CHAPTER 1

Aptitudes, Skills, and Proficiencies

David Lubinski
Iowa State University

René V. Dawis
University of Minnesota

This chapter deals with the role of human abilities or response capabilities as determinants of work behavior. Some definitions are offered for conceptualizing labels used to classify these attributes (e.g., abilities, achievements, aptitudes, and skills) and indices thereof. Particular attention is devoted to cognitive/intellectual abilities and to the optimal utilization of contrasting dimensions generated by factor analytic research. Throughout this treatment, the criterion of scientific significance is employed to evaluate both well-known and contemporary ability parameters. Although the scientific significance of general intelligence and its central role in industrial and organizational psychology as well as psychology in general is quite robust, it is without question that multiple ability dimensions are worthy of both applied and theoretical attention. The importance of setting expectations on how much predictive power to expect solely from ability attributes is addressed. We also suggest how new assessment procedures may be compared and evaluated against existing techniques in terms of an empirically based form of competitive support. In this vein, the importance of using multiple criteria for assessing performance (including the aggregation of distinct criteria) is recommended.

We suggest that personal qualities not typicaly considered in conventional treatments of human abilities (e.g., personality dimensions) may be construed as instrumental response capabilities. The causal status of these entities as determinants of proficient work behavior is developed, and suggestions are offered on how

Lubinski, D., & Dawis, R. V. (1992). Aptitudes,skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.) The handbook of industrial/organizational psychology (pp. 1-59) (Second Edition). Palo Alto: Consulting Psychologists Press.

these attributes might be incorporated into future research. Meta-analytic studies of validity generalization are reviewed. And two topics concerning group differences in abilities are discussed in detail. First, a new methodology is offered for predicting group differences in performance; and second, the importance of assessing group differences in variability (ability dispersion) is examined. The chapter closes by explicating the importance of achievement as opposed to topographical accounts of behavior for measuring human attributes and for future developments in both applied and theoretical psychology. It is suggested that the role of normal science and the importance of systematically accumulating knowledge (using existing techniques) might be underappreciated.

Introduction

THE WORLD IS a continually changing place. Familiar situations recur, but all situations contain components that are new. Familiar tasks change as they take on unfamiliar dimensions. The demands people encounter are becoming more complex. To function successfully in this world, people must be able to change behaviorally in effective ways. To deal with change effectively, old behavior patterns must be modified and new, more sophisticated behaviors must be learned. In a real way, life is a series of actions and reactions to varying degrees of novelty in a variety of contexts. But because future events often share important features with past circumstances, people are able to adapt to new situations—if they have profited from their experience. In many ways, what we call *talent* or *competence* is the ability to produce more effective and more efficient behavior in novel situations. In this sense, every instrumental act is, quite literally, a creative happening.

The way people cope with change reflects both history and disposition—their tendencies to assimilate and accommodate to ever changing environmental demands. The behavior people manifest tells us how they may have acted and reacted in the past and how facile they are likely to be at dealing with unfamiliar tasks in the future. Because behavior is repeated, even when modified to suit the changing context, people can be characterized in terms of their behavioral attributes.

The present chapter is about an important class of behavior attributes: human capabilities. In industrial and organizational psychology, human capabilities are assessed in order to predict work performance. Assessing human capabilities involves sampling the behavioral repertoire for classes of behavior that are related to work performance. These behavior classes also factor into other important facets of a person's psychological world. We will devote particular attention to human capabilities manifested in careers and occupations, but our discussion will necessarily touch on other domains of human conduct.

To understand the human capabilities that make for career and occupational effectiveness, we have also to understand the settings in which behavior takes place—in particular, the school and work settings. It is useful to construe such settings as *environmental ecologies*, following Brunswik (1956), wherein various contingencies operate to structure the behavior classes that are required to achieve prescribed goals. School and work settings are loosely standardized environmental ecologies. Because the behaviors, contingencies, and goals are often highly specialized, if one is to perform successfully in these ecologies, the requisite behaviors must be in place.

This, basically, is why we assess human capabilities—to find out if individuals have the required behaviors in their repertoires.

Whether an individual's current level of behavioral effectiveness is linked primarily to personal gifts or environmental privilege may be an interesting question, but applied psychologists are more interested in the behavioral phenotype—what it is that a person can do. Furthermore, to predict future behavior—whether it be the likelihood that currently available behavior will be manifested or the probability that new behavior will be acquired—assessment is always based on behavior. The significance of these points will become clearer as our discussion unfolds.

Definitions

Assessment means collecting information for a purpose, and the purpose of psychological assessment is to make behavioral predictions. *Psychological assessment* is, fundamentally, classifying and quantifying the current status of selected behavior classes, selected because of their social significance. This is true of all important classes of behaviors—interests, needs, values, and personality traits, as well as abilities (Lubinski & Thompson, 1986; Meehl, 1986a).

Assessment of abilities is usually done (a) to predict behavior individuals are capable of displaying and (b) to evaluate individuals' readiness to acquire other behavior. Both purposes involve assessing the current status of the behavior repertoire, which includes the cognitive, motor, and perceptual ability domains. The latter distinctions denote different emphases rather than discreteness of category, inasmuch as all molar behavior involves blends of all three.

Regarding abilities, psychologists are interested in normative assessment, or what a person is capable of doing compared with others. Assessment of this kind has been referred to as *aptitude*, *ability*, or *skill assessment*. The distinctions implied by these terms are not always clear. As with the categories of ability domains, these categories of assessment more appropriately denote contrasting points of emphasis.

Aptitudes, abilities, and skills all represent categories of behavior classes. Abilities and skills are often considered conceptually similar, but abilities are the broader, more molar category. *Aptitude* is a more elusive term, often defined as the *potential* for acquiring additional or subsequent skills. Thus, the MCAT assesses a premed student's potential for success in medical school, or the LSAT assesses a prelaw student's potential for success in law school. But in both cases, what is being assessed is the current status of a behavioral repertoire. Even the proficiencies that achievement tests test represents behavior classes (although typically they are arbitrary classes, established by instructional curricula).

If these distinctions are useful (i.e., ability/skill vs. aptitude), they are useful for distinguishing different *purposes* for assessment, as opposed to denoting *what it is* that is being assessed. Aptitude assessment is conducted for the purpose of *forecasting* the likelihood of certain behavior. (Given certain test scores and a certain GPA from a particular university, what is the probability that Ms. Jones will graduate from engineering school and how much engineering knowledge will she assimilate over the course of five years of engineering training?) By contrast, ability and skill assessment is typically (but not always) for the purpose of *evaluating the current status* of an individual's behavior repertoire. Although aptitude and ability assessment may stress different purposes, to say that qualitatively different attributes are being indexed (e.g., *potential* vs. *actual*) is simply not accurate or theoretically sound. *Potential* and *actual* are both being indexed by *any* procedure used to assess human capability.

