Preface

MODELS, METHODS, AND MEASUREMENT

Dreams of mathematical laws in psychology have persisted for well over a century in the fields of perception and psychophysics. In modern times, mathematical models have been prominent in judgment–decision and in learning. Mathematical models have also been proposed in social–personality, including integration models for attitudes (Chapter 7) and fairness-unfairness models in deserving theory (Chapter 2).

To realize these dreams required radical conceptual reorientation—a shift from structure of the external world to structures of the internal world (*The Dual Worlds: Internal and External* in Chapter 7). Almost miraculously, the internal world of thought and action follows simple mathematical integration laws in most areas of human psychology. These empirical laws are the foundation of Information Integration Theory.

The long-standing obstacle of true psychological measurement was central in establishing these psychological laws. Conceptual resolution was obtained by making the laws themselves the base and frame for measurement. Empirical resolution was obtained with methods discussed in this chapter. These mathematical laws of information integration, established by many investigators in many countries, are the foundation of psychological measurement theory.

Popular pitfalls invalidate not a few publications. Among these are comparing relative importance of different variables, meaninglessness of much "statistical interaction," illusion of "statistical control," and meretricious *p* values (*Faulty Methods* and *Measurement Pitfalls*).

Methodological issues that arise in planning and interpreting experiments are also discussed, especially ***extra****statistical inference*.

Chapter 6

MODELS, METHODS, AND MEASUREMENT

Empirical data, rooted in empirical conditions and given meaning within one's conceptual framework, are the ground of science. The psychological laws of information integration, applied to moral cognition in this book, rest on such ground. This chapter summarizes a number of issues of method in this theoretical development.

Some issues are technical, as with single person design and with understanding how (and how not) to test an algebraic model. Some are largely empirical, as with pilot work and writing instructions. Most are empirical–conceptual blends: assessing generality of results, pitfalls of confounding, and the treacherous problem of comparing importance of different variables. Most basic are integration laws that make possible true psychological measurement.

EXPERIMENTAL METHOD

Clean data are invaluable in science. The time spent in gathering data in typical experiments is only a fraction of the total time to decide what issue to study, read relevant literature, formulate one's questions or hypotheses, develop experimental procedure and design through successive pilot stages, analyze and interpret the data, relate your work to the literature, and write a report for publication. The worth of all this work rests on the worth of your data—and the worth of your experimental design.

Problems of experimental method have been discussed elsewhere (Anderson, 1974a,b,c, 1979, 1981a, 1982, 1996a, 2001, 2002, 2008). Some important issues are noted in this chapter. These discussions owe much to previous workers in many areas. They are presented as contributions to continuing development of scientific method that need consideration from other perspectives, especially with respect to the many practical problems of social–moral betterment of person and society (Note 0).

EXTRASTATISTICAL INFERENCE

Extrastatistical inference, and *statistical inference,* are twin pillars of science. Techniques of statistical inference can help control variables, avoid certain confounds, increase reliability of results, and assess reliability, as with confidence intervals. Statistical inference is essential to provide reasonable confidence that one has a real result—within the specific empirical situation (see *Random Assignment* below*)*.

Statistical inference is severely limited. Except with random sampling from some population, which is rare, statistical inference applies only to the specific samples at hand: task, stimulus materials, and participants. But nearly all samples are *handy samples*; random assignment allows statistical inference only to the handy sample.

Extrastatistical inference is essential to generalize beyond one's specific situation. All who write articles for publication hope and believe their results will have some generality beyond their handy sample of participants and specific experimental conditions. This belief rests squarely on extrastatistical inference based on empirical understanding.

Extrastatistical inference has paramount importance.

This fact is largely neglected in standard statistics texts; they give the impression that significance tests are be-all and end-all. Extrastatistical inference is hard to discuss, of course, since much depends on empirical, situation-specific understanding. Further discussion of extrastatistical inference is given in *Empirical Direction in Design and Analysis* (Anderson, 2001), hereafter referred to as *Empirical Direction*.

INSTRUCTIONS

Experiments with humans usually rely on instructions to define the task, especially the meaning of the response. Workers in the field of tests and measurements have amply documented how easily people can misunderstand instructions. Pilot work should aim to detect misunderstandings and eliminate them. One help is to have participants summarize their understanding of the task in their own words (see also next section). This may be supplemented by directed, post-experimental questioning, especially in the pilot work, about how they answered specific questions. Two examples are cited in Anderson (2008, pp. 395f).

Instructions are especially important in moral psychology. In studies of forgiveness, as one example, different investigators define forgiveness in different ways. Comparing results of different investigators is thus uncertain. Few present evidence on what their participants had in mind

when they made their judgments. One writer calls this a "pernicious" problem, a fitting adjective (see *Algebra of Forgiveness,* Chapter 7). This problem of psychological meaning of to-be-published data requires special attention in the planning and pilot stages of investigation.

PILOT WORK

Pilot work is important for clean data. Pilot work should pretest the experimental procedure to uncover shortcomings and seek improvements. Sensitive post-task interviewing should seek to understand the experimental task from each participant's view. Too often, investigators assume that common words, such as forgiveness and attitude (Chapter 7) have common meaning.

MENTAL SCHEMAS

Mental schemas underlie each operation in the Integration Diagram. They deserve careful attention in the instruction period. The valuation schema is most important, of course, for it determines the meaning of the data. Also, a mental schema for integration must be constructed by each participant to embody their understanding of the task (see *Integration Processes*, pp. 64-69, Anderson, 1996a).

One example comes from judgments of each of performance, motivation, and ability, given levels of the other two. Empirically,

Performance = Motivation × Ability;
Motivation = Performance – Ability;
Ability = Performance – Motivation.

The Performance schema was diagnosed from a linear fan pattern, the other two from patterns of parallelism (see Anderson, 1996a, pp. 159ff).

The pilot work had revealed that some participants followed a summative schema for the last two judgments, in line with the natural correlation between ability and motivation, failing to realize the inverse constraint imposed by given Performance. This was resolved by including in the instructions *choices* between cases with equal Performance but unequal Ability or Motivation.

CLEAN DATA: ONE PERSON AT A TIME

Testing people in groups is an invitation to unclean data. Their attention is poorly controlled; misunderstandings or carelessness are almost inevi-

table. Some prominent mistaken claims have resulted from group experiments (e.g., representativeness heuristic, Anderson, 1996b). Group data can be useful for pilot work but have limited place in serious science.

It may be feasible to run two participants at a time, which would cut running time by 50%. Three at a time, if warrantable, would yield an additional savings of only 17%.

Group settings are natural in some situations, of course, as in studies of classroom education or in group discussion of moral issues. In such cases, however, each group may correspond to only a single unit (df) in the statistical analysis (see Section 15.1 in *Empirical Direction*).

The Internet offers cheap data but quality control is lacking and response rate is low (Van Acker, Theuns, Hofmans, & Mairesse, 2007). It is hard to put much faith in such data except as pilot work. Buttressing such data with controlled experiments could provide faith.

GENERALITY

Generality of results is always a concern. Any one investigation is narrowly limited to specific conditions (see *Extrastatistical Inference* above). We presume our results will have some generality. We seek to validate this presumption through our experimental procedure.

Random Assignment. *Handy samples* of participants are usual in experimental analysis. Randomness, the essential requirement for statistical inference, may be obtained with Fisher's invention of randomly assigning participants to experimental conditions. The blundering of early studies of the Head Start program, criticized by Campbell and Boruch (1975), began with failure to use random assignment.

Random assignment allows statistical inference only to the handy sample—a statsig result gives confidence that it was not an accident of which participants were assigned to experimental conditions. To go beyond our handy sample, as we desire, rests entirely on *extrastatistical inference* as discussed above.

Random assignment is not possible with variables such as level of education or socioeconomic status. Some investigators attempt to match groups on some relevant variable, often falling into the "black pit of the *regression artifact"* (*Empirical Direction*, p. 582).

Some investigators apply "statistical control" to deal with observational data. These statistical procedures are prone to deadly artifacts (see *Illusion of "Statistical Control"* below). Some uncontrolled variables, however, may be incorporated in the design as a stratified variable with

random assignment within each stratum, and included in the statistical analysis. Of course, this still allows only correlational conclusions for the stratified variable (*Empirical Direction,* Section 14.2.4).

Stimulus Samples. Running an experiment with a single set of stimulus materials is thin ice design. Results may not generalize to other stimulus materials. It is often advisable, therefore, to use more than a single set of stimulus materials to assess this aspect of generality.

Stimulus generality has special importance in moral cognition. Verdi's study of conflict of obligation in Chapter 7 used three kinds of obligation to assess generality of the adding-type law. Other examples include my 1962 integration study of person cognition, which used six sets of personality trait adjectives, Armstrong's studies of wife–husband interaction (Figure 3.2), and blame judgments of criminals by Przygotski and Mullet (1993; see Chapter 4).

Multiple Variables. Multiple variables are the norm in everyday life. The importance of multiple variables was emphasized by Aronson, Wilson, and Brewer (1998, p. 135), who conclude their handbook chapter, *Experimentation in Social Psychology*, by calling for a new synthesis:

> assessing the relative importance of several variables, which all influence an aspect of multiply-determined behavior, rather than on testing to see if a particular variable has a 'significant' impact.

Just such synthesis of multiple variables had been the concern of workers on Information Integration Theory over the previous 30-odd years. Their work had shown that "multiply-determined behavior" obeys mathematical laws in most areas of psychology, from social–personality to learning/memory and judgment–decision. In particular, "relative importance," usually assessed with invalid methods, had been given the conceptual clarification that it needs (see *Measurement Pitfalls* and *Measuring Importance* below) together with valid methods to measure importance.

TWO KINDS OF VALIDITY: PROCESS AND OUTCOME

Outcome validity and process validity are both important but often incompatible in any one investigation. *Outcome validity* is mainly concerned with observable outcomes; *process validity* is mainly concerned with underlying process.

An experiment to test whether showing audience children a film of actor children's group discussion of bullying transferred to playground

behavior of the audience children would be mainly concerned with outcome validity. Success may be ample reward even though it may say little about operative cognitive processes, which could be important for improving the discussion procedure.

In legal psychology, to take a second example, many workers have assumed that customary laboratory experiments could produce results relevant to real jury deliberation. Much of this work has little social relevance. To seek social outcome relevance, psychologists need to work within the legal system, emphasized by Ebbesen and Konečni in their experimental–field studies of bail setting and sentencing (see Chapter 4; see similarly Gerbasi, Zuckerman, & Reis, 1977, on mock juries).

Process validity is concerned with underlying cognitive processes. The blame law of Chapter 3 is one example as are the associated studies of apology and extenuation. These integration processes are basic cognitive capabilities with substantial generality in everyday judgment.

Unfortunately, outcome validity and process validity often require rather different experimental design. Aiming for both may compromise both. Social attitude research gives instructive examples (e.g., Anderson, 2008). Further discussion is given in *Validity*, pages 8-16 in *Empirical Direction*.

PREDICTION AND UNDERSTANDING

Algebraic models may be used for two purposes—*prediction* and *understanding*. Prediction and understanding are both important but often involve quite different design and analysis. Methods useful for one purpose can be a mistake for the other.

Prediction. Prediction usually concerns applied situations in which there is an accuracy criterion of success, as with college admissions and job suitability. The most common prediction model is multiple linear regression which has done remarkably well in many situations using rough and ready values of predictor variables. Indeed, such approaches typically outperform expert judgment at much less cost (e.g., Grove & Meehl, 1996; Swets, Dawes, & Monahan, 2000).

But for understanding cognitive processes that underlie behavior, such prediction models are generally poor and misleading. Although high correlations are generally satisfactory for prediction, they may seriously misrepresent underlying process. Among the examples discussed under *Weak Inference*, Section 4.1 in Anderson (1982), additive models

gave correlations of .99 for multiplicative data and .98 for data with an anti-additive crossover.

Weak inference with linear models (Anderson & Shanteau 1977, p. 1133) analyzed published articles from three popular areas in research on judgment–decision:

> In each of these examples from decision theory, high correlations provided initially compelling support for the model in question. This led to further, often intensive, research on the psychological processes that were presumed to underlie these models. . . . these correlations did more to obscure than to reveal underlying process. As a consequence, much labor came to nothing.

To my knowledge, no one has disputed this analysis. Yet weak inference remains common. Prediction has high importance but understanding requires a different framework, both conceptual and methodological.

Understanding. A model of behavior conveys understanding when it reveals cognitive processes that underlie that behavior. For an algebraic model, this often requires measurement of the psychological values of the variables that were functional in the behavior.

The parallelism theorem of Chapter 1 provided a base for such functional measurement. Observed parallelism supports the validity of the response measure by virtue of benefit 2. True measures of the stimulus variables are similarly given by benefit 3.

Matters were by no means so simple. The opposite effects phenomenon, for example, raised doubt that any algebraic model could account for the data. Twelve such theoretical issues had to be resolved, some of which are discussed in Chapter 1 (see Anderson, 2008, pp. 54-68). Fortunately, these and other difficulties were nicely resolved with the averaging model, which has done well in most areas of human psychology, as in the initial empirical chapters.

Overall, these studies by many dedicated investigators have revealed a general cognitive algebra. This cognitive algebra can help with further elucidation of cognitive process.

OBSERVATION AND EXPERIMENT

Observation and experiment are the two basic sources of data. Observation is arguably more basic as the primary source of meaning of what is measured. Meaning of some concepts may seem sufficiently clear from everyday communication, as with taste or warmth, and in some cases with fairness or blame. Many concepts, however, may have multiple qualities or dimensions that deserve separate analysis, as with gratitude, forgiveness, or attitudes toward women. Other concepts such as personal

well-being and marriage satisfaction present greater difficulties of definition (see also *Response Quality* and *Profile Measures* below).

Workers in tests and measurements have contributed invaluable knowledge about wording of questions and problems of reliability and validity. For constructing tests, factor analysis offers one means to deal with complex concepts that can be very useful for prediction.

The Achilles heel of observational data, of course, is causal analysis. The clinical psychologist Meehl (1990, pp. 229f) concluded that the usual tests of correlational predictions in clinical psychology are subject to ten obfuscating factors such that the "usual research review is well-nigh uninterpretable"—"*a bunch of nothing*" (see *Empirical Direction*, Section 11.4.3). Similar cautions hold for much of current research on health psychology, positive psychology, character education, and other important issues of everyday life. "Statistical control" of nonexperimental data is usually an illusion (see below).

Experimental analysis—with randomized assignment—is usually necessary for causal conclusions in field situations. Among these are comparing effects of different programs of moral education in the schools and someday in the family as well. Observational studies and nonrandomized experiments can make vital contributions—if they understand and lead toward experimental analysis (see *Field Science*, Section 15.5 in *Empirical Direction*).

EXPERIMENTAL–FIELD SCIENCE

Experimental analysis embedded in field situations has high importance for both goals of moral science: cognitive theory and social betterment. Both goals face the problem, noted by workers in many areas, that laboratory experiments can hardly hope to reproduce the variables operative in real life.

This limitation is prominent with social–moral cognition. In equity theory, for example, standard experiments mainly used abstract situations and asked for ideal judgments of fair shares (Chapter 2). In real life, of course, different persons have different values of relevant variables that are ignored in such ideal judgments. As one consequence, the important motivation of unfairness was neglected (*Unfairness Paradox* in Chapter 2).

A stellar illustration of experimental–field analysis is the work by Ebbesen and Konečni (1975) on bail setting in the courts. Judges' ideals, assessed experimentally in their chambers using IIT, took sensible account of case variables. But these ideal judgments had virtually zero

relation to their judgments from the bench. The justice system was thus shown to be systematically *un*just (see Chapter 4).

The limitation of laboratory experiments reappears with problems of contextual variables specific to each social situation. In educational psychology, as Cronbach and Snow (1977) emphasized, classroom context includes specific teachers and socioeconomic status of students that may have major influence. Results from one school may have little generality (see *Field Science,* Section 15.5 in *Empirical Direction*).

Many issues require experiments randomized across a population of social groups: families, schools, playgrounds, churches, factories, ethnic subgroups, or other social units. Randomized experiments are now common in medical science and are gaining popularity in clinical psychology (Kazdin, 2011). Our educational system deserves nothing less, not only for moral education but no less for standard school subjects (see *Adaptive Transfer* under *Education*, Chapter 7).

The psychological laws provide one useful foothold on field experiments. They have unique capability with analysis of multiple determinants. They have substantial nomothetic generality together with idiographic capability with personal values. Context effects can, moreover, be treated as exact units in some cases by virtue of Cognitive Unitization, as with Armstrong's wife–husband study of blame (Figure 3.2). Experimental studies of cognitive theory can thus aid social betterment.

EXPERIMENTAL DESIGN

Experimental design is fundamental in psychological science. Good design can increase reliability, validity, and generality. One class of designs, dealing with persons as design units, is considered here.

BETWEEN PERSON DESIGN

Between person design uses different persons for each experimental treatment. Such design may be necessary when any treatment produces carryover effects that would confound effects of a following treatment. Comparing school programs on moral education or adaptive transfer are examples. Such studies need to be situated in actual social settings to allow a realistic mélange of context variables.

But between person design suffers two difficulties. First, the statistical error term includes individual differences which are generally large, negatively impacting power. Second, it averages out individual differences that may be important.

Studies with natural groups face the additional problem that the scores within each group may be correlated through dependence on common context. In a classroom program on moral impact of courses on U. S. history, for example, all persons in a given class would usually reflect influence of a single instructor and a single set of curricular materials as well as interaction among class members. One class may thus be just one df in statistical analysis. A single group could provide useful pilot experience, but multiple groups would be needed for social outcome validity (*Natural Groups*, Section 15.1 in *Empirical Direction*).

