Psychological Bulletin
1983, Vol 94, No. 1,18-38

Behavioral Development and Construct Validity: The Principle of Aggregation

J. Philippe Rushton, Charles J. Brainerd, and Michael Pressley
University of Western Ontario, London, Ontario, Canada

Many important variables in behavioral development are presumed to be unrelated because of repeated failures to obtain substantial correlations. In this article, we explore the possibility that such null findings have often been due to failures to aggregate. The principle of aggregation states that the sum of a set of multiple measurements is a more stable and representative estimator than any single measurement. This greater representation occurs because there is inevitably some error associated with measurement. By combining numerous exemplars, such errors of measurement are averaged out, leaving a clearer view of underlying relationships. We illustrate the usefulness of this principle in 12 major areas of developmental research in which the issue of negligible correlations figures prominently: (a) the validity of judges' ratings, (b) the cross-situational consistency of moral character and personality, (c) the longitudinal stability of personality, (d) the coherence of stages of cognitive development, (e) metacognition, (f) the attitude—behavior relationship, (g) the personality—behavior relationship, (h) the role-taking/altruism relationship, (i) the moral-judgment/altruism relationship, (j) the legitimacy of the construct of attachment, (k) the existence of sex differences, and (1) the assessment of emotionality in animals. In a final section, we also discuss the implications of the principle of aggregation for conducting experimental research.

Imagine how unreliable assessing undergraduates' course performance with a single multiple-choice item would be, or measuring a person's IQ with a single item from a standardized mental test, or making inferences about a population on the basis of results from a single subject. Yet these methods are not unlike what is done in much construct-validation research, especially within the study of behavioral development.

Generally speaking, this article is a methodological critique of the way certain relationships have been investigated in developmental research. Our specific concern is to argue that some lines of developmental research have been hampered by persistent failures to take account of the principle of aggregation. We begin with a brief discussion of this principle in qualitative language. We then review 12 influential areas of developmental research from the perspective of aggregation. Finally, we consider the implication of the principle of aggregation for experimental research. The areas of research we cover are (a) the validity of judges' ratings, (b) the cross-situational consistency of moral character and personality, (c) the longitudinal stability of personality, (d) the coherence of stages of cognitive development, (e) metacognition, (f) the attitude—behavior relationship, (g) the personality—behavior relationship, (h) the role-taking/altruism relationship, (i) the moral-judgment/altruism relationship, (j) the legitimacy of the construct of attachment, (k) the existence of sex differences, and (1) the assessment of emotionality in animals. Evidence is presented that the weak statistical relationships routinely observed in these literatures may be consequences of failures to aggregate.

A preliminary version of this paper was presented at the Merrill-Palmer Society, Detroit, May 1982. We would like to thank Jack Block, John Borkowski, Douglas Jackson, and Linda Siegel for discussions of the issues involved in this paper.

Requests for reprints should be sent to J. Philippe Rushton, Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2.

Aggregation

According to the principle of aggregation, the sum of a set of multiple measurements

is a more stable and unbiased estimator than any single measurement from the set. One reason is that there is always error associated with measurement. When several measurements are combined, these errors tend to average out, thereby providing a more accurate picture of relationships in the population. Perhaps the most familiar illustration of this effect is the rule in educational and personality testing that the reliability of an instrument increases as the number of items increases (e.g., Gulliksen, 1950; Lord & Novick, 1968). For example, single items on the Stanford-Binet IQ test only correlate about .15; subtests based on 4 or 5 items correlate around .3 or .4, but the aggregated battery of items that make up the Performance subscale correlates around .8 with the battery of items that make up the Verbal subscale.

One of the earliest illustrations of the principle of aggregation is the so-called "personal equation" in astronomy. In 1795, Maskelyne, the head of the Greenwich observatory, discharged an otherwise capable assistant because he recorded transits of stars across a vertical hair line in the telescope about half a second "too late." Maskelyne estimated the error of his assistant's measurements by comparing them with his own observations, which he naturally assumed to be correct. An account of these facts in a Greenwich observatory report was noted by a German astronomer, Bessel, some decades later, and led him to test astronomers against each other, with the result that no two agreed precisely on the time of a transit. Clearly, the only sensible estimate of a star's transit across the hairline was some average of many observations, not one.

Personal-equation phenomena were well known in psychology during the nineteenth century. Both psychophysicists and early students of reaction time took considerable pains to remove measurement errors attributable to idiosyncratic differences between subjects. Many early averaging techniques were formulated for this express purpose (Stevens, 1951; Woodworth & Schlosberg, 1939). In everyday life, similar averaging techniques are used in subjective decision making situations. For example, the reliability of decisions about whom to award prizes for cooking, handicrafts, wine making, physical beauty, and so on is enhanced by averaging the decisions of several judges. This procedure is also routine in forms of athletic competition where performance criteria are partially subjective (e.g., diving, gymnastics). When gradations in qualities to be discriminated are fine, the only fair procedure is to obtain many judgments.

Researchers in the psychometric tradition have made similar points. For example, an early paper by Spearman (1910) on the proper use of correlation coefficients contains the following observations:

> It is the superposed accident (measurement error) that the present paper attempts to eliminate, herein following the custom of all sciences, one that appears to be an indispensable preliminary to getting at nature's laws. This elimination of the accidents is quite analogous to, and serves just the same purpose as, the ordinary process of "taking means" or "smoothing curves."
>
> The method is as follows. Let each individual be measured several times with regard to any characteristic to be compared with another. (pp. 273-274)

Unfortunately, Spearman's advice has rarely been taken in many types of developmental research. Although it has been known for decades in both the experimental and differential "halves" of psychology (Cronbach, 1957), it is lost on much of contemporary developmental psychology. Psychologists interested in behavioral development have often assessed constructs such as altruism, role-taking ability, stage of cognitive development, metamemory, attachment, sex differences, and many others using only a single measure. It is not surprising, therefore, that relationships involving these constructs have been weak. When multiple measures of each construct are used, relationships become more substantial.

The Utility of Judges' Ratings

One traditionally important source of data in developmental psychology has been the judgments and ratings of children made by their teachers and peers. In recent years, judges' ratings have been much maligned on the ground that they are little more than "erroneous constructions of the perceive?' (e.g., Kenrick & Stringfield, 1980, p. 88). This view has led to much disenchantment with the use of ratings. The main empirical reason that is cited for rejecting rating methods is

that judges' ratings only correlate, on the average, .20 to .30. However, it is questionable that correlations between any two judges' ratings are stable and representative. The validity of judgments is likely to increase as the number of judges becomes larger.

One early demonstration of the principle of aggregation as it applies to judges' ratings was provided by Knight (1921). Knight's subjects estimated the temperature of a classroom. Although individual estimates were reasonably accurate, ranging from 60° to 80°, the mean rating of 72.4° was virtually the same as the temperature indicated by a thermometer (72°). Shortly thereafter, Gordon (1924) had subjects rank order a series of objects by weight. The rank orderings were then correlated with true ranks, using average rank orderings from different numbers of subjects as the predictor variable. When the number of subjects increased from 1 to 5 to 50, the corresponding validity correlations increased from .41 to .68 to .94.

Similar findings are available in connection with judges' ratings of social variables (Stevens, 1972). For example, Eysenck (1939) gathered judges' estimates of both the weights of objects and the aesthetic value of pictures. He showed that the more estimates that were averaged, the higher was the correlation with the true score for both variables. As Figure 1 illustrates, the function relating judgmental validity to number of judges is essentially the same for highly subjective estimates of the aesthetic value of pictures using an independent criterion (Curve A) and judgments of the weight of objects, for which there was an objective criterion (Curve B).

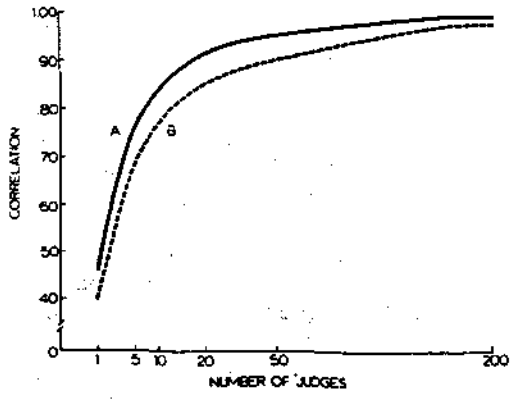

Figure 1. Relation between number of judges (square root) and correlation of their pooled judgments with independent criterion. (A = aesthetic judgment; B = weight judgments. After Eysenck, 1939).

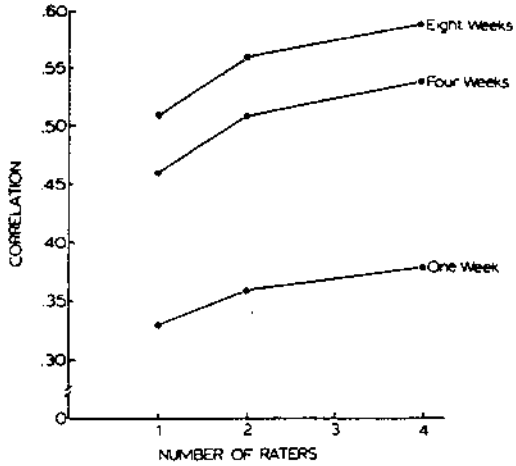

Figure 2. Convergent validity coefficients for dominance based on varying numbers of weeks of observations and varying numbers of raters. (After Moskowitz & Schwarz, 1982).

Another study demonstrating the aggregation effect for judges' ratings of behavioral variables was reported by Moskowitz and Schwarz (1982). These authors varied the number of raters judging a target and the number of observation periods allowed prior to the prediction of actual behavior counts of dominant behaviors (Figure 2). The findings are clear: validity increases directly as the number of raters increases and as the number of groups of observations increases.