The frequently encountered terms, *latent talent* and *hidden ability*, often confuse the issue because these terms imply that what is being assessed is different from manifest behavior. But the assessment of talent *always* involves behavioral evidence. It may be appropriate, for example, to characterize a gifted athlete from a distant country who has phenomenal quickness, running speed, and strength, but is ignorant of the game of football, as having latent talent for football. But this characterization is based on a lot of behavior, a lot of *manifest* behavior. For it to be scientifically meaningful to say that someone has exceptional talent for anything, *exceptional* behavior must be evinced.

Talent is public and the manifestation of talent is quantifiable. Furthermore, to the extent that the behavior observed shares communality with the (predicted) behavior of interest (which would be true in our gifted athlete example), the likelihood of a valid inference about "latent talent" is increased. This, of course, also pertains to aptitude assessment of all other kinds of human capability.

For example, for over 20 years, the Study of Mathematically Precocious Youth (SMPY) has demonstrated that approximately one to two percent of 7th graders are ready for college courses in mathematics (Benbow, 1988; Stanley, 1983). These students are selected as follows: If students score in the top one to two percent in their grade level on standardized achievement tests, they are given the opportunity to take the SAT-M. For these students, their SAT-M scores reproduce the same distribution found for 12th-grade college-bound students (Benbow, 1988). They have the *aptitude* for college courses. But this diagnosis is based on the *current* level of their quantitative ability (albeit a great deal of exceptional behavior). As we will show below, to be gifted in any given domain means that you have *more* behaviors in the domain in question or that your behaviors are more effective or more efficient. Quite literally, to be intellectually gifted means that you have more intellectual behaviors in your repertoire than others.

Just before the last *Handbook* appeared, Cleary, Humphreys, Kendrick, and Wesman (1975) analyzed the dimensions typically used to distinguish tests of human capability. They stressed that there are no differences *in kind* between intelligence and achievement, or between aptitude and achievement. There are, instead, four dimensions appropriate to the description of tests and the behavior repertoires they sample:

- *Breadth of material sampled.* Most intelligence and aptitude tests are broad, whereas most achievement tests are more circumscribed, in terms of the range of behaviors sampled.
- *Curriculum represented.* Achievement tests tend to be tied to a particular curriculum or educational program. Aptitude tests, on the other hand, reflect both formal and informal instruction.
- *Recency of learning sampled.* In general, achievement tests sample *new* learning, that is, behaviors recently acquired, whereas aptitude tests sample older learning.
- *Purpose of assessment.* Aptitude tests are used for predictive purposes, but the same items, if used to assess current knowledge, are thought of as constituting an achievement test (see Cronbach & Snow, 1977).

All tests of human capability can be characterized by these dimensions. Tests that purport to measure either aptitude or achievement have much in common. This reality has been stressed in Standard 8.3 of the most recent edition of *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 1985): "It should not be assumed that, because the words 'aptitude' or 'ability' are used in the title of a test, it measures a construct distinct from what is measured by an achievement test" (p. 52).

Indeed, Cronbach (1990), in his most recent measurement text, defines *aptitude test* as a "measure intended to predict success in a job, educational program, or other practical activity. Usually an ability test" (p. 701). Just as Dunnette and Borman (1979) have argued that the three traditional kinds of validities—content, criterion, and construct—should not be considered wholly distinct, but rather should be combined in common purpose, so do we assert that the three categories of human capabilities—aptitude, achievement, and ability—should be combined into a *generic behavior class* if we are to understand better the nature of our measuring instruments and the proficiencies they measure.

Psychological Meaningfulness and the Criterion of Scientific Significance

The psychological literature on human capability reveals a large number of purportedly distinct abilities, especially within the cognitive domain. Guilford's (1967) model, for example, proposed 120 factors (more recently, 180 factors), which motivated Carroll (1989a) to remark, "I like to say that Guilford fell victim to 'hardening of the categories' about halfway through his project" (p. 45). And, indeed, most people in the field feel that there is quite a bit of redundancy in Guilford's model. But the specific number of abilities still remains unadjudicated. Are there seven, as Thurstone (1938) suggested, or only one dominant dimension, as Spearman (1904) would have us believe? And what classes of abilities are most important to assess?

There is an important concept in the experimental literature that has been neglected in recent years. It was introduced by Spence (1948) as *the criterion of scientific significance.*

The basic idea is this: For a psychological construct to be scientifically significant, measures of the construct must not only display a respectable degree of internal consistency and replicability, but they must also relate to an array of meaningful psychological criteria. That is, they must display relationships with external criteria that we are interested in predicting and understanding. There are many measures of ability that meet the first requisite, but relatively few that meet the second. We believe that consideration of Spence's criterion would result in fewer trivial constructs and more scientifically significant measures within the ability domain.

To be sure, this idea is not new. The importance of linkage to external criteria has been discussed some by factor analysts and other methodologists. Vernon (1961), for example, stressed the need for factors to relate to a "capacity of daily life" (p. 27), although he admitted that what constitutes this criterion is somewhat subjective. Humphreys (1962) has suggested that the construct of general intelligence (*g*) should be splintered only when differential or incremental validity is displayed across meaningful external criteria. Snow (1980) has used the term *worldly significance* in the same vein, whereas Kelley (1939), McNemar (1964), and Stanley (personal communication, 1991) prefer *social utility*. We will employ this criterion to evaluate both the concepts and the measures of familiar, as well as newer, abilities.

Ability measures may appear to index different constructs, if we go by their labels. If, indeed, they measure distinct behavioral functions, they should have distinct correlates and show incremental validity when added to established ability measures. For example, in discussing advances in information processing approaches, Sternberg (1981) noted, "the incremental validity of information processing scores over psychometric scores has yet to be demonstrated for interesting external criteria" (p. 1186). To be scientifically significant and not simply to be rebuilding existing measures,

information processing instruments must do something meaningful that goes beyond what conventional psychometric measures currently offer. Achieving high degrees of reliability in experimental laboratory preparations is not enough.

Dunnette (1966) captured the essence and the importance of this idea years ago under the heading "The Delusions We Suffer."

