*Memory & Cognition*
*1980, Vol. 8 (4), 313-321*

# Some problems in the study of differences in cognitive processes

JONATHAN BARON and REBECCA TREIMAN
*University of Pennsylvania, Philadelphia, Pennsylvania 19104*

To test whether groups differ in a particular ability, researchers often compare their performance on two tasks: an experimental task that is sensitive to the ability of interest and a control task that measures other influences on the experimental task. A group difference will be reflected in a differential deficit, a greater difference between groups in experimental task performance than in control task performance. Before concluding from such a result that the groups differ in the ability of interest, three methodological problems must be faced. First, a differential deficit may be an artifact of task differences in discriminating power. That is, the experimental task may be more sensitive than the control task to group differences in abilities other than the one of interest. Second, a differential deficit may be an artifact of group differences in familiarity with the stimuli or the task. Third, a group difference in one ability may be due to a difference in some other ability that is more, or less, general than the first. These problems affect research in a number of areas, including cognitive development, psychopathology, learning disabilities, and the theory of intelligence. We discuss some possible solutions to these problems.

Cognitive psychologists have repeatedly been urged to apply their experimental methods to the study of individual and group differences (Cronbach, 1957; Underwood, 1975). This advice is currently being taken in the study of schizophrenia (e.g., Oltmanns, 1978), dyslexia (e.g., Vellutino, Steger, DeSetto, & Phillips, 1975), reading speed (e.g., Jackson & McClelland, 1979), intelligence (e.g., Hunt, 1978; Lyon, 1977; R. Sternberg, 1977, 1979), intellectual development (e.g., Chi, 1978), and other areas. Such studies, ideally, will elucidate the nature and causes of individual differences and will help to answer theoretical questions in cognitive psychology. But, as we shall show, a number of serious methodological problems must be solved before this potential can be realized.

An advantage of modern experimental methods is that they define processes or processing parameters in terms of differences between two conditions. Such comparisons permit us to make inferences about processes that cannot be observed directly in a single task. For example, the time to search a memorized list for a given item can be measured by comparison of the time to search a four-item list with the time to search a one-item list, thus controlling (it is assumed) for the time required to identify the stimulus, produce the response, and so on (S. Sternberg, 1969). A difference between conditions is a face-valid indicator of the existence of the process or parameter in question, since the process or parameter is reasonably defined in terms of such a difference. Moreover, the processes and parameters measured tend to be of a more theoretical, and thus potentially more powerful, character than measures derived from traditional psychometric studies. For example, where traditional IQ tests measure "digit span," a modern cognitive psychologist might be interested in "primary memory capacity," a variable whose role in intellectual functioning is more easily understood.

The methods of cognitive psychology cannot, however, be applied straightforwardly to the study of group differences without attention to methodological problems peculiar to this application. Chapman and Chapman (1973a, 1973b, 1974, 1978) have pointed out some of these problems, with special attention to the study of schizophrenic thought disorder. They show that most studies of thought disorder are methodologically flawed and that when the studies are repeated with proper methodology, the conclusions originally drawn do not hold. We shall argue that the methodological problems discussed by Chapman and Chapman arise in a wide range of studies of group differences and individual differences. We shall also point to other methodological problems that can arise in studies of such differences, and we shall suggest some solutions. We do not undertake to catalogue the methodological errors in extant literature, but we do believe that these errors are both widespread and serious.[1]

## AN ILLUSTRATIVE EXAMPLE

Suppose we want to test the hypothesis that psychologists are more distractable than mathematicians (who, after all, are said to walk into unopened doors while lost in thought). We give mathematicians and

We thank M. Foard, J. Persons, D. Reisberg, E. Spelke, and R. Sternberg for comments on earlier drafts. The work was supported by PHS Grant MH29543 (Jonathan Baron, principal investigator).



0090-502X/80/040313-09$01.15/0

psychologists a choice reaction time task in which they must press a different key in response to each of four digits. On each trial, a randomly chosen digit is presented, and the subject must press the appropriate key. While performing the task, the subject hears (on tape) an AM radio in one ear and a jackhammer in the other. (Ears, groups, and order of tasks are appropriately counterbalanced.) We might be tempted to compare the performance of the two groups on the experimental task (Task E), but we must admit that mathematicians might perform better simply because they are better at the choice task, not because they are less distractable. Thus, we need a control task (Task C) to measure proficiency at the choice task. The obvious control is the same task done in a soundproof booth. We now hypothesize an interaction between groups and tasks. That is, we hypothesize that mathematicians will show a smaller distraction effect (Task E reaction time minus Task C reaction time) than psychologists, or, equivalently, that the superiority of mathematicians to psychologists will be greater in Task E than in Task C. Chapman and Chapman (1973a) call this result a differential deficit.

