subjects, who were asked to record instances where they consciously applied one of the phrases to a thinking task.
3. Subjects solved practice problems and discussed the methods of thinking used in the problems and elsewhere. Care was taken to avoid overlap with Arithmetic and Logic tasks in terms of the specific strategies that oculd be used and to avoid specific training in any such strategies. (This is in contrast to Whimbey and Lockhead, 1980, which teaches tricks for specific types of problems.)
4. Hypothetical situations from classroom or home were discussed.
5. The trainer modeled talking aloud while solving problems, and the subjects imitated her.
6. Subjects were instructed in mnemonic devices.
7. Subjects discussed an article about the nature of good thinking, written for them by Gaskins. They then wrote an essay of their own about good thinking.
8. Subjects solved problems in pairs, according to the method of Whimbey and Lockhead (1980).

It is possible that any success that the training had can be ascribed as much to the relationship between the trainers and each student as to the training methods themselves. These relationships became especially strong during the individual meetings in the last half of the training. The trainers took the students' side, praising them for doing well, and counseling them about difficulties. It was in this "mentor" role that the trainers were, it seemed to them, most able to deal with the idiosyncratic factors that affect good thinking. (Further details of the training are given by Gaskins and Baron, in press.)

Posttest Method

The posttest consisted of the same laboratory tests as the pretest, and the same teacher questionnaire, except for the following changes.

The lists of problems in the Arithmetic and Logic tasks were revised and reordered. To eliminate differences among problems resulting from the possibility of being correct by guessing, Logic problems were reworded so that the answer was always "yes," "no," or "can't tell." These three alternatives were printed on the screen below each problem. Arithmetic problems that permitted yes or no answers were reworded so that they did not. Finally, the problems were reordered (with the help of pre-test data, when possible) with a view to increasing the strength of relationship between difficulty and order in the list.

The time required for the Arithmetic and Logic tasks was reduced to 7 minutes each. However, instead of timing each problem from the time it was presented (and subtracting an estimate of the reading time), the time was computed from the point at which the experimenter finished reading a problem. He pressed the "space" on the computer at this point. Thus, only the time spent working on the problem after it had been read was counted. The effect of these changes was to create a slight positive correlation between impulsiveness and the total time spent working. The more impulsive subjects answered every question almost immediately and thus had to do many problems to use up their seven minutes.

The questionnaire measure was the same, except that the teacher was asked to put a plus in a blank after each item if the student had improved in that item over the course of the year and a minus if the student had gotten worse. This change allowed us to derive an improvement score (which could be negative) for each of the three scales. We decided in advance to rely primarily on this measure, because we had noticed that questionnaire responses tended to reflect presumed stable traits, so that first impressions have a disproportionate influence. For example, drastic behavior changes (apparent to all) resulting from stimulant medicine were not reflected in the original questionnaire responses. The specific request for information about changes in each item called the teacher's attention to such changes, if they occurred.

In addition, nine experimental and 12 control subjects graduated from Benchmark at the end to the study. Follow-up data on judged performance in new schools were available for 7 of the experimentals and 9 of the controls in this group. These data are described later.

Results and Discussion

Individual differences in the measures. We first asked whether our laboratory measures met the traditional criteria for measure of impulsivity as position on the speed-accuracy tradeoff relative to other subjects. As we argued in the first section, if they do, they may also measure impulsiveness relative to the optimum (unless the observed impulsivity differences are explained by differences in the optimum itself). In particular, we expected latencies (on errors) to correlate positively with each other across tasks, and negatively with accuracy. Accuracy measures should also correlate with each other; however, we might have expected accuracy to be affected by ability in the specific areas as well as by style. Finally, the laboratory measures should correlate with relevant questionnaire measures.

For the Arithmetic and Logic pretests, pretest error latencies were first corrected by subtracting the time it took the experimenter to read each problem. (These were estimated from a single run of simply reading the problems aloud. Greater accuracy was not necessary, because the corrected measures correlated .99 with the uncorrected measures, across subjects.) For the matching task, the latency on correct "same" responses was used. The log of each subject's latency for each task was used as the actual measure to create a more symmetrical distribution of latencies. (One subject did not complete the Logic test, and was assigned a log latency that gave the same z score as his log latency on the Arithmetic test.) Correlations among (log) latency measures for the pretest were .50 for Arithmetic and Logic, .26 for Arithmetic and Matching, and .43 for Logic and Matching; correlations for the posttest were .52, .16, and .45, respectively. A composite measure of latency was computed by adding the log latencies for the three tasks; thus, all three tasks were weighted equally, assuming that impulsiveness affects latency by a constant proportion in different tasks. This composite correlates .30 (p = .01) with IQ (Table 5.1), despite the fact that long response times are penalized in some subtests of the IQ test used, the WISC-R. (Sternberg, 1984, reviews similar results.) In sum, the latency measures seem to be valid measures of a general trait of impulsiveness, because they correlate with each other and with IQ.

We note here that examination of correct-response latencies generally supported our decision (based on the findings of Smith et al., 1984) to use only error-response latencies. Although correct latencies showed the same pattern of correlations with other variables as did error latencies, the former correlations were almost never as high as the latter and often were not significant. One exception was the correct latency for the logic task, which correlated with IQ as highly (.27) as did the error latency (.23). Correct responses in the logic task may have been guesses for the impulsive subjects, or they might have been extensively checked for the others.