> A common delusion seems to arise out of the early recognition that gathering data from real people emitting real behaviors in the day-to-day world proves often to be difficult, unwieldy, and just plain unrewarding. Thus, many retreat into the relative security of experimental or psychometric laboratories where new laboratory or test behaviors may be concocted to be observed, measured, and subjected to an endless array of internal analyses. These usually lead to elaborate theories and taxonomies, entirely consistent with themselves but lacking the acid test of contact with reality. Last year, McNemar (1964) summarized once more for us the pathetic record of factor analytically derived tests for predicting day-to-day behavior: Psychologists who choose to partake of the advantages of the more rigorous controls possible in the psychometric or experimental laboratories must also accept responsibility for assuring the day-to-day behavioral relevance of the behavioral observations they undertake. (p. 346)

The importance of coupling meaningful external correlates with internally consistent measurement operations has resurfaced in contemporary discussions. In a recent symposium on abilities, motivation and methodology, Jenkins (1989) remarked, "I hope that in 20 years when this group meets again, we will have really important things to say about psychological structures of organism-in-the-real-world. That will truly be a test of success of this whole endeavor"(p. 490). We recommend the proceedings of this symposium (Kanfer, Ackerman, & Cudeck, 1989), especially Jenkins' chapter.

The Nature and Organization of Cognitive Abilities

Contemporary, well-established measures of cognitive abilities, if we stay with those predictive of criterion behaviors in industrial and organizational settings, had their origins in factor analytic findings dating back to Spearman (1904). Although Thurstone (1938) gave us the term, it was Spearman who first called attention to the *positive manifold* (the positive intercorrelations displayed among all cognitive ability tests), and formulated as its explanation *g*, the underlying factor common to all forms of cognitive functioning and manifested in systematic sources of individual differences across all cognitive ability tests. Spearman theorized that *g* represented *mental energy*. Although the construct is still prominent in contemporary work on cognitive abilities, the surplus meaning of mental energy that Spearman attached to this entity is no longer seriously entertained.

Spearman (1927) thought mental energy was the fuel that ran the content-specific engines for the many different kinds of cognitive tasks. This idea provided the foundation for Spearman's two-factor theory (Spearman, 1914), with which many of Spearman's contemporaries (and subsequent factor theorists as well) took issue. The idea of a single regnant cognitive functioning capacity appeared too simplistic, and many tried to decompose *g* into its constituent components. These more molecular factors were viewed as the *primary abilities*. Early discussions of these primaries appeared reminiscent of faculty psychology. They were viewed as the psychological building blocks of cognitive functioning, akin to atoms in matter. Individuals who pioneered this approach were, most notably, Thurstone (1938), Cattell (1971), and Guilford (1967).

Other more behaviorally oriented theorists approached the problem differently, finding the explanation of the positive manifold in intimately intertwined stimulus-response bonds (e.g., Ferguson, 1954; Thomson, 1951; Thorndike, 1926; Tryon, 1935). These bonds were not thought of as directly linked to separate cognitive components that underlie cognitive functioning. Rather, the underlying structure for cognitive functioning was conceptualized as overlapping neural mechanisms, functioning collectively as a biological system.

Interestingly, this latter group's thinking is compatible with Skinner's (1969) ideas about the nature of intelligence.

> To say that intelligence is inherited is not to say that specific forms of behavior are inherited. Phylogenic contingencies conceivably responsible for "the selection of intelligence" do not specify responses. What has been selected appears to be a susceptibility to ontogenetic contingencies, leading particularly to greater speed of conditioning and the capacity to maintain a large repertoire without confusion. (p. 183)

These, then, were the two traditions that were brought to bear on the problem of explaining the covariation among cognitive ability tests. One was aimed at uncovering the basic components thought to underlie cognitive functioning, whereas the other conceived of cognitive behavior as resulting from multiple mechanisms operating in concert. The former tradition is found in contemporary cognitive psychology (Sternberg, 1984, 1985, in press), whereas the latter tradition is found in more behaviorally based treatments of cognitive abilities (Humphreys, 1979, in press). We will return to these two different ways of conceptualizing cognitive abilities, but for now, what is central is that both groups were addressing the same phenomenon, the positive manifold, and the question of how best to conceptualize its psychological significance in the most parsimonious way.

The Nature and Structure of the Positive Manifold

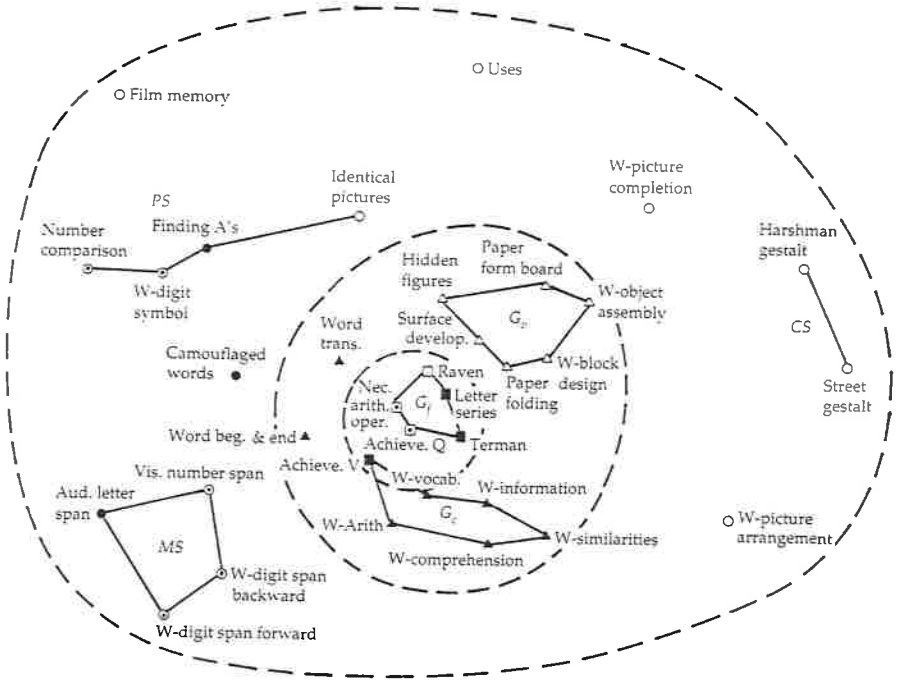
The consensus is growing among several contemporary investigators that cognitive abilities can be arranged hierarchically (Ackerman, 1989; Carroll, 1989a, 1989b; Cronbach, 1984; Cronbach & Snow, 1977; Gustafsson, 1984, 1988; Humphreys, 1979, 1982, 1985; Lohman, 1989; Snow, Kyllonen, & Marshalek, 1984). Concurrent with this perspective is the view that cognitive abilities can be arranged, pictorially, in a *radex*. This idea, illustrated in Figure 1, is, we believe, a most useful approach to the problem of how best to depict the domain of cognitive abilities.

Guttman (1954) had originally suggested this way of organizing cognitive abilities and, as it turns out, this framework is highly compatible with hierarchical schemes, starting with Burt (1949) and Vernon (1947, 1950; see also Ackerman, 1987, 1989; Humphreys, 1962, 1979; Lohman, 1989; Marshalek, Lohman, & Snow, 1983). The radex reveals rather clearly that there are at least two ways in which cognitive abilities, and measures thereof, can be psychologically close (or can covary): by sharing either *content* or *complexity*. Content is held constant as complexity varies along dimensions radiating from the centroid of the radex to its periphery, whereas complexity is held constant as content varies around circular bands defined by the radii of differing distances from the centroid of the radex.