Judges' ratings of various dimensions, when assessed reliably by aggregating across judges, have considerable predictive utility. For example, Eron (1980) found that average peer ratings of aggressiveness at age 8 years correlated .43 with the average of a different set of peer ratings of aggressiveness at age 19. Also, those who had been rated as aggressive at age 8 were three times as likely to have police records by age 19 than those not so rated. This suggests that perceptions of personality can be both stable over time and predictive of behavior. Another example is a study by Rushton, Murray, and Paunonen (1983), in which it was found that ratings of university professors, made by their faculty

peers on 29 different personality traits, agreed an average of .56 with an independent set of ratings made by the professors' students. Of even more importance is that these ratings of personality traits allowed significant prediction of the professors' research and teaching effectiveness as measured by independent criteria. Effective researchers, for example, were found to be ambitious, enduring, seeking definiteness, dominant, showing leadership, aggressive, independent, nonmeek, and nonsupportive. Effective teachers, on the other hand, were found to be liberal, sociable, showing leadership, extraverted, nonanxious, and supporting. In summary, if judges' ratings are aggregated over numerous judges, the aggregated score is predictive of behavior and, presumably, more reflective of psychological reality.

Moral Character in Childhood: The Question of Cross-Situational Consistency in Personality Development

For several decades, there have been two opposing viewpoints on the question of whether human behavior is consistent across situations. The classic study of this problem is the enormous "character education inquiry" carried out by Hartshorne and May in the 1920s and published in three books (Hartshorne & May, 1928; Hartshorne, May, & Mailer, 1929; Hartshorne, May, & Shuttleworth, 1930). These investigators gave 11,000 elementary and high school students some 33 different behavioral tests of-altruism (referred to as the "service" tests), self-control, and honesty in home, classroom, church, play, and athletic contexts. Concurrently, ratings of the children's reputations with teachers and classmates were obtained. Altogether more than 170,000 observations were collected. Scores on the various tests were correlated to discover whether behavior is specific to situations or consistent across them. This study will be discussed in some detail because it is the largest study of the question ever undertaken, it raises most of the major points of interest, and it has been seriously misinterpreted by many investigators, as noted by Burton (476), Eysenck (1970), and Rushton (1980). The various tests administered to the children are summarized in Table I.

We first consider the results based on the measures of altruism. Any one behavioral test of altruism correlated, on the average, only .20 with any other test. But when the five behavioral measures were aggregated into a battery, they correlated a much higher .61 with the measures of the child's altruistic reputation among his or her teachers and classmates. Furthermore, the teachers' and peers' perceptions of the students' altruism were in close agreement *(r = .80).* These latter results indicate a considerable degree of consistency in altruistic behavior. In this regard Hartshorne et al. (1929) wrote:

> The correlation between the total service score and the total reputation score is .61. . . . Although this seems low, it should be borne in mind that the correlations between test scores and ratings for intelligence seldom run higher than .50. (p. 107)

Similar results were obtained for the measures of honesty and self-control. Any one behavioral test correlated, on average, only .20 with any other test. If, however, the measures were aggregated into batteries, then much higher relationships were found either with other combined behavioral measures, with teachers' ratings of the children, or with the children's moral knowledge scores. Often, these correlations were on the order of .50 to .60. For example, the battery of tests measuring cheating by copying correlated .52 with another battery of tests measuring other types of classroom cheating. (See, for example, Vol. 1, Book 2, pp. 130-133, Tables 97-99; Vol. 2, Book 1, p. 104, Table 20; Vol. 2, Book 2, pp. 351-352; Vol. 3, pp. 166, Table 19). Thus, depending on whether the focus is on the relationship between individual measures or on the more representative relationship between averaged groups of behaviors, the notions of situational specificity and situational consistency are both supported. Which of these two conclusions is more accurate? •

Hartshorne and colleagues (1928, 1929, 1930) focused on the small correlations of .20 and .30. Consequently,. they argued for a doctrine of specificity:

> Neither deceit nor its opposite, 'honesty' are unified character traits, but rather specific functions of life situations. Most children will deceive in certain situations and not in others. Lying, cheating, and stealing as measured by the test situations used in these studies are only very loosely related. (1928, p. 411)

Table 1
Some of the Measures Used in the Studies in the Nature of Character Investigation

Tests	Nature and scoring of the task
Service tests	
Self-or-class test	Whether the student chose to enter a competition to benefit him- or herself or the class.
Money voting test	Whether the student voted to spend class money on him- or herself or charity.
Learning exercises	Whether the student learned material when performance increments led to money going to the Red Cross.
School-kit test	Number of items donated to charity from a pencil case given to child.
Envelopes test	Number of jokes, pictures, etc. collected for sick children in an envelope provided.
Honesty tests	
Copying technique	Whether student cheated on a test by copying answers from the person next to him or hen
Duplicating technique	Whether student cheated on a test by altering answers after his or her paper had been duplicated without his or her knowledge.
Improbable achievement	Whether student cheated as indicated by an improbably high level of performance on a task.
Double testing technique	Whether students' scores on an unsupervised test (e.g., number of pushups) decreased when a retest was supervised.
Stealing	Whether students stole money from a puzzle-box.
Lying	Whether students admitted to having cheated on any of the tasks.
Self-control tests	
Story resistance tests	Time students persisted in trying to read the climax of an exciting story when words ran into each other.
Puzzle mastery tests	Time spent persisting at difficult puzzles
Candy test	The number of pieces of candy not eaten in a "resisting temptation" paradigm.
Tickle test	The ability to keep a "wooden face"' while being tickled by a feather.
Bad odor test	The ability to keep a "wooden face" while having a bad odor placed under the nose.
Bad taste test	The ability to keep a "wooden face" while tasting unrefined cod liver oil.
Knowledge of moral rules	
Cause—effect test	Agreement to items such as "Good marks are chiefly a matter of luck."
Recognitions test	Agreement that items such as "Copying composition out of a book but changing some of the words" constituted cheating.
Social—ethical vocabulary	Picking the best definition of words denoting moral virtue (e.g. *bravery, malice).*
Free-response foresight test	Students wrote out consequences for transgressions such as "John accidentally broke a street lamp with a snowball."
Probability test	Students ranked the probability of various outcomes for such behaviors as "John 'started across the street without looking both ways."
Reputational ratings	
Recording of helpful acts	For 6 months teachers recorded helpful acts performed by students.
The "guess who" test	Children wrote names of classmates who fitted very short descriptions (e.g., Here *is* someone who is kind to younger children . . .).
Check list	Teachers rated each child on adjectives such as kind, considerate, and stingy.

Their conclusions and data have often been cited in the subsequent literature as supporting situational specificity. For example, Mischel (1968), in an influential review, argued for specificity on the ground that the average correlation between behavioral instances of a "trait" is .20 to .30. According to Mischel (1973), people exhibit "discriminative facility" between situations.

The specificity hypothesis has been useful because it emphasizes that contexts are important and that people have different methods of dealing with different situations. Unfortunately, it has sometimes been interpreted as meaning that cross-situational consistency does not exist, or at least that it does not exist in sufficient quantity to make the concept of traits very useful. This, how-

ever, is quite wrong. By focusing on correlations of .20 and .30 between any two measures, a misleading impression is created. A more accurate picture is obtained by using the principle of aggregation and examining the predictability achieved from a number of measures. To reiterate, this effect occurs because the randomness in any one measure (error variance) is averaged out over several measures, leaving a clearer view of what a person's true behavior is like. Correlations of .50 and .60 based on aggregated measures support the view that there is cross-situational consistency in altruistic and honest behavior.

Further evidence for this conclusion is found in Hartshorne's and May's data. Examination of the relationships between the battery of altruism tests and batteries concerned with honesty, self-control, persi stence, and moral knowledge suggest there may be a general moral character factor (see, e.g., Hartshorne et al., 1930, p. 230, Table 32). Mailer (1934) was one of the first to note this. Using Spearman's tetrad difference technique, Mailer isolated a common factor in the intercorrelations of the character tests of honesty, altruism, self-control, and persistence. Subsequently, Burton (1963) reanalyzed the Hartshorne and May data and found a general factor that accounted for 35%-40% of the common variance.

Since the pioneering work of Hartshorne and colleagues (1928, 1929, 1930), many other studies have provided data that speak directly to the specificity versus consistency of moral behavior. As other reviewers have noted (Burton, 1976; Rushton, 1980), the typical correlation between any two behavioral indices is about .30. Combining measures, on the other hand, normally leads to greater predictability. Failures to ' take account of this fact have led to the widespread and erroneous view that moral behavior is almost completely situation specific. This, in turn, has led students of moral development to neglect research concerned with the origins of general moral "traits." The fact that, judging from the aggregated correlational data, such traits exist, and, moreover, that they seem to appear early in life, poses a considerable challenge to moral development research (Rushton, in press). The argument presented for the existence of moral traits, of course, also applies to the existence of other personality traits (Epstein, 1979, 1980; Eysenck, 1970, 1981; Rushton, Jackson, & Paunonen, 1981).

Longitudinal Stability of Personality

Block (1981) has pointed out that the question of cross-situational consistency becomes a question about longitudinal consistency when the time dimension is introduced. To what extent, over both time and situation, do a person's behaviors stem from enduring traits of character? Currently, this is a controversial question (Brim & Kagan, 1980; Rubin, 1981). A prerequisite for answering it is reliable and valid measurement. When studies measure personality at different ages by aggregating over many different assessments, longitudinal stability is usually found. But when single measurements or other less reliable techniques are used, the longitudinal stability of personality is less marked (Block, 1971; Rubin, 1981). In Block's (1971, 1981) work, for example, where the principle of aggregation has been strictly adhered to, substantial coherence of personality has been found over several decades.