The search for a differential deficit characterizes much research on group and developmental differences. As another example, consider the use of training procedures to study memory strategies in retardates (Brown, 1974) and young children (Flavell, 1970). Typically, in such studies, children who show no evidence of rehearsal (e.g., no lip movements) in a serial memory task are compared with (older or brighter) children who do apparently rehearse. Both groups are then trained to rehearse and are then retested on the original task. Often, the nonrehearsers improve, both in memory performance and in overt rehearsal, more than the rehearsers, and it is concluded that the nonrehearsers were deficient in spontaneous use of a rehearsal strategy. Again, we have an interaction between groups and tasks. Here, the "experimental" task is the memory task before training is given, and the "control" task is the memory task after training is given. The nonrehearsers show a differential deficit, that is, a larger impairment in the experimental task than in the control task.

We shall now point out some of the problems inherent in differential deficit studies—problems that, if not solved, can prevent researchers from drawing the conclusions they wish to draw. One class of problems has to do with whether there is truly a differential deficit. Differences in the "discriminating power" of experimental and control tasks may produce an interaction between measures and groups even though no true differential deficit exists. Another class of problems concerns the interpretation of a differential deficit. A differential deficit may result from group differences in familiarity with the stimuli or the task. A deficit may be wrongly attributed to a specific ability (e.g., ability to ignore distraction) when it is actually due to a difference in a more general one (e.g., mental resources). Conversely, a deficit may be due to a specific cause (e.g., ability to ignore auditory distraction) when a more general one is sought. We shall discuss these problems and some possible solutions, and we shall conclude with a comment on the study of intelligence, an area in which all the problems come to roost in a flock.

## THE PROBLEMS

### Discriminating Power

Discriminating power is a potential problem whenever we seek a differential deficit of the sort we have described, an interaction between group membership (or some measure of individual differences) and experimental vs. control task. The problem arises most clearly when experimental and control tasks are both sensitive to many variables. (Such group differences in extraneous variables are particularly to be expected in studies of any group with manifest deficiencies, such as schizophrenics, retardates, the very young, the very old, or those with specific disabilities or diseases. In essence, as we shall discuss, it is difficult to pick a control group that is matched to the group of interest in all extraneous variables, so we settle for a control task that measures these variables.) When an interaction between groups and tasks is found, it is possible that both tasks measure the same individual differences variables, but Task E (the experimental task) is more sensitive to them than Task C (the control task). If this is so, we say that Task E has more discriminating power, more power to discriminate individual differences in the abilities that affect the tasks. The larger group difference in E than in C would be due to this difference in discriminating power rather than to differences in the variable of interest. In our example, a spurious differential deficit would be found if mathematicians and psychologists were equally distractable, but mathematicians were better at some other component of the task, and the experimental task was more sensitive to this other component than was the control task.

Differences in discriminating power can arise for different reasons, depending to some extent on the measure of performance used. When the measure is percent correct or mean reaction time and when the tasks differ in difficulty according to the measure, the problem is often one of scaling (Loftus, 1978). By "scale," we mean the hypothetical function that relates a measure to the ability it measures. A problem arises when the slopes of these functions for the experimental and control tasks differ in the ranges of interest. For example, a difference between 250-msec and 300-msec reaction times might reflect a substantial ability difference, whereas a difference between 850 and 900 msec may reflect only a small difference in ability. Similarly, a difference between 50% and 60% correct may not mean the same thing as a difference between 80% and

90%. Reaction time and percent correct are really arbitray measures of underlying abilities. Usually, we can assume that these measures are related only monotonically to the abilities they measure. If Fechner, acting as the deity, had commanded that we should use only arcsin percent and log reaction time, interactions now found would disappear and those not found would appear. If Task E shows a larger group difference in reaction time than Task C, but a smaller difference in log reaction time (or some other transform), what should we conclude? Are we interested in the time itself or in the underlying variable it measures? (Answer: The underlying variable.) "Ceiling effects" and "floor effects" are specific types of scaling problems in which the measure becomes completely insensitive to what it measures in a certain range (e.g., 0% or 100% correct). But the fact that some subjects are not quite at ceiling, so that there is still a little room for improvement, does not show that a measure is as sensitive for these subjects as it is for subjects in a lower range.

Scaling problems are easily solved if certain assumptions can be made and if certain results are found. One assumption (already made) is that the scales for both tasks are monotonic; in this case, "crossover" interactions are interpretable. For example, if psychologists perform better than mathematicians in the control task but worse in the distraction task, we can safely conclude that they are more distractable.

A second assumption is that each scale has positive (or negative) slope throughout its range, in other words, that there are no ceiling and floor effects. In this case, we can interpret an interaction in which two groups are truly equal on one task but unequal on the other. For example, if all scores from an experiment are between 30% and 70% correct, and if other experiments have shown more extreme scores on the same measures, this assumption may be reasonable. However, the usual caution is required in accepting the null hypothesis of no difference between groups on one task.