The accuracy measure for the matching task was the percentage of "different" responses given a "different" stimulus minus the percentage of "different" responses given a "same" stimulus, essentially a correction for guessing. For the Arithmetic and Logic tests, the measure of accuracy was a measure of how far the subject got in the list of problems, which were arranged in order of increasing difficulty. For the pretest, the measure was the highest problem number the subject solved, such that he was correct on 80% of the problems lower than that number. (This measure was unavailable for the Logic test for five subjects; these were omitted

TABLE 5.1
Correlations relevant to the reliability, stability, and validity of the measures.

A. Correlations based on sum of pretest and posttest measures

Age							
−.27^{d}	IQ						
.20	.30^{c}	latency composite (sum of logs)					
.15	.52^{c}	.57^{c}	accuracy composite (sum of z scores)				
−.07	−.03	−.30^{c}	−.24^{a}	questionnaire impulsivity			
.25^{d}	−.26^{b}	−.06	−.33^{c}	.40^{c}	questionnaire nonpersistence		
.04	.06	.09	−.05	.21	−.03	questionnaire rigidity	
−.10	−.14	−.18	−.20	.40^{c}	.15	−.21	task rigidity

B. Results relevant to stability of measures: correlations of pretest and posttest.

Laboratory measures: pretest, fall 1981; posttest, spring, 1982.

pretest latency					
.53^{c}	pretest accuracy				
.10	.07	pretest task rigidity			
.36^{c}	.43^{c}	.06	posttest latency		
.26^{a}	.49^{c}	.24	.50^{c}	posttest accuracy	
.22	−.07	.31^{c}	.08	.31^{c}	posttest task rigidity

Questionnaire scales: pretest, spring 1981; posttest, spring 1982

pretest impulsivity.					
.28^{b}	pretest nonpersistence				
.04	−.05	pretest rigidity			
.36^{c}	.07	.05	posttest impulsivity		
.19	.41^{c}	−.12	.54^{c}	posttest nonpersistence	
.02	−.11	.47^{c}	.44^{c}	.15	posttest rigidity

All results are for experimental and control subjects only (i.e., excluding the rest of the school).
Significance levels: a, $p<.05$ 1-tailed; b, $p<.025$, 1-tailed; c, $p<.01$, 1-tailed; d, $p<.05$, 2-tailed.

from analysis of accuracy.) For the posttest, 90% was used instead of 80%. These two different cutoffs were chosen because they led to closely equivalent means (see Table 5.2), variances, and ranges, and they both avoided ceiling and floor effects. (Note that these measures would, in principle, correlate with ability even if the problems were randomly ordered; thus, correct ordering of problems by difficulty is helpful, but not crucial.)

The correlations among accuracy measures were .23 for Arithmetic and Logic, −.12 for Arithmetic and Matching, and .19 for Logic and Matching in the pretest, and .40, .20, and −.06 respectively in the posttest. We suspect that the Matching measure was simply unreliable; for some subjects it was based on only a few trials. The correlations between Arithmetic and Logic were both significant, but we suspect that these tests measured somewhat different traits. Pretest Arithmetic accuracy correlated with age (r = .32, p < .01), but Logic did not (r = −.02). On the other hand, pretest Logic correlated with IQ (r = .34,

TABLE 5.2
Results Relevant to Effects of Training.

A. Correlations relevant to effects of training. The group variable is coded as 1 for experimentals, 0 for controls. The group-by-IQ variable, as explained in the text, is the z score of IQ for experimentals, the negative z score for controls; a significant correlation indicates an interaction between group and IQ. All measures listed at the left are change scores expressed as improvement.

	Group	*Group-by-IQ*
Composite latency	.51***	.12
Composite accuracy (z's)	.13	.26**
Questionnaire, all items	.06	−.04
Retrospective questionnaire, all	.38***	.10
Restrospective impulsivity	.33***	.05
Restrospective nonpersistence	.31***	−.07
Retrospective rigidity	.35***	.25*

*, $p<.05$; **, $p<.025$; ***, $p<.01$

B. Group means of measures or major interest, before and after training. For time measures, values presented are antilogs of means of logs. For restrospective questionnaire measures, values are mean number of items changing for the better minus items changing for the worse.

		Experimental		*Control*	
		pretest	*posttest*	*pretest*	*posttest*
Arithmetic latency (sec.)		9.8	18.6	12.1	14.7
Logic latency		6.2	10.2	7.4	6.9
Matching latency		13.8	27.1	16.0	21.5
Arithmetic accuracy (problems)		22.2	24.9	20.8	22.3
Logic accuracy		11.6	18.4	14.3	20.4
Matching accuracy (% correct)		33.7	45.4	41.4	47.0
Retrospective questionnaire	impulsivity		3.83		1.87
	nonpersistence		3.63		1.70
	rigidity		4.07		1.43

$p < .01$), but Arithmetic did not ($r = .19$). (Arithmetic, of course, is taught and can therefore be expected to improve with age. Although logic might be thought to develop naturally, it seems that little development occurs in the ages used here. Some subjects performed remarkably well.) Even though the three measures differed in these ways, there was no reason to think that they would not be affected similarly by training, so a composite accuracy measure was formed using z scores (because the Matching score was not otherwise comparable to the two others). Thus, z scores were computed for each of the three measures for the pretest and for each of the three measures for the posttest. (Each z score is the subject's mean score minus the mean score for all subjects, all divided by the standard deviation across all subjects.) The composite score was the sum of the three accuracy z scores. This composite (summed for pretest and posttest) correlated .52 with IQ and .57 with the composite latency measure (Table 5.1). Thus, accuracy seemed to be affected both by impulsiveness and by general abilities.