Block analyzed extensive personality data from about 170 individuals. Data were first obtained in the 1930s when the subjects were in junior high school. Further data were gathered when the subjects were in their late teens, in their mid-30s, and in their mid-40s. The archival data so generated were enormously wide-ranging and often not in a form permitting of direct quantification. To systematize the data, Block employed clinical psychologists to study individual dossiers and to rate the subject's personality using the Q-sort procedure—a set of descriptive statements such as "is anxious," which can be sorted into piles that indicate how representative the statement is of the subject (Block, 1961). To ensure independence, the materials for each subject were carefully segregated by age level, and no psychologist rated the materials for the same subject at more than one time period. The assessments by the different raters (usually three for each dossier) were

found to agree with one, another to a significant degree, and they were averaged to form an overall description of the subject at that age.

Using this careful procedure of aggregated ratings based on diverse items of information, Block (1971, 1981) found personality stability across the ages tested. Even the simple correlations between Q-sort items over the 30 years between early adolescence and the mid-40s provided evidence for stability. Correlations indicating stability were, for example, for the male sample: "genuinely values intellectual and cognitive matters," .58; "is self-defeating," .46; "has a high aspiration level," .45; and "has fluctuating moods," .40; for the female sample, "is an interesting, arresting person," .44; "pushes and tries to stretch limits to see what she can get away with," .43; "esthetically reactive," .41; and "is cheerful," .36. When the whole range of variables for each individual was correlated over 30 years, the mean correlation was .31. These are lower bound estimates, uncorrected for the inevitable presence of unreliability of measurement. Even more substantial relationships could be expected to occur if individual items were aggregated to create typologies (see Block, 1971, 1981). Block's (1981) results, therefore, demonstrate that when personality is measured adequately, longitudinal stability is found. Moreover, this stability has important implications for other psychological phenomena (Eichorn, Clausen, Haan, Honzik, & Mussen, 1981). As Block (1981) points out, the problem is to identify conditions that lead to consistency and conditions that lead to change. This, like the previous section on moral character, poses a considerable challenge to current theorizing (Rushton, in press)—a problem that does not even exist unless aggregated measures are used.

Stages of Cognitive Development

In Piagetian theory, cognitive development is described in terms of a sequence of qualitatively distinct states called stages. Each stage is defined by a unique set of cognitive structures, the so-called *structures d'ensemble* **principle.** The prototypical conceptual skills of each stage are said to be generated by the structures in a manner that is not unlike the way that a mathematician derives theorems from a set of axioms (Brainerd, 1978a). In the case of the concrete-operational stage, for example, such familiar concepts as conservation, seriation, classification, horizontality, and the like are all thought to be generated by eight quasi-algebraic entities known as grouping structures (Piaget, 1941, 1949).

To students of cognitive development, it has seemed almost self-evidently true that these assumptions demand strong correlations between same-stage concepts. The rationale is straightforward. If two same-stage concepts are both monotonically related to the same structure, then they should be monotonically related to each other. This prediction was explored in several early studies concerned with the concrete-operational stage (e.g., Dodwell, 1961; Elkind, 1961). As a rule, the design of these studies consisted of administering tests for a few concrete-operational concepts plus a standardized ability test of some sort (e.g., Stanford-Binet Intelligence Scale). The general idea was that correlations between tests for same-stage concepts, which share a common conceptual base (e.g., grouping structures), should be higher than correlations between these tests and standardized measures. This pattern failed to emerge. Instead, the typical result was that all correlations, whether between same-stage concept tests or between these tests and standardized measures, were, as Elkind remarked in connection with one of his studies, "sometimes significant, and usually low" (1961, p. 44). It was even found in some studies that same-stage measures correlated better with standardized ability tests than with each other (e.g., Dodwell, 1961). The bulk of this early stage-validation literature was reviewed by Flavell (1963, Chapter 11).

The situation has not changed appreciably during the intervening years. For example, Berzonsky (1971) reported a study in which three types of measures were administered to a sample of elementary school students: (a) tests of various concrete-operational concepts, (b) tests of various formal-operational concepts, and (c) a standardized intelligence test. The obvious predictions from Piagetian theory are, first, that the measures in Category a should correlate more highly with each

other than with the measures in Categories b and c, and second, that the measures in Category b should correlate more highly with each other than with the measures in Categories a and c. In the event, however, the correlations within category were generally low. Moreover, when the data were subjected to factor analysis, a five-factor solution was obtained that was not related to the test's stage classifications in any sensible way.

More recently, Ford (1979) has challenged the validity of the egocentrism construct (concrete-operational stage) on the basis of observed correlations between different measures of egocentrism. Ford pointed out that correlations among tests for three types of egocentrism (affective, cognitive, and spatial) have usually been "low and often nonsignificant" (p. 1169). In place of Piagetian stages and structures, Ford argued that such small correlations "are more parsimoniously interpreted as the result of other explanatory constructs, such *as* those referring to the general level of cognitive, perceptual, or linguistic-development of the child" (p. 1185). The results of a subsequent study of this problem by Scheffman (1981) appeared to bear out Ford's argument. Scheffman administered tests of four types of egocentrism (spatial, cognitive, affective, and communicative,) to a sample of 3- to 5-year-olds. Scheffman's data are unpublished; we reproduce them in Table 2. The correlations range from a low of –.34 to a high of .72, with the mean correlation being .06. At first glance, these results seem to argue strongly against the position that the tests measure a common theoretical construct.

In accord with the view being presented here, however, there is considerable motivation for reexamining the Piagetian stage literature in the light of the principle of aggregation. The typical design has involved administering a single test for each target concept rather than multiple tests for each. If, for example, a hypothetical study is concerned with transitivity, seriation, and conservation (concrete-operational stage), then the standard design would consist of administering one transitivity test, one seriation test, and one conservation test to a sample of children. Worse, following the Piagetian tradition of assigning children to stages, performance on individual tests has simply been scored pass/fail in most studies (cf. Brainerd, 1977a, 1977b). From the standpoint of the aggregation principle, the chances of obtaining high correlations between single tests scored as pass or fail are not promising. In fact, the low correlations that have been re-

Table 2
Correlations Based on ***Nonaggregated Measures*** *of Four Types of Egocentrism (After Scheffrnan, 1981)*

	Egocentrism variables									
	Spatial		Cognitive			Affective		Communicative		
Egocentrism variables	1	2	3	4	5	6	7	8	9	10
Spatial										
		41	17	01	05	04	22	—16	05	12
2			16	18	21	12	—05	—30	—19	—17
Cognitive										
3				11	14	—17	14	—15	10	—12
4					24	03	—11	02	—03	—04
5						00	—01	—34	—04	—16
Affective										
6							05	—04	—26	22
7								29	29	31
Communicative										
8									OS	10
9										72
l0										

Note. Decimal points have been omitted.

Table 3
Correlations Based on Aggregated Measures of Four Types of Egocentrism (After Scheffman, 1981)

Egocentrism variable	Egocentrism variable			
	Cogni- tive	Affec- tive	Commu- nicative	Total scale
Spatial	.22	.16	—.24	.58
Cognitive		.13	.12	.49
Affective			.34	.61
Communicative				.05

ported in most articles are about all that one has a right to expect.

• Do same-stage correlations increase when test scores are aggregated across two or more measures of individual concepts? Unfortunately, it is not possible to make confident statements on the basis of the extant literature because, as we mentioned, only one test per concept has been administered in most studies. Moreover, of those few studies with multiple measures per concept (e.g., Hooper, Swinton, & Sipple, 1979; Hooper, Toniolo, & Sipple, 1978), aggregate correlations have typically not been reported in published versions of the research. Nonetheless, there is some evidence available that suggests that same-stage correlations may be much larger for aggregated measures than for individual ones. We present two illustrations.

First, we reconsider the Scheffman (1981) study of egocentrism. In this study, two tests of spatial egocentrism, three tests of cognitive egocentrism, two tests of affective egocentrism, and three tests of communicative egocentrism were administered. Aggregate correlations for these measures appear in Table 3. These aggregate correlations are of two types. First, correlations between the combined scores for tests of particular forms of egocentrism are reported in the first three columns. Note that these correlations are two to five times larger than the .06 average in Table 2. Second, correlations between combined scores for each form of egocentrism and combined scores for all 10 tests are reported in the fourth column. Here, the aggregation effect is quite dramatic: The correlations for three types of egocentrism (spatial, cognitive, affective) account for a quarter to a third of the variation. The important point is that the theoretical conclusions that emerge from the individual correlations (Table 2) and the aggregate correlations (Table 3) are completely different. On the basis of Table 2, one would be likely to conclude, in line with Ford's (1979) review, that these egocentrism tests do not measure any common ability. But on the basis of the fourth column of Table 3, one would probably conclude that tests for three types of egocentrism tap a common ability, whereas tests for the fourth type of egocentrism (communicative) tap some unrelated ability.

Second, we consider some data from a longitudinal, construct-validation study of the concrete-operational stage by Hooper and his associates (e.g., Hooper, Brainerd, & Sipple, 1975; Hooper & Toniolo, 1974). In this study, tests designed to measure the composition and reversibility operations of Piaget's eight grouping structures were administered to subjects in the kindergarten to Grade 6 range. The test for each grouping structure was composed of eight subtests, four subtests measuring the relevant composition operation and four subtests measuring the relevant reversibility operation, each of which was scored pass/fail. A key prediction under investigation was that the subtests in each block should be highly correlated because they presuppose the same structure. However, low positive correlations were observed in most cases. The picture was different when performance on individual subtests was correlated with the aggregated scores for their structure. The average same-structure correlation rose to .55 for Piaget's four classificatory structures (grouping 1-4) and to .58 for his four relational structures (grouping 5-8). Once again, the message from the aggregate correlations is that tests that do not correlate very well with each other nevertheless may measure common traits.