REPEATED MEASUREMENTS DESIGN

In repeated measurements design, multiple treatments are given each person. This design type was standard in experiments on deserving and blame in Chapters 2 and 3. The main effect of individual differences is factored out, yielding an error term usually several times smaller than between person design (e.g., Figure 6.1). Smaller error yields shorter confidence intervals and greater power.

Individual analysis is also possible because each individual has been in multiple conditions. Individual error terms are obtainable by presenting some or all conditions two or more times or even by pooling higher order interactions. Perhaps individual analysis should be standard with repeated measurements design (see also *Cluster Analysis* below).

Repeated measurements design can suffer from order and carryover effects. Response to later treatments may be influenced by practice, for example, or by carryover from previous treatments. To reduce such order and carryover effects, it is usually desirable to adapt persons to the task in the initial instruction stage. Latin square design (see below) can help balance order effects and assess their magnitude (see index entries for *Order effects* in *Empirical Direction*).

SINGLE PERSON DESIGN

In single person design, each person is tested with multiple treatments and a complete analysis is made for each person (Anderson, 2002). Single person design is an ideal for process theory because the locus of process is in each individual person. In particular, the functional values of the stimulus informers are those of that person, not some agglomerate of unknown different values of different other persons.

Single person design may be undesirable in outcome studies, at least in initial stages. With moral education, for example, developing proce-

dures that will be effective with most children would be well worthwhile. Preliminary information on individual differences might still be obtained by, for example, stratifying participants on some pertinent variable.

PERSONAL DESIGN

Personal design embeds the experiment within each person's life space. Personal design was used in a study of divorced wives' marriage satisfaction. In an initial session, each wife recollected satisfactory and unsatisfactory incidents from her marriage. These incidents were used as stimulus levels in an Affection × Appreciation design. The response was a judgment of satisfaction with a week of married life characterized by a pair of incidents. This personal design helped extend the averaging law into this basic area of social life (Anderson, 1990, 2008).

Personal design could be useful in marriage and family counseling. A personal integration graph could help participants come to grips with their own needs and problems. Personal design could also help liberate current trait conceptions of personality from dependence on group data to study individual persons in their personal life space (*Person Science and Personality* in Chapter 7).

INDIVIDUAL DIFFERENCES

Individual differences can pose difficult problems in every field of psychology. The common hope for general nomothetic law often runs aground on individual differences.

The laws of information integration, however, unify nomothetic and idiographic using within person design. These laws employ individual values of stimulus informers, yet the laws themselves have substantial generality across individuals. Three other issues of individual differences in experimental design and analysis are noted briefly here.

Stratification. Participants may be stratified on some pretest, with strata included as a factor in the design. Error variance can be decreased; individual differences associated with strata are fractionated out of the error yielding shorter confidence intervals.

More important is whether the strata differ in reactions to the experimental variables. One would expect, for example, that conservatives and liberals would differ in valuation of such variables as deserving and need of applicants for family assistance. Similarly, initial moral profiles of children could help improve design of programs on moral education.

Individual Design and Analysis. Individual analysis comes naturally when using repeated measurements design. This approach was used to good advantage in Leon's studies of the blame law discussed in Chapter 3 (see also *Repeated Measurements Design* above).

Individual analysis is explicitly planned with single person design, especially personal design. Individual analysis is valuable for process validity because cognitive processes operate separately in each individual. Cognitive algebra represents generality in the *integration* process while allowing for large individual differences in *valuation*.

Cluster Analysis. Cluster analysis sorts individuals into clusters that are similar in some way. A striking example is the substantial minorities of Always Forgivers and Never Forgivers found in France by Girard and Mullet (1997) discussed under *Algebra of Forgiveness* in Chapter 7. An impressive application to functional theory of pain is given by Oliveira, de Sa′ Teixeira, Oliveira, Breda, and da Fonseca (2007); see also averaging law for phenomenal causality (Schlottmann & Anderson, 1993). The insight and clarification available with cluster analysis is revealed in numerous studies by Etienne Mullet and associates.

An important extension of cluster analysis was made in innovative work by Hofmans and Mullet (2013). Individuals may differ in more than one way, not only in their psychological values, but also in their integration rules. With integration data, four kinds of similarity are possible, each illustrated with published data.

Clustering on values is a simple case. All persons are assumed to obey an adding-type model and to use a linear response scale. For each person, the stimulus values measured by marginal means of any variable will also be a linear scale. These are difficultly comparable across persons, however, because zero and unit will differ across persons.

To make persons comparable, subtract the lowest value from each value and divide by the highest value, separately for each person. This yields a zero-one scale that is comparable across persons. Single values are not comparable across persons, of course, but closeness in values is. Hence cluster analysis may be applied to cluster persons with similar value spectra. This method of finding clusters does not require an integration design; it may be applied to responses to a single variable.

SIMPLER DESIGNS

Integration designs study joint influence of multiple variables. This is important for social cognition, in which multiple variables are typically

operative. Studying one or two at a time may be misleading. But multiple variables can lead to large, cumbersome integration designs that present all possible combinations of each of several variables.

Simpler designs are often possible and will often be more effective than complete integration designs. Simpler designs may be essential with doctors, judges, politicians, and other professionals, who may balk at inroads on their time. Studies with children, families, or people in the street may also benefit from simpler designs.

Field experiments must often use smaller designs than laboratory experiments, often with a need to include larger numbers of variables. The field experiment of bail setting by Ebbesen and Konečni (1975) used a four-variable design with 36 conditions, and still did not include the fifth variable of severity of crime. To get judges' cooperation, they felt it necessary to ask each to judge only 8 conditions. To allow Anova, they adopted the device of assigning these 8 conditions to judges at random, but obtaining 4 judges per condition. This allowed a reasonable statistical analysis.

More effective designs have been developed by statisticians. Thus, all five of the foregoing variables could be studied in a 2^5 design with 32 conditions but using a 1/4 fractional replication that requires only 8 conditions. This would measure all 5 main effects with 2 df for selected interactions. Moreover, it would provide sensitive error terms based on within-person variability (see below).

Main Effects. Main effects of variables usually have primary importance. With several variables, balanced designs that provide equal information on each main effect are usually desirable. Two types of balanced designs are discussed in the next two subsections. Both achieve their goal by sacrificing information on statistical interactions, which become confounded with main effects. This may not be serious, especially when previous work indicates that interactions are small.

Of course, statsig interactions may occur with unequal weighting in the averaging model, as with the negativity effect, source reliability, or other variation in amount of information. Even when real, however, an interaction may require little or no qualification of main effects (see *Understanding "Interactions*" below).

Latin Square Design. Great reduction in design size is possible with designs of the Latin square type. In the 3 × 3 square below, the 3 levels of each of 3 variables are denoted {A, B, C}, {a, b, c} and {1, 2, 3}. Note the balance: each level of each variable is paired with each level of each other variable exactly once to yield 9 experimental conditions:

Aa1	Bb2	Cc3
Bc1	Ca2	Ab3
Cb1	Ac2	Ba3

This square reduces the full 3^3 design from 27 to 9 conditions. Remarkably, a fourth variable could be added, reducing 81 conditions to 9 (Note 1). A 4 × 4 square could reduce $4^3 = 64$ or $4^4 = 256$ conditions to 16. Latin square design may be especially useful in preliminary work to get an overview of main effects. An experiment from attitude theory that included a fourth variable is given in *Empirical Direction* (p. 420).

Latin square design has notable potential to reduce design size in moral science. Smaller designs could be essential with professionals as already noted. Using such designs in pilot work could help familiarize a useful tool (see Section 14.3 in *Empirical Direction*). Standard Latin square design yields no information on interactions. However, any specific interaction that deserved consideration could be assessed by including a specific supplementary design.

A seeming limitation is that Latin squares require all variables to have the same number of levels, 3 in the above example. If some variable had fewer levels, it could be replicated to equalize the number. Thus, if the first variable in the above example had only two levels, A and B, one level could be replicated to yield {A, B_1, B_2} where B_1 and B_2 are identical. Supplementary tests would be needed allowing unequal numbers of observations for A and B.

Fractional Replication. Main effects of all variables can be measured with a fraction of a complete design. In a study of obligation, for example, 6 variables, each at 2 levels, could be studied using a 1/8 fraction, which would reduce design size from 64 conditions to 8. This design leaves 1 df to study a selected interaction. An experimental example from judgment–decision is given in Figure 15.1, page 455, in *Empirical Direction*.

A clear, simple exposition of fractional replication is given by Cochran and Cox (1957), together with an appendix of specific designs (see also Montgomery, 2001). These designs are straightforward if all variables have 2 levels but more complex with more levels. A variable with more than two levels could still be handled with 2-level fractional replication. Thus, a 4-level variable could be treated nominally as two 2-level variables. A 3-level variable, {A, B, C}, might be treated as two 2-level variables, {A, B} and {B, C}. Supplementary analysis to test all given levels could be needed in either case. As with Latin squares, such replication may entail some loss of power.

Test runs with artificial data are highly advisable; mistakes are easy to make, hard to rectify. If some interaction not specifically allowed in the design seems a potential problem, it could be included in the artificial data to see whether it would trouble the results. Any specific interaction could be assessed with a specific supplementary design.

METHODS OF INFORMATION INTEGRATION THEORY

Present discussion of methods is mainly concerned with the two premises of the parallelism theorem of Chapter 1 (see further Anderson, 1982). Premise 1, additivity, is fundamental, being a substantive base for IIT. This additivity premise has been supported in many applications in nearly every field of human psychology.

For the ubiquitous averaging process, however, additivity depends on equal weighting, in which all levels of each separate stimulus variable have equal importance weight. Equal weighting depends on experimental procedure: each level of a variable should convey the same amount of information. The list of 555 personality trait adjectives was screened to represent approximately equal importance for judgments of likableness (see *Batteries of Stimulus Materials* below). It is also necessary, of course, to equalize attention to each stimulus level (Note 2).

Premise 2, response linearity, depends on experimental procedure, discussed in the next section for the rating method. Response linearity has general importance—then the observable pattern in an integration graph is a faithful image of the pattern in the underlying response. Linear response can also help study interaction and configurality that produce deviations from parallelism (see *Configurality* and *Response Generality* discussed below).

METHOD OF FUNCTIONAL RATING

The method of functional rating was developed to eliminate the well-known nonlinear biases suffered by rating methods in common use. These biases, it may be noted, prevented the discovery of the simple additive model despite widespread use of analysis of variance (see also *Understanding "Interactions"* below).

Rating Schema. The rating method seems simple but actually involves two complexities. One is its relative nature; the rating of each stimulus depends on its relation to the other stimuli. The end anchors and practice are intended to set up this stimulus–rating correspondence.

The other complexity is that the rating scale is an abstract quantification of some specified quality. The extensive network of evidence for rating linearity constitutes a major achievement of the many investigators of IIT. Rating linearity represents a fundamental cognitive-motor capability (see *Metric Cognition* in Chapter 7).

Method of Functional Rating. The two main procedures of functional rating are end anchors and preliminary practice. *End anchors* are stimuli a little more extreme than the regular experimental stimuli. They are used in the instructions to define the ends of the response scale. End anchors begin the process of establishing the frame of reference for the response. They can also eliminate *end bias*, the tendency of people to simplify by using scale endpoints for highest and lowest stimuli.

Preliminary practice familiarizes participants with the task and helps firm up the frame of reference for the response. Preliminary practice is generally necessary because ratings are relative judgments. Early work used extensive practice but later work indicates that the valuation process stabilizes fairly quickly (Anderson, 1996a, pp. 92f).

The ideal scale is a continuous graphic scale. Category scales, such as 0–10 or 1–20, have been widely used with satisfactory results. They risk category preferences, however, especially with few categories. Graphic scales can minimize effect of previous responses as well as category preferences. Graphic scales seem essential in studies of children and may be essential for cross-cultural generality.

THEORY OF FUNCTIONAL RATING

Functional rating rests on a mathematical model. This is an application of the decision averaging law, in which rating of any stimulus is located between the two end anchors in proportion to its relative similarity:

$$R = \frac{Sim_U}{Sim_U + Sim_L},$$

where Sim_U and Sim_L are the similarities of the given stimulus to the upper and lower end anchors. Sergio Masin and his students have extended this model to study psychological structure of the judgmental representation of the end anchors. This model did well in several experiments (Dai Prà, 2007; Masin & Busetto, 2010) that ruled out three alternative theories, including Parducci's (1995) range-frequency theory.

Linearity of functional rating has surprised many, especially those who have insisted on choice data in psychological measurement theory

(see Appendix). For graphic rating, linearity is considered to derive from accuracy of motor movement in local space. The common category scale is considered a more symbolic internalization of motor response.

SELF-MEASUREMENT THEORY

Self-measurement refers to procedures in which persons quantify personal values for each single stimulus informer. Validity is the basic issue. People readily report these self-measures when asked—but are they faithful measures of their underlying reality (Note 3)?

Valid self-measures of integrated *response* can be provided with functional rating (benefit 2 of the parallelism theorem). But can people give valid self-measures of the separate *stimulus* informers? One important form of invalidity is the halo effect discussed below.

Valid self-measures of stimuli are essential in many applications of multiattribute analysis in judgment–decision theory. However, the several methods in common use (e.g., tradeoff, point allocation, part-worth, magnitude estimation, rating) disagree with one another, often markedly. Which is valid—if any (see Note 3)?

This validity question can be answered. The algebraic laws can provide valid measures of the stimuli (e.g., benefit 3 of parallelism theorem). These are validity criteria for self-measures. Further discussion of self-measurement is given in Anderson (1982, Section 6.2, *Self-Estimated Parameters*), Anderson and Zalinski (1991), Surber (1985), and Zalinski and Anderson (1989, 1991). Zhu and Anderson (1991 found that the once-most popular method—allocating 100 points among the attributes to measure their importance—was the worst.

Of special importance, self measures allow analysis of situations that do not admit factorial-type integration designs. In a study of attitude change in group discussion, for example, each of three persons received a different biographical paragraph about some U.S. president. They discussed one another's information and their own attitude and then each separately judged the president on statesmanship. Finally, they judged the polarity value and importance weight of their own information and of the discussions of each other group member on their own final attitude. These self-measures yielded good accounts of their final attitude (Anderson & Graesser, 1976).

In such group discussion, factorial-type design is difficultly applicable. Indeed, exact analysis of uncontrolled discussion might seem utterly impossible. Exact analysis was possible, however, by virtue of Cognitive Unitization (see *Addition Law* below). A remarkable example of self-

measurement with females' judgments of dates was given by Shanteau and Nagy (1976; see Figure 1.24, p. 76, in Anderson, 1981a).

The part-worth method, which requires judgment of the total contribution of each stimulus informer to the integrated response, showed promise in Surber (1985) and in Zhu and Anderson (1991). Part-worth corresponds to the total effect, weight × value, of an informer stimulus. This is simpler than getting separate estimation of the two parameters, and could be especially useful in applied multiattribute analysis.

Part-worths would be appropriate for adding models, in which weight and value operate jointly as a single unit. Marginal means thus estimate part-worths on a common scale for each separate stimulus variable. Analogous procedure applies to equal-weight averaging models (Note 22).

As yet, however, self-measurement of stimulus informers is not well-developed. A number of studies have shown promise but systematic analysis is needed (see further Anderson, 1982, Section 6.2; 1991a, pp. 165-178; 1996a, pp. 343f, 391f, Note 14; 2002, 2008, pp. 391-393).

PERCENTILE STIMULUS METRICS

Quality and *quantity* of stimulus informers may be confounded. In the original task of person cognition, for example, participants judge likableness of hypothetical stimulus persons described by trait adjectives such as *sociable* and *punctual*. Each adjective must be valuated both for its polarity *value* and for its importance *weight* with respect to the response dimension of likableness. The trait *sociable*, however, is a location on the dimension of sociableness, and hence a composite of quality and quantity.

Percentile stimulus metrics may be able to separate quantity and quality. A paragraph could be used to explicate the quality of sociableness, emphasizing that people show substantial differences but without implicating any specific quantity of sociableness. Each stimulus person could then be characterized as, say, sociable-30, sociable-60, sociable-90. Similarly for other stimulus informers.

Percentile quantification may also provide simple comparison of importance weight. In the personality trait task, suppose that *sociable* and *punctual* are both quantified at 30, 60, 90 in a 3 x 3 design. If the participant quantifies both at equivalent values, then the main effects are comparable measures of importance weight. Difficulties of using the Average program are bypassed. If a less important factor has a smaller actual response range, as seems plausible, the main effects are still a valid com-

parison of relative importance. This method of percentile stimulus metrics may help to avoid the insidious concept–instance confounding that vitiates so many attempts to compare importance of different variables (see *Confounding* and *Measuring Importance* below).

Another advantage of percentile quantification is with complex stimuli, such as family life, school life, job satisfaction, or alternatives in moral dilemmas. Such complex stimuli could be defined initially with paragraphs that describe its various components, emphasizing that each component can vary from low to high. Social reality could be increased by requiring participants to summarize the stimulus description in their own words before beginning the experiment. With quality established, experimental stimuli could be quantified as family life-30 (-60, -90). Similarly for other stimulus variables.

Graphic quantification may be preferable to numeric. Quantity could thus be represented by a mark on a line or length of a stick rather than a number. This graphic format would be usable with young children and with persons unused to numerical quantification.

BATTERIES OF STIMULUS MATERIALS

Integration experiments typically require multiple responses from each person. Batteries of stimulus materials are needed for many such experiments, especially for single person analysis.