To solidify this concept, Figure 2, panel *a*, provides a theoretical skeleton of a "perfect" radex (containing 10 tests); panel *b* contains a four-factor hierarchical solution (to be elaborated on below), depicting hypothetical

FIGURE 1

Radex Organization of Human Abilities



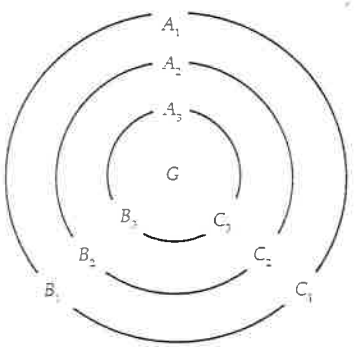
Note: Each point in the diagram represents a test. These tests are organized by content and complexity. Complex, intermediate, and simple tests are indicated by squares, triangles, and circles, respectively. Distinct forms of content are represented as black (verbal), dotted (numerical), and white (figure-spatial). Clusters of abilities that define well-known factors are indicated by a G. G_f = fluid ability, G_c = crystallized ability, G_v = spatial visualization. Tests having the greatest complexity are located near the centroid of the radex.

From "The Complexity Continuum in the Radex and Hierarchical Models of Intelligence" by B. Marshalek, D. F. Lohman, and R. E. Snow, 1983, *Intelligence, 7*, p. 122. Copyright 1983 by Ablex Publishing Corporation. Reprinted with the permission of Ablex Publishing Corporation.

degrees of association between the tests; and finally, panel *c* organizes these tests hierarchically. By studying the geometric distances (panel *a*) and hypothetical values (panel *b*), the two ways in which psychological tests can be close are clearly revealed. These three schemes are simply different ways chosen by different investigators to organize the

FIGURE 2

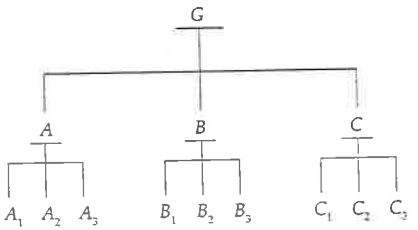
Parallelism Between the Radex and the Hierarchical Factor Model



(a) Radex Scaling for 10 Tests

	G	A	B	C
A_1	1	3	0	0
A_2	2	4	0	0
A_3	3	3	1	1
B_1	1	0	3	0
B_2	2	0	4	0
B_3	3	1	3	1
C_1	1	0	0	3
C_2	2	0	0	4
C_3	3	1	1	3
G	4	2	2	2

(b) A Corresponding Hierarchical Factor Matrix



(c) The Associated Hierarchical Factor Diagram

Note: This is a hypothetical example illustrating the degree of overlap between the radex and the hierarchical factor model.

From "The Topography of Ability and Learning Correlations" by R. E. Snow, P. L. Kyllonen, and B. Marchalek in *Advances in the Psychology of Human Intelligence* (Vol. 2, p. 61), R. J. Sternberg, Ed. (1984), New Jersey: Erlbaum. Copyright 1984 by Erlbaum. Reprinted by permission.

psychometrics of the radex. All are highly compatible.

The radex is composed of an indefinite number of simplexes and circumplexes. Simplexes are revealed by showing that along the complexity dimensions, which extend from the centroid to the periphery, correlations between tests diminish as a function of their distance from one another. Circumplexes, on the other hand, are revealed by traveling along a circular band defined by a constant degree of complexity. As tests diverge from one another

within a circular band, they change in content but not complexity, and their correlations diminish. These correlations continue to decrease, reaching their theoretical minimum at a 180° arc, when the tests are exactly opposite one another on the circular band. (A line drawn between two such tests, passing through the centroid, traces the diameter of the circular band.) To the extent that tests covary, they are close to one another within this two-dimensional space. At least two dimensions, then—content and complexity—are necessary to triangulate a test's specific location.

A number of regions in Figure 1 are labeled to denote groupings of abilities of special theoretical importance. For example, Cattell's (1971; Horn & Cattell, 1966) distinction between G_f (*fluid*) and G_c (*crystallized*) abilities are illustrated in Figure 1. As initially proposed, G_f was thought to reflect a physiological parameter of the individual—raw potential, or the capacity to learn. G_c, in contrast, was construed as indexing the acquisitions of learning or the cultural products of experience. Just as *g* has lost the surplus meaning Spearman originally attached to it (viz., *mental energy*), G_f and G_c are currently thought of more as just differing clusters of abilities and less as biological potential versus stored experience. Cronbach (1977) has noted that "fluid ability is itself an achievement.... [It is] the residue of indirect learning from varied experience" (p. 287; for similar views but for different reasons, see Horn, 1985; Humphreys, 1981; Scarr & Carter-Saltzman, 1980; Snow & Yalow, 1982).

The G_f, G_c, and G_v (*spatial visualization*) clusters of Cattell and Horn correspond to Vernon's (1950, 1961) *g* (*general intelligence* or the *general factor*), *v:ed* (*verbal-numerical-educational*) and *k:m* (*practical-mechanical-spatial*), respectively. These are the most common splinterings (into roughly two halves) of the positive manifold or radex. Further parallels between Cattell/Horn and Vernon are given by Gustafsson (1984, 1988) and Lohman (1989). Other investigators, like Thurstone (1938), preferred to subdivide these broad factors into more circumscribed abilities at uniform levels of abstraction and common content. In doing so, smaller clusters are formed at greater distances from one another as well as from the centroid of the radex. Some of these areas are also labeled on Figure 1.

Much of the early factor analytic work was aimed at mapping the covariation displayed among the cognitive abilities with the fewest dimensions possible. How many continua are necessary to extract psychological significance from the communality shared among these tests has been a matter of much debate: one dominant dimension, as Spearman (1904) proposed? a few dimensions—"seven primaries"—as Thurstone (1938) advocated? or as many as 120 to 180, as Guilford (1985) has hypothesized? The covariation that these theorists wished to explain involved the same space, but the number of dimensions needed to map this space remains the issue. Some comments about factor analysis might help to focus this discussion.

Factor Analysis

The goal of factor analysis is to condense the information in the covariation of n variables to a smaller set of m factors. The amount of covariation among the variables in a matrix defines the common variance. Common variance can be estimated by computing the squared multiple correlation between each variable and all of the other variables and then summing these (n) R^2s. This sum, divided by the number of variables, estimates the common variance proportion of the total variance that the variables share.