To avoid misinterpretation, we would like to stress that we are not disposed to use these aggregation effects as grounds for arguing that Piaget's stage construct ought to be resurrected. The list of logical and empirical objections to stages is far too long to be dismissed this easily (see papers by various commentators in Brainerd, 1978b, 1979). Our only contention is that, by and large, the pre-

diction that most investigators regard as the minimum requirement of stage models, high covariation between same-stage traits, has not been fairly tested in the case of Piaget's theory.

The Relationship Between Cognition and Behavior

A great deal of research in psychology has sought to predict behavior from knowledge of hypothesized processes thought to be operating within the organism. Five current examples are metacognition, attitudes, personality trait, role taking ability, and level of moral judgment. In the main, accurate behavioral prediction from measures of internal constructs has rarely been achieved. We believe that this lack of success may be due to a failure to measure a wide enough (aggregated) sampling of either the internal predictor variables or the behavioral criteria. We review literature that supports this contention.

Metacognition as an Explanatory Construct

Metamemory is usually defined as *everything* that one might know or come to know about memory, including what one can do to enhance memory, characteristics of people who are good and poor memorizers, and what task variables enhance or inhibit memory (e.g., Flavell & Wellman, 1977). A primary hypothesis of metamemory- theory is that metamemory may mediate memory behavior (Flavell & Wellman, 1977). In more familiar information processing terms, metamemory has been hypothesized to be the "executive" that directs memory functioning. Most investigators would probably agree that before causally oriented research on metamemory–memory connections can be carried out, substantial correlations between metamemory and memory behaviors must be demonstrated. The search for such correlations initially produced little success (Pressley, Borkowski, & O'Sullivan, in press). Our review of the research indicates that the aggregation principle was largely ignored in early metamemory research and that it may have accounted for the lack of metamemory-memory correlations in those studies.

The standard design was to administer one or a very few metamemory measures and correlate performance on these measures with a single memory measure, such as strategy usage or amount remembered. Not surprisingly, from our perspective, when these single metamemory measures of low reliability (Kurtz, Reid, Borkowski, & Cavanaugh, 1982) were used to predict single memory measures (of unknown reliability), few substantial correlations occurred (e.g., Brown & Campione, 1977; Kelly, Scholnick, Travers, & Johnson, 1976; Salatas & Flavell, 1976).

Even when multiple measures of metamemory have been available, researchers have chosen to analyze individual rather than composite measures. One study is particularly notable in this connection. Cavanaugh and Borkowski (1980) administered a metamemory test originally devised by Kreutzer, Leonard, and Flavell (1975) that was composed of 13 items. The elementary school children in the study were subsequently administered three memory tasks. Performance on each of the 13 metamemory items was separately correlated with performance on each of the memory measures. The mean correlation produced in this analysis was..1 with the highest correlation being .38. We have no way of knowing what the relationship would have been had a metamemory aggregate measure and/or a memory–behavior aggregate measure been used in the analysis because no such analysis was conducted. What is known is that the reliabilities of the individual metamemory items were quite low: The mean reliability of individual items was found to be .38 in a later study (Kurtz et al., 1982). Borkowski (Note 1) reported that by aggregating several metamemory variables, higher metamemory–memory correlations are obtained than those reported by Cavanaugh and Borkowski (1980). In the only published paper in which this was done (Kurtz et al., 1982), the two metamemory-memory behavior correlations were .39 and .26, which are not large but are comparable to the highest correlations obtained in the Cavanaugh and Borkowski (1980) study.

As alluded to earlier, research on metacognitive development encompasses more than metamemory. One recent study of

metacognitive reading skills provides compelling support for our aggregation arguments. Forrest (1980) administered several measures of knowledge of reading and language to elementary schoolers. When individual measures were correlated with actual reading scores, the correlations were very low with a mean of .17. Forrest also aggregated her scores into two categories (e.g., knowledge of decoding, knowledge of comprehension skills) and correlated aggregated scores with actual reading achievement. The correlations were impressively larger *(M = .40),* and thus, Forrest concluded that a relationship between metacognition and reading exists.

One argument against combining metamemory tests into composite measures has been that composite measures are not conceptually meaningful; the aggregate is a hodgepodge of diverse intellectual processes (see Cavanaugh & Borkowski, 1980). However, this argument does not preclude the construction of tests consisting of several items measuring individual metamemory skills. It is to be hoped that some attention will be given to the development of such tests. At a minimum, however, the search for metamemory—memory correlations must be better informed about the dependence of such correlations on aggregation effects. Forrest's (1980) work on metacognition and reading provides a model of how multi-item metacognitive scales for particular cognitive processes can be constructed and related to actual behavior, that metamemory researchers should find instructive.

Attitude—Behavior Relationships

In a well-known review of attitudes and attitude change, McGuire (1969) concluded that "the person's verbal report of his attitude has a rather low correlation with his actual behavior toward the object of the attitude" (p. 156). Fishbein and Ajzen (1974) subsequently explained this result in terms of the principle of aggregation. They pointed out that although a great deal of effort had gone into refining techniques for measuring attitudes, relatively little consideration had been given to the adequacy of measurements on the behavioral end of the attitude—behavior relationship. Hence, whereas attitudes were often measured by multi-item scales, the behavior to be predicted was usually a single act.

Fishbein and Ajzen (1974) proposed that multiple-act criteria be used on the behavioral side to see whether, under these circumstances, attitudes would predict behavior better. Using a variety of attitude scales to measure religious attitudes and a multiple-item self-report behavior scale to measure religious behaviors, they found that attitudes were related to multiple-act criteria but had no consistent relationship to single-act criteria. Whereas the various attitude scales had a mean correlation with single behaviors ranging from .14 to .19, their correlations with aggregated behavioral measures ranged from .70 to .90. Fishbein and Ajzen's (1974) paper provides another dramatic example of the principle of aggregation in action.

Predicting Social Behavior from Personality Traits

In a similar paper to Fishbein and Ajzen's, Jaccard (1974) examined why personality traits rarely correlate better than .3 with social behavior. He suggested that, as in the attitude—behavior literature, whereas much thought had been given to the creation of multiple-item personality scales, relatively little attention had been given to measurement of the to-be-predicted behaviors. Usually, it was a case of a highly reliable, multiple-item personality scale being correlated with a single item of behavior of unknown reliability and validity. Jaccard (1974) therefore carried out an investigation to determine whether the dominance scales from the California Psychological Inventory (Gough, 1957), and the Personality. Research Form (Jackson, 1967), would predict self-reported dominance behaviors better in the aggregate than they would at the single-item level. The results were in accord with expectations. Whereas both personality scales had a mean correlation of .20 with individual behaviors, the aggregated correlations were :58 and .64.

The Role-Taking/Altruistic-Behavior Relationship

Earlier, we mentioned research on the construct of egocentrism (Piaget, 1932). A topic

that is closely related to egocentrism is the notion of role taking ability. Essentially, it has been argued that around the age of 7 children "decenter" and henceforth are able to "take the role of another." There has been extensive research on the relationship between role taking and social behaviors such as altruism. Typically, the correlations have been low (e.g., .2 or .3), and specific relationships have often proved unreplicable. This has led reviewers to question both the adequacy of the egocentrism construct (Ford, 1979) and the adequacy of the methods used to measure it (Krebs & Russell, 1981; Rubin, 1978).

In Ford's (1979) review, as we have already discussed, it was noted that role taking tasks have had either low or nonsignificant correlations with each other. This led Ford to conclude that there was little or no support for the construct validity of egocentrism. He proposed, instead, an alternative interpretation based on task-specific and response-specific cognitive constructs. In Krebs and Russell's (1981) review of the same literature, a different conclusion was advanced. These authors proposed that an improved theory of egocentricity was needed. They stated "It is largely due to the absence of an overriding theory, we contend, that studies have failed to conduct valid tests of the relationship among measures (p. 148)." They implied that with a better theory, improved measurement techniques would follow.

We suggest a third and more straightforward interpretation. When using role taking as a predictor variable, only one item has usually been administered. These items vary widely in reliability (Enright & Lapsley, 1980). Moreover, one item has normally been used for each to-be-predicted variable. Under such circumstances, it would be surprising to find correlations much higher than .2. But suppose that the role taking tasks were combined into a battery and an aggregated, composite score of role taking ability were used. A study of this sort has been reported by Elder (1982).

Elder administered batteries of four role taking tasks, five prosocial moral judgments, and several indices of altruism (including 5 laboratory measures) to 89 seven-year-old children. Teacher ratings and measures of play behavior were also obtained. The important finding for our purposes is that whereas the various correlations of each role taking task with each laboratory measure of altruism had a mean of .10, aggregated role taking scores correlated an average of .45 with an aggregated altruism score. Once again, spurious and misleading conclusions flow from a literature of single-measurement studies.

Additional support for our interpretation comes from the results of a meta-analysis of this literature by Underwood and Moore (1982). Meta-analysis refers to aggregating over independent studies rather than, as we have discussed so far, over independent measures within the same study. The basic principle, of course, is the same (i.e., that errors associated with any one measurement average out when measurements are combined). Underwood and Moore (1982) used a meta-analytical technique to assign exact probabilities to the results of a series of studies. On this basis, they concluded that reliable relationships are to be found between role taking and altruism.

The Moral-Judgment/Altruistic-Behavior Relationship

Finally, we might note that whereas the studies concerned with the relationship between role taking and altruism present a confused picture, those concerned with the relationship between moral judgment and altruism appear to be a little more reliable (Blasi, 1980; Rushton, 1980).. It also happens that the researchers who examine the relationship between moral judgment and altruism have typically combined responses to the several moral measures into an overall aggregate score.