Personality Trait Adjectives. A much-used stimulus battery is the list of 555 personality trait adjectives, each with its mean likableness value and variability (Anderson, 1968a). This list is reproduced in Appendix B of Anderson (1982), with demarcation of four ranges of 32 words each of High, Medium-high, Medium-low, and Low value.

This personality adjective task provided the original base for IIT (Anderson, 1962a). Participants judged likableness of persons described by lists of trait adjectives. Analogous person judgments, based on diverse stimulus informers, are a basic personality function of each of us.

One advantage of this list is that most trait adjectives have approximately equal importance weight on the response dimension of likableness. This is ordinarily facilitated with instructions that each adjective was contributed by a different acquaintance who knew the person well. Equal weighting allows simple parallelism analysis. Other response dimensions, such as honesty or industriousness, could require screening to select adjectives with approximately equal weight.

A special advantage of this adjective list is that the same trait adjective may be used in describing different persons. A third advantage is that each stimulus informer is a single word, easily assimilated.

Note that individual value differences must be expected for any trait adjective. Single person design and analysis may thus need to prescreen adjectives for each person.

Witness Testimony. Another battery consists of summarized testimony of 6 prosecution and 6 defense witnesses from the Hoag bigamy trial of Figure 4.3. This experiment gave what seems the first definite evidence of basal–surface structure of attitudes. These witness testimonies are reproduced in Hommers and Anderson (1991).

Marriage. Marriage studies in IIT have made good use of stimulus batteries, including personal design based on incidents from each person's marriage (see *Marriage as an Investigational Setting*, Section 4.5.2 of Anderson, 1981a). Margaret Armstrong's (1984) PhD thesis includes 81 pages of ingenious stimulus materials used in her several experiments (see also Anderson & Armstrong, 1989). Two sets of experimental stimuli used in marriage experiments are given in Anderson (1991f, Appendices A and B). A general-purpose battery based on common marital conflicts and negotiations could be useful.

Attitudes Using Within Person Experiments. Within person design has been extremely rare in attitude research because of carryover effects. Unlike the personality adjectives, a typical attitude message can seldom be used a second time because of memory carryover from the first time. The dominating concern of social psychologists with persuasion and changing attitudes led to between person design and diverted the attitude field away from functional theory (Anderson, 2008, pp. 82ff).

President Paragraphs. The president paragraphs were developed to allow within person experiments on attitudes. Participants judged statesmanship of a president based on one or two such biographical paragraphs (Figure 6.1). Thus, a High value paragraph about Andrew Jackson could be paralleled with a High paragraph about Woodrow Wilson. Within design can be much more sensitive than standard between design. In Anderson (1973), use of these president paragraphs required fewer than one-tenth the number of participants than a corresponding between person design (Anderson, 1981a, p. 27; *Empirical Direction*, p. 420). Within design has the additional advantage of allowing individual analysis. One experimental study is shown in Figure 6.1 below.

This stimulus battery consists of 220 short biographical paragraphs, 8 or 16 about each of 17 U. S. presidents with end anchors of Washington, Lincoln, and Harding, together with a brief historical overview (Anderson, Sawyers, & Farkas, 1972). These were based on biographies to yield paragraphs of four graded values from low to high, reproduced in Anderson (1982, Appendix C). Almost every president had such a range of events in his administration. Here are one high and one medium-low paragraph about Theodore Roosevelt (1901-1909).

> President Theodore Roosevelt was the first national leader to be concerned with the problem of conservation on a large scale. He took many measures to halt the destruction of the country's wilderness areas. During his two terms as President, the National Forest Service was established, and acreage for national forests was greatly increased. In addition, 5 additional parks and 13 national monuments were opened. The first federal bird reservation was established by Roosevelt, with 50 opened before he left office. Fervently believing in conservation, President Theodore Roosevelt publicly stated: "As a people we have a right and a duty, second to none, to protect ourselves and our children against the wasteful development of our natural resources.

> Theodore Roosevelt was a skilled politician. However, this characteristic is not always necessarily good in a national leader. One example occurred as the time for Roosevelt's reelection drew near. In order to secure enough votes for himself at the national convention, Roosevelt found it necessary to give a public office to a man whom he had justly denounced as an enemy of the civil service system at an earlier time. Roosevelt excused this action, saying, "In politics we have to do a great many things that we ought not to do."

Besides usefulness in experimental analysis, stimulus batteries can help improve the moral level of society. The president experiments, for example, provided a small but meaningful learning experience about U. S. history. A similar study of remarkable American women is given by Simms (1978).

Cognitive Unitization. The importance of Cognitive Unitization may be reemphasized with these president paragraphs. Each paragraph requires a complex valuation process by each subject. Yet this complex processing is treated as a cognitive unit in the integration process. Such unitization underlies the parallelism of Figure 6.1 below as with the two paragraphs just quoted about Theodore Roosevelt.

Even stronger unitization is illustrated in Armstrong's (1984) studies of wife–husband interaction. The success of the integration model implied that the entire discussion of each spouse functioned as a cognitive unit in the integration process for the revised attitude.

Moral Judgment. Batteries of stimulus materials for moral judgment would be a valuable contribution. Conflict situations from everyday

life—marriage adjustments, parenting, growing up, broken promises, getting even, fractured friendships, obligation, and life goals are among the many issues. Cooperative work by investigators at different institutions would be invaluable.

EXTENDED INTEGRATION DIAGRAM

Goal pursuit usually involves a sequence or hierarchy of subgoals, each of which may be separately represented by the Integration Diagram. Fair shares division, for example, may require preliminary valuation–integration for each person. Learning experiments provide another class of examples (see e.g., Figure 8.3).

Learning. The Integration Diagram has straightforward application to learning. Each trial in a learning experiment involves valuation of a given stimulus informer and its integration into the response being learned. On each successive trial, therefore, this response is updated by valuation/integration of the stimulus information on that trial.

This learning process was illustrated with the learning curves for witness testimony in the Hoag bigamy trial (Figure 4.3). These curves revealed two-component structure of learning: an enduring *basal* component and a labile *surface* component. This basal-surface structure is important in functional learning theory (see e.g., Figure 8.3).

This integration analysis also led to a functional conception of learning substantially different from traditional learning paradigms. What is learned usually represents construction of goal-oriented meaning that need have no objective relation to the stimulus informers (see *Functional Theory of Memory* in Chapter 8; see also Williams, 2001). Traditional reinforcement, of course, has narrow relevance; *reinforcer* is replaced by *informer*.

Integration Learning Design. New capability for learning theory is available with the integration laws. Treat some trials as a separate factor with two or more levels. The influence of such trials on later responses can then be measured.

This capability extends standard learning curves whose theoretical analysis often depends on some assumption that all trials have similar effect, as with analysis of sequential dependencies (Anderson, 1956, 1959a). Other examples are given in the 1959 jury trial experiment of Figure 4.3 and the age effect in children's learning of Figure 8.2. Both experiments revealed basal–surface structure of learning.

Internal Stimulus Informers. Internal stimuli require explicit consideration in some situations. *Prior state* is one example. This represents an attitude or feeling the person brings to a situation that is integrated together with external information. Mood is another example (see *Mood Is Information*, Chapter 7).

Knowledge Systems. Knowledge systems are what are learned in IIT (Anderson, 2008, pp. 68ff). Knowledge systems represent a constructionist conception: they include integrals of goal-directed values that reflect situational context. Knowledge systems differ markedly from the associationist conceptions of traditional learning theories (e.g., Mowrer & Klein, 2001).

One advantage of this constructionist view may be illustrated with learning of attitudes, traditionally conceptualized as "readiness to respond" on a one-dimensional, good-bad scale (see *Response Quality* below). This traditional conception is far too narrow to deal with attitude function in everyday life (see *Functional Theory of Attitudes*, Chapter 8). Similar liberation of learning theory applies in every area of psychology.

Little is known about structure of knowledge systems. They are necessary, however, to deal with the ubiquity of multiple determination, the importance of context, and the goal-oriented nature of thought and action (see also *Profile Measures* below).

INTEGRATION DATA VS. QUESTIONNAIRE-TYPE DATA

Integration studies embody a conceptual shift away from the questionnaire framework that underlies much current social–personality. The most obvious need is capability to study joint influence of multiple variables. Such capability, essential for understanding goal-oriented function, is available with the integration laws. No less important is the need to study response structure (see *Profile Measures* below).

This functional focus of IIT is needed in personality theory which has been struggling to free itself from the traditional trait-typological framework to deal with situational context. The laws of information integration can help study personality function, especially construction of situation-dependent, goal-oriented values. Functional measurement of these values provides a foundation that can help study person–situation interaction (see *Analytic Context Theory*, Chapter 7).

Current theory of social attitudes also rests on questionnaire-type data, egregiously so in its typical conception of attitude as one-dimensional, good-bad reaction. Reliance on such questionnaire-type

data, perhaps only a single question, roadblocked development of functional theory of attitudes. In sharpest contrast, attitudes are conceptualized as knowledge systems in Information Integration Theory (see *Functional Theory of Attitudes*, Chapter 8).

Attitude integration theory recognizes that attitudinal reactions depend on multiple determinants (see *Integration Diagram*). The need for methods that can deal with combined effect of multiple variables in the earlier quote from Aronson, et al. (1998) was repeated by Wilson, et al (2010, p. 79) in their call for a new synthesis that can assess "relative importance of several variables." Their discussion, however, reflects the general ignorance of how to handle the treacherous problem of measuring "relative importance" (see *Measuring Importance* below). Such synthesis of "multiply-determined behavior" was already well underway theoretically and empirically, with the three laws of information integration (e.g., Anderson, 1974a,b,c, 1981a,b).

INTEGRATION LAWS

The three basic laws of information integration are summarized briefly in the following sections. A two-variable integration task, A × B, is assumed. Extension to more variables is mostly straightforward. Other discussion is given in *Empirical Direction*, Chapters 20-21 (see also Anderson, 1982, Sections 3.3, 3.4). Additional detail on testing a law and estimating parameters is given in the final Note 22.

ADDITION LAW

The addition model for a two-variable design may be written

$$\rho_{jk} = \psi_{Aj} + \psi_{Bk}. \qquad (1)$$

Here ρ_{jk} is the internal psychological response to stimulus combination $\{S_{Aj}, S_{Bk}\}$ in row j, column k of the integration design, with respective psychological values of ψ_{Aj} and ψ_{Bk} (see Integration Diagram, Figure 6.2 below; see also Note 4).

Testing even this simple model might seem impossible; it involves three nonobservables: ρ, ψ, and +. Fortunately, it suffices to measure ρ_{jk} by virtue of parallelism analysis.

Parallelism Analysis. If the addition model is true, the row × column graph of ρ_{jk} will be parallel. Of course, this graph of ρ_{jk} is unobservable.

However, if your measured R is linear ($R_{jk} = c_0 + c_1 \rho_{jk}$), then its row × column graph will also be parallel. If you have a linear response measure, therefore, you need only test whether your observed graph is parallel. You need know nothing at all about the stimulus values, ψ_{Aj} and ψ_{Bk}.

Fortunately, adding-type models hold in many empirical situations. The first empirical demonstration (Anderson, 1962a) succeeded because it used the method of functional rating (see above), which avoids nonlinear biases of common rating methods. Since then, parallelism analysis has done well in most fields of human psychology (Note 5).

Benefits of Parallelism. Observed parallelism supports multiple benefits noted in Chapter 1 that are repeated here.

Additive Integration. Parallelism supports an adding–type model, either averaging with equal weights or strict adding.

True Response Measurement. The observed R_{jk} is a linear measure of the unobservable ρ_{jk}. This benefit has special value because of the wide applicability of the method of functional rating. With such method, moreover, pattern of nonparallelism in an integration graph is a valid picture of configurality in the nonobservable response (see *Response Generality* below).

True Stimulus Measurement. The true stimulus values, ψ_{Aj} and ψ_{Bk}, are estimated by row and column means of the data table (see Note 4). This stimulus measurement holds for individuals.

Meaning Invariance. Each stimulus informer has constant value, regardless of which other stimulus it is paired with. Still-popular claims about interactive change of meaning were shown to be invalid.

Cognitive Unitization. Complex stimulus fields function as cognitive units in an algebraic law. Functional measurement can finesse all complexity of the valuation operation to yield the functional value of a complex stimulus field (Anderson, 1981a, Section 1.1.5).

Unitization is invaluable for psychological theory. Complex stimulus fields are common, but the integration laws can treat them as units. The psychological laws justify Cognitive Unitization, a unique tool for studying cognition. As one example, an integration law such as Blame = Responsibility + Consequences implies that all three terms are unitary cognitive constructs at the level of judgment.

Cognitive Unitization seems a general-purpose capability. Hence it may hold even with nonsimple integration processes. An excellent explication with reference to face cognition and pain is given by Oliveira, Silva, Viegas, Teixeira, & Gonçalves, 2012).

MULTIPLICATION LAW

Some variables are expected to multiply. Subjective Expected Value = Subjective Probability × Subjective Value is the classic example. This model may be written formally as

$$\rho_{jk} = \psi_{Aj} \times \psi_{Bk}. \quad (2)$$

This SEV model had been widely conjectured but the first valid test seems to be that given with the linear fan analysis introduced by Anderson and Shanteau (1970). This test supported the SEV model.

This multiplication model predicts a fan of straight lines in the integration graph when the ψ_{Bk} are spaced at their functional values on the horizontal. These functional values are estimated by the column means of the integration data table (see Note 22).

This linear fan analysis has done well empirically. Besides expected value in children and adults, other applications include motivation, predictions of behavior, and language (see e.g., Anderson, 1981a, Figures 1.13-1.19; *Empirical Direction*, pp. 711-716).

Matching Law. Functional measurement theory allows a new look at Herrnstein's "matching law" (see Herrnstein's collected papers in Rachlin and Leibson, 1997). The matching law asserts that each of two choice alternatives is chosen in proportion to the relative frequency of its reinforcement:

$$\frac{R_1}{R_2} = \frac{r_1}{r_2} = r_1 r_2^{-1}. \quad \text{(behavioral matching law)}$$

where R and r denote response rate and reinforcement rate, respectively.

But this behavioral matching law cannot deal with reinforcers of different quality as with different foods. This limitation was denied by Rachlin (1971) and Killeen (1972), who argued that it implicitly defines the concept of reinforcement. Their argument is incorrect; the matching law can be supported or disproved with linear fan analysis (Anderson, 1978, Note 1). Linear fan analysis is markedly superior to the tests that have been used in studies of the matching law (Anderson, 1978, p. 375, 1996a, p. 330).

This behavioral matching law can be paralleled by a psychological matching law using psychological ψ values in place of the observable r values.

$$\frac{R_1}{R_2} = \frac{\psi_1}{\psi_2} = \psi_1\psi_2^{-1}. \quad \text{(psychological matching law)}$$

This psychological matching law predicts a linear fan pattern—if the response ratio on the left is a true proportional scale. Such linear fan patterns were obtained by Farley and Fantino (1978) who used food and shock as the two "reinforcers," (see Figure 11.10, p. 312 in Anderson, 2001. This result is remarkable since it implies that the observed response ratio was a true proportional scale and measured food and shock in equivalent terms. Other applications of linear fan analysis are cited in Section 21.5 of *Empirical Direction*).

AVERAGING LAWS

The averaging model makes explicit allowance for importance *weights* that multiply polarity *values*. Analogous to Equation 1,

$$\rho_{jk} = \frac{\omega_{Aj}\psi_{Aj} + \omega_{Bk}\psi_{Bk}}{\omega_{Aj} + \omega_{Bk}}. \qquad (3)$$

Weight and Value. The distinction between importance weight (ω) and value polarity (ψ) deserves comment. *Weight* refers to amount of information constructed from a stimulus informer; *value* refers to its polarity on the dimension of response.

As a concrete illustration of this weight–value distinction, consider the task of judging the proportion of women to men in some group. Samples of 3 women and 1 man have the same *value* (.75) as samples of 6 women and 2 men. But the larger sample contains more information and so has greater *weight*.

Equal Weight Averaging. *Equal weight* means that all ω_{Aj} are equal and so also all ω_{Bk} in Equation 3. The denominator of Equation 3 is then constant so the model is formally equivalent to the addition model. All benefits of the addition model listed above apply to equal weight averaging.

The simplicity and power of parallelism analysis mean that experimental procedures to produce equal weighting may be well worthwhile. All row stimuli should thus contain equal amounts of information and so also all column stimuli. Equal attention by participants to all levels of each stimulus variable is similarly desirable (Note 2).

Equal weighting may be adequately approximate in some tasks. This holds for the standard personality adjective task—for the standard re-

sponse of likableness. For judging honesty, in contrast, the traits *reliable* and *humorous* would have unequal importance weights. The many empirical findings of near-parallelism indicate usefulness of empirical procedures to facilitate equal weighting within each variable separately.

Unequal Weight Averaging. With unequal weights within any one variable, the averaging model will generally produce nonparallelism. The pattern of nonparallelism may reveal the pattern of weighting, as with the negativity effect (greater weight of more negative stimuli).

The averaging model has the notable property of making it possible to measure importance *weight* separately from polarity *value*. Exact measurement requires the Average program (Zalinski & Anderson, 1989, 1991). Special cases can yield a linear scale or rank order of importance (Anderson, 1982, p. 97). A simpler approach may be possible when weight can be expressed as a simple function of value as with the negativity effect (see Note 22).

Adding Versus Averaging. What produces averaging rather than adding is puzzling. Insightful work with children by Schlottmann, Harman, and Paine (2012) found an averaging law when the task required an inference from the sample to some underlying property but an adding law when the inference rested on the sample itself.