The latency and accuracy composites both correlated significantly with questionnaire impulsiveness, as shown in Table 5.1(A). We know of no other demonstration of a correlation between a laboratory measure and a questionnaire measure of impulsiveness. The accuracy composite also correlated with persistence, perhaps because persistence is judged in part by the tendency to complete a task accurately.

Table 5.1(B) shows the results concerning the stability of measures over the 8 months between pretest and posttest (despite the effects of training). The accuracy measure, the latency measure, and the positive correlation between them were all stable over the 8 months between pretest and posttest. The three questionnaire scales were individually stable as well, although the correlations among them were not (perhaps because of restriction of range in the total score).

The laboratory measure of rigidity was the number of times the subject failed to change an error response when asked to "rethink," divided by the total number of error responses, for Arithmetic and Logic combined. Changes to a different error response or to a correct response were included; in both cases, the subject followed the instruction to rethink. This measure did not correlate with the rigidity questionnaire scale, as shown in Table 5.1, but it did correlate with the impulsiveness scale. It seems that the laboratory measure of rigidity and perhaps the questionnaire measure were invalid; discussion of reasons for this would take us beyond the scope of the present chapter.

Training effects

Means of the various scores for the two groups are shown in Table 5.2. Of primary interest are the composite measures of latency and of accuracy. The main results may be expressed as (point biserial) correlations between group membership (experimental vs. control) and relevant change measures. (These correlations are equivalent to tests of group by time-of-testing interaction.) The change measure for latencies was simply the difference between the posttest and pretest scores. Because these scores were based on logs, this measure corresponds to the log of the ratio of posttest to pretest latencies; it is thus a measure of proportional change rather than absolute change. This measure correlated highly ($t = 4.58$, $p < .001$) with Group (experimental versus control). In terms of actual latencies, from the pretest to the posttest, the experimental group increased their latencies by 84%, and the control group, by 16%. This result was significant for all three error latency measures ($t = 3.25$ for Arithmetic, 4.22 for Logic, and 2.13 for Matching). Thus, the training was highly successful in making the subjects take time while solving problems, even in a situation removed from the training itself. (Correct latencies were also increased by training, but this effect was significant only for the logic task; t = .65, 3.24, and 1.37 for the three tasks, respectively.)

Training did not have a significant effect on accuracy, although the effect was in the predicted direction. Three additional analyses (in fact, the only three performed) looked for moderating variables that might influence the effect of training on accuracy. First, we asked whether (pretest) questionnaire impulsiveness would predict who would benefit from training. If it would, there would be an interaction between questionnaire impulsiveness and group membership in predicting improvement in accuracy. To test this, we formed a composite variable, which we called Group-by-Imp. The subjects' group membership was coded as 1 for experimental and − 1 for control. Impulsiveness on the pretest questionnaire was converted to a z score across all subjects; thus, subjects with impulsiveness below the mean for the whole sample received negative scores. The Group-by-Imp composite was the product of these two scores. It was thus high for impulsive experimentals and nonimpulsive controls and low for impulsive controls and nonimpulsive experimentals. If this measure correlates positively with improvement in accuracy, we can conclude that the training improves accuracy more for impulsives than for nonimpulsives. The correlation was in fact .28 (t = 2.37, p < .025, 1 tailed). (Corresponding correlations for rigidity and nonpersistence were .09 and − .07, respectively, both nonsignificant. The correlation between Group-by-Imp and the latency change was .07, which was not significant. The training seems to have induced the nonimpulsive subjects to slow down, even though doing so did not improve their accuracy.) In addition, questionnaire impulsiveness correlated with accuracy change in the experimental group (r = .34, p < .05) but not in the control group (r = − .23, N.S.).

A second possible moderating variable is training section. There was a significant accuracy effect for one of the two training sections, the 16 subjects taught by Gaskins, relative to controls (t(39) = 2.08, p. < .025), but not for the other group. This difference might have been the result of differences in instruction. However, the sections had not been assigned at random, and Barnes's section turned out to have lower I.Q.'s. (t(28) = 3.39, p < .01, two tailed). In fact, there are two reasons to think that IQ might moderate the effect of training on accuracy. First, the high IQ subjects might be better learners, might therefore learn more effectively, or in a more transferrable way, from the training. Second, the (relatively) low-IQ subjects might be unable to make use of the extra time effectively, once they learned to slow down. That is, the low-IQ subjects may not have been truly impulsive, taking normative considerations into account. Hence, the third analysis examined the role of IQ as a moderating variable.