Attachment

In the early 1970s there was great pessimism about whether there is a unique type of bond between mothers and their infants—that is, whether the intuitively appealing construct of infant–mother attachment was valid. This pessimism grew out of the results of studies that included a number of measures presumed to assess mother–infant attachment (e.g., visual regard toward mother by infant, vocalizing by infant to mother, touch-

ing of infant by mother). The general finding was that these "attachment" behaviors were not stable and did not correlate well with each other (e.g., Maccoby & Masters, 1970; Masters & Wellman, 1974). The conclusion that was drawn was that there is little evidence for attachment as a construct. The results were also interpreted as consistent with social learning theories of socialization. According to the social learning framework, individual attachment behaviors develop in interaction with specific environmental contingencies, contingencies that vary from behavior to behavior and from mother—infant pair to mother—infant pair. Thus, neither great consistency within mother—infant pairs nor between mother—infant pairs would be expected.

In contrast to this earlier research, recent studies of mother—infant attachment have relied on composite, categorical measures of mother—infant bonding (e.g., Ainsworth, Blehar, Waters, & Wall, 1978; Waters, 1978, 1981). Categories that have been used include proximity and contact seeking, contact maintaining, resistance, avoidance, search, and distance interaction, with each category composed of measures on several variables. For instance, the resistance category has included measures of "pushing away, throwing away, dropping, batting away, hitting, kicking, squirming to be put down, jerking away, stepping angrily, and resistance to being picked up or moved or restrained" (p. 350, Ainsworth et al., 1978). The reliabilities of these aggregate categorical variables are quite high. For example, in a comparison of the correlations between 12- and 18-month time samples of discrete mother—infant attachment behaviors versus the correlations between 12- and 18-month ratings on the mother—infant interactive categories, Waters (1978) reported a mean correlation of .12 for the discrete behaviors and a mean of .44 for the aggregated behavior categories.

By examining different patterns of outcomes, it has been possible to assign infants to reliable attachment categories. For example, a securely attached infant is high in proximity seeking, high in contact maintaining, low in proximity avoiding, and low in contact resisting. The development of this reliable classification scheme has greatly increased the volume of attachment research, including work on neonatal differences and subsequent attachment (e.g., Crockenberg, 1981; Waters, Vaughn, & Egeland, 1980), attachment and early maltreatment (e.g., Egeland & Sroufe, 1981), effects of day care on attachment (e.g., Anderson, Nagle, Roberts, & Smith, 1981; Vaughn, Gove, & Egeland, 1980), and the relationship between attachment and behavioral competence in later childhood (e.g., Arend, Gove, & Sroufe, 1979; Cohen & Beckwith, 1979; Matas, Arend, & Sroufe, 1978; Waters, Wippman, & Sroufe, 1979).

Substantial relationships have been reported in all of these investigations. However, the Arend et al. (1979) and Waters et al. (1979) studies are particularly noteworthy in the present context. In both of those studies, attachment classifications obtained during infancy were correlated with ego measures (interpersonal conduct scales) taken later in development. The ego scales consisted of multiple items with aggregate scores derived following the methodological recommendations discussed here. Waters et al. (1979) found substantial relationships between quality of attachment during infancy and interpersonal competence at age 31/2 years. Arend et al. (1979) also found relationships between attachment classification and behaviors of children 3 years later (at age 4 to 5 years). These results are especially striking when contrasted with the consistent finding of instability in attachment behaviors that emerged from research based on individual measures. As Masters and Wellman (1974) summarized, there was little stability whether "the intervening time was three minutes, one day, three months, four months, or longer" (p. 224).

Sex Differences: An Example of Aggregating Over Studies

In a major review of the sex difference literature, Maccoby and Jacklin (1974) concluded that the only sex differences that are fairly well established are (a) girls excel in verbal ability, (b) boys excel in visual-spatial ability, (c) boys are superior in mathematical ability, and (d) males are more aggressive. However, Block (1976) subsequently argued

that this review was biased against finding sex differences due to inappropriate methods of combining data. Specifically, Block argued that many of the individual studies reviewed 'used single-item dependent variables of unknown reliability, and hence, they were potentially insensitive to sex differences. To examine this possibility, Block, after specifying the units to be combined, aggregated over studies to determine the proportion favoring males or females in higher mean score on each dimension. We present here a new tabulation based on Block's (1976) reanalysis (see Table 4).

Block's meta-analysis led her to rather different conclusions from Maccoby and Jack-Tin's: Block (1976) concluded that males are not only higher on spatial and quantitative abilities and aggressiveness, but also are

> better on insight problems requiring restructuring, and more dominant and have a stronger, more potent self-concept, are more curious and exploring, more *active.* and more impulsive. (p. 307)

In addition, she suggested that females not only score higher on tests of verbal ability hut also

> express more fear, are more susceptible to anxiety. are more lacking in task confidence. seek *mo*re help and reassurance, maintain greater pros..... y to friends, score higher on social desirability, and, at the younger ages at which compliance has been studied, are more compliant with adults. (p. 307)

Table 4
Proportions of Studies Demonstrating Sex Differences Based on Block's (1976) Reanalysis of Maccoby and Jacklin's (1974) Literature Review

	Ratio of significant comparisons to total number of comparisons			
	Girls and women significantly higher		Boys and men significantly higher	
Behavior assessed	Ratio	Proportion	Ratio	Proportion
Cognitive dimensions				
Verbal abilities	45/160	.28	18/160	.09
Spatial abilities	5/100	.05	35/100	.35
Quantitative abilities	6/35	.17	14/35	.40
Analytic impulsivity	6/80	.08	22/80	.28
Breaking set-responses to "insight" problems	0/14	.00	**12/14**	.86
Anagrams-breaking up words to form new words	4/10	.40	0/10	.00
Descriptive, analytic sorting style	0/6	.00	1/6	.17
Auditorially oriented	6/26	.23	2/26	.08
Social dimensions				
Aggressiveness	**5/94**	.05	52/94	.55
Fear, timidity, anxiety	36/79	.46	0/79	.00
Activity level	**6/109**	.06	39/109	.36
Competitiveness	**6/50**	.12	14/50	.28-
Dominance.	**4/89**	.05	35/89	.39
Compliance & rule following	26/51	.51	**1/51**	.02
Nurturance, maternal behavior, helping, donating and sharing	10/58	.17	**7/58**	.12
Sociability	60/215	.28	36/215	**.17**
Suggestibility	36/125	.29	8/125	.06
Achievement orientation	5/23	.22	4/23	.17
Dependency	**28/88**	.32	**10/88**	**.11**
Curiosity and exploration	8/50	.16	. 20/50.	**.40**
Social desirability	7/9	.78	0/9	.00
Self-concept				
Strength and potency of self concept	**0/8**	.00	7/8	**.88**
Low self-esteem	**20/84**	.24	13/84	.16
Confidence **on** task performance	0/33	.00	25/33	.76
Other				
Tactile sensitivity	5/13	.38	**0/13**	.00

Other investigators have also aggregated studies of sex differences (Cooper, 1979; Hall, 1978; Hoffman, 1977). For example Hoffman (1977) investigated whether females demonstrate greater empathic responsivity than males to the suffering of another person. He examined 16 such studies. Although only a few of the 16 studies reported a statistically significant sex difference, 16 of the 16 found a mean difference in favor of greater female responsivity, a result highly improbable by chance. In Cooper's (1979) meta-analysis, males appeared to conform less to social pressure than females.

These conclusions have implications for both theory and research in child development (Rushton, in press). Once again, the appropriate use of aggregation (this time over separate studies) establishes clear relationships otherwise obscured by single measurement methodology. Similar conclusions have recently been advanced by authors of meta-analyses of experimenter-expectancy research (Rosenthal, 1978) and psychotherapy research (Smith & Glass, 1977).

The Assessment of Emotionality in Animals

Many psychologists working on the causes of emotional development have studied laboratory animals because of the control this gave over experimental variables. Classic among these studies are those examining the effects of early experiences (Scott, Stewart, & De Ghett, 1974). Levine (1969), for example, examined the effects on later behavior of differing amounts and types of early stimulation. At first he simply handled newborn rats briefly during their first days and noticed that such early handling subsequently led the animals to be less emotionally reactive and to engage in more exploratory behavior than the nonhandled animals. In subsequent studies, Levine discovered that rats given mild electric shock early in their lives were less emotionally disrupted when subsequently exposed to stressors than were rats not given shock; the shocked rats even proved more resistant to disease.

The research strategy used by Levine involves "behavior matching," in which it is assumed that direct parallels exist between the behavior of other animals and those of humans. Specifically, situations are studied that elicit emotional behaviors assumed to be "the same" (i.e., under the control of the same underlying mechanism) in both animal and human subjects. One advantage of such a strategy is that a match can be made between the verbal reports provided by humans and the cytological, endocrinological, and neurological data obtained from animal subjects (Gray, 1982; Panksepp, 1982; Plutchik, 1980). The ultimate fruitfulness of such a strategy is, of course, open to dispute (Vanderwolf & Goodale, 1982). For our purposes, however, the interesting thing about such research is that it illustrates that it is as important to take account of aggregation when dependent variables are behavioral and the subjects are animals as it is in paper and pencil research with humans.

The emotional trait most often researched in the rat, for example, is fear, and the most commonly used measures of fear include defecation, ambulation (a measure of exploratory behavior presumed to be negatively related to defecation), and rearing (also a measure of exploratory behavior). If an emotional trait of fear exists in rats, then correlations across individual differences on these measures should be high (Gray, 1971). Some researchers, however, have challenged both the usefulness of the concept of fear in the rat and the particular indices used, on the ground that the different assessments only correlate around .2 or .3. Ossenkopp and Mazmanian (Note 2), however, have recently demonstrated that quite substantial correlations among measures ensue if the indices are aggregated over time.