A third assumption is that both tasks share the same scale and that this scale has increasing slope. For example, we might assume that a fixed difference in ability at a task would correspond to a small difference in reaction times if reaction times were short but a large difference if they were long. This assumption would lead us to expect that the task with slower times would show larger group differences. Interactions in the opposite direction could not be due to scaling problems under this assumption. For a real example, Jackson and McClelland (1979) found that fast and slow readers showed a large difference in reaction time to decide whether two letters (in different type cases) were identical but a small difference in a time to decide whether two complex dot patterns were alike, a task that produced longer reaction times for both groups.

The patterns of results we have described, even under the best of assumptions, do not absolve the researcher of the need to test an interaction statistically. A significant group difference in one task coupled with a nonsignificant result on another does not imply a significant interaction. (This fact would seem hardly worth stating except that it is still frequently ignored.)

Another way to handle scaling problems is to match Tasks E and C so that the measures of performance fall in the same range for both tests. For example, we might use a three-alternative reaction time task in the distraction condition and a four-alternative task in the control condition. This manipulation might bring the two tasks into the same range of reaction times, so that an interaction between groups and tasks could no longer result from a scaling problem. However (as pointed out by Traupman, 1976), such a change might subvert the effort to make the tasks equally sensitive to all variables except the one of interest (distractability). For example, the four-alternative (C) task might be more sensitive to ability to memorize stimulus-response pairs than the three-alternative (E) task with distraction. If psychologists were better memorizers than mathematicians, they would do well on Task C relative to Task E, the same interaction that would be found if psychologists were more distractable.

Chapman and Chapman (1974) would rule out this artifact by showing that a differential deficit disappears when a manipulation of interest (e.g., distraction) is replaced by some other manipulation (e.g., sunglasses). While such solutions may be possible in principle, we feel that there are simpler solutions to problems of scaling and, more generally, problems of discriminating power.

Our proposed solution involves looking for differences in correlation coefficients. We suggest testing the hypothesis that the correlation between Task E performance and group membership is higher than that between Task C and group membership (taking into account the correlation between Tasks C and E). A difference between correlations cannot be due to scaling problems, since the correlation is unaffected by changes in scale.[2] In our example, we would compute the point biserial correlation between group membership (psychologists = 0, mathematicians = 1) and performance in Task E, that between group and Task C, and that between Tasks E and C across all subjects. If the first correlation is higher than the second, then psychologists must be more distractable than mathematicians, since a group difference in distractability is the only factor that can effect the correlation between Task E and group but not that between Task C and group (assuming that Task C is as reliable a measure as Task E, as we shall explain). Acceptance of this hypothesis amounts to rejection of the null hypothesis that the groups differ only in abilities measured by both tasks. By the null hypothesis, the two tasks measure the same variable, which has some correlation with group membership. (There is a substantial literature on the problem of comparing dependent

correlations. Cohen and Cohen, 1975, p. 53, give the most useful formula; Dunn and Clark, 1971, compare various formulas; Williams, 1959, discusses the theory.)

When we compare dependent correlations, we must ask whether the measure of performance for Task C is as reliable as the measure for Task E. That is, we must ask whether the correlation between the obtained score and the true score is as high for Task C as for Task E. (The true score is the expectation of the obtained score over parallel tests given under identical conditions; see Lord and Novick, 1968). If not, the null hypothesis might still be true. In particular, Tasks E and C might measure the same variable, and this variable might correlate with group membership, but Task E might correlate more highly with this variable than does Task C. Thus, Task E might show a higher correlation with group membership, even though there is no real differential deficit. We can rule out this possibility, once we have shown a difference between correlations, by showing that the measure of performance in Task E is no more reliable than the measure in Task C. Task differences in reliability thus represent a second source of task differences in discriminating power.

Comparison of dependent correlations and reliabilities is, we feel, a promising solution to problems of discriminating power. However, other possibilities have been suggested. S. Sternberg (1969) has suggested a strategy for the study of group differences based on his "additive factor" model of reaction time. By this model, reaction time may be decomposed into a series of component processes or stages, whose times add together to yield the total reaction time. We can manipulate factors that affect only a single stage. For example, in a choice reaction time task, sunglasses might affect a perceptual stage, number of alternatives might affect a decision stage, and use of the toes instead of the fingers to make the response might affect a motor stage. If sunglasses increased the reaction time by 100 msec and toes increased it by 200 msec, then sunglasses and toes together should increase it by 300 msec. Such a finding both validates the time scale and supports the hypothesis that the factors affect different stages.