Decision Averaging Model. The decision-averaging model may apply when the task involves compromise between two alternatives. The two alternatives correspond to values, and these may be set at 1 and 0. Fair sharing between two persons, A and B, is one example. Deserving of each person corresponds to the weights. Equation 3 then becomes

$$\rho_{jk} = \frac{\omega_{Aj}}{\omega_{Aj} + \omega_{Bk}} \tag{4}$$

(see Anderson, 1981a, Section 1.6.4).

This decision averaging model has the same ratio form as a popular Bayesian model for two–choice tasks. The same ratio form also appears in Luce's (1959) choice model. These Bayesian and choice models, however, apply only to response probability, whereas the decision averaging model of IIT applies generally to metric response (Anderson, 1981a, Section 1.7.4). This metric ratio model has done well empirically.

Opposite Effects. The averaging model makes a counterintuitive prediction—the same stimulus informer may have opposite effects illustrated in Figure 6.1, which shows attitudes toward U. S. presidents based on

one or two biographical paragraphs. The near-parallelism of the three solid curves based on two paragraphs supports an add-ave model. But the dashed curve, based on just one paragraph, crosses over the two-paragraph Med curve. This rules out strict adding, supports averaging—the Med paragraph averages up the Lo paragraph, averages down the Hi paragraph. Opposite effects also rules out the sure-thing axiom, once seen as foundation for utility theory (Anderson, 1996a, pp. 322ff).

Prior State. People often have some prior opinion about any judgment situation. In the personality adjective task, for example, prior opinion corresponds to belief or expectation about people in general. Prior state acts as an additive constant in the adding model and also in the averaging model with equal weights. Hence it can be ignored for many purposes. The parallelism theorem still applies, with its listed benefits (Note 6).

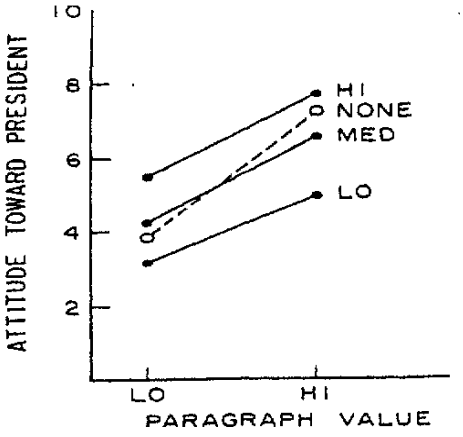

Figure 6.1. Parallelism of the three solid curves is strong support for an add-ave model. Crossover of dashed and medium curves eliminates addition/summation models, supports averaging theory.

Each of 48 person served in all eight experimental conditions; to get equivalent power using between person design would have required 510 participants. This within person design was made possible with the standardized president paragraphs (*Batteries of Stimulus Materials*, above).
(After Anderson, 1973.)

With unequal weights in the averaging model, however, prior state cannot be neglected in estimating parameters. In Equation 3, the term $\omega_0 \psi_0$ would be added in the numerator, the term ω_0 in the denominator. These terms would need to be included in the data analysis which can be done using the Average program (Zalinski & Anderson, 1991, pp. 377ff). Impressive study of pain descriptors is given by Oliveira, et al. (2007).

Qualitative Tests. Unequal stimulus weights yield predictable deviations from parallelism in the averaging model. To illustrate an often useful qualitative test, suppose more serious Consequences have greater importance weight in the blame law: Blame = Responsibility + Consequences. Then the Responsibility curves in the integration graph will be closer together for more serious Consequences. Three simple qualitative

tests are discussed in Anderson (1981a, pp. 65f). Equal weighting on one variable allows rank ordering of weight for levels of other variables (Anderson, 1982, p. 97). Valid comparisons of weight of different levels of another variable would thus be available directly from the integration graph, without the Average program.

QUALITATIVE INTEGRATION MODELS

Qualitative integration models can portray main trends in the data without claim to be exact. Qualitative models have not yet been much used. Previous work has been focused on exact models in part for their substantive interest, in part as a base and frame for establishing methods for true measurement of response and stimulus. For some purposes, however, approximate quantitative models may be satisfactory.

One qualitative integration model is given in the amnesty study in Togo (see Note 6 in *Algebra of Forgiveness*, Chapter 7). Analogous qualitative models may be expected in multiattribute judgment–decision. The usefulness of qualitative models depends in good part on using a response measure that is approximately linear.

CONFIGURALITY

The method of functional rating has the invaluable property of response generality (see *Response Generality* below). Hence deviations from parallelism suggest some configural component. The negativity effect was discovered in this way (Anderson, 1965; Hendrick & Costantini, 1970). Moreover, discounting from stimulus inconsistency was found to be much smaller than expected from the many claims that had been made (Anderson & Jacobson, 1965). Howe (1991) found that court judges down-weighted justification for a harmful action when the harm was small whereas students used an adding-type rule (Chapter 4).

An important configurality—*imputations* about missing information—was discovered by Leon (1980). Important work on imputations has been done by Singh (1991, 2011) and by Colleen Surber Moore on self-reports about medication acceptance from different perspectives (Wills & Moore, 1994, 1996; see also Oliveira, et al., 2012).

An interesting configurality in judgments of anticipated quality of life was reported by Muñoz Sastre and Mullet (2012). Anticipated quality of life was an additive function of five troubles such as anxiety/depression and pain/discomfort with a one-point exception: with no troubles, anticipated quality of life was higher than the additive model

predicted. This finding illustrates the power of integration models for configural analysis.

Michael Birnbaum (see Birnbaum & Zimmerman, 1998) has reported studies of averaging theory with a special form of configural weighting: the weight parameter of any stimulus informer depends on the rank of its value in relation to the other levels of that variable. The second highest stimulus, in particular, would get reduced weight.

This rank-weighting assumption is limiting. As one limitation, the second highest stimulus might contain more information in several different ways and so have greater weight. The highest stimulus, for example, might come from a source of lower reliability and so have lower weight (see also Anderson, 1996a, pp. 133f, Note 4). Other models with rank-dependent weighting are discussed by Weber (1999).

A different conceptual view is indicated by averaging theory: the weight parameter of a stimulus informer depends on the amount of information constructed from it. Rank-dependence may merely reflect weight–value correlation, as with the negativity effect. This alternative could be tested by manipulating weight independently of rank value, as with source reliability or amount of information.

MULTIATTRIBUTE MODELS

Multiattribute models have been much used in judgment-decision theory for choosing among alternative courses of action. An early example was given by Benjamin Franklin (Note 7). The basic idea is simple cost-benefit analysis. Represent each alternative by a set of independent attributes, assign a polarity value and an importance weight to each attribute. Choose that alternative with the largest weighted sum.

We all do something of this sort in making choices although seldom in a systematic way. Bok's (1999, pp. 49f) *Lying* follows a similar mode of evaluating justifications for lies as "questions of benefit and harm," although repeatedly pointing out the social–moral complications from overemphasizing personal benefits and underemphasizing harm to others and to self.

Cost-benefit analysis is often used in industry, as in selecting location for a new plant or store. Among the attributes would be labor supply, tax breaks from government, roads, transportation, and so on. Measurement of values and weights often depends on expert opinion.

A critical obstacle is that the several methods for measuring expert opinion (e.g., tradeoff between pairs of attributes; distributing 100 points among the attributes; magnitude estimation; rating; part-worth) give dif-

ferent results, often quite different. The best alternative using one measurement method may be poor using another. Although books have been written on multiattribute analysis, this critical measurement problem has been left unresolved, largely slurred over.

This measurement problem can be resolved with functional measurement. Run an integration experiment on tasks such as job satisfaction that obey the averaging law. This law will provide correct measures of each attribute (Zalinski & Anderson, 1991). These measures provide validity criteria for those obtained by standard methods such as tradeoff or point allocation. Both of these methods did poorly in Zhu and Anderson (1991) whereas part-worth showed promise (see also Wang & Yang, 1998). The main goal, of course, is to use this more tedious method to develop valid methods for self-measurement in the many tasks in which it is needed (see *Self-Measurement Theory* above).

GOODNESS OF FIT

Perfect fits between model and data are not expected. Discrepancies from model predictions will result from normal response variability. Are the observed discrepancies any more than normal response variability? *This is what it means to test goodness of fit of a model* (Note 22).

Analysis of Variance. Standard analysis of variance can give optimal tests of goodness of fit for all three integration models. This is simple with the addition model. The statistical interaction in analysis of variance provides a test of deviations from parallelism in a two-variable graph. This test also holds for the averaging model with equal weights.

Statsig deviations may result from nonadditive integration or from nonlinear response bias. The method of functional rating is expected to eliminate response bias so that statsig deviation may reasonably be considered real nonadditivity.

For the multiplication model, the linear × linear component of the statistical interaction should be statsig. The residual interaction should be nonstatsig. Weiss (2006) includes a disc with a computer program. More detailed discussion is given in Anderson (1982, pp. 72-85).

The averaging model with unequal weights can be analyzed using the Average program. This is not simple and is not considered here (Zalinski & Anderson, 1991; *Empirical Direction*, p. 732, Note 21.4.3a).

How Not To Test A Model. Standard correlation–regression analysis is generally invalid for testing models of psychological process. Correlations are invalid in principle; they fail to test the *discrepancies* between

model predictions and data. Examples in which additive regression models yielded correlations from .977 to .996 for severely nonadditive data are shown in Anderson (1982, Figures 4.2-4.5). Correct analysis is possible with functional measurement (Anderson, 1982, Section 4.3, *Regression Analysis*; see similarly Blanton & Jaccard, 2006c).

This issue is related to the process–outcome distinction discussed previously. *Prediction* of behavior is an *outcome* concern for which high correlations are desirable and usually sufficient.

But *understanding* behavior is a *process* concern that usually requires analysis of discrepancies from prediction. Extremely high correlations can easily be obtained from models that seriously misrepresent cognitive process (Anderson, 1962a; Anderson & Shanteau, 1977; Birnbaum, 1976; Parker, Casey, Ziviar, & Silberberg, 1982).

Relevance of this process–outcome distinction to model analysis may need emphasis because of common focus on prediction to neglect of understanding. This issue needs consideration in the planning stage. Otherwise a valid test of the model may not be available. The eight models of attitude ambivalence discussed in Anderson (2008, pp. 140-145) were earnest attacks on an important problem that accomplished nothing at all owing to failure to use simple appropriate methods.

Interpreting Statsig Deviations. Statsig deviations from model prediction do not necessarily mean the model should be abandoned. The rating scale, for example, may be troubled by residual number preferences or end bias. The process envisaged in the model may be valid.

Alternatively, the deviations may result from some additional process not included in the model. The negativity effect (greater weight for more negative information) was discovered in this way as a deviation from parallelism (Anderson, 1965; see *Negativity Theory* below). Statsig deviations should be taken seriously but they need not be fatal. Indeed, they may reveal something new, as with the negativity effect (see e.g., *Interaction and Configurality*, pp. 357-364, Anderson, 2008).

UNIQUENESS

We are so familiar with numbers that have objective value, as with hours, miles, and dollars that we tend to take all numbers at face value. This is usually a mistake in psychology. “Interaction” in factorial design is a common example. Empirical reality of a statistical interaction depends on the assumption that the response measure is a linear (equal interval) scale. Without this assumption, which is rarely given justifica-

tion, statistical interaction may be empirically meaningless (see *Understanding "Interactions"* below).

Comparison of Stimulus Values. Estimates of ψ_{Aj} and ψ_{Bk} from the marginal means of the integration design are measured in terms of the response. Each set of estimates is thus on a linear scale.

Can a row mean be compared with a column mean? Not thoughtlessly; they likely have different zero points. However, both have the same unit as the response. Hence a difference between two row means can be compared with a difference between two column means because their zero points cancel out in the differences. One example appears in the relative range index below. Comparison of positive and negative values of a single variable is discussed in Note 22.

The averaging law has the remarkable property that it can separate weight from value and measure weights that are on a common proportional scale and thus properly comparable across different variables. Detailed discussion of uniqueness is given in Anderson (1982, Chapter 2).

Conflicting Goals. Interpersonal comparisons of value appear throughout everyday society in terms of fairness and justice. Elster (1995) gives an illuminating discussion of many such situations that involve integration of conflicting determinants, as with allocating scarce medical resources to those who will benefit most or to those in most need but with poor prospects (see further contributors to Elster, 1992).

Kahneman and Varey (1991) give a perceptive discussion of psychological considerations in judgments of fairness and justice that are neglected in the standard objectivist view in decision science. One is that people adapt to present circumstances and are poor at predicting their future hedonic state.

PSYCHOLOGICAL MEASUREMENT

True measurement of psychological quantities had been sought for over a century. True response measurement means that our observed response is a linear function of the unobservable quantity, R and ρ, respectively. For an adding-type model, similarly, the marginal means of an integration design should be linear functions of the stimulus ψ values. The roadblock is that ρ and ψ are unobservable, as emphasized in the Integration Diagram of Figure 6.2 following.

Most response measurement in psychology is monotone (ordinal), in which R and ρ have the same rank order. Monotone measurement is

widely useful for testing whether some variable has an effect and gives a rough idea of effect size. But it is not true linear measurement.

Empirical laws of information integration are the foundation for psychological measurement. Functional measurement theory is grounded on this base and frame.

> A guiding idea of functional measurement is that measurement scales are derivative from substantive theory (Anderson, 1970, p. 153).
> The logic of the present scaling technique consists of using the postulated behavior laws to induce a scaling on the dependent variable (Anderson, 1962b, p. 46).

The potential of this functional approach was illustrated with the six benefits of the parallelism theorem (Chapter 1). Such laws must have empirical reality for these benefits to be real. Such empirical laws are the foundation for theory of psychological measurement.

Algebraic laws had been widely conjectured, of course, as with the equity models of Chapter 2 and with Subjective Expected Value = Subjective Probability × Subjective Value. But without capability for psychological measurement, these conjectures remained conjectures. Indeed, the multitudinous empirical applications of analysis of variance throughout psychology failed to reveal even the simple adding law.

Using functional measurement, however, the initial 1962 study of person cognition supported an adding-type law in single person design and analysis. Later applications of functional measurement have done well throughout human psychology.

The three integration laws also showed that psychological measurement theory differs conceptually from what were and remain common conceptions.

THE NATURE OF PSYCHOLOGICAL MEASUREMENT

Psychological measurement differs fundamentally from standard conceptions, largely derivative from physical science. The nature of psychological measurement is implicit in the Integration Diagram, repeated here from Chapter 1. Metrication occurs primarily in the valuation operation, which transmutes informer stimuli, S, into goal-oriented values, ψ. This metrication is continued with the next two operations, integration and action. Metric value is not in the stimuli themselves, as with length and gram weight in physics. Instead, metric value is constructed by the organism—in relation to the operative goal. The same stimulus informer may thus have different values relative to different goals.

Metrication derives from purposiveness, especially the approach–avoidance nature of goals and action (see *Metric Cognition*, Chapter 7). Metrication originates in goal-directed thought and action.

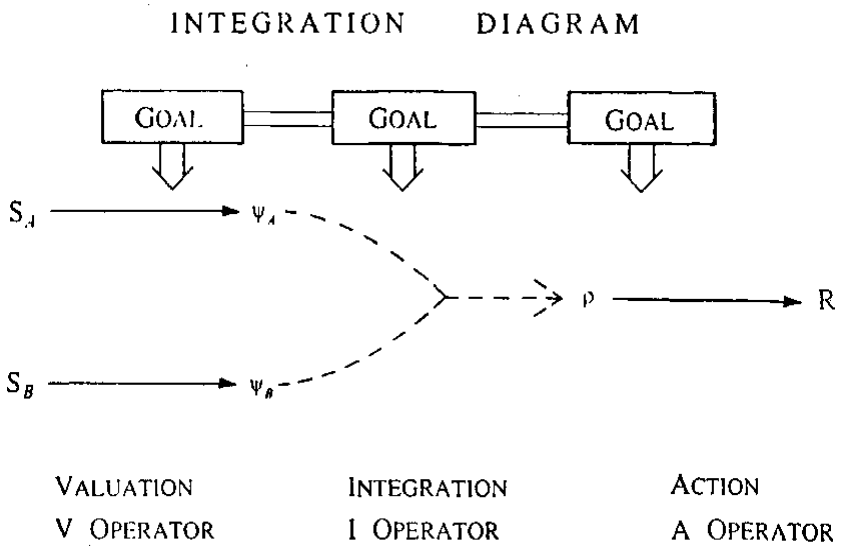

Figure 6.2. Information integration diagram. Chain of three operators, **V – I – A,** leads from observable stimulus field, {S}, to observable response, R.
Valuation operator, **V**, transmutes stimuli, S, into subjective representations, ψ.
Integration operator, **I**, transforms subjective field, {ψ}, into internal response, ρ.
Action operator, **A**, transforms internal response, ρ, into observable response, R.
(After N. H. Anderson, *Foundations of information integration theory*, 1981a.)

True measurement of ψ and ρ becomes possible by virtue of the algebraic laws of information integration (benefits 3 and 2 of the parallelism theorem). Monotone (ordinal) scales are widely useful to assess whether a given stimulus variable has an effect. But monotone scales have limited value with fundamental problems of information integration (see e.g., *Measuring Importance* and *Understanding "Interactions"* below). Empirical integration laws are the true foundation of psychological measurement theory.