In Ossenkopp and Mazmanian's study, ambulatory, rearing, and defecating behaviors of rats were measured in The Open Field Test over different time periods: 1-minute periods (4 minutes per test session), and sessions (four sessions at 48-hour intervals). The results showed that whereas the mean correlations across measures on a minute by minute basis averaged .17, the mean correlation was .37 on a session by session basis, and the correlation was .55 over the total test period (see Table 5).

Experimental Studies of Social and Intellectual Development

Finally, we consider an illustration of how failures to aggregate are capable of producing conclusions about the relative contributions of learning to different areas of development that may be incorrect. The areas in question are social development, on the one hand, and intellectual development, on the other. With respect to social development, it is considered well established that observational learning from models has powerful effects on both the development of aggression (Bandura, 1973) and the development of altruism (Rushton, 1980). These findings, but especially the former, have prompted governmental concern about possible inadvertent learning from television (United States Department of Health and Human Services, 1982). Concerning intellectual development, it is equally well known that intervention programs designed to boost children's intelligence, some of them employing observational learning, have achieved only modest success (e.g., Detterman & Sternberg, 1982).

The apparent difference in the relative malleability of social and intellectual development has been explained in various ways by theoreticians. One leading interpretation is that intellectual development is controlled by variables that are "structural" and, therefore, minimally susceptible to learning, whereas social development is controlled by variables that are "motivational" and, therefore, more susceptible to learning (Bandura, 1977; Endler, 1981; Mischel, 1968; Rushton & Endler, 1977). An analysis of the dependent variables used in the two types of studies, however, suggests a much simpler interpretation based on the aggregation principle.

In modeling studies of children's aggression and altruism, a single dependent variable is typically used to measure the behavior, for example, the number of punches delivered to a Bo-Bo doll in the case of aggression (e.g., Bandura, 1973) or the number of tokens donated to a charity in the case of altruism (e.g., Rushton, 1980). In intellectual training studies, however, multiple-item dependent variables such as standardized educational tests, standardized achievement tests, and standardized intelligence tests are typically used (Detterman & Sternberg, 1982). Throughout this article we have stressed that the low reliability of nonaggregated measures can mask strong underlying relationships between variables. In the case of learning studies, it can have essentially the opposite effect: It is always easier to produce a change in some trait as a consequence of learning when a single, less stable measure of the trait is taken than when more stable, multiple measures are taken. This fact may explain why modeling studies of social development have generally been more successful than training studies of intellectual development.

Table 5
Effects of Aggregating Open Field Scores for Ambulation, Defecation, and Rearing in Rats (After Ossenkopp Mazmanian, Note 2)

Open field measure	Mean *r*	Range
Ambulation with defecation		
Minutes	—.17	.01 to —.31
Sessions	—.39	.01 to —.56
Total	—.59	
Rearing with defecation		
Minutes	—.16	—.01 to —.32
Sessions	—.35	—.01 to —.63
Total	—.49	

It is important to add here that we are not claiming either (a) that social and intellectual development are equally malleable or (b) that observational learning does not have a major impact on social development. Concerning the former, our contention is merely that no conclusion is possible until the relative degrees of aggregation in the two types of studies have been more precisely equated. Concerning the latter, there is no denying the importance of observational learning for social development in an absolute sense (Bandura, 1977). The effects of television, for example, have now been demonstrated from several methodological perspectives (USDHHS, 1982).

The implications of the principle of aggregation for experimental research are as important as those for correlational studies. The nongeneralizability of experimental findings to similar situations and, sometimes, their nonreplicability have been sources of concern (Greenwald, 1975). We suggest that if

the principle of aggregation is attended to in the construction of dependent variables, stronger empirical generalizations will occur. The necessity to sample a variety of items in the dependent variables of experiments is an argument made long ago by Brunswik (1947) and reiterated more recently by Epstein (1980). Recently, some investigators have taken to gathering multiple dependent variables and then analyzing them using multiple analysis of variance (MANOVA) statistical designs. This, however, should not be confused with what we recommend. We are suggesting instead that researchers aggregate dependent variables prior to conducting their statistical analyses.

Concluding Remarks

We have reviewed 12 influential areas of developmental research in which there is a substantial but erroneous volume of opinion either that there is little consistency in target behavior or that hypothesized mediating constructs do not predict behavior. On the basis of our review, we conclude that these opinions often turn out to be too pessimistic when the focus is shifted from single dependent variables to more reliable aggregated measures.

Fortunately, there seems to be a growing realization that the principle of aggregation applies to behavioral measures as well as to paper-and-pencil tests. Others who have recently noted that single behavioral measures have lower predictive power than the average of many include Hogan, DeSoto, and Solano (1977), Green (1978), Epstein (1979, 1980), Jackson and Paunonen (1980), Rushton, Jackson, and Paunonen (1981), and Wiggins (1981).

Despite the more optimistic picture of construct validation research in developmental psychology that emerges from this review, our position is a conservative one. Essentially, we are saying (and illustrating) only what others have said before in other contexts (e.g., Campbell & Fiske, 1959), namely that due consideration must be given to the reliability of one's measurements. Single measures are typically less reliable than multiple measures, and using less reliable measures necessarily attenuates empirical relationships. Although these points have long been recognized in paper-and-pencil research, they have generally been ignored when dependent variables are behavioral.

Of course, there are occasions when aggregation is unnecessary, and even harmful. In clinical contexts, for example, when highly robust phenomena such as reflexes and overlearned habits are being studied, it may often be more useful to focus on a specific response than on general characteristics of the person (Bandura, 1969; Mischel, 1968). Similarly, aggregation may not strengthen empirical relationships when potent, ego-involving events are being investigated, as in Epstein and Fenz's (1965) study of anxiety in sport parachuting. Likewise, there may be little gain from aggregating self-ratings or ratings by others when these are based on impressions gathered over several past observations (Epstein, 1980). Finally, there is inappropriate aggregation, as when one aggregates over unreliable or poor items or studies. In this respect, for example, Eysenck (1978) characterized Smith and Glass's (1977) meta-analysis of psychotherapy outcome studies as an exercise in "mega-silliness" because it aggregated over a hodge-podge of methodologically poor studies.

In closing, we reiterate that whenever there is the possibility of unreliability of measures, then aggregation becomes a desideratum. This is true whether one is measuring a construct or behaviors that might be predicted by the construct. Aggregation can be accomplished through multiple measures of the same variable or measurement of multiple variables that are believed to be related. Aggregation can also be accomplished by combining the ratings of various observers or combining the results of various behavioral or paper-and-pencil assessments, as well as by combining the results of a number of studies directed at the same issue. Examples of all of these approaches have been provided herein. Of course, not all types of aggregation are equally appropriate for all purposes. For example, if one is testing for a trait it is not sufficient to aggregate over the same item many times (as recently done by Mischel & Peake, 1982). To provide an index of a hy-

pothesized trait, it is necessary to aggregate alternative assessments of the same underlying concept.

The question of what to aggregate is always an issue. In this review, we were limited to published data that were amenable to analysis from both single and multiple measurement perspectives. In future work, the logical step would be to move in the direction of batteries of tests for particular constructs. Pilot work concerned with the psychometric properties of measurement has to be accepted as part of development research, as it has long been in the field of testing. This requires sampling a variety of items and stimuli and then aggregating over only those that demonstrate desired properties. Among these are high test-retest reliabilities, high item-total correlations, and both convergent and divergent validity. In short, as developmental researchers, we must become more concerned with the construct validity of our measures (Anastasi, 1982; Campbell & Fiske, 1959).

Reference Notes

I. Borkowski, J. G. Personal communication to M. Pressky, January 1982.

Ossenkopp, K.-P., & Mazmanian, D. *The principle of aggregation in psychobiological correlational research: An example from the open field test.* Unpublished manuscript, University of Western Ontario, 1983. (Available from De K-Peter Ossenkopp, Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2)

References

Ainsworth, M. D. S., **Blehae M. C., Waters, E., & Wall,** S. *Patterns of attachment. New* York: Erlbaum, 1978.

Anastasi, A. *Psychological testing* (5th ed.). New York: Macmillan, 1982.

Anderson, C. W., Nagle, R. J., Roberts, **W. A.,** & Smith, J. W. Attachment to substitute caregivers as a function of center quality and caregiver involvement. *Child Development,* 1981, *52,* 53-61.

Arend, R. A., Gove, F. L, & Sroufe, L A. Continuity of individual adaptation from infancy to kindergarten: **A** predictive study of ego-resiliency and curiosity in preschoolers. *Child Development,* 1979, *50,* 950-959.

Bandura, A. *Principles of behavior modification. New* York: Holt, Rinehart & Winston, 1969.

Bandura, A. *Aggression: A social learning analysis.* Englewood Cliffs, NJ.: Prentice-Hall, 1973.

Bandura, A. *Social learning theory* Englewood Cliffs, NJ.: Prentice-Hall, 1977.

Berzonsky, M. **D. Interdependence of Inhelder and Piaget's model of logical thinking.** *Developmental Psychology,* 1971, 4, 469-476.

Blasi, A. Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin,* 1980, 88. 1-45.

Block, J. *The Q-sort method in personality assessment and psychiatric research.* Springfield, Ill Charles C Thomas, 1961.

Block, J. *Lives through time.* Berkeley, Calif.: Bancroft Books, 1971.

Block, J. Some enduring and consequential structures of personality. In A. I. Rabin, J. Aronoff, **A. M.** Barclay, & R. A. Zucker (Eds.), *Further explorations in personality. New* York: Wiley, 1981.