To test this effect, we formed a Group-by-IQ variable, exactly analogous to the Group-by-Imp variable just described. This variable correlated significantly with accuracy change (r = .26, t = 2.01, p < .025, 1 tailed). In addition, IQ correlated .41 with accuracy improvement in the experimental group (p < .02, 1 tailed) but did not correlate in the control group (r = − .04). The difference between the two trainers can be explained in termd of this IQ effect. When IQ is partialled from the correlation between trainer and accuracy change, the correlation falls to .14. (Neither the IQ effect nor the effect of training group can account for the moderating effect of impulsiveness described in the first analysis. Group-by-Imp and Group-by-IQ were correlated .03, and I.G's group was actually slightly less impulsive [3.8] than B.B.'s [5.9].) The possible mediating role of IQ in the effect of training on accuracy provides one way of explaining some of the conflicting results of earlier training studies concerning such an effect.

Which of the two explanations accounts for the IQ effect? Are high-IQ subjects more easily trained, or do they make better use of the extra time they are taught to take? One way of asking this question is to look for a comparable IQ interaction in the latency change measure. Group-by-IQ correlates .12 with latency change. The low value of this correlation supports the view that high-IQ subjects make better use of extra time, but it surely does not establish this.

The results for the questionnaire data are shown in Table 5.2. The retrospective measure (the teachers' indications of the direction of change, if any, on each item), showed significant effects of training for all three scales. Thus, in the eyes of the teachers (which were not blind to training condition, we note), the training was successful for all three styles we tried to improve. It is of interest that Group-by-IQ did not correlate with the retrospective change scores for impulsiveness and nonpersistence but did correlate significantly for rigidity. These results suggest an interpretation of the correlation of Group-by-IQ with the laboratory accuracy measures. Specifically, the high-IQ subjects were originally inaccurate because they were rigid, i.e., unwilling to critize their own initial ideas. The training made them less rigid, hence more accurate. The low-IQ subjects may have remained rigid despite the training. Alternatively, their inaccuracy was not primarily the result of their rigidity. Possibly they were not really impulsive at all in the normative sense. Although they stopped thinking sooner than other students, they may not have done so irrationally, because it was true that, given their knowledge and skill, extra time would not have helped them improve their accuracy. By this account, if they are to be helped, it would not be by training to end impulsiveness, but rather by training in more specific skills and methods.

In summary, we found that our measures of impulsiveness, particularly the latency and questionnaire measures, are valid. We also found evidence that cognitive styles can be modified in general, that is, in a way that transfers to new situations. If we think of impulsiveness as a composite of accuracy and latency (in which impulsiveness decreases with increasing latency or with increasing accuracy), then the increased latency that our training produced in the laboratory tests is real evidence of a change in style. The retrospective questionnaire data provide further evidence of improvement in all three styles we tried to train (although the raters were not blind to group membership). However, for subjects who had low IQ's or who were not rated as impulsive initially, the change in impulsiveness was merely technical, as the increased latency was not accompanied by increased accuracy. Although some low-IQ subjects may appear to be impulsive, they may not be when impulsiveness is defined as stopping too soon. These subjects appeared to be impulsive because they worked quickly, but in fact, taking longer did not do them any good because they could not make good use of the extra time. Likewise, subjects not initially perceived by their teachers as extremely impulsive were not helped by the training; they too slowed down but did not become more accurate. These results call attention to the need to select subjects carefully, and on the basis of normative considerations, for training of the sort we gave. Teachers' ratings are useful in this regard (except possibly for low-IQ subjects, where it is possible that teachers cannot easily distinguish impulsiveness from a student's rational response to lack of ability in the task).

Follow up

Twenty one subjects graduated from Benchmark at the end of the study. As part of the standard procedure at Benchmark, the new teachers of these students were asked to fill out questionnaires concerning the students' behavior and performance. Questionnaires were returned from at least one teacher for 16 students, seven experimentals and nine controls. (On the average, questionnaires were returned from 2.9 classes per student.) Of these, four experimentals and five controls attended private college-preparatory schools, two in each group attended parochial schools, two controls attended public schools, and one experimental attended a private school for students with special needs. The questionnaires asked specific questions about academic performance in each class, e.g., for science, "demonstrates understanding of material through daily work," and "manipulates basic scientific apparatus carefully and appropriately." Each question was answered on a six-point scale from "always or almost always" (scored as 5) to "never or hardly ever" (scored as 0), or from "excellent" (5) to "unacceptable" (0). To obtain a measure of judged academic performance for each student, the mean value of these scores for each class was calculated, and then the mean for each student was calculated across the classes available. For the experimental group, the mean performance score was 3.11 (roughly, "frequently" or "good") and for the controls, it was 2.56 (between "frequently" and "occasionally", or between "good" and "fair"). This difference was not quite significant (t(14) = 1.68). However, when IQ was used as a covariate, the difference between groups was significant (t(13) = 1.97, $p < .05$). (The difference was also significant when the one student at a school for special needs—the student who had the lowest IQ, 87 when last tested—was removed, even without taking into account the IQ of the rest.) In sum, instruction appears to have a lasting effect on judged academic performance, amounting to about half a grade on the six-point scale, despite a move to different schools.