S. Sternberg (1969) has suggested that group membership may be considered a factor (as done by Wishner, Stein, & Peastrel, 1978). If sunglasses had equal effects on psychologists and mathematicians, we could conclude that these groups did not differ in the perceptual stage. If psychologists were less affected than mathematicians by use of toes (i.e., if there were an interaction between groups and use of toes), we could conclude that group membership affects the motor stage. A series of such results, in which group membership consistently interacted with factors affecting a certain stage but not with factors affecting other stages, would indicate that the groups differ in the stage in question. Such a result could not stem from differences in the power of the measure (time) to discriminate differences in general performance, since such power differences would show up as interactions with all factors. (These arguments apply only under the assumptions of the additive-factor method. For a critique of these assumptions, see McClelland, 1979).

Another possible solution to problems of discriminating power is to match groups of subjects on possibly confounded variables. For example, if we could match psychologists and mathematicians on Task C performance, we would not need to worry about differences in discriminating power of the two tasks. A group difference in Task E would have to be due to factors other than those that affect Task C. However, this solution is more difficult to implement than one might first think. To match subjects on a control task, it is not sufficient to pick subjects who get the same score on an initial administration of the task, since such subjects can be expected to regress toward their respective group means on a second administration. (If the reliability of Task C is known, it is possible to compensate for this effect, however; see Lord and Novick , 1968). Furthermore, matching may lead to the selection of subjects atypical of each group with respect to the hypothesis of interest. Mathematicians who are as slow as psychologists at Task C might be the least distractable of all mathematicians, since success in mathematics might require either speed or lack of distractability.

A common solution to this set of problems is to use some sort of statistical control, such as partial correlation or analysis of covariance. Performance on Task C might be partialed out from the group difference in Task E (or, equivalently, used as the covariate in an analysis of covariance). While these techniques are commonly used, they are inadequate unless Task C has perfect reliability (Cohen & Cohen, 1975; Lord, 1969). To see why this is so, assume that the correlation between Task E and group membership, r(E,G), equals .60, that r(C,G) = .60, and that r(E,C) = .36. Further, suppose that Tasks E and C both have reliability .36. Then the partial correlation r(E,G/C) will be positive. But Tasks E and C may still measure the same variables with the same reliability and validity, and these variables may be perfectly correlated with group membership. While it is possible to correct for the unreliability of a measure in partial correlation (Cohen & Cohen, 1975), procedures for inferential statistics on such corrected values are unknown (at least to us and to Cohen & Cohen, 1975). The procedure we recommended earlier, comparison of the correlations r(E,G) and r(C,G), taking r(C,E) into account, may be the only available way to "remove" the effects of group differences in C.

Often, groups are matched on variables other than performance on a control task itself. If successful, such matching would equate the groups on all variables that would affect performance on any Task C that might be used. For example, suppose we are interested in the correlates of reading ability as distinct from other

abilities. We may match good and poor readers on a composite test of other abilities, such as an IQ test or an achievement test. A problem with such matching is that the composition of our groups will depend on the mix of abilities measured by the composite test. For example, groups matched on a nonverbal IQ test will probably differ in many verbal abilities, but groups matched on a vocabulary test will probably not. (This is because reading ability is more highly correlated with other verbal abilities than with nonverbal abilities.) We can thus come to different conclusions about the correlates of reading ability, depending on the test we use for matching. And the choice of this test is ordinarily arbitrary. Worse still, if the matching measure is not a perfectly valid measure of the variables we want to match on, true group differences in these variables may remain after matching. (In the extreme, imagine use of head size as a measure of intelligence. While head size is a reliable measure, it is somewhat invalid, and good and poor readers matched in head size would probably still differ in abilities other than reading.) A solution to these problems is to use a "control" measure to define the variables we are not interested in. Then we ask whether this measure correlates as highly with an experimental measure as does the ability of interest. For example, if we are interested in whether some measure E correlates with a test of reading comprehension (R), we might use a test of comprehension of spoken language (S) as a control. We would look for a higher correlation between E and R than between E and S. [Baron (1979), Baron and Treiman (1980), and Treiman and Baron (1980) use this technique to study ability to use spelling-sound rules.] To make sure that such differences between correlations are not spurious, we must show that the reliability of the control measure is as high as that of the task of interest.

A solution to the problem of differential discriminating power might also be found in the theory of latent traits and item-characteristic curves (Lord & Novick, 1968). We mention this only to call attention to this theory; in fact, we suspect that it is not yet applicable to data from the small samples usually used in research of the kind we are discussing.

In general, we feel that the comparison of dependent correlations and reliabilities is at present the most promising solution to the problem of discriminating power. Techniques are not yet available for significance tests on disattenuated partial correlations.[3] The additive-factor method requires extensive testing of the assumptions behind it for a given application, so it is best confined to domains in which such testing has been done. Likewise, the effort to match tasks in discriminating power requires additional checks, which may be unnecessarily time-consuming. On the other hand, we should point out that the comparison of correlations is conservative, since it may fail to show a differential deficit when there is one. In particular, if group membership is highly correlated with Task C, there may be a lower correlation between group and Task E, even though there is a real group difference in the ability tapped by Task E. The best protection against this problem is to choose groups that do not differ much on Task C.