Block, J. H. Issues, problems, and pitfalls in assessing sex differences. A critical review of *The Psychology of Sex Differences. Merrill-Palmer Quarterly,* 1976, *22.* 283-308.

Brainerd, C. J. Cognitive development and concept learning: An interpretative review. *Psychological Bulletin,* 1977, *84,* 919-939. (a)

Brainerd, C. J. Response criteria in concept development research. *Child Development.* 1977, *48,* 360-366. (b)

Brainerd, C. J. *Piaget's theory of intelligence.* Englewood Cliffs, NJ.: Prentice-Hall, 1978. (a)

Brainerd, C. J. The stage question in cognitive-developmental theory. *The Behavioral and Brain Sciences,* 1978, *I.* 173-213. (b)

Brainerd, C. J. Continuing commentary. *The Behavioral and Brain Sciences,* 1979, *2,* 137-154.

Brim, 0. G., lc, & Kagan, J. (Eds.). *Constancy and change in human development.* Cambridge, Mass.: Harvard University Press, 1980.

Brown, A. L, & Campione, J. C. Training strategic study time apportionment in educable retarded children. *Intelligence,* 1977, *1,* 94-107.

Brunswik, E. *Systematic and representative design of psychological experiments.* Berkeley, Calif.: University of California Press, 1947.

Burton, R. V. Generality of honesty reconsidered. *Psychological Review,* 1963, *70,* 481-499.

Burton, R. V. Honesty and dishonesty. In T. Lickona (Ed.), *Moral development and behavior. New* York: Holt, Rinehart & Winston, 1976.

Campbell, D. T., & Falk, D. W. Convergent and discriminant validation by the **multitrait-multimethod matrix.** ***Psychological Bulletin,*** **1959,** ***56.*** **81-105.**

Cavanaugh, J. C., & Borkowski, J. G. Searching for metamemory-memory connection= **A developmental study.** ***Developmental*** *Psychology,* 1980, *16,* 441453.

Cohen, S. E., & Beckwith, L Preterm infant interaction with the caregiver in the first year of life and competence at age *two. Child Development. 1979, 50.* 767-776.

Cooper, H. M. Statistically combining independent studies: Meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology* 1979, 37, 131-146.

Crocicenberg, S. B. Infant irritability, mother responsiveness, and social support influences on the security of infant-mother attachment. *Child Development,* 1981, *52,* 857-865.

Cronbach, L .1. The two disciplines of scientific psychology. *American Psychologist,* 1957, *12,* 671-684.

Detterman, D. K., & Sternberg, R. J. (Eds.), *How and how much can intelligence be increase&* Norwood, NJ.: Ablex, 1982.

bodwell, **P. C.** Children's understanding of number concepts: Characteristics of an individual and a group test. *Canadian Journal of Psychology,* 1961, *15,* 29-36.

Wand, B., & Sroufe, L A. Attachment and early maltreatment. *Child Development,* 1981, *52,* 44-52.

Eichorn, D. H., Clausen, **J. A.,** Haan, N., Honzik, M. P., & Mussen, P. H. (Eds.), *Present and past in middle life* New York: Academic Press, 1981.

Eller, *J. The relationship between role-taking skills, prosocial reasoning and prosocial behavior.* **Unpublished** doctoral dissertation, University of Western Ontario, 1982.

Elkind, D. The development of quantitative thinking: A systematic replication of Piaget's studies. *Journal of Genetic Psychology,* 1961, *98,* 37-46.

Endler, N. S. Persons, situations, and their interactions. In **A. L** Rabin, J. Aronoff, A. M. Barclay, & R. A. Zucker (Eds.), *Further explorations in personality.* New York: Wiley, 1981.

Enright, **R. D.,** & Lapsley, D. K. Social role-taking: A review of the constructs, measures, and measurement properties. *Review of Educational Research.* 1980, *50,* 647-674.

Epstein, S. The stability of behavior 1. On predicting mast of the people much of the time. *Journal of Per- - sonality and Social Psychology,* 1979, 37. 1097-1126.

Epstein, S. The stability of behavior: II. Implications for psychological research. *American Psychologist.* 1980, *35,* 790-806.

Epstein, S., & Fenz, W. D. Steepness of approach and avoidance gradients in humans as a function of experience. *Journal of Experimental Psychology,* 1965, *70.* 1-12.

Eron, L D. Prescription for reductions of aggression. *American Psychologist.* 1980, *35,* 244-252.

Eysenck, H. J. The validity of judgments as a function of the number of judges. *Journal of Experimental Psy**dsology*** 1939, *25.* 650-654.

Eysenck, H. .1. *The structure of human personality* ***New*** **York Macmillan, 1970.**

Eysenck, H. J. An exercise in mega-silliness. *American Psychologist.* 1978, 33, 517.

Eysenck, H. J. (Ed.), *A model for personality* New York: Springer, 1981.

Fishbein, M., & Ajzen, I. Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review,* 1974, *81,* 59-74.

Flavell, J. H. *The developmental psychology of Jean Piart.* Princeton, NJ.: Van Nostrand, 1963.

Flan* J. H., & Wellman, H. M. Metamemory. In **R. V.** Kail & **J. W. Hagen (Eds.),** *Perspectives on the development of memory and cognition.* **Hillsdale, NJ.: Etibaum, 1977.**

Ford, M. E. The construct validity of egocentrism. *Psychological Bulletin,* 1979, *86,* **1169-1188.**

Forrest, D. L Cognitive and meta-cognitive aspects of reading (Doctoral dissertation, University of Waterloo, 1980). *Dissertation Abstracts International,* 1980, *41,* 1134-B.

Gordon, K. Group judgments in the field of lifted weights. *Journal of Experimental Psychology,* 1924, 7. 398-400.

Gough, H. G. *California Psychological Inventory Manual.* **Palo Alto, Calif.: Consulting Psychologists Press, 1957.**

Gray, J. A. *The psychology of fear and stress.* London: Weidenfeld and Nicholson, 1971.

Gray, J. A. *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. New* York: Oxford University Press, 1982.

Green, B. F., Jr: In defense of measurement. *American Psychologist.* 1978, *33,* 664-670.

Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychological Bulletin,* 1975, *82,* 1-20.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Hall, A. Gender effects in decoding nonverbal cues. *Psychological Bulletin,* 1978, *85.* 845-857.

Hartshorne, H., & May, M. A. *Studies in the nature of character: Vol. I. Studies in deceit.* New York: Macmillan, 1928.

Hartshorne, H., May, M. A., & Mailer, J. B. *Studies in the nature of character: Vol. 2. Studies in self-control.* New York: Macmillan 1929.

Hartshorne, H., May, M. A., & Shuttleworth, F. K. *Studies in the nature of character: Vol. 3. Studies in the organization of character.* New York: Macmillan, 1930.

Hoffman, M. L Sex differences in empathy and related behaviors. *Psychological Bulletin,* 1977, *84, 712-722.*

Hogan, R., DeSoto, C. B., & Solano, C. Traits, tests, and personality research. *American Psychologist,* 1977, *32,* 255-264.

Hooper, F. H., Brainerd, C. J., & Sipple, T. S. *A representative series of Piagetian concrete operations tasks.* Madison: Wisconsin Research and Development Center for Cognitive Learning, 1975.

Hooper, F. H., Swinton, S. S., & Sipple, T. S. Logical reasoning in middle childhood: A study of the Piagetian concrete operations stage. In H. J. Klausmeier & Associates (Eds.), *Cognitive learning and development Piagetian and information-processing perspectives.* **Cambridge, Mass.: Ballinger,** 1979.

Hooper, F. H., & Toniolo, **T. A.** *A longitudinal analysis of logical reasoning relationships: Conservation and transitive inference.* Madison: Wisconsin Research and Development Center for Cognitive Learning, 1974.

Hooper, F. H., Toniolo, T. A., & Sipple, T. S. **A** longitudinal analysis of logical reasoning relationships: Conservation and transitive inference. *Developmental Psychology.* 1978, *14,* 674-682.

Jaccard, J. J. Predicting social behavior from personality traits. *Journal of Research in Personality* 1974, 7, 358-367.

Jackson, D. N. *Personality Research Form Manual.* **Goshen, N.Y.. Research Psychologists Press, 1967.**

Jackson, D. N., & Paunonen, S. V. Personality structure and assessment. **In** M. R. Rosenzweig & L W. Porter (Eds.), *Annual Review of Psychology* **(Vol. 31). Palo Alto, Calif.: Annual Reviews, 1980.**

Kelly, M., Scholnick, E. F., Travers, S. H., & Johnson, J. W. Relations among memory, memory appraisal,

and memory strategies. *Child Development.* 1976, 47, 648-659.

Kenrick, D. T., & Stringfield, D. Q Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review,* 1980, *87,* 88-104.

Knight, H. C. *A comparison of the reliability of group and individual judgments.* Unpublished master's thesis, Columbia University, 1921.

Krebs, D. L., & Rukcell, C. Role-taking and altruism: When you put yourself in the shoes of another, will they take you to their owner's aid? In J. P. Rushton & R. M. Sorrentino (Eds.), *Altruism and helping behavior: Social, personality and developmental perspectives.* Hillsdale, N.J.: Erlbaum, 1981.

Kreutzet M. A., Leonard, C., & Flavell, J. H. An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development,* 1975, *40(1,* Serial No. 159).

Kurtz, B. E., Reid, M. K., Borkowski, J. G., & Cavanaugh, J. C. On the reliability and validity of children's metamemory. *Bulletin of Psychonomic Society,* 1982, *19.* 137-140.

Levine, S. Infantile stimulation: A perspective. In J. A. Ambrose (Ed.), *Stimulation in early infancy.* New York: Academic Press, 1969.