In the context of factor analysis, the variance for each variable has three components: common, specific, and error variance. Common and specific variance compose the variable's reliable variance (viz., r_{xx}, with the complement, $1 - r_{xx}$, being error variance); whereas specific and error variance constitute the variable's unique variance (viz., $1 - h^2$, where h^2 = common variance by conventional factor-analytic notation).

Therefore, the decomposition of the total variance in a correlation matrix into common and unique components can be expressed as:

$$
\begin{array}{ll}
R^2_{1\cdot 2,3,4,\ldots,n} & 1 - R^2_{1\cdot 2,3,4,\ldots,n} \\
R^2_{2\cdot 1,3,4,\ldots,n} & 1 - R^2_{2\cdot 1,3,4,\ldots,n} \\
R^2_{3\cdot 1,2,4,\ldots,n} & 1 - R^2_{3\cdot 1,2,4,\ldots,n} \\
\vdots & \vdots \\
R^2_{n\cdot 1,2,3,4,\ldots,n-1} & 1 - R^2_{n\cdot 1,2,3,4,\ldots,n-1}
\end{array}
$$

$$\frac{\sum R^2}{n} = \text{common variance}$$

$$\frac{\sum (1 - R^2)}{n} = \text{unique variance}$$

common variance + unique variance = total variance = 1.00

Factor analysis strips away the unique variance in a correlation matrix (the specific and error variance of each variable) and focuses on the common variance or *communality* (h^2) among the variables. The psychometric objective of factor analysis is *not* so much to understand the *nature* of the variables but rather to account for the *covariation* among the variables. What is at issue is how best to map this covariation with the fewest number of dimensions (factors). To the extent that the communality of a matrix is appreciable—as it is for heterogeneous collections of cognitive ability tests—the covariation of the variables can be expressed by fewer dimensions than there are variables in the matrix. This task, with respect to cognitive abilities, has occupied the professional careers of many prominent psychometricians (Carroll, 1989a; Cattell, 1971; Guilford, 1967, 1985; Horn, 1986; Humphreys, 1962; Kelley, 1928; Spearman, 1927; Thurstone, 1938; Thurstone & Thurstone, 1941; Vernon, 1950, 1961).

To the extent that a group of variables covary appreciably with one another, their covariation may collectively be characterized by one dimension or a *general factor*. If the situation is more complex—for example, if certain pockets of variables form distinct covarying subgroups—factor analytic solutions result in multiple factors, one for each subgroup. Through this procedure, a finite set of (m) factors—appreciably smaller than the number of variables under analysis—can characterize most of the common variance. (If desired, factor scores can be generated to estimate individual differences on each factor; Harman, 1976; Rummel, 1970.) This technique is a very useful tool—it can often reduce the number of variables in a matrix to a much smaller subset, thereby dispensing with the redundancy that runs through individual measures. A few factors can replace a long list of variables!

The final product of a factor analysis is a factor matrix, with as many rows as there are variables and as many columns as there are factors (see Radex solution, Figure 2, panel *b*). Factors are simply dimensions or axes "placed" within the multidimensional space that represents a correlation matrix. Factor loadings are correlations between the variables and the factors.

The extent to which factors and variables are correlated can also be expressed geometrically. Variables (or factors) can be represented as geometric vectors, and the cosine of the angle formed at their intersection is numerically equal to their correlation. Variables are independent when their vectors cross at 90° (cosine 90° = .00, hence r_{xy} = .00). Conversely, they correlate at unity when their vectors lie on the same line, in the same or opposite direction (cosine 0° and cosine 180°, hence r_{xy} = 1.00 and −1.00, respectively). Cosines of angles between 0° and 90° take on positive values between 1.00 and 0.00 (positive correlations), whereas those between 90° and 180° take on negative values between 0.00 and −1.00 (negative correlations).[2]

One reason for going through all of this is to demystify the nature of psychometric factors. They are simply *arbitrary* dimensions placed into a space representing correlations. The goal of factor analysis is to characterize

TABLE 1

Table of Intercorrelations Defining Two Factors

	1	2	3	4
1		90	75	75
2			75	75
3				90
4				

Three Factor Solutions

	Orthogonal		*Oblique*		*Hierarchical*		
	I	II	I	II	*g*	I	II
1	837	447	949	000	865	390	000
2	837	447	949	000	865	390	000
3	447	837	000	949	865	000	390
4	447	837	000	949	865	000	390

Note: The correlation between the oblique factors is .831.

From "The Construct of General Intelligence" by L. G. Humphreys, 1979, *Intelligence*, 3, pp. 107–108. Copyright 1979 by Ablex Publishing Corporation. Adapted with the permission of Ablex Publishing Corporation.

the covariation between (*empirical*) variables with the smallest possible number of (*theoretical*) factors. And there are different ways of placing factors in a correlation-matrix space to account for the common variance.

Because factor matrices contain correlations between variables and factors, the principles of multiple regression apply. When for each factor, the factor loadings are squared, summed, and the sum divided by the amount of common variance in the matrix (which is the sum of the communalities), the result is the proportion of common variance accounted for by each factor. When all these proportions are summed (for factor solutions yielding independent or uncorrelated factors), the sum is the proportion of common variance accounted for by the factor solution.

Orthogonal (independent factors) and *oblique* (correlated factors) solutions are well-known methods of factoring, but *hierarchical methods* (featuring orthogonal solutions) are rapidly gaining popularity. (We hope they soon will appear in conventional statistical packages.) One of the more attractive hierarchical techniques is the Schmid and Leiman (1957) solution, which transforms an oblique factor solution into a pattern of orthogonal factors arranged in two (or more) tiers. Descriptions of the Schmid-Leiman approach, with computational examples, may be found in Ackerman (1987), Carroll (1985, 1989b), and Humphreys (1962, 1982, 1984).

Table 1 compares the results of these three kinds of factor solutions for a hypothetical correlation matrix. The intercorrelation of four variables is shown. Variables 1 and 2 might be measures of the same ability, say, verbal ability, whereas variables 3 and 4 are measures of, say, mathematical ability. Three kinds of factor solutions are illustrated. The orthogonal solution shows two factors that are hard to interpret. The oblique solution appears to be better than the orthogonal in the Thurstonian sense of simple structure, but the factors are correlated (r = .831). The hierarchical factors are uncorrelated and highly interpretable but do require more factors than the other solutions. However, when the correlation matrices being factor analyzed are large, hierarchical solutions can be just as parsimonious as the other solutions.

Hierarchical solutions place the most comprehensive factor (i.e., the general factor) at the vertex. With respect to cognitive abilities, the general factor may be interpreted as the overall level of the cognitive repertoire, or the general factor, or general intelligence, essentially, Spearman's g but without the surplus meaning. Individual differences on this factor reflect complexity or sophistication of the repertoire. The branches and twigs that splinter off this base organize the remaining common variance along circular bands of content in the radex. All of this description is found conceptually in Vernon's (1961) early work on the structure of human abilities (Figures 3, 4, and 5), which parallels the theoretical model given earlier (Figure 2, panels *a*, *b*, and *c*, respectively). Vernon's construction of cognitive abilities clearly is consistent with and overlaps the radex.