**Familiarity**

Suppose we find a true differential deficit, one that cannot be ascribed to problems of discriminating power. Before we conclude that we have found an ability that distinguishes our two groups, we must ask whether there are other explanations of the results. One common alternative explanation is that the groups differ in familiarity with the stimuli (or procedures) used in an experiment and that familiarity, in turn, affects the process or parameter of interest. Familiarity with materials seems to affect reasoning (Johnson-Laird, Legrenzi, & Legrenzi, 1972), conservation tasks (Cole & Scribner, 1974, p. 152ff.), memory tasks (Baddeley, 1976; Chi, 1978), choice reaction time (Conrad, 1962), perceptual comparison (LaBerge, 1975), and many other tasks. Familiarity may also affect measure derived from comparison of two tasks. For example, familiarity with the stimuli used in our control task may affect the ability to ignore distraction, as well as the reaction time on the control task itself. Since mathematicians may be more familiar with digits than psychologists are, the mathematicians may appear to be less distractable as a result.

The problem of familiarity is especially pernicious in comparisons of younger and older children. Older children are almost by definition more familiar with everything. We suspect that many results thought to show something about the nature of intellectual development can be accounted for by familiarity effects. For example, increased familiarity with the stimuli used in memory tasks might account for the increased use of memory strategies with age (Flavell, 1970). Familiarity might free resources from identification of the stimuli and allow these resources to be used to decide on and implement a strategy. (One might argue that there is no doubt that strategies develop with age, so our objection has no force. Our reply is this: If there is no doubt, why do experiments?)

The problem of familiarity cannot be solved by equating subjects for exposure to the stimuli: We are concerned not with mere exposure but with the effects of that exposure. Thus, retardates might be less likely than normal controls to rehearse in a memory task because they are "less strategic."

Nor is the problem solved by use of stimuli with which all subjects are highly familiar. Available evidence (Fitts & Posner, 1967) suggests that there may be no measurable asymptote for effects of long-term practice (i.e., familiarity with a task) on reaction time. Even if there is a measureable asymptote, we cannot assume that

all subjects have reached it. And even if they have, the number of resources necessary to do the task may continue to decrease (LaBerge, 1975), and resources left over from one stage of processing may affect the speed of some other stage or process. For example, ability to ignore distraction may depend in turn upon resources left over from stimulus identification, and these resources may depend upon familiarity with the stimuli.

We might attempt to solve the problem of differential familiarity by selecting material totally unfamiliar to all subjects. However, practically any stimulus can be related to something experienced before. And people may differ in ability to find and use such relations (Baron, 1978), as well as in familiarity with old stimuli to which new ones may be related.

A more promising way to solve the problem of familiarity is to find an independent measure of its effects and then to show that this cannot account for the results. For example, Baron, Freyd, and Stewart (1980) found a difference between graduate students (selected for intelligence) and control subjects on a recognition memory test. Two types of items were used: "strong cues," consisting of words presented during an incidental learning phase, and "weak cues," words with most letters missing (e.g., A____E for ANYONE). Graduate students were more likely than controls to recognize the weak cues as parts of words they had seen, but they were less likely than controls to recognize the strong cues. This interaction was taken to show that the students were superior at use of weak cues for retrieval. To rule out an explanation in terms of familiarity, the effects of word frequency on the two types of recognition items were determined. In fact, frequency had an (equal) negative effect on both, so it was argued that frequency could not account for the students' superior performance with weak cues. (The students' inferior performance with strong cues may have been due to their greater familiarity with the words.)

A second way to remove effects of familiarity is to find a measure that is demonstrably independent of such effects. Asymptotic reaction time, if we could estimate it, might be such a measure. For example, Baron et al. (1980) assumed that the time to read a list of words declined as an exponential function of the number of times the list had been read before. They fit exponential functions to each subject's reading times and used the best-fitting functions to estimate asymptotes for each subject at each of five levels of word frequency. For students, the estimated asymptotes turned out to be independent of word frequency, although other parameters of the curves (starting point and decay rate) were strongly affected by frequency.

Alternatively, we might find some transform T of reaction time that could be applied to Task C times and Task E times so that $T(E) - T(C)$ was demonstrably unaffected by practice. In order to use such a transform, we would have to show that the variance across all subjects of (the true scores of) T(C) was as great as that of T(E). If the variance of T(E) was greater, a greater group difference in T(E) than in T(C) could be an artifact of T(E)'s being a more sensitive measure of the variables that affect both tasks.

### The Generality of Abilities

When we find a group difference in a single measure of an ability, we must ask how best to describe the difference. The difference may be due to a more specific ability than the one of interest. For example, psychologists and mathematicians may not differ in general distractability, but rather in susceptibility to auditory distraction. Alternatively, groups may differ in a more general ability, or simply a different ability, than the one of interest. For example, the ability to ignore distraction might be determined by available mental resources, which, in turn, might affect many measures other than measures of distractability.