Lord, R. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Maccoby, E. E, & Jacklin, C. J. *The psychology of sex differences.* Palo Alto, Calif.: Stanford University Press, 1974.

Maccoby, E. E., & Masters, J. C. Attachment and dependency. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 2, 3rd ed.). New York: Wiley, 1970.

Mailer, J. B. General and specific factors in character. *Journal of Social Psychology,* 1934, *5.* 97-102.

Masters, J. C., & Wellman, H. The study of human infant attachment A procedural critique. *Psychological Bulletin,* 1974, *81,* 218-237.

Matas, L., Arend, **R. A.,** & Sroufe, L. A. Continuity of adaptation in the second year: The relationship between quality of attachment and later competence. *Child Development,* 1978, *49,* 547-556.

McGuire, W. J. The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *The handbook of social* ***psychology*** (Vol. 3, **2nd ed.). Reading, Mass.: Addison-Wesley, 1969.**

Mischel, W. *Personality and assessment.* New York: Wiley, 1968.

Mischel, W. Toward a cognitive social learning reconceptualization of personality. *Psychological Review.* 1973, *80,* 252-283.

Mischel, W., & Peake, P. K. Beyond deja vu in the search for cross-situational consistency. *Psychological Review* 1982, *89,* 730-755.

Moskowitz, **D. S.,** & Schwarz, J. C. Validity comparison of behavior counts and ratings by knowledgeable informants. *Journal of Personality and Social Psychololgy.* 1982, *42,* 518-528.

Panksepp, J. Toward a general psychobiological theory of emotions. *The Behavioral and Brain Sciences.* 1982, 5, 407-467.

Piaget, J. *The moral judgment of the child.* London: Routledge & Kegan Paul, 1932.

Piaget, J. *Classes, relations et nombres: Essai sur les -groupements" de la logistique et la reversibilite de la pensee.* Paris: Vrin, 1941.

Piaget, J. *Traite de logique.* Paris: Colin, 1949.

Plutchik, R. *Emotion: A psychoevolutionary synthesis.* New York: Harper & Row, 1980.

Pressley, M., Borkowski, J. G., & O'Sullivan, J. Children's metamemory and the teaching of memory strategies. In D. Forrest-Pressley, G. E. MacKinnon & T. G. Waller (Eds.), *Mew-cognition, cognition, and human performance.* New York: Academic Press, in press.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin,* 1978, *85.* 185-193.

Rubin, K. H. Role-taking in childhood: Some methodological considerations. *Child Development,* 1978, 49, 428-433.

Rubin, Z. Does personality really change afer 20? *Psychology Today*, 1981, *15,* 18-27.

Rushton, J. P. *Altruism, socialization, and society.* Englewood Cliffs, NJ.: Prentice-Hall, 1980.

Rushton, J. P. Sociobiology: Toward a theory of individual and group differences in personality and social behavior. In J. R. Royce (Ed.), *Annals of theoretical psychology,* (Vol. 2). New York: Plenum Press, in press.

Rushton, J. P., & Endler, N. S. Person by situation interactions in academic achievement. *Journal of Personality*. 1977, 45. 298-309.

Rushton, J. P., Jackson, D. N., & Paunonen, S. V. Personality: Nomothetic or idiographic? A response to Kenrick and Stringfield. *Psychological Review,* 1981, *88.* 582-589.

Rushton, J. P., Murray, H. G., & Paunonen, S. V. Personality, research creativity and teaching effectiveness in university professors. *Scientornetrics,* 1983, 5, 93-116.

Salatas, H., & Flavell, J. H. Behavioral and meta-mnemonic indicators of strategic study behavior under remember instructions in first grade. *Child Development,* 1976, 47, 80-89.

Scheffman, J. *The effects of individual and group play experience on preschoolers' perspective taking, referential communication and number conservation abilities.* Unpublished doctoral dissertation, University of Western Ontario, 1981.

Scott, J. P., Stewart, J. M., & De Ghett, V. J. Critical periods in the organization of systems. *Developmental Psychobiology.* 1974, 7, 489-513.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Pspzhologist,* 1977, *32,* 752-760.

Spearman, C. Correlation calculated from faulty data. *British Journal of Psychology,* 1910, 3, 271-295.

Stevens, S. S. *Experimental Psychology*. New York: Wiley, 1951.

Stevens, S. S. *Psychophysics and social scaling.* Morristown, NJ.: General Learning Press, 1972.

Underwood, B., & Moore, B. Perspective-taking and altruism. *Psychological Bulletin,* 1982, *91.* 143-173.

United States Department of Health and Human Ser-

vices. *Report. Television and behavior: Ten years of progress and implications for the eighties* (Vol. 1, *Summary Report;* Vol. 2, *Technical Reviews).* Washington, D.C.: United States Government Printing Office, 1982.

Vanderwolf, C. H., & Goodale, M. A. Does introspection have a role in brain-behavior research? *The Behavioral and Brain Sciences.* 1982, 5. 448.

Vaughn, B. E., Gove, F. L., & Egeland, B. The relationship between out-of-home care and the quality of infant-mother attachment in an economically disadvantaged population. *Child Development.* 1980, *51,* 1203-1214.

Waters, E. The reliability and stability of individual differences in infant-mother attachment. *Child Development.* 1978, *49,* 483-494.

Waters, E. Traits, behavioral systems, and relationships: Three models of infant-adult attachment. In K. Smmelman, **G. W.** Barlow L. Petrinovich, & M. Main (Eds.), *Behavioral development.* London: Cambridge • University Press, 1981.

Waters, E., Vaughn, B. E., & Egeland, B. R. Individual differences in infant-mother attachment relationships at age one: Antecedents in neonatal behavior in an urban, economically disadvantaged sample. *Child Development,* 1980, *51.* 208-216.

Waters, E., Wippman, J., & Sroufe, L A. Attachment, positive affect, and competence in the peer group: Two studies in construct validation. *Child Development,* 1979, 50, 821-829.

Wiggins, J. S. Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychology Review,* 1981, *1,* 3-18.

Woodworth, R. S., & Schlosberg, H. *Experimental psychology*. New York: Holt, 1939.

Received July 1, 1982
Revision received January 21, 1983 ■

Editorial Consultants for This Issue: Review Articles

John R. Aiello
Max Allen
Terry W. Allen
Thomas Andre
Hymie Anisman
Mortimer H. Appley
Albert F. Ax
Fern J. Azima
Mark A. Barnett
Ira Belmont
Leonard Berkowitz
Phyllis Berman
Karen Linn Bierman
Milton R. Blood
Milton L Blum
Arthur P. Brief
Ann L. Brown
Roger B. Burton
Robert B. Cairns
Joanne R. Cantor
Richard Carlson
John B. Carroll
Charles S. Carver
Werrett W. Charters, Jt
Louis D. Cohen
Harris M. Cooper
David S. Cordray
Charles G. Costello
John L. Cotton
James E. Cowden
Jennifer Crocker
Michael Davis
Richard A. Dienstbier
Emanuel Donchin
Carl Eisdorfer
Robert F. Eme
Ora Fagan-Simcha
David Fay
Herman Feifel
Jack M. Feldman
Norma D. Feshback
Fred E. Fiedler
Donald W. Fiske
Jerome D. Frank
Russell E. Glasgow
Charles J. Golden
Gerald Goldstein
Sanford Golin
Gerald Goodman
Gerald Gratch
William W. Grings
Douglas T. Hall
Joseph T. Hart
Joseph B. Hellige
Richard Hirschmann
William C. Hirst
Larry A. Hjelle
David S. Holmes
James S. House
Robert J. House
Frederick Mark Jablin
Durand Frank Jacobs
Arthur G. Jago
Herbert M. Jenkins
Jerome Kagan
Rabindra N. Kanungo
Stanislav V. Kasl
Daniel Katz
Herbert Kaufman
Verner M. Knott
Eric S. Knowles
Gerald P. Koocher
Janet Tracy Landman
Ellen J. Langer
Herbert C. Lansdell
John T. Lanzetta
Gary P. Latham
Paul R. Latimer
Barry Ledwidge
Herschel W. Leibowitz
Mark R. Lepper
Daniel J. Levinson
Jerre Levy
Peter M. Lewinsohn
Jane Loevinger
Raymond P. Lorton
Charles Lowe
Dorothy I. Marquart
George W. McConkie
Joseph E. McGrath
Douglas Media
Sarnoff A. Mednick
Donald H. Meichenbaum
Manfred J. Meier
Mary Ann Metzger
Herbert Meyer
Beth E. Meyerowitz
Lance A. Miller
Thomas I. Miller
Theodore Millon
John B. Miner
John T. Monahan
Alan Monat
Bennet B. Murdock, Jr.
Gregory L. Murphy
Helmer R. Myklebust
Bernice L. Neugarten
Allan Newell
Warren T. Norman
Charles A. O'Reilly III
Marlene Oscar-Berman
H. McDvaine Parsons
Miles L Patterson
Thane S. Pittman
Robert Plutchik
Dean G. Pruitt
Julian Rappaport
Stephen K. Reed
Lance J. Rips
Jeffrey Z. Rubin
Glenn Sanders
Neal W. Schmitt
Carmi Schooler
Donald P. Schwab
Alan Searleman
Bernard Segal
Mark Seidenberg
William R. Shadish, Jr.
John W. Shaffer
Stefanie Shattuck-Hufnagel
Gordon Shulman
Ervin Staub
Eugene Frank Stone
Lowell H. Storms
Anne E. Sutherland
Samuel Sutton
John A. Swets
George P. Taylor
E. Paul Torrance
Leonard P. Ullman
Robert Vallerand
Herbert J. Walberg
Mary A. Weber
Robert F. Weiss
Alan F. Williams
Michael Wogan
Stephen R. Yussen