Person cognition, because of its ubiquity in everyday life, is a prime area for theory of psychological measurement. In the basic trait adjective task, the value of *critical*, for example, is not in the adjective itself. Instead, it arises from goal-oriented valuation. Its value will differ if the goal is to judge research potential or friendship. Judgments of blame, to take a second example, may require preliminary metrication of harm

from an action that involves integration of a complex stimulus field including potential harm that did not actually occur (Chapter 3).

Influence of context is a basic problem discussed by many writers from Aristotle through the Gestalt psychologists to Person × Situation personality theorists. The integration laws provide a base for context theory because they can measure a complex context as a goal-oriented cognitive unit (see *Analytic Context Theory*, Chapter 7).

The internal field elicited by a given stimulus may contain substantial metric components from previous learning. This is the case with many attitudes, social–moral attitudes especially. These components may be partly responsible for the common misconception of attitudes as one-dimensional, good-bad evaluations ("readiness to respond"), which imposed a misleading conceptual framework on the attitude field.

This integration-theoretical conception of psychological measurement applies also in psychophysics. Unfortunately, traditional psychophysics induced a conception of measurement derived from physical science, in part because many sensory stimuli have physical metrics and many psychophysical sensations seem one-dimensional. The present integration-theoretical conception of measurement is necessary, however, as may be illustrated with the size-weight illusion, in which felt heaviness of a given gram weight depends also on its visual appearance. In the experiment of Figure 7.2, weight was measured in grams and size by the centimeter side of cubical blocks. But "size" could be varied in irregular shapes and influenced by visual cues such as hue. The integration law will still apply (no doubt), allowing a true *psycho*physics.

True measurement in psychology rests squarely on our inestimable good fortune of the three algebraic laws of information integration.

These empirical laws are the foundation of true measurement.

RESPONSE GENERALITY

A response method that has proven linear in a variety of situations may reasonably be expected to be linear in others without requiring specific evidence. Such response generality holds for the method of functional rating—including the practical precautions already discussed. This method may thus be useful in situations in which parallelism validation may be overly demanding or not possible (*Interaction and Configurality* below). A similar approach may be feasible more generally with behavioral responses such as bar press rate by rats (Anderson, 2002).

RESPONSE UNITIZATION

The unitization principle for informer stimuli (benefit 5 of the parallelism theorem) also holds for many responses. Response unitization reflects the general goal–directedness of living (Axiom of Purposiveness). This twofold stimulus–response unitization underlies the algebraic integration laws.

Many responses, of course, include more than a single quality. Examples range from forgiveness (Chapter 7) and attitudes toward women (see third following section) and to taste (Marks, et al., 2007; McBride & Anderson, 1990). It may thus be desirable to develop profile measures for separate response qualities.

RESPONSE QUALITY

Although *quality* of what is being measured has paramount importance, quality has been relatively neglected. Most empirical investigators take for granted that their instructions to judge blame, attitude, well-being, and so on, elicit a unitary concept, the same for different persons. This approach is undoubtedly useful but it has limitations and dangers.

Measurement theory in psychology has fixated even more strongly on quantity and neglected quality. Nearly all these theories have been constructed in the image of physical science, reinforced by physical metrics that underlie common sensations such as loudness and heaviness.

This neglect of quality reflects a widespread tendency to impose one-dimensional views on complex entities. The great usefulness of one-dimensional measures has obscured real limitations. A fateful misuse appears in the traditional definition of attitude as one-dimensional evaluative reaction. This misconstrues the phenomena (see *Functional Theory of Attitudes* in Chapter 8).

INTERACTION ANALYSIS

Functional measurement can help study integrations that do not follow simple algebraic rules. If the response measure is linear (equal interval), as with the method of functional rating, the pattern in the response will give a true picture of pattern in the underlying cognition (see Note 7* in Chapter 3). Statistical interactions may then have substantive meaning.

MULTIPLEX MEASURES

A single dimension is inadequate to represent many qualities. This issue may be illustrated with a well-known scale of attitudes toward women. Participants made separate positive/negative judgments of statements about women in multiple contexts, including workplace, home, personal conduct, and so on. Factor analysis was used to enforce a single factor, in line with the prevalent conception of attitudes as one-dimensional, good-bad evaluations. Such measures can be useful for some purposes, such as assessing population stereotypes or work opportunities.

But attitudes toward women have multiple aspects, as this scale implicitly recognized. A person strongly in favor of gender equality in work might still be averse to gender equality in the home and shrink from female obscenity. Multiplex measures may thus be desirable, with separate measures for different social contexts (see *Profile Analysis*, pp. 185f, in Anderson, 2008).

Multiplex response technique may be useful. In legal psychology, duplex response showed that punishment responses may have distinct components of restitution and retribution (Hommers, 2007; Hommers & Anderson, 1991, pp. 122f). In deserving theory, Farkas showed that praise and money were distinct components of outcome, both of which obeyed averaging theory (Figure 2.5). In each case, the usual single response would obscure quality of the underlying feeling.

An obvious approach to multiplex analysis would ask for joint judgment of several qualities. Judgment of gratitude, for example, could ask for separate judgments of positive feeling toward the benefactor, personal benefit from the benefaction, obligation to and cost to the benefactor. A similar approach could be used with other qualities such as forgiveness (see *Defining and Measuring Forgiveness*, Chapter 7; see also Notes 20, 21).

Multiplex measures may be especially desirable to represent the conflict and ambivalence that are so common in moral cognition. Separate measures of positive and negative have particular importance. Examples include attitudes about wife–husband–child interactions, about lying, forgiveness, moral dilemmas, and other common conflict situations discussed in the next chapter.

COMPARISON OF PERSONS OR GROUPS

Some difficulties of comparing different persons or groups are well recognized, as shown in attempts at statistical control (see *Illusion of "Sta-*

tistical Control" below). True measurement is possible for each individual with an algebraic law but individual differences in zero and unit trouble comparisons of their responses.

Rating responses, in particular, are often compared across different groups of persons. Examples include comparison of subjective well-being or moral judgment across gender, groups, or cultures. Such comparisons are common but their validity is often taken for granted. Actual evidence for validity, however, hardly exists (see also *Uniqueness* above). Indeed, the meaning of validity is often unclear.

Thus, a rating of "7" may represent rather different underlying feelings for different people. One example of this problem appears in the spate of articles claiming that people's self-reports of their well-being or happiness are often surprisingly little affected by their material circumstances. "Objective measures of health, diet, and working conditions, and especially numerous negative affects, may be more meaningful. And more useful for improving society." (Anderson, 2008, p 291).

Comparison across different cultures is a special problem about which little is known. An instructive beginning using Information Integration Theory to study Quality of Life (Well-being) in three different cultures (Belgium, Poland, Algeria) is presented by Theuns, Baran, Van Vaerenbergh, Hellenbosch, and Tilouine (2012).

Comparison of *patterns* of response can be meaningful with linear response measures (e.g., Figure 1.2). Cogent demonstration of pattern comparison is discussed above under *Cluster Analysis*.

INTEGRATION PSYCHOPHYSICS AND PERCEPTION

Relations between the dual worlds, internal and external, are central in psychophysics and perception. The main approach has been to use structure of the external world to study structure of the internal world. A polar opposite is provided by the three integration laws of IIT. These laws understand the internal world in its own terms.

Misconception of Psychophysical Law. The conception of psychophysical law was inspired by physical laws such as Galileo's law of falling bodies and Newton's law of universal gravitation. These physical laws inspired dreams of *psycho*physical laws relating the internal world of sensation to the external world of physical stimuli.

A notable proposal for psychophysical law was Fechner's ingenious idea of using just noticeable differences as additive units, quite like centimeters on a measuring tape. This led to his logarithmic law—which

was overwhelmed in modern times by Stevens' power law based on numerical ratio judgments of sensation magnitude. But the critical prediction that the power exponent should be constant, independent of the standard, failed its first experimental test (Stevens, 1956) and failed continuously ever since until it was finally abandoned leaving no validational test (e.g., Anderson, 1981a, pp. 340ff).

An obvious bias undercuts Stevens' method of magnitude estimation: diminishing returns in the use of numbers, as though the difference between 85 and 86 is less than between 5 and 6. One illustration of this bias appeared in the power function for lightness (grayness): positively accelerated for magnitude estimation and negatively accelerated for functional measurement (see Anderson, 1996a, Figures 3.1, 9.3, 9.8, and 9.9). Stevens work with magnitude estimation echoes the Sheldon–Stevens (1942) long-discredited somatotype theory, that body type and personality were very closely related (see p. 222 in *Empirical Direction*).

Direct Perception. Information Integration Theory differs fundamentally from Gibson's theory of direct perception which is grounded on correct perception of the external world and must explain illusions by "supplementary assumptions" (1966, p. 287). IIT, in contrast, treats illusion as normal information processing illustrated in following sections.

Nor can Gibson's theory recognize internal senses such as taste and pain, both of which have exhibited algebraic integration laws. These same laws have shown promise in other areas also outside the scope of Gibson's framework, as with learning of perceptual-motor skills.

Affect. The functional nature of psychophysics is manifest in approach–avoidance senses. Integration laws for taste/odor/flavor were established in extensive work by McBride (e.g., 1993; McBride & Anderson, 1991) and by Marks, et al. (e.g., 2007). As one example, McBride established a dominant component model for total intensity of sucrose-citric acid solutions using functional measurements (see Figure 9.6, p. 294 Anderson, 1996a).

A new foundation for pain theory has been developed by Armando Oliveira and his associates based on integration laws (e.g., Oliveira, et al, 2007; de Sà Teixeira & Oliveira, 2007). This allowed extension of three categorical scales of pain to exact metric values. Of special interest, they have replaced the common holistic conception of facial expression of pain, important in clinical cases, with an integration law for facial action units. Other integration studies of pain have been done by Algom and associates (e.g., Algom, Raphaeli, & Cohen-Raz, 1986).

Nonconscious Sensation. Nonconscious processing, fundamental in thought and action, becomes analyzable with IIT. One example appears with the size-weight illusion of Figure 7.2 of the next chapter. Notably, this nonconscious measure of heaviness of visible lifted weight agreed with the conscious measure obtained from a collateral test of judging average heaviness of two unseen lifted weights (see 9-point itemization of results on the size-weight illusion in Anderson, 1996a, pp. 286-290).

Analysis of phenomenology is severely limited without ability to assess nonconscious processing. One example is the analysis of the phenomenal change of meaning interactions in person cognition (see *Science of Phenomenology*, Chapter 7).

Signal Detection Theory (SDT). Functional measurement provides a notable extension of SDT to use metric response. Algebraic integration structure replaces the assumption of normal distribution. The SDT claim of constancy of the d' measure of signal intensity appears as the parallelism property of adding models (see *Integration Decision Model*, Anderson, 1982 pp. 129-134).

Metric response provides more power than choice response. Moreover, combined effects of multiple variables is readily studied.

Illusion Integration Theory. In psychophysical illusions, some ostensibly irrelevant stimulus is integrated into perception of a focal stimulus. A unifying framework with analytical power may be possible with the integration laws. One example was the size-weight illusion just discussed.

An integration-theoretical interpretation of the striking Müller-Lyer illusion (two equal line segments, one directly above the other, one ending with inward pointing arrows, the other ending with outward pointing arrows, seem manifestly unequal) was suggested by Massaro and Anderson (1970, 1971). They found that 3-dimensional versions disagreed with R. L. Gregory's interpretation in terms of transfer from 3-dimensional corners of rooms.

The integration-theoretical interpretation assumed that the apparent length was an average of actual length and the shorter/longer distance between the flanking arrows. This interpretation agrees with findings of similar illusion with flanking figures of semicircles, cited by Gordon (2004) as evidence against Gregory's interpretation. Experimental analysis would be possible by varying size or angle of the flanking arrows in a two-factor design.

Geometrical illusions have excited attention for over a century, as with the above Müller-Lyer illusion and frequently appear in introductory texts. Virtually all these theories have relied on a single process: either

assimilation-type or contrast-type. But all these theories have had very limited success.

A new approach, two-process illusion theory, was proposed by John Clavadetscher (1977, 1991) based on integrated action of both assimilation and contrast. This approach became testable using methods of information integration theory. It had good success with the Ebbinghaus illusion (apparent size of a center circle surrounded by a concentric ring of smaller or larger circles). A key implication was that assimilation decreased rapidly with distance between the focal and context components whereas contrast decreased slowly. His data allowed an estimate of 10 to 1 in decay rates (Anderson, 1996a, p. 303).

FAULTY METHODS

Reliance on faulty methods has undercut much earnest effort. Some common faulty methods are discussed here.

ILLUSION OF "STATISTICAL CONTROL"

Comparisons across groups of people who differ preexperimentally suffer from confounding with correlated variables that undercut causal interpretation. Examples include family background in moral development, personality variables in marriage, school and aptitude variables in education programs, a large proportion of results cited in health psychology, and many, many others.

Some investigators attempt to eliminate such confounds by employing "statistical control" such as multiple regression, partial correlation, quasi-experimental design, analysis of covariance, or causal, structural equation models. These are nearly always statistical illusions.

> Many writers assert that multiple regression "controls," or "holds constant," or "partials out" uncontrolled variables. Such phrases seem to justify some causal interpretation. It would be wonderful if this were true, but it is false. Regression equations **do not** *control or hold constant* in any substantive sense. (*Empirical Direction*, p. 501.)

Two confoundings are critical for such "control." *Missing variable confounding* refers to a variable that is operative in the empirical situation but not measured, directly or indirectly, for inclusion in statistical analysis. Hence it is impossible to "control." Statistical "control" can, as various writers have shown, seriously misrepresent importance of variables. Importance weights can even change sign.

Missing variables are almost unavoidable. One major cause is that the usual measured variables are one-dimensional whereas the corresponding psychological variables are usually multidimensional, as with attitudes towards women.

Person–variable confounding refers to the assumption that variable effects are equal across different individuals. This improbable assumption further vitiates claims to control uncontrolled variables (*Empirical Direction*, Sections 13.2 and 16.2).

Quasi-experimental design suffers in both these ways in its reliance on analysis of covariance to "control" uncontrolled variables. This artifact has since been acknowledged by the primary inventor of quasi-experimental design (Campbell, 1978). Causal, structural equation analysis suffers similar problems plus another of its own (see references in *Empirical Direction*, pp. 505f).

The popularity of these attempts to "control" what is not controlled is due in large part to textbooks written by persons who lack understanding of relations of statistical analysis to empirical reality. These methods can be valid in certain narrow circumstances. Establishing these is no easy task, however, that is usually glossed over or just ignored.

> I sympathize with persons who have been taught that multiple regression is a means of "statistical control." It took me longer than I care to think to see through the smoke and mirrors. Disbarring such terms would prevent much confusion. (*Empirical Direction*, p. 514.)

MISUSE OF CORRELATION

Field investigators love to use correlation to imply causation. This misuse of correlation is ubiquitous in health psychology and in current positive psychology (see *Positive Psychology*, Chapter 7). It is often disguised by calling it "association" and/or by leading the reader to infer causality. Here is one example taken from the editors' introduction to *Understanding marriage* (Noller & Feeney, 2002, p. 1):

> Waite and Gallagher (2000) present empirical support for the proposition that there are several major advantages to marriage. First, married men and women tend to enjoy better mental and physical health than the unmarried. Second, married men and women are likely to have more assets and income than the unmarried, with marriage even being described as a 'wealth-enhancing institution' (Hao, 1996). Third, married people have more and better sex than the unmarried. Fourth, children of married parents also enjoy a number of advantages, including better physical and mental health, and higher levels of education and career success . . . . Marriage has clear implications for individuals' general sense of well-being.

This statement is intended to lead the reader to think that these benefits are *caused* by marriage—as the final sentence makes explicit. But the opposite direction of causality seems at least as likely; the married may have been better before they were married. People with ill health or psychological problems, for example, may be less likely to get married, to get well-paying jobs, or enjoy more and better sex. Much of what is written about observational data represents similar double talk.

Correlations can be valuable clues to causality. Finding poorer communication skills in less happy marriages, for example, might help develop experimental methods to improve such skills, together with valid tests of their efficacy. What is not appropriate is language that misleads readers to infer causality from correlation (Note 8).

I wish to add that the contributors to *Understanding marriage* were chosen for their interest in family interaction. Their chapters can be helpful for further study of this important social problem. This, however, requires better understanding how experiments can be useful for studying and bettering wife–husband interaction. Couple experiments offer one opportunity, as in Armstrong's work illustrated in Figure 3.2.

CONFOUNDING

Confounding is a major concern in empirical analysis. An experimental treatment intended to produce one causal process may be effective by virtue of another process that it also produces. The medical placebo effect, in which suggestion masquerades as medicinal effect, is the classic example (see *Confounding*, Chapter 8 in *Empirical Direction*). Besides the confounds of the two previous sections, two other important confounds are noted here.

Concept–Instance Confounding. Concept–instance confounding is common but often not recognized. Experiments that seek to manipulate some concept usually do so with specific stimulus instances. The concept is thus confounded with the specific instances.

Concept–instance confounding undercuts the frequent practice of comparing importance of two variables by comparing their main effects or statistical effect sizes. One of many examples is attempts to compare relative importance of objective and subjective variables (e.g., physical damage and intent) in judgments of blame (Chapter 3). Each main effect is completely confounded with its specific instances; opposite conclusions could generally be obtained with different instances. Using regression analysis only makes matters worse (see *Invalidity of Regression*

Weights below).

Correct analysis is possible when the averaging law holds; this law allows valid measures of importance weight. A simple application was given in ingenious work by Surber (1982, 1985). Other simple applications are possible with *Qualitative Tests* discussed above. Also, comparing main effects may be valid when variables cover some natural range (see *Measuring Importance* below*).*

Response Confounding. Researchers often take their instructions for granted—that when they tell participants, "judge X," participants do judge X. This practice seems plausible with fair division and with blame. Some moral judgments, however, may have more than one component and/or may mean different things to different people.