Experiment 2

The following year, the cognitive training program was implemented in a single math class consisting of ten students judged to be performing at the 6th grade level in reading but at only the 5th grade level in math. The purpose of this study was to determine whether effective cognitive-style training could be integrated into the regular curriculum. Effectiveness was once again determined by the use of laboratory measures unrelated to any specific content that was trained (although within the general area of mathematics). The class was taught by Gaskins and another teacher. The methods used were drawn from those used in Experiment 1, but all training was integrated with the material taught in the course. The students were told that they were being taught how to think about math, and they were also told that the methods would be useful elsewhere. The same three principles—take time to think, stick to it, and consider alternatives—were emphasized throughout at every available opportunity. The control group consisted of two different math classes, one matched in reading level (but a year ahead in math) and one matched in math level (but a year behind in reading).

To evaluate the effects of the training, a new laboratory measure was devised, with the idea that it might be sensitive to a number of different cognitive styles, not just impulsiveness. The task consisted of a series of problems of increasing difficulty, presented on a computer. The subject worked alone to minimize social influences. The task involved using three bars of different lengths to make up a fourth bar of given length. The subject could add together any of each of the three bars, and the result of this additon was displayed as the subject typed keys corresponding to each of the three bars. For example, a subject would have to make up a bar 59 units long by adding together bars of 3, 11, or 15 units. The three bars could be added by typing the key J, K, or L, respectively. Possible solutions to this problem were four K's and an L (in any order), that is, $4 \times 11 + 15$, or three L's, a K, and a J. These problems have much in common with the water-jug problems used extensively since the work of Luchins (1942). In Newell and Simon's (1972) terms, they have a well-defined task environment.

Method

Subjects. The ten experimental subjects ranged from 11.2 to 13.9 years old at the beginning of the study, and their IQ's ranged from 90 to 139 (mean = 111, s.d. = 14). The 22 control subjects ranged from 8.9 to 13.9 years old, and their IQ's ranged from 94 to 129 (mean = 111, s.d. = 9). There was one female among the experimentals and eight females among the controls.

Procedure. On each trial of the task, the subjects saw a display with four horizontal bars. The first three were labelled J, K, and L, respectively, from the top. The letter J, K, or L, appeared to the left of each line, separated by a space. The full display read:

USE
- J (first bar) (length of bar)
- K (second bar) (length of bar)
- L (third bar) (length of bar)

TO MAKE
- (target bar) (length of bar)
- (blank space for subject's bar) (length of bar)
- USE 'DEL' to BACK UP,
- USE ? to GIVE UP

Each bar was 4.5 mm wide, which was also the height of the letters. The bars were separated by 4.5 mm from each other or from the text. The number representing the length of each bar, including the bar made by the subject, was printed at its right. All bars were drawn to scale, which varied from problem to problem so that the target bar would fit comfortably within the screen, which was 160 mm wide (excluding an unusable border).

When the subject typed the keys J, K, or L, the corresponding length was added to the bar in the blank space. Pressing "del" removed lengths in the reverse order from which they were typed. Thus, a subject could work on the problem by trial and error, adding lengths, erasing any number of them, and then adding more from that point. If the total length made up by the student exceeded the width of the screen, the end of the bar was printed in gray (a checkered character) instead of white; the actual total was still printed to the right of the bar. If the length of the subject's bar equalled that of the target bar, the trial ended and a pleasing display (an explosion of hearts) appeared. The subject could also end the trial by pressing "?" on the keyboard. In either case, the next problem appeared immediately. The program stored the full sequence of key presses, the time spent examining each problem before any keys were pressed, and the total time spent working on the problem. Subjects were told simply how the trials worked; no special instructions were given about speed, method of scoring, etc. The subjects were told that the experimenter would leave and then return to end the experiment in 15 minutes.

The first four problems were counted as practice. The experimenter watched the subject to make sure that the instructions were understood. Following is a sample of the remaining problems, to give an idea of the range of difficulty:

5. use 1, 19, 25 to make 26
10. use 11, 13, 14 to make 26
15. use 3, 11, 15 to make 59
20. use 9, 15, 19 to make 95
25. use 9, 19, 51 to make 116.

There were a few harder problems, but no subjects get any further than this.

This test was given to each subject once in September and once in May.

Results and discussion

Data from one experimental subject and six controls were completely lost due to a faulty tape recorder; in addition, due to a programming error, only time data were available for three controls. This left nine experimentals (including one female) and 13 controls (including three females) for full analysis, 16 controls (including five females) for analysis of times. Because of the small number of subjects, no analyses of interactions with IQ or questionnaire impulsiveness were attempted.

Three performance measures were derived for problems attempted by each subject in both pretest and posttest, excluding the first four (practice) problems:

- Erasures—the total number of letters deleted (problems with "?" in either pretest or posttest were deleted from both unless these trials contained more erasures in the session with the "?" than in the session without it);
- Perfects—the number of problems solved without any erasures;
- Questions—the number of question marks (giving up without a solution).

A composite performance score was formed by adding the z scores on these three measures, with perfects weighed positively and the other two measures weighed negatively. The first two measures assess the solving of problems "in one's head" and therefore the extent of advance planning, or, we might say, the extent of thinking as opposed to trial-and-error. The question measure asseses persistence; with enough time all problems could eventually be solved by trial and error.