We could repeat our experiments using a variety of types of distraction (e.g., visual as well as auditory distraction). We would hope to find differential deficits in all these measures. But even given this kind of consistency, we still cannot be sure that the groups differ in a single ability, as opposed to a number of different abilities that just happen to be described the same way. Ability to ignore flashing lights and ability to ignore radios can be described similarly, but we still do not know whether there is a common underlying ability that influences both.

What should we mean when we say that two measures measure the same ability? The traditional answer to this question has been that the same ability is involved to the extent to which the correlation between measures is high (and lower than the correlations with other measures). But this sort of result is a sign of generality, not a definition. It is possible that measures of two different abilities could be highly correlated and that measures of the same ability could show a low correlation (as we shall explain).

We suggest that a definition of generality for abilities should be based upon consideration of how general-ability differences come to exist. There are two ways in which such differences can arise: through learning or through biological limits. (If some abilities are affected by both learning and biological limits, then our arguments for both cases apply to these abilities.)

Note that it is easy to be wrong about the origin of a particular ability. For example, some developmental differences in memory tasks were once thought to be due to developmental changes in biological limits but are now thought to be due to changes in the use of learned strategies (Belmont & Butterfield, 1971; Brown, 1974; Flavell, 1970). Conversely, group differences in strategy use may be due to differences in other unlearned abilities, such as the ability to benefit from practice (affecting the ability to learn strategies) or in available mental

resources (affecting the tendency to use already learned strategies).

In the case of learned abilities, we suggest (following Baron, 1973, 1978) that two abilities are the same to the extent to which one was learned with the help of transfer of learning or transfer of practice from the other. When transfer occurs, the learner must often recognize that the two applications of the ability are similar. That is, he must recall an earlier application when deciding what to do in the new situation. The learner might therefore include the memory of a previous application of an ability in the representation of subsequent applications. Learned abilities are thus the same to the extent to which they have common representations in memory, and such common representations probably arise through transfer.

According to this argument, experiments in transfer are necessary to find out if a learned ability is general—experiments in transfer of learning for new abilities and in transfer of practice for abilities already learned. For example, if we think the ability to ignore distraction is learned and general, we might attempt a transfer-of-practice experiment. If practice at ignoring flashing lights transfers to ignoring radios, we could conclude that a common ability underlies both. (Such an experiment has been attempted by Reisberg, Baron, and Kemler, 1980, who found no evidence for generality.)

Because transfer may be imperfect, the correlation between two measures of the same ability may be low. But if there is any transfer at all, we can still say that the abilities measured are the same, in the sense we have defined.

Transfer experiments may be particularly valuable in studying the development of strategies such as rehearsal. In such experiments, training might be designed to mimic naturally occurring experiences, such as repeating a message. If the strategy transfers to other tasks, we can conclude that a general strategy will be learned from such experiences. This, in turn, may allow us to conclude that the strategy develops naturally. A transfer experiment of this sort might be the best we can do to show that general strategies develop, given the objections made above to other sorts of demonstrations. Ironically, the strongest claims about development may be made without comparison of different ages at all. Even if the assumption that the training procedure is similar to natural experiences proves invalid, we can still conclude that the general strategy is teachable. This, too, is no small conclusion.

Unlearned abilities become general not through transfer, but rather from common biological influences. For example, the hippocampus may affect storage of many kinds of memories. One of the brain's "arousal" systems might have something to do with mental energy or effort (Kahneman, 1973). Thus, the most direct way to find out that an unlearned ability is general is to find its physiological basis. This may not be a pipe dream, for we need not understand the physiology in detail in order to study effects of physiological manipulations on psychological tasks. For example, there are now several cases in which a drug seems to affect one mental ability but not another (e.g., MacLeod, Dekaban, & Hunt, 1978). The next step is to study the generality of such drug effects. Just what is the class of abilities affected by a certain drug? If this class is the same as one that accounts for some kind of group difference (e.g., between schizophrenics and normals), we might have converging evidence for a certain description of the nature of the difference.

## DEFINITION AND MEASUREMENT OF INTELLIGENCE

Our arguments have implications for the study of intelligence. Before we discuss these implications, we must state more clearly what we mean by intelligence. First, we are interested in individual and group differences, not in the (quite legitimate) use of "intelligence" to characterize what is common to all human mental activity. Second, we are interested in general intelligence, that is, in those abilities that affect performance regardless of the content of a particular task (e.g., verbal or spatial). Note that these abilities may be learned (Baron, 1978) or unlearned. We are concerned with performance on tasks that involve acquisition and use of knowledge, both in unanticipated situations, which we might call "problems," and in anticipated ones. And we are particularly concerned with research aimed at identification of the general abilities that affect performance on these tasks. Typically, such research involves comparison in experimental tasks of groups thought to differ in intelligence.