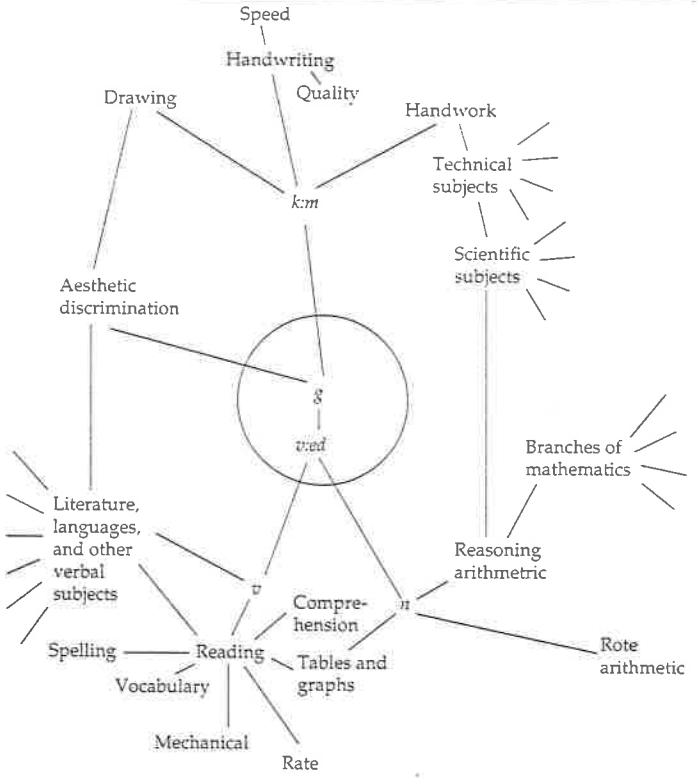
There are other advantages to the Schmid-Leiman hierarchical approach. First, independent factors can stand alone as reference dimensions, whereas the correlation of oblique factors requires interpretation. Second, the nature of the loadings in the hierarchical solution can often provide insight into how best to arrange or order tests in terms of their content (the more circumscribed tests tend to be identified more with particular forms of content, e.g., linguistic, numerical, or figural). And finally, major and minor group factors can be evaluated for incremental validity in terms of external correlations over and above the contribution of the general factor to ascertain whether they account for additional psychologically meaningful variance or whether they are merely pieces of common variance having little or no psychological significance.

It should be stressed that whatever the factor solution, this is only half the picture with respect to the criterion of scientific significance. Once a factor solution is obtained, the predictive efficacy of each factor must be assessed in order to address the other half of the criterion of scientific significance. The external correlates of each factor must be evaluated for their psychological meaningfulness. Factor analysis does *not* reveal what proportion of the common variance in a correlation matrix is responsible for useful external correlates. It only tells us what proportion of the reliable variance of each variable is shared with other variables in the matrix and how much of this communality is accounted for by the *m* factors. It does *not* tell us which factors have meaningful correlates.

In our view, it is not essential for the factors in a factor analysis to account for all of the common variance in a correlation matrix. There is no reason to suppose that all of the common variance in a correlation matrix should be represented in the factor matrix. In fact, there is good reason to suppose that this should *not* be the goal of factor analysis. Some of the common variance undoubtedly represents correlated method variance that we should gratefully discard. This idea is not new, although its extension to the problem of determining the number of factors is. Nonattribute but reliable variance is contained in all psychometric measures. These components function as nuisance variables that distort predictor-criterion relationships. These sources of variance have been discussed as *method variance* (Campbell & Fiske, 1959), *construct irrelevancies* (Cook & Campbell, 1976), *systematic bias* (Humphreys, 1976), *constant error* (Loevinger, 1954), systematic *ambient noise* (Lykken, 1968), and *crud* (Meehl, 1990b). It is the type of bias that Cronbach and his colleagues have tried to minimize via their generalizability

FIGURE 3

Vernon's Conceptualization of Human Abilities: A Spatial Representation



From *The Structure of Human Abilities* (2nd ed., p. 47) by P. E. Vernon, 1961, New York: Routledge, Chapman, & Hall. Copyright 1961 by Routledge, Chapman, & Hall. Adapted by permission.

methodology (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963). It is a well-known concern.

It is also important not to be misled by the distinctive content of supposedly contrasting measures. Ostensibly distinct scales might be tapping the same source of individual differences—but in content-unique ways—through converging operations. Spearman was aware of this possibility early in his career, when he formulated the concept of the *indifference of the indicator*—that is,

FIGURE 4

Vernon's Conceptualization of Human Abilities: An Empirical Solution

	Tests	Group Factors					
		g	*k:m*	*v:ed*	*v*	*n*	h^2
0	Progressive Matrices	.79	.17				.65
	Dominoes (nonverbal)	.87					.75
	Group test 70, Pt. 1	.78	.13				.62
4	Squares	.59	.44				.54
8	Assembly	.24	.89				.85
2	Bennett mechanical	.66	.31				.54
25	Verbal	.79		.29	.45		.90
	Dictation	.62		.54	.48		.90
14	A.T.S. spelling	.68		.41	.43		.82
21	Instruction	.87		.23	.09		.82
3A	Arithmetic Pt. I	.72		.49		.39	.91
	Arithmetic Pt. II	.80		.38		.16	.82
23	A.T.S. arithmetic	.77		.36		.32	.82
	Variance percent	52.5	8.7	8.4	6.9		76.5

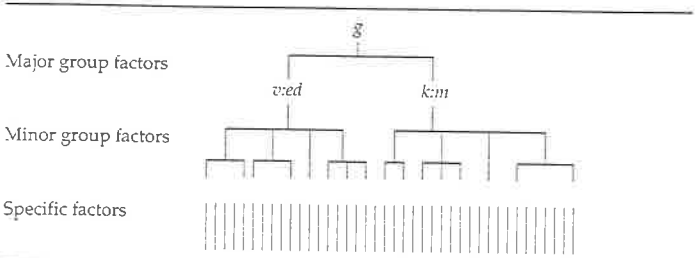
From *The Structure of Human Abilities* (2nd ed., p. 23) by P. E. Vernon, 1961, New York: Routledge, Chapman, & Hall. Copyright 1961 by Routledge, Chapman, & Hall. Adapted by permission.

that all cognitive ability tests index *g* to varying degrees. Later in his career, he commented on the lack of appreciation of this concept:

> What really does manifest a dominant influence upon the conception of the mental processes involved in tests is the suggestion emanating from language. When once a test has received a name, this has generally been accepted as expressing its sole essential character. Few are the analysts who have pushed their examination beyond this merely popular stage. (Spearman & Jones, 1950, p. 143)

FIGURE 5

Vernon's Conceptualization of Human Abilities: A Hierarchical Representation of Factors at Contrasting Levels of Generality



Note: Minor group factors correspond to Thurstone's primaries. Specific factors refer to individual tests.