The experimentals outperformed the controls in terms of the change in the composite measure from pretest to posttest ($t(20) = 2.05$, $p < .05$). The changes were significant for two of the components of the composite as well. The erasure score declined from a mean of 61 in the pretest to 52 in the posttest for the experimentals, but increased from 63 to 100 for the controls ($t(20 = 1.91$, $p < .05$ for the group difference in the change from pretest to posttest). The perfect score increased from 8.0 to 9.6 for the experimentals but decreased from 8.8 to 8.5 for the controls ($t(20) = 2.01$, $p < .05$ for the group difference in the change). However, the question score did not decline any more in the experimental group than the control group ($t = 0.40$); the experimentals went from a mean of 1.9 to 1.7, the controls from 1.2 to 1.2. The failure to find an effect here may be indicative of a failure of the training to affect persistence. More likely, it was due to the fact that most subjects were near the floor on this measure in both sessions and therefore could not show much change: five experimentals and eight controls produced 0 or 1 "?" responses in both sessions combined.

We might ask whether the increased tendency of the experimental group to solve the problems in their head was the result of improved ability at math rather than a change in cognitive style. One way to examine this is to look at the latency from the time the problem was presented to the time the first key was pressed. This latency decreased from 23.8 to 13.6 sec. for controls, but from 22.9 to 22.1 for experimentals (based on means of log times). Again, the group difference in the amount of change (based on log times) was significant ($t(23) = 1.98$, $p < .05$). Thus, the controls took less time before responding in the posttest than in the pretest, but the experimentals did not, despite the familiarity of the test. It thus appears that there was a true change in style, and it is reasonable to assume that this change contributed to the change in performance.

In conclusion, we have found in two studies that general training of a sort that could reasonably be provided either as a supplement or as part of a regular school curriculum appears to remedy style deficits in impulsiveness, and probably also in nonpersistence and rigidity. These results (and others like them) indicate that the characteristic postion of children on the speed-accuracy tradeoff is malleable. We have also shown that these styles and the improvements in them can be measured in a variety of ways, including problem tasks in which error latency and accuracy are measured, teacher questionnaires, and more complex problem-solving tasks. Of particular interest is the finding that improvement in accuracy can be predicted from a teacher-questionnaire concerning impulsiveness; this finding suggests that the questionnaire measure is sensitive to impulsiveness normatively defined. We hope that others will be encouraged to experiment with training programs of the sort we used here, in which the emphasis is on the student's conduct of thinking, and with measures of the sort we have used. The results of the follow up of Experiment 1 suggest that such training will ultimately enhance, rather than interfere with, instruction in academic content or in domain–specific thinking heuristics. At the same time, the results of Experiment 1 also suggest that style training should be adapted to the needs of individual students. Further research needs to be directed at the identification of individual difficulties in cognitive style (in the normative sense), and specific remediation of the difficulties identified.

ACKNOWLEDGMENT

The first section is based on a review written by Jonathan Baron and Peter Badgio. The second section is based on a collaborative study undertaken jointly by Irene Gaskins and Jonathan Baron. Several people contributed to that study. Edith "Bonnie" Barnes was the second trainer in Experiment 1. Linda Six was the second math teacher in Experiment 2. Peter Badgio, Edith Barnes, Rebecca Hemphill, Craig Stevens, and Thorne Elliot made helpful comments on an earlier report of that study. We also thank the parents who went out of their way to bring in their children for pretesting.

REFERENCES

Ault, R. L. (1973). Problem-solving strategies of reflective, impulsive, fast-accurate, and slow-inaccurate children. *Child Development, 44*, 259–266.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215.

Baron, J. (1977). Use of orthography in reading and learning to read. In D. LaBerge & S. J. Samuels (Eds.), *Basic Processes in Reading: Perception and Comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron, J. (1979). Orthographic and word-specific mechanisms in children's reading of words. *Child Development, 50*, 60–72.

Baron, J. (1985a). What kinds of intelligence components are fundamental? In J. W. Segal, S. F. Chipman, and R. Glaser (Eds.), *Thinking and learning skills, Vol. 2* (pp. 365–390). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron, J. (1985b). *Rationality and intelligence*. New York: Cambridge University Press.

Baron, J., & Treiman, R. (1980a). Some problems in the study of differences in cognitive processes. *Memory and Cognition, 8*, 313–321.

Baron, J., & Treiman, R. (1980b). Use of orthography in reading and learning to read. In J. F. Kavanagh & R. L. Venezky (Eds.), *Orthography, reading, and dyslexia*. Baltimore: University Park Press.

Baron, J. Freyd, J., & Stewart, J. (1980). Individual differences in general abilities useful in solving problems. In R. Nickerson (Ed.), *Attention and Performance VIII*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron, J., Treiman, R., Freyd, J., and Kellman, P. (1980). Spelling and reading by rules (pp. 763–778). In U. Frith (Ed.), *Cognitive Processes in Spelling*. London: Academic Press.

Barstis, S. W., & Ford, L. H., Jr. (1977). Reflection-impulsivity, conservation, and the development of ability to control cognitive tempo. *Child Development, 48*, 953–959.

Blackman, S., & Goldstein, K. M. (1982). Cognitive style and learning disabilities. *Journal of Learning Disabilities, 15*, 106–115.

Block, J., Block, J. H., & Harrington, D. M. (1974). Some misgivings about the Matching Familiar Figures Test as a measure of reflection-impulsivity. *Developmental Psychology, 10*, 611–632.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In P. H. Mussen (Ed.), *Handbook of Child Psychology, Vol. III*. (Vol. Eds., J. H. Flavell & E. M. Markman). New York: Wiley.