Most of the problems in the study of group differences are present in the study of group (or individual) differences in intelligence. First, it is unlikely that these general abilities can be measured by a single test. Instead, we must use the approach developed in this paper: the use of two tasks, E and C, that differ in sensitivity to the ability we want to measure. This approach allows us to control for extraneous abilities (e.g., specific perceptual and motor skills) that could affect performance in Task E given alone.

Second, we must make sure that individual or group differences do not arise spuriously from differences in the discriminating power of our experimental and control tests. To solve this problem, we may want to compare correlations and reliabilities between group membership and tasks. We would need to show that the experimental task correlates more highly with group membership than does the control task (and that the experimental task is not more reliable than the control task). (See Baron et al., 1980, for an example in which this approach has been used successfully.)

Third, we would want to make sure that differences

in familiarity with stimulus materials do not account for observed differences. To measure unlearned abilities, we might be able to estimate asymptotic performance or to transform performance measures in ways known to make differences independent of familiarity. To measure learned strategies, we could equate subjects on some measure of task performance (e.g., percent correct) by manipulation of stimulus familiarity and then seek differences in a measure of strategy use in the task. If we found group differences in strategy use, we might be able to conclude that these differences were not accounted for by familiarity. In the case in which our control Task C does not require use of the strategy of interest at all, we might be able to show that differences in strategy use are unaffected by extensive practice on Task C alone (which will familiarize the subjects with the stimuli, the procedures, etc.).

Fourth, we would want to show that the differences we find are in abilities that are general. We would want to show that our results hold for several different measures of the same ability. Also, depending on the type of ability in question, we would do transfer experiments, or we would look for biological manipulations that affect our measures.

## CONCLUSION

Many of the methodological problems we have discussed have been pointed out by others. These problems have been widely ignored, in part because they were thought to be insoluble and, in any case, immaterial. We acknowledge that for some purposes our criticisms do not apply. For example, when intelligence tests are used to select people for special opportunities, it would be unrealistic to compare performance on an experimental task and a control task. Testees could (once the word got out) intentionally lower their performance on the control task to achieve a high "score." However, when we want to develop a theoretical understanding of the abilities we measure—a goal of current research—the problems we have discussed are relevant. We think these problems are also soluble and that fruitful research on the nature of group differences in mental abilities is a real possibility.

## REFERENCES


BADDELEY, A. D. *The psychology of memory*. New York: Basic Books, 1976.

BARON, J. Semantic components and conceptual development. *Cognition*, 1973, **2**, 299-317.

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

BARON, J. Orthographic and word-specific knowledge in children's reading of words. *Child Development*, 1979, **50**, 60-72.

BARON, J., FREYD, J., & STEWART, J. Individual differences in general abilities useful in solving problems. In R. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, N.J: Erlbaum, 1980.

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

BELMONT, J. M., & BUTTERFIELD, E. C. Learning strategies as determinants of memory deficiencies. *Cognitive Psychology*, 1971, **2**, 411-420.

BROWN, A. L. Strategic behavior in retardate memory. In N. R. Ellis (Ed.), *International review of research in mental retardation* (Vol. 7). New York: Academic Press, 1974.

CHAPMAN, L. J., & CHAPMAN, J. P. *Disordered thought in schizophrenia*. New York: Appleton-Century-Crofts, 1973. (a)

CHAPMAN, L. J., & CHAPMAN, J. P. Problems in the measurement of cognitive deficit. *Psychological Bulletin*, 1973, **79**, 380-385. (b)

CHAPMAN, L. J., & CHAPMAN, J. P. Alternatives to the design of manipulating a variable to compare retarded and normal subjects. *American Journal of Mental Deficiency*, 1974, **79**, 404-411.

CHAPMAN, L. J., & CHAPMAN, J. P. The measurement of differential deficit. *Journal of Psychiatric Research*, 1978, **14**, 303-311.

CHI, M. T. H. Knowledge structures and memory development. In R. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, N.J: Erlbaum, 1978.

COHEN, J., & COHEN, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, N.J: Erlbaum, 1975.

COLE, M., & SCRIBNER, S. *Culture and thought: A psychological introduction*. New York: Wiley, 1974.

CONRAD, R. Practice, familiarity, and reading rate for words and nonsense syllables. *Quarterly Journal of Experimental Psychology*, 1962, **14**, 71-76.

CRONBACH, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, **12**, 671-684.

DUNN, O. J., & CLARK, V. Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association*, 1971, **66**, 904-908.

FITTS, P. M., & POSNER, M. I. *Human performance*. Belmont, Calif: Brooks/Cole, 1967.