From *The Structure of Human Abilities* (2nd ed., p. 22) by P. E. Vernon, 1961, New York: Routledge, Chapman, & Hall. Copyright 1961 by Routledge, Chapman, & Hall. Adapted by permission.

Just as there has been a problem in experimental psychology of conflating hypothesis testing with theory testing (Meehl, 1967, 1978), so has there been an equally detrimental problem in correlational psychology of considering tests with different content and different labels to be measuring different abilities (and even different underlying processes). This is, of course, the problem of reification of factors, a problem with a long history in factor analytic research. To be sure, factors can represent real entities and distinct underlying processes *if* internal and external relationships justify it. But the factor analysis of cognitive abilities appears to show, by and large, a structure indicative of continua of complexity and content rather than of discrete breaks. Although robust major group factors can be extracted reliably, from a hierarchical point of view they are better construed as more focused or more concentrated behavioral repertoires developed with differing emphases on distinct symbol systems. As linguistic, numerical, and figural kinds of content merge within the radex along concentric bands of uniform complexity, we find sectors with symbol compositions that tend to reflect blends of linguistic-numerical, numerical-figural, or figural-linguistic hybrids.

To summarize, the products of factor analysis should be required to have meaningful external correlates. This is an important requirement because, although all common variance is reliable variance, not all common variance is scientifically significant variance. Scientific significance is never established via factor analysis alone. Scientific significance requires the establishment of external relationships; it requires going beyond the internally consistent statistical relationships among indicators. Otherwise, we might simply have little more than common variance due to method sources of variance that function systematically to create reliable factors and distort the interpretability of other factors, but contain little substance outside the testing situation. Kelley (1939) had a name for such factors; he called them "mental factors-of-no-importance."

Assessing the Scientific Significance in Factor Structures at Different Levels of Generality

There are several conventional test batteries that assess multiple factors at uniform levels of generality (cf. Carroll, 1989b, pp. 165–166). Many studies are conducted with these instruments each year to uncover the "latent structure" of different predictor-criterion relations. Consider, however, the following.

Suppose that multiple regression equations using moderately correlated group factors as predictors can account for slightly more variance than that achieved by the general factor. It is still possible that the successive R^2-change increments could have been contributed by the common variance of the group factors (i.e., the variance they share not only with each other but also with the general factor), rather than by their specific variances. For example, conventional measures of spatial and mathematical ability share common variance, but not all of their common variance is shared with verbal ability. In a two-predictor regression equation with verbal ability entered first, the incremental validity contributed by mathematical ability might be due to that proportion of its common variance shared with spatial ability. Hence, the R^2 increment is actually due to the fact that the two-predictor variate (verbal plus mathematical) is now a better estimator of the common variance in cognitive tests—that is, the general factor—and not, as one might assume, due to the contribution of the specific variance associated with either group factor. Yet it is to the specific variance of each group factor that we typically attribute the incremental validity shown by multiple predictors.

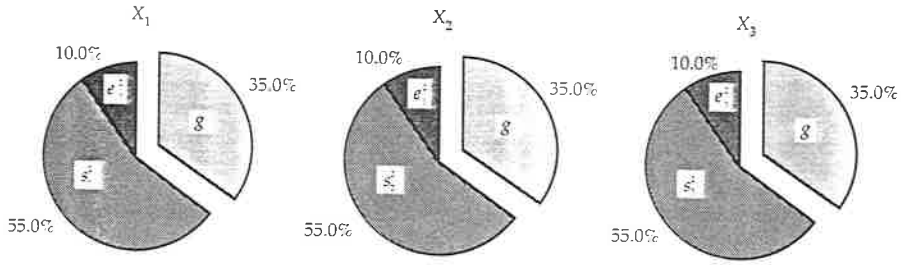
If it is the general factor cutting across cognitive tests that shares variance with the criterion of interest, then successive R^2 increments might simply mean that the aggregate variate, with the addition of each contributing variable (e.g., verbal + mathematics + spatial), provides a better estimator of the general factor via successive approximation. This possibility should be considered when multiple predictor equations are being evaluated (see "The Criterion Problem" below).

Furthermore, sample-specific bias compounds the problem. Because multiple predictors enhance the likelihood of capitalizing on sample specific variance, thus inflating R^2s, it is imperative to compare multiple regression equations using major or minor group factors with simple regression equations using the general factor via their comparative cross-validated R^2_{cv}s (see below). It may be that, for the more intellectually demanding professions, the valid variance of multiple predictor batteries (i.e., the predicted variance or R^2 that holds up on cross-validation), is simply *g* variance that is being better estimated by successive approximation with the use of the multiple predictors. And the reason regression coefficients shrink more on cross-validation when multiple predictors are used, in contrast to the situation for a single predictor that measures the general factor, is that more complex variates are more likely to generate more sample-specific noise.

With a single global measure of general intelligence, however, most of the reliable variance is common with the general factor. The proportion of specific variance in the total variance is minimized. (There may be chunks of specific variance sprinkled throughout the global measure of general intelligence, but the aggregation of common variance attenuates their contribution to the composites' total variance.) In a regression equation that uses a global measure of general intelligence, most of the measure's reliable variance is attribute variance associated with the communality that is found in all cognitive tests. If criterion performance is also *g*-saturated, then the regression coefficient will not shrink much on cross-validation. If samples are large and representative of the population of applicants, the structural

FIGURE 6

Three-variable Variate



Note: Three hypothetical variables having the same amount of common, specific, and error variance. As individual components of a predictor variate, most of the variance of each component is specific variance.

parameters b or beta will tend to capture most of the predictor-criterion covariation from sample to sample.

This contrasts sharply with the situation for multiple predictors. With multiple predictors, there is more noise attached to each regression coefficient because only a little common variance enters into the valid (cross-validated) predictor-criterion correlation. There is more slippage from sample to sample because several structural parameters are being estimated. This can be illustrated quantitatively, using the following formula and the hypothetical values found in Figure 6.