Gratitude is a simple example. One may be personally grateful to one's benefactor; one may be thankful for the benefit; one may feel under some obligation. These have rather different quality and deserve separate measurement which has yet to be done (see *Gratitude and Ingratitude* in Chapter 7). Multiple meanings are not uncommon, as with blame and forgiveness. *Profile Measures* may be useful (see above).

HALO EFFECT: A CONCEPTUAL PITFALL

Halo means that an overall impression of a whole influences judgments of some part of that whole. Halo effects have long been a concern in applied psychology. Supervisors' ratings of subordinates' performance are likely influenced by their likableness which may have little relation to actual performance. This work, however, rests largely on correlations. Rigorous experimental analysis is possible with IIT (see *Halo Integration Theory*, pp. 55-58 in Anderson, 2008).

Halo effects should be suspected whenever judgment is required about any component of an integrated whole. The classic example is the recurrent claim that trait adjectives change their meaning when integrated into a person cognition. This was found to be a halo artifact (see *Foundations of Person Cognition*, Chapter 3 in Anderson, 2008). An unresolved case from legal judgment is cited in Chapter 4. The halo effect illustrates the need for cognitive theory to validate introspection (see *Science of Phenomenology* in Chapter 7).

"BIAS" AND FUNCTIONAL THEORY

Bias refers to a deviation from some standard of accuracy or correctness. Some writers misuse the term by referring to a standard that is inappropriate or even nonexistent. Thus the *negativity effect* (greater importance weight of stimuli with more negative value) is popularly referred to as "bias." But often, perhaps typically, more negative stimuli carry more information and so *should* have greater importance weight (e.g., Anderson, 1972b). This is not bias but a sensible effect.

"Bias" is often misused in the attitude literature. The scholarly volume of Eagly and Chaiken (1993, p. 680) echoes a common view in saying that "Individuals resist influence through multiple cognitive processes that are *biased* in favor of their initial attitude" (italics added). Functional theory, in sharp contrast, conceptualizes attitudes as knowledge systems that have a proper function of utilizing past learning in present thought and action (Anderson, 1991b, pp. 210-227).

To illustrate, consider the carefully designed experiment of Lord, Ross, and Lepper (1979). Participants received a matched pair of alleged research reports, one arguing that capital punishment acted to deter murder, the other arguing the opposite. Participants judged the research report that supported their prior attitude as more convincing and better conducted. This outcome was interpreted as "biased processing" and is widely cited as a definitive demonstration.

This "bias" interpretation fails to understand the functional nature of attitudes. They are knowledge systems that help us get along in an uncertain world of limited information and personal goals. The widespread acceptance of "bias" interpretations reflects basic misconceptions about nature and function of cognition.

"Bias" interpretations rest on some assumption, often implicit, of some "correct" response. In this example, the pro and anti research reports were carefully constructed to be equal and opposite in face value. To a person with neutral prior attitude, the pro and anti messages would seem equally well done and equally convincing. But the proper function of prior attitudes is to aid present judgment. This is a vital function even though it will usually rest on insufficient information and often be controversial. Thus, the perpetual political disagreements between conservatives and liberals stem from different attitudes that often have a sensible base not recognized by the other side (see also *Juror Bias*, Chapter 4).

Often a correct standard is simply assumed, as in many studies of ethnic prejudice or social stereotypes. Rightly speaking, the "bias" is in

this assumption. Such beliefs may be socially undesirable, but calling them "biased" is a claim for being a God of truth.

Misuse of "bias" also pervades judgment–decision. Much such "bias" represents deviations from optimal behavior prescribed by normative models. Once it is realized that normative framework is largely foreign to cognitive process, the "bias" vanishes (e.g., Anderson, 1968b, p. 392, 1996a, pp. 344-351; Shanteau, 1978). The once-much-studied "conservatism bias," which arose from mispressing normative Bayesian statistical theory into a descriptive, psychological mode, is a prime example (Anderson, 1982, p. 333). Once the cognitive irrelevance of the normative Bayesian model is realized, conservatism is seen to be a "non-effect."

Why is "bias" so popular? One answer is that it attributes psychological reality to the "bias," making it seem that the data mean something, that they reflect operation of real cognitive processes, as in the above quotation from Eagly and Chaiken and with "conservatism." "Such terms as . . . *bias* . . . are attractive because they invoke the image of dynamic, interactive psychological processes. At present, however, they are largely free-floating theory" (Anderson, 1982, p. 336). Often, they obfuscate the real problem of understanding the phenomena.

The basic question concerns choice of fruitful research issues, of which "bias" is a minor example. More general issues come from the general functional conception that underlies Information Integration Theory, which has led to different conceptual foundations in many areas, including memory, learning, social attitudes and moral science (see sections in Chapter 8; see also *Achievement*, pp. 365-373, in Anderson, 2008). This general issue of fruitful choice of research issues deserves discussion from many points of view.

MEASUREMENT PITFALLS

Some popular methods depend on untested and unlikely assumptions about measurement. One of these concerns *statistical* interaction in analysis of variance, widely confused with everyday English meaning of interaction. A group of others concern measurement of importance. Two others relate to regression analysis and arbitrary metrics.

UNDERSTANDING "INTERACTIONS"

"Interactions" in analysis of variance are often meaningless. Statistical interaction is *defined* as a deviation from additivity in observed data.

Any deviation from parallelism in a two-variable graph is thus an interaction. But such deviation may be merely artifact of a nonlinear response measure, completely meaningless (Anderson, 1961; *Empirical Direction*, Table 7.1, Section 7.6.4, p. 193).

To illustrate this issue of nonlinear response, consider a task for which either *time* or *speed* may be used to measure performance. Both have often been used since shorter time and faster speed both signify better performance. But speed = 1/time. Hence an additive law for time would yield a nonadditive interaction for speed. And vice versa; an additive law for speed would yield a nonadditive interaction for time. Analogous ambiguity is common in the literature. Unless evidence for linear response is given, reported interactions may be completely meaningless (Anderson, 1961; *Empirical Direction*, Table 7.1, p. 193).

Statistical "interaction" may also be produced by applying standard analysis of variance to data from a nonadditive model, for example, a multiplication model such as Subjective Expected Value = Subjective Probability × Subjective Value. This interaction merely reflects use of the wrong integration model, not from any interaction that changes values of the stimuli themselves. If the correct model is used (and if the response scale is linear), the "interaction" will be properly represented in the linear × linear component.

Most statistics texts reify statistical interactions as though they had substantive reality (see e.g., Anderson, 2001, p. 208). This is one symptom of failure to appreciate importance of psychological measurement theory. Statistical interactions always deserve attention but they are often unreal. Even when real, moreover, they are often unimportant.

Besides *Empirical Direction*, no other statistics text I know of tells how to assess response linearity on which meaningfulness of "interaction" depends. Most texts seem unaware of this problem (see Chapter 7, *Understanding Interactions,* in *Empirical Direction* which gives a thorough discussion). Much confusion could be avoided if "interactions" were called by their correct name—*residuals* from an arbitrary, Procrustean additive model to use Tukey's apt adjective.

True linear response measures can be established with functional measurement theory. Statistical interactions may then be psychologically meaningful (Note 9).

Functional measurement can help study stimulus interaction and other processes that do not follow key algebraic laws. With a linear (equal interval) response, pattern in an integration graph is a true picture of underlying response. Statistical interactions would then have substan-

tive meaning. The negativity affect (greater weight of more negative stimuli) was discovered in this way (Anderson, 1996, Note 5, p. 134).

MEASURING IMPORTANCE

Many investigators desire to assess relative importance of two variables to show meaning and worth of their results. Satisfying this desire, which might seem straightforward, turns out to be treacherous and difficult. It is possible, however, with averaging theory using the Average program (Zalinski & Anderson, 1989, 1991; see also *Comparison and Measurement of Importance*, Sections 2.3.2, 2.3.3, and 6.1 in Anderson, 1982; see also *Regression Models* below and *Scale Types* in Appendix).

Main Effects In Factorial Design. In this popular method, the importance of each variable in a factorial design is measured by its main effect, specifically, by the difference between its largest and smallest levels. This is not generally valid because it depends on arbitrary choice of stimulus levels of each variable. Different choices could give opposite conclusions (see *Concept–Instance Confounding* above).

Main effects can be valid measures of importance when the levels of a variable cover a natural ecological range. Then its main effect represents its ecological importance. This may be properly compared with another such variable. An impressive application to face cognition was given by Oliveira, et. al., (2007; see Note 10).

Relative Range Index. The relative range index is a ratio of main effects. With two variables, A and B, the range R_A of variable A is defined as the difference between effects of its largest and smallest levels, and similarly for R_B. The relative range index for R_A is thus

$$RR_A = \frac{R_A}{R_A + R_B}.$$

When A and B each cover some ecologically natural range, RR_A measures relative ecological importance of variable A (e.g., Note 10). Detailed discussion of this and other proposed indexes is given in Anderson (1982, Section 6.1). If the range of R_A or R_B is arbitrary, the index suffers similar arbitrariness. It may still be useful, as in the following marriage example.

Relative Range Index in Marriage. An instructive use of the relative range index was included in extensive work by Armstrong (1984) on marital interaction. In one part of one experiment, spouses made

independent judgments of deserved blame for a child's misdeeds characterized by the two variables of damage and intent. Each spouse was given a relative range score, Damage/(Damage + Intent). The wife-husband correlation of these scores was .56.

Of potential value, three of the four couples who had distressed marriages showed widely different range indexes, a likely source of parental disharmony. A battery of such indexes covering important areas in marital interaction could be a simple diagnostic tool. Given as part of preparation for marriage, such a battery might help iron out differences before they became daily discords (see *Group Dynamics*, Chapter 7).

Meretricious *p* Values. The *p* value of a statistical test has only one valid use—as an indication whether the null hypothesis may provisionally be rejected.

The *p* value is a bad measure of effect size. With large samples, tiny effects can be "highly significant." And large effects may miss significance with small samples or high variability.

The *p* value is not a measure of power. If $p = .05$, a best guess of the power of an exact replication of the experiment is approximately .50 (*Empirical Direction*, p. 104).

The *p* value is not a measure of the probability that the null hypothesis is false. It is invalid for this purpose, not only numerically, as in the previous paragraph, but also conceptually, as Bayesian statisticians have repeatedly emphasized.

The primitive ritual of *p* value worship evident in articles bespangled with *, **, and *** stems from misappreciation of empirical science and misunderstanding of statistics. This statistical clutter can be avoided by saying once that cited effects are statsig at $p = .05$ (or .01). Any reader who really desires the exact *p* value can get it from the given F, t, or r. This practice helps focus attention on what is important—the actual data (*Empirical Direction,* Section 2.4.3).

Confidence intervals are much better than *p* values. They not only provide significance tests but also estimates of the size and variability of the effect (*Empirical Direction*, Sections 2.4, 18.1, and 18.3).

Statistical Effect Sizes. Some statistics texts seek to go beyond tests of significance to present effect sizes. One useful index of effect size is a confidence interval. This estimates the true effect together with an estimate of its reliability. Power effect size is also useful, almost routine in designing an experiment.

Except for the confidence interval, however, indexes of effect size are often merely statistical window dressing that obfuscate rather than

illuminate. Effect size is basically an extrastatistical, empirical matter (see *Size and Importance of Effects*, pp. 551-559 in *Empirical Direction*; see similarly Blanton and Jaccard, 2006a).

Negativity Theory. *Negativity*—greater importance of negative stimuli—seems clear to common sense and has generated a minor literature. Much of this work, however, rests on misconceptions about psychological measurement. Most investigators have recognized that demonstrating negativity requires separating importance from value. One common attempt was to preselect positive and negative stimuli that were "equal and opposite" in value and then show that their combined effect was negative. But this "equal and opposite" method, which is critical, rested on mere hope; instead, it requires grounded theory of measurement. Indeed, it appears to be incorrect as shown by Oliveira, et al. (2006).

Valid negativity theory is possible with functional measurement. The averaging law can assess importance weight separately from value. A qualitative application was used in what seems the first definite evidence for negativity (Anderson, 1965). The averaging law is notably more general than the "equal and opposite" condition; it can compare importance of only negative or only positive stimuli. Detailed discussion is given in *Negativity Theory* in Anderson (2008, pp. 349-356).

Definition and Measurement of Importance. The need to measure *importance* unconfounded with *value* was recognized in the "equal and opposite" method just discussed. This need is recognized in the averaging law with the dual concepts of importance weight, ω, and polarity value, ψ. Empirical success of the averaging law thus provides a conceptual foundation for both value and importance, together with a means to measure both.

Valid Importance Weights. Valid measurement of importance is possible with the averaging law; this law can separate importance weight, ω, from value polarity, ψ. This removes the confounding suffered by regression coefficients (see also *Self-Measurement Theory*).

Estimation of importance weights may require the Average program. This requires suitable experimental design (Zalinski & Anderson, 1991).

REGRESSION MODELS: PREDICTION VS. UNDERSTANDING

For *prediction*, standard regression analysis has two remarkable advantages. It can utilize convenience values of predictor variables. And it automatically allows for intercorrelation among these predictors, as is

typical in prediction tasks. No less remarkable, regression models outpredict experts in nearly every field, from clinical psychology to personnel selection (see Sections 16.1 and 16.2 of *Empirical Direction*).

But for *understanding* cognitive process, regression analysis suffers treacherous pitfalls. Detailed discussion of regression analysis, including tests of multiplication models, is given in Anderson (1982, Section 4.3).

Invalidity of Regression Weights. Weights of regression variables are sometimes interpreted as measures of importance. That such weights demonstrate "the relative importance placed on" the regression variables is a common misconception. In fact, regression weights are typically invalid as measures of relative importance of variables.

A major source of invalidity is that each regression weight is confounded with the unit of its scale (see *Uniqueness* above). Celsius and Fahrenheit scales, for example, are both linear scales of temperature but their regression coefficients would differ by 5 to 9. Hence they could yield opposite conclusions about relative importance.

Some writers have thought to avoid this unit confounding by standardizing the values of each variable. This makes matters worse; it further confounds the scale unit with the range of values. Detailed discussion is given in Anderson (1982, pp. 262-265; see Note 11).

Integration Analysis. Standard additive regression models are usually invalid and misleading for analysis of integration *process*. Correlations higher than .97 can easily be obtained with severely nonadditive models of cognitive process (see Chapter 4 of Anderson, 1982).

Nor does statsig deviation imply a nonadditive model. Deviation may result because values of the regression variables are inaccurate estimates of the true values, or because the response measure is nonlinear.

Analysis of variance, in contrast, avoids the first pitfall because it does not rely on prior values of the stimulus variables. It can also avoid the second pitfall by using an integration law to develop a method for linear response (see also *Response Generality*).

Nonlinearity and Interaction. Claims for nonlinear or interactive relations based on regression analysis require explicit justification but this is rarely given. Observed nonlinearity may reside entirely in the arbitrary predictor variables and/or in a nonlinear response measure.

Process Analysis. For analysis of cognitive process, standard regression analysis is extremely limited as just discussed. As noted in *Empirical Direction*, page 514:

> Most texts do warn that substantive theory is generally prerequisite for process analysis with multiple regression. These warnings, however, are obscured un-

> der the mass of statistical detail, by such phrases as "statistical control," and especially by students' implicit assumption that what is being taught must be worth learning. At best, these warnings are little help because little is said about what constitutes adequate substantive theory.

Brunswik's Lens Model and "Policy Capturing." The lens model of Brunswik (1956) has been employed by a number of writers (e.g., Brehmer & Joyce, 1988; Hastie & Dawes, 2001) without recognition of its fatal flaws. Brunswik's lens model is a superficial form of the Integration Diagram. The "lens" is a superficial analogy in which influences are imagined to radiate from variables in the environment and impinge on an imaginary "lens" which imaginarily "focuses" [integrates] them into a response. This lens analogy is helpless with the integration problem.

Integration has usually been handled with an arbitrary assumption of multiple regression; the regression weights are considered to "capture" the person's "policy." Fatal flaws of this "policy capturing" have long been known.

One fatal flaw is that the regression coefficients are generally invalid measures of importance of variables (see *Invalidity of Regression Weights* above). Of course, this vitiates claims for "policy capturing" (Anderson, 1982, Note 6.1.3a; 2008, Note 2, pp. 396f). No less fatal, the regression model is generally invalid for integration (Note 12).

NONARBITRARY METRICS WITH FUNCTIONAL MEASUREMENT

Functional measurement can help develop "nonarbitrary" measures of psychological qualities, measures that reflect functional relevance.

Arbitrary Metrics. The issue of importance of a measure was emphasized in clear, cogent articles by Blanton and Jaccard (2006a,b). Many measures are "arbitrary," they say, in the specific sense of being uninformative about their importance in a person's thought and action. For illustration, they used the Implicit Association Test (IAT) advocated by Greenwald, Nosek, and Sriram (2006), a reaction time measure to variables such as race and gender. Blanton and Jaccard (p. 35) rightly point out that this method lacks both internal and external validity:

> The arbitrary nature of the IAT metric and the fact that the diagnoses have not been linked to any observable acts of automatic preference suggest that researchers have no way of gauging the true magnitude of the implicit preference expressed by a given IAT score.

This is the problem. To say that one person's IAT race reaction time is

larger than that of another says nothing about the role of race in either person's thought and action. It could be unimportant for both.