FLAVELL, J. H. Developmental studies of mediated memory. In H. W. Reese & L. P. Lipsett (Eds.), *Advances in child development and behavior* (Vol. 5). New York: Academic Press, 1970.

HUNT, E. Mechanics of verbal ability. *Psychological Review*, 1978, **85**, 271-283.

JACKSON, M. D., & MCCLELLAND, J. L. Processing determinants of reading speed. *Journal of Experimental Psychology: General*, 1979, **108**, 151-181.

JOHNSON-LAIRD, P. N., LEGRENZI, P., & LEGRENZI, M. S. Reasoning and a sense of reality. *British Journal of Psychology*, 1972, **63**, 395-400.

KAHNEMAN, D. *Attention and effort*. Englewood Cliffs, N.J: Prentice Hall, 1973.

LABERGE, D. Acquisition of automatic processing in perceptual and associative learning. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V*. London: Academic Press, 1975.

LOFTUS, G. R. On interpretation of interactions. *Memory & Cognition*, 1978, **6**, 312-319.

LORD, F. M. Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 1969, **72**, 336-337.

LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley, 1968.

LYON, D. R. Individual differences in immediate serial recall: A matter of mnemonics? *Cognitive Psychology*, 1977, **9**, 403-411.

MACLEOD, C. M., DEKABAN, A. S., & HUNT, E. Memory impairment in epileptic patients: Selective effects of phenobarbital level. *Science*, 1978, **202**, 1102-1104.

MCCLELLAND, J. L. On the time-relations of mental processes: A framework for analyzing processes in cascade. *Psychological Review*, 1979, **86**, 287-330.

MOSTELLER, F., & TUKEY, J. *Data analysis and regression*. Reading, Mass: Addison-Wesley, 1977.

OLTMANNS, T. F. Selective attention in schizophrenia and manic

psychoses: The effect of distraction on information processing. *Journal of Abnormal Psychology*, 1978, **87**, 212-225.

REISBERG, D., BARON, J., & KEMLER, D. G. Overcoming Stroop interference: Effect of practice on distractor potency. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, **6**, 140-150.

STERNBERG, R. J. *Intelligence, information processing, and analogical reasoning.* Hillsdale, N.J: Erlbaum, 1977.

STERNBERG, R. J. The nature of mental abilities. *American Psychologist*, 1979, **34**, 214-230.

STERNBERG, S. The discovery of processing stages: Extensions of Donder's method. In W. G. Koster (Ed.), *Attention and performance II.* Amsterdam: North Holland, 1969.

TRAUPMAN, K. L. Differential deficit: Psychometric remediation is not acceptable for psychometric artifact. *The Quarterly Newsletter of the Institute for Comparative Human Development*, 1976, **1**, 2-3.

TREIMAN, R., & BARON, J. Segmental analysis ability: Development and relation to reading. In T. G. Waller & G. E. MacKinn (Eds.), *Reading research: Advances in theory and practice* (Vol. 2). New York: Academic Press, 1980.

UNDERWOOD, B. J. Individual differences as a crucible in theory construction. *American Psychologist*, 1975, **30**, 128-134.

VELLUTINO, F. R., STEGER, J. A., DESETTO, L., & PHILLIPS, F. Reading disability: Age differences and the perceptual deficit hypothesis. *Child Development*, 1975, **46**, 487-493.

WILLIAMS, E. J. *Regression analysis.* New York: Wiley, 1959.

WISHNER, J., STEIN, M. K., & PEASTREL, A. L. Information processing stages in schizophrenia. *Journal of Psychiatric Research*, 1978, **14**, 35-44.

## NOTES

1. Furthermore, the problems we raise sometimes arise in kinds of research other than those we discuss. For example, designs based on multiple regression, factor analysis, or analysis of covariance matrices may conceal these problems rather than solve them.

2. Instead of looking for differences between correlations, we might attempt to solve scaling problems by using z scores instead of raw scores for each task. We might try to show that the z score for Task E performance minus the z score for Task C performance is correlated with group membership. While we could find no well-known techniques to assess the significance of such a correlation, a possible technique for this is the "jackknife" method (Mosteller & Tukey, 1977). To use this method, we would compute r', the correlation between z-score differences and group, for the whole sample of N subjects. Then we would delete one subject at a time, Subject i, and compute r'(i), the correlation for the sample excluding the ith subject. Then we would compute "pseudovalues," r(i), of the correlation, as if the correlation of interest were the mean of N pseudovalues, one for each subject, so $r(i) = Nr' - (N - 1)r'(i)$. A t test across the r(i) might tell us whether the correlation of interest is significant.

3. This problem might be solvable by use of the jackknife method (Mosteller & Tukey, 1977). The reliability (see Lord & Novick, 1968) must be recomputed for each of the subsamples resulting from deletion of a subject, however. The computing time may be prohibitive for large samples.

(Received for publication September 24, 1979;
revision accepted January 16, 1980.)