Butkowsky, I. S., & Willows, D. M. (1980). Cognitive-motivational characteristics of children varying in reading ability: Evidence for learned helplessness in poor readers. *Journal of Educational Psychology, 72*, 408–422.

Cegalis, J. A., & Ursino, A. (1979). Cognitive style and recognition memory in young adults. *Journal of Research in Personality, 13*, 119–126.

Chapin, M., & Dyck, D. C. (1976). Persistence in children's reading behavior as a function of N length and attribution retraining. *Journal of Abnormal Psychology, 85*, 511-515.

Chi, M. T. H., & Gallagher, J. D. (1982). Speed of processing: A developmental source of limitation. *Topics in Learning and Learning Disabilities, 2*, 23–32.

Corbett, A. T. (1977). Retrieval dynamics for rote and visual image mnemonics. *Journal of Verbal Learning and Verbal Behavior, 16*, 233–246.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Denney, D. R. (1972). Modeling effects upon conceptual style and conceptual tempo. *Child Development, 43*, 105–119.

Denney, D. R. (1973). Reflection and impulsivity as determinants of conceptual strategy. *Child Development, 44*, 614–623.

Denney, D. R. (1974). Relationship of three cognitive style measures to elementary reading abilities. *Journal of Educational Psychology, 66*, 702–709.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology, 36*, 451–462.

Diener, C. I., & Dweck, C. S. (1980). An analysis of learned helplessness: II. The processing of success. *Journal of Personality and Social Psychology, 39*, 940–952.

Drake, D. M. (1970). Perceptual correlates of impulsive and reflective behavior. *Developmental Psychology, 2*, 202–214.

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology, 31*, 674–685.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology, 2*, 312–329.

Edwards, W., & Slovic, P. (1965). Seeking information to reduce the risk of decisions. *American Journal of Psychology, 78*, 188–197.

Egeland, B. (1974). Training impulsive children in the use of more efficient scanning techniques. *Child Development, 45*, 165–171.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: the appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 552–564.

Flavell, J. H. (1977). *Cognitive Development*. Englewood Clifs, NJ: Prentice-Hall.

Freyd, P., & Baron, J. (1982). Individual differences in acquisition of derivational morphology. *Journal of Verbal Learning and Verbal Behavior, 21*, 282–285.

Fried, L. S., and Peterson, C. R. (1969). Information seeking: Optional versus fixed stopping. *Journal of Experimental Psychology, 80*, 525–529.

Galotti, K. M., Baron, J., & Sabini, J. P. (in press). Individual differences in syllogistic reasoning: Deduction rules or mental models. *Journal of Experimental Psychology: General.*

Gaskins, I. W., & Baron, J. (in press). Teaching poor readers to cope with maladaptive cognitive styles: A training program. *Journal of Learning Disabilities.*

Hall, V., & Russell, W. (1974). Multitrait-multimethod analysis of conceptual tempo. *Journal of Educational Psychology, 66*, 932–939.

Hood, J., & Kendall, J. R. (1975). A qualitative analysis of oral reading errors of reflective and impulsive second graders: a follow-up study. *Journal of Reading Behavior, 7*, 269–281.

Janoff-Bulman, R., & Brickman, P. (1982). Expectations and what people learn from failure. In N. T. Feather (Ed.), *Expectations and actions: Expectancy-value models in psychology* (pp. 207–237). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kagan, J. (1965). Reflection-impulsivity and reading ability in primary grade children. *Child Development, 36*, 609–628.

Kagan, J., & Kogan, N. (1971). Individual variation in cognitive processes. In P. Mussen (Ed.), *Carmichael's manual of child psychology*. Vol. I. New York: Wiley.

Kagan, J., Lapidus, D. R., and Moore, M. (1978). Infant antecedents of cognitive functioning: A longitudinal study. *Child Development, 49*, 1005–1023.

Kagan, J., Moss, H. A., and Sigel, I. E. (1963). Psychological significance of styles of conceptualization. *Monographs of the Society for Research in Child Development, 28* (2, serial no. 86), 73–111.

Kagan, J., Pearson, L., & Welch, L. (1966). Conceptual impulsivity and inductive reasoning. *Child Development, 37*, 583–594.

Kagan, J., Rosman, B. L., Day, D., Albert, J., and Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs, 78*, Whole No. 578.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, 430–454.

Kemler, D. G. (1982a). Classification in young and retarded children: The primacy of overall similarity relations. *Child Development, 53*, 768–779.

Kemler, D. G. (1982b). The ability for dimensional analysis in preschool and retarded children: Evidence from comparison, conservation, and prediction tasks. *Journal of Experimental Child Psychology, 34*, 469–489.

Kemler, D. G. (1983). Exploring and reexploring the issues of integrality, perceptual sensitivity, and dimensional salience. *Journal of Experimental Child Psychology, 36* 365–379.

Klahr, D., and Wallace, J. G. (1970). An information processing analysis of some Piagetian experimental tasks. *Cognitive Psychology, 1*, 358–387.

Kogan, N. (1976). *Cognitive styles in infancy and early childhood*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kogan, N. (1980). Cognitive styles and reading performance. *Bulletin of the Orton Society, 30*, 63–78.