A pertinent example comes from the study of forgiveness in Lebanon. It seems a safe conjecture that the IAT would have revealed clear prejudice between Christians and Muslims. But the forgiveness judgments showed no sign of this; forgiveness of Christian and Muslim gunmen was virtually identical for Muslim and Christian participants (see Figure 7.5 in *Algebra of Forgiveness*, Chapter 7).

Nonarbitrary Metrics With Information Integration Theory. This problem of arbitrary metrics had been explicitly recognized in the earlier criticism of stereotype research in Anderson (1981a, p. 248):

> Notably lacking in studies of . . . stereotypes is the use of tasks that require information integration . . . If another piece of information that had solid relevance to the judgment were to be included, then a meaningful relative importance could be determined.

The common practice of eliminating everything besides the stereotype information thereby eliminates the possibility of determining *importance,* at best a treacherous problem (see *Measuring Importance* above).

This criticism of arbitrary metrics was repeated in *Stereotype Theory*, Chapter 5 in Anderson (1991b, p. 232):

> Integration designs seem essential for stereotype theory. Most stereotype studies use simplified tasks that contain only stereotype information. This can yield artificially large effects because the subject has little else on which to base a judgment. Such effects might be relatively weak.

These comments suggested that stereotype experiments should routinely use integration designs with a pertinent, nonstereotype variable as standard to assess importance (Anderson, 2008, pp. 197, 331).

Integration theory also implies that the concept of "true magnitude" of stereotype scores such as IAT is simplistic; it fails to recognize that expression of any stereotype depends on operative goals. The small effect of religion in the cited forgiveness judgments does not mean it would have small effects in social interaction. The IAT embodies the dominant misconception of social attitudes as one-dimensional, good–bad reactions (see *Functional Theory of Attitudes* in Chapter 8).

THEORY AND METHOD ARE ONE

The unity of theory and method lies at the heart of Information Integration Theory. Revealing the three laws of information integration depended on true measurement of the psychological values of the stimulus in-

formers that are integrated. These values themselves are derived from the laws (see e.g., *Parallelism Theorem* of Chapter 1). These laws solve the long crux of true measurement—both of *response* and of *stimulus*.

Success of this approach depended on Nature's beneficence in making these three integration laws organic to thought and action. To reveal these laws also depended on experimental methods discussed above, especially the method of functional rating.

Serious shortcomings of still-current statistics texts were revealed by the integration laws. Foremost is the basic importance of *extra*statistical inference in empirical research. Also important is the basic ambiguity of statistical interactions in analysis of variance. Interactions depend critically on the assumption that the response measure is a linear (equal interval) scale, which is rarely verified.

Such shortcomings remain prominent despite a half-century of exposure (*Understanding Interactions*, Chapter 7 in *Empirical Direction*, Anderson, 2001). Such shortcomings arise from teaching statistics as statistics, whereas it should be taught as organic to empirical method.

ACHIEVEMENT

All of us strive for achievement. We hope that, when our lives draw to their close, our teaching and research will have left behind some worthy contribution to the benefit of our students and the progress of our field (see also *Achievement*, pp. 365-371, Anderson, 2008).

Method is one guide to achievement, a concern of this chapter. Much earnest effort is being wasted owing to reliance on faulty method. Multiple determination is one example. The importance of multiple determination is widely recognized but the common methods of analysis of variance and multiple regression have serious pitfalls as discussed above with “interactions” and with measurement of importance.

Respect for phenomena is a second guide to achievement. Much of past progress in psychology has consisted of recognizing new phenomena that enlarge and enrich our conceptual horizons. The 20-some issues of the next chapter are grounded on respect for phenomena. Moral considerations pervade everyday life. Moral science constitutes a conceptual framework that can unify our fragmenting field (Chapter 8).

APPENDIX: MEASUREMENT THEORY

True measurement of psychological quantities has been actively sought since 1860. This measurement problem appears in the Integration Diagram: how can we tell whether our observed response, R, is a true measure of ρ? This might seem impossible— ρ is unobservable.

An obvious approach is to ask people to give numbers to represent the magnitude of their sensations or feelings. But are these response numbers valid—linearly related to their unobservable feelings?

Metric response methods have been considered invalid by most persons who have sought to develop psychological measurement theory. Instead, they have relied on ordinal judgments of greater than/less than, as in Thurstone's pair comparisons and in conjoint measurement. Solid ground for this denigration of metric response appeared in the nonlinear rating biases noted by Thurstone and reemphasized by the large difference between the rating method and the once-popular, now-defunct method of magnitude estimation (see e.g., Anderson, 1970, 1972a, 1981a, Section 5.4, 1996a, Chapter 3; Note 13).

Fechner (1860) proposed a clever assumption for psychophysics:

just noticeable differences in sensation are psychologically equal.

Hence just noticeable differences may be used as units to measure sensory value, just like centimeters on a meter stick. Fechner's plausible assumption has continually eluded proof.

Thurstone's (1927, 1959) method of pair comparisons finally allowed a proper test of Fechner's assumption for the special case of psychophysical sensations, such as heaviness and loudness, that can be varied continuously for each individual (see also Link, 1994).

Thurstone's big claim, however, was that his method also applied with discrete stimuli that are the norm in social–moral judgment, as with seriousness of his list of criminal offenses and especially with general social attitudes. Thurstone's claim was not justified because it made illegitimate use of individual differences (Anderson, 1981a, Sections 5.3.1 and 5.3.2, 1996a, pp. 85f, 2008, p. 186).

A new approach came in the 1960s with realization that a two-variable additive law could provide a firm foundation. Two variables would seem to complicate the matter because it is then necessary to take account of the two unobserved ψ values in the Integration Diagram as well as the unobserved response, ρ, as in Equation 1 above. Two variables, however, can provide enough mathematical constraint to find a best monotone transformation to additivity and still retain degrees of freedom to test nonadditivity. Two qualitatively different proposals were made to

capitalize on additivity structure: *conjoint measurement* (Luce & Tukey, 1964) and *functional measurement* (Anderson, 1962a,b).

FUNCTIONAL MEASUREMENT

Empirical laws of information integration are the foundation for psychological measurement. Functional measurement theory is grounded on this base and frame.

> The logic of the present scaling technique consists of using the postulated behavior laws to induce a scaling on the dependent variable (Anderson, 1962b, p. 46).

> A guiding idea of functional measurement is that measurement scales are derivative from substantive theory (Anderson, 1970, p. 153).

The potential of this functional approach was illustrated with the benefits of the parallelism theorem listed above. But, such laws must have empirical reality for these benefits to be real. Such empirical laws are the foundation for theory of psychological measurement.

Algebraic laws had been widely conjectured, of course, as with the equity models of Chapter 2 and with Subjective Expected Value. But without capability for psychological measurement, these conjectures remained conjectures. Using functional measurement, however, the initial 1962 study of person cognition supported an adding-type law in single person design and analysis. Later applications of functional measurement have done well throughout human psychology.

The three integration laws also showed that psychological measurement theory differs conceptually from what were and remain common preconceptions. Six conceptual differences deserve consideration (see also Anderson, 1982, pp. 101-104).

Metric Response. Much cognition is metric, a consequence of the goal-directedness of approach–avoidance in the external world (*Metric Cognition*, Chapter 7). Linear metric response should thus be a prime goal of psychological measurement. Metric responses can suffer nonlinear biases, however, as with common ratings, so virtually all other attempts to develop measurement theory condemned metric response and grounded themselves on nonmetric, choice response.

Fortunately, experimental procedures introduced with rating in the initial 1962 experiment have been generally successful in eliminating nonlinear rating biases. This *method of functional rating* can provide true linear response scales even with young children and nonliterate persons.

Metric response was a key to the psychological laws. And thereby a key to true psychological measurement.

Response Generality. Metric response methods have the invaluable potential of *generality*. A method that has yielded a linear scale across a number of empirical situations, as functional rating has done, may reasonably be expected to do the same more generally.

Metric response is important for situations that do not obey a simple algebraic law. Many such situations are known. With a linear response, pattern in an integration graph will be a faithful image of pattern in underlying response—regardless of the integration process.

Metric response has central importance in psychological science.

Interaction and Configurality. Metric response is invaluable for studying configural integration. With a linear response, deviations from parallelism are clues to understanding interaction and configurality, as with the negativity effect and inconsistency resolution (see *Interaction and Configurality*, Anderson, 2008, pp. 357-364).

Goal and Context. *Stimulus values always depend on goal and context*. The same stimulus informer may have very different values relative to different goals. This value dependence is recognized by GOAL in the Integration Diagram (Figure 6.2). This dependence of value on goal is explicit in the valuation operation. This goal dependence of value seems unrecognized in other measurement theories.

Weighted Average Model: Importance *Weight* and Polarity *Value*. An essentially new conception of psychological measurement theory is entailed by the averaging law. Most tasks that might have been expected to follow an addition law have instead followed the averaging law. One consequence is that *importance weight* and *value polarity* become coequal measurement parameters.

This two-parameter, weight–value representation emerged from the averaging law with unequal weighting across levels of a single variable. To illustrate, consider judging proportion of blue balls in an urn of red and blue balls. Random samples of 3 red/1 blue and 6 red/2 blue have the same *value*, .25. The larger sample carries more information, however, and so has greater *weight*. Unequal weighting would thus result if sample size was varied across one variable in an integration design.

This unequal weighting was initially disagreeable because it does not follow the simple parallelism theorem and so cast doubt on what parallelism had been obtained. It was a blessing in disguise, however, because it accounts for several observed phenomena such as source reliability and

negativity/positivity effects. Moreover, it makes possible measurement of the weight parameter separately from value.

Strict adding models, it may be emphasized, predict parallelism even with unequal weighting. Indeed, the weight parameter is not generally separable from the value parameter in such models.

Self-Measurement. Self-measurement has basic importance in functional measurement theory. Self-measurement of *response* rests on success of an integration model, as with benefit 2 of the parallelism theorem. These response data can be used to derive valid *stimulus* measures, as with benefit 3 of the parallelism theorem. These stimulus measures may then be used as validational criteria to develop valid methods of stimulus self-measurement.

Self-measurement needs to be extended to handle situations that may not allow formal integration designs or that do not obey an integration law. Much work on multiattribute analysis is of this type (see *Self-Measurement Theory* and *Response Generality* above).

Fundamental Measurement. Functional measurement is *fundamental measurement*—no prior measurement is necessary to establish true linear scales. Qualitative, rank-order data suffice.

> With two or more stimulus variables in factorial-type design, sufficient constraint is potentially available to solve both measurement problems and to provide the necessary test of goodness of fit. No auxiliary assumptions are needed. No prior scales are needed. All that is at issue is the algebraic structure of the model. That provides the base and frame for measurement that is scale-free, or fundamental, not dependent on prior measurement.
> (Anderson, 1982, p. 207.)

This monotone parallelism theorem requires only a monotone, rank-order response such as may be obtained with choice data. The first step is to estimate a best-fitting additive response (e.g., Kruskal, 1965). The critical problems of testing goodness of fit and measuring response and stimuli have been resolved and successfully applied to empirical data (Anderson, 1982, Chapter 5, *Monotone Analysis*; see Note 14).

Behavioral metric responses, such as response rate or amplitude, may thus be validatable as true metrics, which can greatly facilitate experimental analysis, especially with infrahumans (Anderson, 2002).

Functional measurement theory implies that measurement in physics also rests on empirical law (Anderson, 1981a, pp. 361f; Masin, 2007).

SCALE TYPES

The functional conception of scale type implied by the Integration Diagram of Figure 6.2 is quite different from the standard conception. Specifically, scale type is defined in terms of the relation between R and ρ, that is, between responses in the external and internal worlds (see *The Dual Worlds*, Chapter 7).

The three common scale types (ordinal, equal interval, and ratio) thus become monotone, linear, and proportional:

monotone: $R_1 > R_2$ if and only if $\rho_1 > \rho_2$;
linear: $R = c_0 + c_1 \rho$, with zero and unit constants, c_0 and c_1;
proportional: $R = c_1\rho$.

The conceptual difference between these two conceptions of scale type may be illustrated by contrasting equal interval scales with linear scales. Equal intervals derives from the conception of measurement scale in terms of additive units in physics, as with additive unit weights or successive marks on a meter stick. This conception of equal intervals entered psychology with Fechner's jnd scale (Note 15) and has been widely accepted. This traditional conception attempted to place the meaning of equal intervals within the scale itself.

The present functional view, in contrast, places scale type in the relation between the external and internal worlds, R and ρ, respectively. This view is needed to recognize that value depends on operative goals. Establishing scale meaning thus depends squarely on empirical integration laws. Thus, linearity of the method of functional rating was established by empirical success of the parallelism theorem (benefit 2).

MONOTONE ANALYSIS

Many behaviors are not amenable to the method of functional rating commonly used to obtain linear response measures in IIT. These include response time and response rate, physiological and neural measures, and response measures with infrahumans. Such measures are widely useful to study directional effects of stimulus variables but are limited for quantitative analysis.

To illustrate, consider response rate and response time, both in common use. However, they are nonlinearly related:

$$\text{response rate} = \frac{1}{\text{response time}}.$$

Hence, at least one must be nonlinearly related to the underlying response measure (ρ in the Integration Diagram).

This obstacle confronts every application of analysis of variance, which explicitly assumes an additive integration, deviations therefrom being interpreted as "interactions." Such statistical interactions are widely assumed to have empirical reality as influence of one variable on another, an assumption fostered by nearly every statistics text. Of course, such statistical interactions may merely reflect nonlinear response, quite devoid of empirical reality (Anderson, 1961, 2001).

Most fortunately, the method of functional rating has been proven linear by the successes of the three integration laws. The hypothesis that these laws are innate (Chapter 7) suggests that similar laws may be found with other responses besides rating, and with infrahumans. This requires capability with model analysis based on rank-order or monotone response measures.

Fortunately, practicable techniques for analysis of monotone data have been developed. These are described with empirical applications in *Monotone Analysis*, Chapter 5 in Anderson (1982). These techniques, it may be emphasized, are fairly demanding. However, they may do good service as a base for developing linear or near-linear measures of behavioral and neural responses, perhaps by using approximate response transformation.

As one example, it is widely thought that bar press rate by rats may be a linear measure (see Note 9, p. 104 in Anderson, 1996a, *Matching Law*, this chapter), although evidence is very limited. It would certainly be desirable to establish linear or approximately linear measures for behavioral and physiological responses. Such measures would open a door to study integration of two or more variables, a fundamental issue in every field of psychology.

CONJOINT MEASUREMENT

Conjoint measurement is grounded absolutely on nonmetric response. Only ordinal (greater than/less than) response is allowed. Such choice response can be trusted; metric responses such as rating scales were well known to suffer nonlinear response biases.

It seemed a triumph, therefore, when Luce and Tukey (1964) proved that ordinal data were, in principle, sufficient to establish an additive

model. This triumph is fruitless, of course, unless their axiomatic base can be used to show that addition laws have empirical reality. Nonexistent laws cannot yield real measurement.

In fact, conjoint measurement never succeeded in establishing any empirical law. Despite extensive mathematical elaboration by persons of high ability, the critical problem of testing goodness of fit was never solved (see Luce, Krantz, Suppes, & Tversky, 1990, p. xiii). Not one single positive empirical application has ever been made—conjoint measurement has been empirically empty (Notes 16-19).

Indeed, Cliff (1992), formerly a strong proponent of conjoint measurement and critic of functional measurement (Cliff, 1973), concluded that conjoint measurement was "the revolution that never happened." Cliff thus reaffirmed earlier evaluations (Anderson, 1974a, pp. 286ff, 1981a, Section 5.5, 1982, Sections 5.4 and 5.5). Indeed, Cliff adopted a stand very like functional measurement. Conjoint measurement has been a blind alley in psychology.

The alternative approach of reliance on an empirical algebraic model, now advocated by Cliff, had already been put on a solid empirico-theoretical base with functional measurement (e.g., Anderson, 1974a,b,c, 1979, 1981a, 1982).

Functional measurement is the revolution that did happen.

Conjoint measurement rested on simplistic preconception of psychological measurement. Various aspects of this misconception appeared in the preceding discussion of functional measurement. Most important are:

- Conjoint measurement is empirically empty. Although it grounds itself on algebraic models, it has been unable to show empirical reality of any model.
- Conjoint measurement missed the fundamental importance of metric response in the approach–avoidance actions of living (see *Metric Cognition* in Chapter 7). Metric response allows response generality which can analyze configural integration outside the scope of conjoint measurement theory (see *Response Generality* above).
- Conjoint measurement cannot handle the ubiquitous averaging model because this model requires two coequal parameters—polarity *value* and importance *weight*. Hence the averaging model can be disordinal (see Figure 6.1 above and *Opposite Effects* in Chapter 1). Disordinal models cannot be handled with the ordinal methods of conjoint measurement.

THE NATURE OF PSYCHOLOGICAL MEASUREMENT

Measurement has fundamentally different nature in psychology and in physics. Typical physical measures are properties of physical entities, as with length and mass. These can be measured in themselves with additive units.

Psychological measures, in sharpest contrast, are goal-oriented constructions of the organism—functional measures. The same external situation may thus yield different measures depending on the goal of the organism. This is clear with social–moral judgment but holds generally, even with vision or emotion.

The integration laws arise out of the goal-directedness of living. And they provide a metric base for deeper analysis of thought and action at the individual level which deserve systematic study (see *Response Quality* and *Profile Analysis* above; Notes 20, 21).

We are fortunate in the extensive validity of the integration laws; they allow true measurement for individuals, both for response and for stimuli (benefits 2 and 3 of the parallelism theorem). These measures provide a priceless foundation for psychological science.

NOTES

Note 0. I should acknowledge that my own practice falls short of ideal. In part, this results from my gradual realization of some of the issues discussed in this chapter (see e.g., *Gratitude and Ingratitude*, Chapter 7; see also Note 7.18a in Anderson, 1982).

Note 1. Serial integration designs (e.g., Figure 8.3) may profit from fractional replication. As one illustration, 6 serial positions could be included in a 1/8 replication of a 2^6 design. This would estimate main effects for all 6 serial positions with 1 df for selected interaction. An example is given in Figure 10.9, p. 340 in Anderson (1996a).

Note 2. Equal attention to each stimulus informer is important for the parallelism predicted by the averaging model. Accordingly, instructions in the personality adjective task stated that each adjective was contributed by a different acquaintance who knew the person well. Timing should be controlled to prevent hasty responding. Previous response may influence present response, one reason for using graphic rating. It may be useful to present end anchors occasionally during the experiment.

Stimulus value ψ might properly be conceptualized as a distribution rather than a fixed number (Anderson, 1982). Studies with the personality adjective task indicated that operative value was influenced by the other adjectives on an initial trial but thereafter remained fixed (Anderson, 1969; Anderson & Clavadetscher, 1976). This is one reason for stimulus familiarization in the preliminary practice. Of course, such value change during the experiment would be expected to violate parallelism (see also *Construction of meaning by association*, p. 138 in Anderson, 2008).

Note 3. The extreme claim that people are *never* aware of the causes of their behavior (Nisbett & Wilson, 1977; Nisbett & Bellows, 1977) would disable self-estimation. And indeed there are multiple lines of evidence for nonconscious cognition including blindsight and posthypnotic suggestion as well as the size-weight illusion of Figure 7.2.

The cited articles, however, claimed that any accuracy of self-estimates derived entirely from shared cultural norms. Empirical disproofs had previously been given by Shanteau (1974) and Shanteau and Nagy (1976); see Table 6.2 and Figure 6.3 in Anderson (1982). A direct empirical disproof was given in Wright's (1995) ingenious double yoked procedure (see Anderson, 1996, pp. 391f).

Note 4. For completeness, the addition model of Equation 1 would include weights for each stimulus variable as well as an additive constant to represent prior state of the participant, denoted with subscript 0:

$$\rho_{jk} = \omega_o \psi_o + \omega_A \psi_{Aj} + \omega_B \psi_{Bk}.$$

This model makes the same parallelism prediction. It is convenient to use the simpler Equation 1, in which ψ actually stands for $\omega\psi$.

Note that importance weight and value polarity are completely confounded in the addition model. They act jointly as the effective value; what is estimated by the marginal means is thus the product, weight × value = $\omega \times \psi$, also called part-worth (Note 22).

Note 5. Observed parallelism is not absolute proof of additivity. It is logically possible that nonadditivity in the integration operator, **I**, is exactly cancelled by nonlinearity in the action operator, **A**, to yield net parallelism. This logical possibility is no longer a serious concern (see Anderson, 1982, p. 71, 1996a, pp. 94-98, 105).

Note 6. Prior State. *Prior state* was originally called *initial impression* in the personality adjective task. It was needed, in particular, to harmonize the averaging model with the set-size effect (more polarized response for more stimuli of equal value) which was verified quantitatively (Anderson, 1967). Prior state is essential in parameter estimation for the averaging model with unequal weights. The new term, prior state, was adopted in place of initial impression by analogy to prior belief in Bayesian theory.

Note that prior state is more general than Bayesian prior belief. It applies to the averaging model, which seems outside the scope of standard Bayesian theory. Also, it applies to metric judgments such as the president attitudes of Figure 6.1.

Note 7. Franklin's simple form of multiattribute analysis for choice between two actions was given in a 1772 letter to his friend, Joseph Priestly, a famous British chemist and a nonconformist minister who opposed his government's policies toward the American colonies. Franklin's method was to list pros and cons in separate columns and cross out combinations that seemed to cancel; what remained would then determine his choice. This short letter, reproduced on pages 253-254 of Franklin (1792/1983), ends by calling his method "*moral* or *prudential algebra*" (see Note 7 in *Moral Philosophy*, Chapter 7).

Note 8. Correlational studies are stock-in-trade in health psychology, but seldom with acknowledgement of their severe limitations. As one of many examples, Taylor's (1994) *Health psychology* rested almost entirely on correlational "associations" with almost no discussion of their many limitations, even less on how to do valid investigations.

Correlations can, of course, be useful clues to causation, but they deserve clear indi-

cation of confounding factors. Wolf's (2011) critique of breast- versus bottle-feeding rightly emphasizes the many confounds, especially diverse mother-infant attitudes.

Note 9. Some writers have advocated systematic study of Anova interactions as a means to deal with the multiplicity of operative variables and context effects. This strategy rests on implicit reification of Anova with no recognition of the dual problems of response linearity and integration model discussed in the text (see Section 7.4, *Interactionist Theories* and especially Section 7.6.4, *Statistics Teaching,* in *Empirical Direction;* see also *"Interaction" as Attitude Integration Theory*, pp. 133ff, Anderson, 2008). A pertinent example is given in Note 2 under *Person Science and Personality* in Chapter 7.

Note 10. *Relative Range Index in Face Cognition.* Pregnant application of the relative range index to measure importance of facial action units in pain perception is presented by Oliveira, de Sá Teixeira, Oliveira, Breda, and da Fonseca (2007; see also de Sá Teixeira & Oliveira, 2007). They used ecological ranges of three muscularly-based facial action units (brow lowering, levator contraction, orbit tightening) with natural-looking synthesized faces. Hence relative range indexes were valid measures of relative importance of these action units. Cogent extension to much-needed theoretical–empirical clarification of holistic processing of faces has been given by Oliveira, Silva, Viegas, Teixeira, and Gonçalves (2011, 2012).

Their approach also has notable generality; they can go beyond face cognition per se to study general social interaction. This involves integration of verbal cues together with nonverbal cues such as gesture and tone of voice, as well as facial cues per se.

Note 11. To illustrate the invalidity of standardized regression coefficients as measures of relative importance, consider the equal weight regression, $R = X + Y$. Assume X and Y both have mean 0, standard deviations of 1 and 2, respectively. The standardized variables are $X' = X'$, $Y' = Y/2$. Regression with these standardized variables is $R = X' + 2Y'$—contradicting the given condition of equal weights.

Note 12. *Ecological Validity and Invalid Ecology.* Brunswik's (1956) lens model rests squarely on the assumption that the environment constitutes an ecology that operates as a validity criterion. This assumption may have seemed reasonable for perception of the physical world, with which Brunswik was originally concerned. Some writers, however, have transposed this approach to the human world, failing to realize that the human ecology can be invalid.

A striking case of invalid ecology appeared with the study of bail setting by Ebbesen and Konečni (1975) discussed in Chapter 4. Judges integrated relevant variables such as family ties sensibly in the privacy of their chambers but ignored such variables when setting actual bail in the courtroom. These actual bail settings constitute the ecology but were invalid.

In this example, as in much of human affairs, the main problem is to change the ecology. Other prime examples include family life, school education, and the social–moral world.

Note 13. Magnitude estimation uses metric response and was once extremely popular. Magnitude estimation repeatedly failed tests of validity, however, and has been virtually abandoned, a blind alley that contributed almost nothing of substance (see e.g., Anderson 1974a, 1981a, Section 5.4, 1982, pp. 19-21, 1996a, Chapter 3, 2008, Note 9, p. 266; see

also *Achievement*, pp. 365-371 in Anderson, 2008).

Note 14. This base in monotone functional measurement justifies treating the parallelism theorem as fundamental measurement. The metric response is what would be obtained with monotone analysis but simpler and more exact.

Note 15. Fechner's just noticeable difference, ΔS, may be defined in terms of that stimulus intensity, $S + \Delta S$, that is judged larger than S on (say) 75% of the trials. Empirically, ΔS is approximately proportional to S over the main range of many physical sensory dimensions, such as heaviness (grams). In physical units, therefore, $\Delta S/S$ is approximately constant (Weber's law).

Fechner's basic assumption was that all ΔSs, although different in physical size, are equal in *psychological* size by virtue of their just discriminability. This plausible assumption yielded Fechner's psychophysical law, $\psi = c \log S$, where c is a unit constant of proportionality, which is approximately correct over the main range of many sensory dimensions, as in the common decibel scale of loudness. But 150 years of determined efforts failed to prove Fechner's ingenious assumption.

The metric response method of functional measurement is simpler and definitive (e.g., Anderson, 1970, 1974a, 1979, 1992a,b, 1993; Carterette & Anderson, 1979; Marks, Elgart, Burger, & Chakwin, 2007; Masin, 2003, 2007; McBride, 1993; McBride & Anderson, 1991; Weiss, 1972, 2006). Of special importance, metric response also applies with discrete stimuli common in social–personality and judgment–decision, but which do not generally allow just discriminable differences.

Note 16. Conjoint measurement is a failure; it has failed to measure anything. It lacks capability for the essential step of testing goodness of fit. As Luce, Krantz, Suppes, and Tversky (1990, p. xiii) admit: "The chapter on statistical methods was not written because the development of statistical methods for fundamental measurement turned out to be very difficult."

In fact, statistical methods for fundamental measurement with choice data had been provided with functional measurement. Monotone functional measurement can provide the needed statistical analysis and has been successfully applied to empirical data (Anderson, 1982, Chapter 5). Advocates of conjoint measurement could have used these nonmetric methods to measure something but they have never done so.

Note 17. This distrust of rating scales was succinctly expressed by Luce and Galanter (1963, pp. 264f):

> To the theorist, however, the whole business is a bit hair-raising. To calculate the means of category *labels*, to plot them against physical measures of the stimuli, and then to discuss the form of the resulting function strikes him as close to meaningless. . . . we do not think that the absolute form of the obtained function using the first *k* integers as labels has any meaning (Luce & Galanter, 1963, pp. 264-265). [But see Note 18 next.]

This quotation missed the biopsychological importance of metric cognition. The common rating scale is an internalization of the goal-directedness of living (see *Metric Cognition*, Chapter 7). Its latent linearity has been actualized with the method of functional rating and especially by the extensive success of the laws of information integration. The meth-

od of functional rating removed this long-standing roadblock to theory of psychological measurement.

Note 18. Luce's position seems to have changed. Nonmetric choice response was taken as fundamental in Luce's choice model and in all work on conjoint measurement.

> An about-face may have appeared in Luce's position on psychological measurement. Whereas functional measurement has been dedicated to continuous response measures. Luce has been dedicated to discrete choice measures, both in his choice theory and in his extensive work on conjoint measurement. About-face appears in Luce, Mellers, and Chang (1993, p. 115), who rely on continuous response measures and conclude that "Choice is viewed as a derived, not a primitive, concept." (*Empirical Direction*, p. 736)

In functional measurement, continuous response measures have always been primitive concepts (Anderson, 1962a,b). Choice has thus been viewed as a derived concept from the beginning.

Note 19. Conjoint *scaling*, sometimes misnamed conjoint measurement, usually relies on metric response analyzed with multiple regression or analysis of variance. The term "conjoint" may be misleading because conjoint scaling differs sharply from conjoint measurement which absolutely disallows metric response.

Practical methods and applications of conjoint scaling are discussed systematically in the text by Louviere (1988). Louviere recommended functional measurement as conjoint scaling; it was the only one that dealt with the critical question of goodness of fit and hence of measurement validity. Functional measurement was thus the only one that could provide true measurement. Moreover, it had actually been successful in numerous experimental tests.

Note 20. Hedonic psychophysics may be a useful domain for studying multi-quality experience. Taste is one example. Different taste qualities such as sweet, bitter, temperature, crunchiness, odor, and so forth seem well defined both physically and psychologically, alone and in various combinations. Although rather afield from social–moral phenomena, their physical manipulability and their experiential qualities in combinations are attractive for experimental analysis and could be used in tandem with social–moral qualities (see further *Hedonic Psychophysics*, Anderson, 2008, pp. 291f; Marks, et al, 2007; McBride, 1989, 1993; McBride & Anderson, 1991).

Taste has an additional advantage of allowing comparative studies with animals, which can employ methods not feasible with humans including extended sessions and motivation produced by deprivation. Four integration studies are reported in Anderson (1978a, 1996a, p. 104) including the notable study of food–shock motivation by Farley and Fantino (1978).

The adding-type law is considered an innate integration capability which suggests it may also be present in lower animals. One possible integration task could study response rate to obtain a plate with varied amounts of two foods of different hedonic value. Evidence from operant studies suggests that response rate may be a linear scale (Anderson, 1996a, p. 104, 2002). Perhaps a graphic rating-type response could be trained.

Note 21. One issue for profile measures concerns the relation between the separate components and the overall judgment. With ambivalent attitudes, for example, is the overall attitude any simple integration of the positive and negative components?

In general, each component quality may contain unique information that warrants separate study, despite commonality with other component qualities. Quality analysis is a

prime field for psychological measurement theory.

Note 22. Goodness of Fit and Parameter Estimation.

Information Integration Theory rests on extensive experiments that have revealed algebraic models of stimulus integration in most fields of human psychology, even in young children and nonliterate persons. Two statistical issues arise with these models: *goodness of fit* and *parameter estimation*. Detailed discussion of these issues is given in Anderson (1982; see also 2001, 2002). A brief overview is given here.

Goodness of Fit. Does the model account for the data? This question requires test whether the deviations from the model are statsig. Standard correlation analysis is invalid for testing goodness of fit as detailed in the main text.

Add–ave models imply that deviations from parallelism in a factorial integration design are not statsig. This implication may be tested with the interaction term from analysis of variance (Anova).

This Anova test requires independent responses except as allowed with repeated measurements Anova. Independent responses requires care in experimental procedure. Thus, the initial practice is intended to stabilize the stimulus values and the frame of reference for the response. Influence from previous response is undesirable, one reason for using a graphic response that does not remain visible. Presenting trials in random order can randomize out possible carryover.

Power may be markedly increased with repeated measurements design (see legend of Figure 6.1) and even more with single subject design. Cluster analysis may be generally useful.

Parameter Estimation. Functional measurement rests on the principle that the model uses–and provides–the stimulus values that functioned in the integration. Estimation of these functional values depends on the model. A linear response measure is assumed in what follows.

Strict Additive Models. The two-variable additive model, error omitted, may be written as follows:

$$R_{jk} = C + \omega_{Aj}\,\psi_{Aj} + \omega_{Bk}\,\psi_{Bk}.$$

The difference between rows 1 and 2 is thus the difference between two part-worths,

$$R_{1k} - R_{2k} = \omega_{A1}\,\psi_{A1} - \omega_{A2}\,\psi_{A2} = \mu_{A1} - \mu_{A2},$$

which may be averaged over the column index, k. These differences are on a proportional scale with a true zero; the same scale holds for the column part-worth differences. Row and column part-worth *differences* are thus directly comparable.

However, comparing row with column means faces the difficulty that $\mu_{Aj} + \mu_{Bk} = (\mu_{Aj} + c) + (\mu_{Bk} - c)$ for any constant c (Anderson, 1982, pp. 69f). This difficulty might be resolvable by including a stimulus with zero value on each variable. Alternatively, include the two one-way designs to obtain separate estimates of each single part-worth. This may require iterative analysis, however, as with the Average program.

Averaging Model: Equal Weight. The two-variable averaging model with equal weights may be written

$$R_{jk} = (\omega_o\,\psi_o + \omega_A\,\psi_{Aj} + \omega_B\,\psi_{Bk}) / (\omega_o + \omega_A + \omega_B).$$

The denominator has the same value in every cell of the design. With equal weighting of each separate variable, therefore, the averaging model is formally an additive model and the above analysis applies.

Averaging Model: Unequal Weights. With unequal weighting across levels of any one variable, the averaging model is nonlinear. Parallelism analysis does not apply but the Average program may be used to estimate parameters and test goodness of fit (Zalinski & Anderson, 1989, 1991). Practice with artificial data before running the experiment is highly advised.

A special case arises when the weight can be expressed as a linear or quadratic function of the scale value, as with the negativity effect. This allows a fairly simple analysis, illustrated with clinical judgment in Anderson (1972).

Multiplying Model: Linear Fan Theorem. The multiplying model may be written

$$R_{jk} = C_o + \psi_{Aj}\,\psi_{Bk}.$$

This equation implies that the row curves will form a diverging fan of straight lines if the column stimuli are spaced at their functional values on the horizontal axis. These functional values may be estimated by the column means. As with the parallelism theorem, an observed linear fan supports three benefits:

Benefit 1: support for the multiplying model.
Benefit 2: support for response linearity.
Benefit 3: linear scales of both stimulus variables.

Statistical analysis is straightforward. The linear × linear SS should be statsig and all other interaction components should be nonstatsig. Weiss (2006) includes a disc with a computer program (see further Anderson, 1982, Section 2.2).

Multivariable Models. The foregoing results generalize fairly simply to integration models with three or more variables. Further discussion is given in Anderson (1982).

Are Averaging and Adding Distinct Mental Operations? The algebraic similarity of averaging and adding suggests they may be related as mental operations. Both might thus operate together. This possibility of joint operation is not ruled out by opposite effects, as in the presidents experiment of Figure 6.1.

Evidence for distinctness of averaging and adding was obtained in a 1967 experiment that used the averaging model to estimate weight as a function of number of informers. If averaging occurs alone, these estimates should be constant, as indeed they were (see Table 2-6, p. 134, in Anderson, 1981). Joint operation of adding and averaging would argue otherwise (p. 135). About their nature, little is known.