Kogan, N. (1983). Stylistic variation in childhood and adolescence: Creativity, metaphor, and cognitive styles. In P. H. Mussen (Ed.), *Handbook of Child Psychology, Vol. III*. (Vol. Eds., J. H. Flavell & E. M. Markman) (pp. 630–705). New York: Wiley.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Humand Learning and Memory, 6*, 107–118.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art. In D. Kahneman, P. Slovic, and A. Tversky (Eds), *Judgment under Uncertainty: Heuristics and Biases*, (pp. 306–334). New York: Cambridge University Press.

Linton, H. B. (1955). Dependence on external influence: correlates in perception, attitudes, and judgment. *Journal of Abnormal and Social Psychology, 51*, 502–507.

Lösel, F. (1980). On the differentiation of cognitive reflection-impulsivity. *Perceptual and Motor Skills, 50*, 1311–1324.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54*, Whole No. 248.

Meichenbaum, D., & Goodman, J. (1971). Training impulsive children to talk to themselves: A means of developing self control. *Journal of Abnormal Psychology, 77*, 115–126.

Meichenbaum, D. (1977). *Cognitive behavior modification: An integrative approach*. New York: Plenum.

Messer, S. B. (1970). Reflection-impulsivity: Stability and school failure. *Journal of Educational Psychology, 61*, 487–490.

Messer, S. B. (1976). Reflection-impulsivity: a review. *Psychological Bulletin, 83*, 1026–1052.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*, 346–354.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Human Learning and Memory, 9*, 676–696.

Salkind, N. J., & Wright, J. (1977). The development of reflection-impulsivity and cognitive efficiency: an integrated model. *Human Development, 20*, 377–387.

Salkind, N. J., Kojima, H., & Zelnicker, T. (1978). Cognitive tempo in American, Japanese, and Israeli children. *Child Development, 49*, 1024–1027.

Shugan, S. M. (1980). The cost of thinking. *Journal of Consumer Research, 7*, 99–111.

Smith, J. D., & Baron, J. (1981). Individual differences in classification of stimuli by dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1132–1145.

Smith, J. D., & Kemler Nelson, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General, 113*, 137–159.

Smith, J. D., Smith, C. A., & Baron, J. (1984). Reflection-impulsivity and task ability: Measurement and determinants of problem-solving performance. Manuscript. The New School.

Smith, L. B. (1979). Perceptual development and category generalization. *Child Development, 50*, 705–715.

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology, 24*, 279–298.

Snapper, K. J., & Peterson, C. R. (1971). Information seeking and data diagnosticity. *Journal of Experimental Psychology, 87*, 429–433.

Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. *Brain and Behavioral Sciences.*

Treiman, R. (1984). Individual differences among children in spelling and reading styles. *Journal of Experimental Child Psychology, 37*, 463–477.

Treiman, R., & Baron, J. (1981). Segmental analysis ability: development and relation to reading. In T. G. Waller and G. E. MacKinnon (Eds.), *Reading Research: Advances in Theory and Practice*, Vol. 2. (pp. 159–198). New York: Academic Press.

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 103–112.

Whimbey, A., and Lockhead, J. (1980). *Problem Solving and Comprehension: A Short Course in Analytical Reasoning*, second edition. Philadelphia: The Franklin Institute Press.

Widiger, T. A., Knudson, R. M., & Rorer, L. G. (1980). Convergent and discriminant validity of measures of cognitive style and abilities. *Journal of Personality and Social Psychology, 39*, 116–129.

Wolfe, R., Egelston, R., & Powers, J. (1972). Conceptual structure and conceptual tempo. *Perceptual and Motor Skills, 35*, 331–337.

Zelnicker, T., & Jeffrey, W. E. (1976). Reflective and impulsive children: Strategies of information processing underlying differences in problem solving. *Monographs of the Society for Research in Child Development, 41* (5, Serial No. 168).

6 Spatial Ability: An Information Processing Approach to Psychometrics

Patricia A. Carpenter
Marcel Adam Just
Carnegie-Mellon University

INTRODUCTION

Psychologists since the time of Binet have been interested in analyzing intelligence into component abilities in order to analyze and measure the components. It is generally accepted that one aspect of intelligence relates to the ability to process language; it is involved when we read a book, understand a poem, solve a riddle, or write a paper. The verbal aspect of intelligence is widely acknowledged and appreciated, especially by educational institutions, but it is by no means the whole story. Another part of intelligence is the ability to process spatial information; it is used when we interpret a picture, read a blue print, visualize a car route, or understand a diagram. Such spatial thinking plays an important part in the daily work of many engineers, architects, designers, mechanics, graphic artists and scientists. This chapter examines some of the processes in spatial thinking. The particular focus is on individual differences in spatial ability; that is, what makes some individuals better than others at solving spatial problems. This chapter discusses three approaches to this question: psychometrics, information processing, and a synthesis of the psychometric and information processing approaches. Finally, it presents a more general theory of spatial information processing and problem solving strategies.

Spatial problems are those that have a significant amount of spatial information in the original presentation of the problem or in the way a person represents it. These problems tend to be solved by generating a mental representation of a two or three dimensional structure and then assessing its properties or performing a transformation of the representation. These transformations may add or delete