The basic formula for the correlation of a composite ($X_1 + X_2 + \ldots + X_N$) with another measure, Y, or a source of variance, is:

$$r_{Y(X_1 + X_2 + \ldots + X_N)} = \frac{\Sigma CYX_i}{[\Sigma \mathrm{Var}(X_i) + 2 \Sigma \Sigma CX_iX_j]^{1/2}}$$

where:

ΣCYX_i = the sum of all the covariances between Y and the constituents, X_i, of the composite

$\Sigma \mathrm{Var}(X_i)$ = the sum of all the X_i variances and

$\Sigma \Sigma CX_iX_j$ = the sum of all the constituent covariances

Figure 6 shows three hypothetical variables. Each shares 35 percent of its variance with a general factor ($r_{gX_i} = .59$), and each has 55 percent specific variance (so for each, $r_{xx} = .90$). Given these values, each variable is correlated with the other (.59 x .59) or $r_{X_iX_j} = .35$. If we aggregate these variables, their overlap with this general factor increases to (see Figure 7):

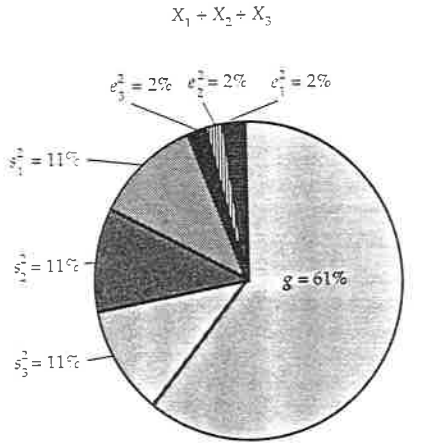
$$r_{g(X_1 + X_2 + X_3)} = \frac{(.35)^{1/2} + (.35)^{1/2} + (.35)^{1/2}}{[1 + 1 + 1 + 2(.35) + 2(.35) + 2(.35)]^{1/2}}$$

$$= .786$$
$$r^2 = .618$$

Further, the correlation of this composite and an individual component of specific variance (of which there are three), is:

FIGURE 7

Three-variable Composite



Note: When the three components found in Figure 6 are aggregated, most of the composite's variance is variance shared with a general factor common to each. Moreover, the influence of any one form of specificity is considerably reduced.

$$r_{g(X_1 + X_2 + X_3)} = \frac{(.55)^{1/2}}{[1 + 1 + 1 + 2(.35) + 2(.35) + 2(.35)]^{1/2}}$$

$$= .328$$
$$r^2 = .108$$

And finally, the correlation of the composite and an individual component of error variance (of which there are three) is:

$$r_{e_i(X_1 + X_2 + X_3)} = \frac{(.10)^{1/2}}{[1 + 1 + 1 + 2(.35) + 2(.35) + 2(.35)]^{1/2}}$$

$$= .14$$
$$r^2 = .0196$$

If we can assume that criterion performance has a large general factor associated with it, then the composite, ($X_1 + X_2 + X_3$), is much more descriptively apt at characterizing the construct-relevant predictor/criterion covariation than the three variables taken in weighted combination (viz., the three-variable variate). There will be less slippage in the cross-validation of the composite than there will be for the three-variable variate because all of the relevant information is contained in the composite and only one structural parameter is being estimated. Aggregation in the composite attenuates the individual components of specific variance so that each component contributes less to the composite's overall variance. Figure 7 shows how the individual components of specific variance are attenuated in the composite. Because they are independent sources of variation, they do not aggregate (covariation is necessary for aggregation!). To summarize, large components of specific variance not only can mask common variance contributions to predictor/criterion covariation, they may also distort estimation by capitalizing on sample specific noise through the estimation of multiple parameters. Aggregation can be used to minimize these undesirable effects.

If, however, components of specific variance are relevant to the criterion, then using a multiple-variable variate will improve prediction of the criterion over the use of the composite. For example, Humphreys (1985) and Vernon (1961) have shown that, for certain criteria, *v:ed* and *k:m* can improve on prediction from *g* alone, and the improvement holds up on cross-validation. This is because relevant specific variance of *v:ed* and *k:m* adds to the general factor variance that both hold in common. Furthermore, McNemar (1964) has shown that it is profitable to splinter the major group factor, *v:ed*, into verbal (*v*) and quantitative or numerical (*n*) minor group factors for the differential prediction of

academic criteria. In such instances, use of the more complex multiple predictor equation is justified.

Criticisms of General Factor

Cronbach (1989) has criticized the dominant dimension approach to defining general intelligence, inasmuch as estimating the centroid of a radex varies as a function of the tests an investigator chooses to employ. How does one know where the desired centroid is or should be? This is a valid criticism. It applies to all sets of tests, and we are unaware of an analytic solution. Furthermore (for reasons he delineates in Cronbach, 1975a, and other writing), the actual centroid for any collection of cognitive ability tests is, to some degree, in a state of flux (and this is true for both interindividual and *intra*individual differences). After all, premiums are placed on different sets of skills, cognitive behaviors develop and change over generations, opportunities vary, and so on, enough to suggest that the patterns of phenotypic traits are dynamic in a manner not unlike that of genetic drift. Finally, if this criticism is valid for the general factor, it is also valid for the major and minor group factors derived factor analytically or built by aggregating minor or specific factors systematically.

However, there are methods with which to assess how conceptually close two measures are. At the least, existing measures can be assessed for their conceptual and empirical interchangeability. Furthermore, because several instruments are available that purport to measure the general factor (and less general factors), these different measures, built by different investigators, provide the opportunity to establish convergent validity and demonstrate constructive replication (Lykken, 1968).

This is an important consideration because measures can correlate in the high .90s and still display different patterns of external correlations (McCornack, 1956). The possible range of correlation that a measure, m_2, can display with a criterion, c, given that we know the correlation of m_2 with an equivalent measure, m_1, and the correlation between m_1 and c, is given by the following formulas for upper and lower limits:

$$\text{Upper limit of } r_{m_2c} = r_{m_1c}\, r_{m_1m_2} + [(r_{m_1c})^2 (r_{m_1m_2})^2 - (r_{m_1c})^2 - (r_{m_1m_2})^2 + 1]^{1/2} \quad (1a)$$

$$\text{Lower limit of } r_{m_2c} = r_{m_1c}\, r_{m_1m_2} - [(r_{m_1c})^2 (r_{m_1m_2})^2 - (r_{m_1c})^2 - (r_{m_1m_2})^2 + 1]^{1/2} \quad (1b)$$

Table 2, taken from Jensen (1980a), reveals the extent to which a (*new*) measure, m_2, having a given correlation with a (*known*) measure, m_1, can vary in its correlation with a criterion variable, c, across different values of $r_{m_1m_2}$. As Table 2 reveals, the possible range of correlation with the criterion is incredible, even for $r_{m_1m_2} = .95$. This shows that two measures ought not to be considered empirically interchangeable, solely on the basis of a high correlation between the two (or their *congruence coefficient*).

A way to evaluate the extent to which two measures are interchangeable is provided by the concept of *extrinsic convergent validation*, an elaboration of the multitrait multimethod matrix method advanced by Fiske (1971).[3] According to Fiske, measures of the same construct should not be considered empirically interchangeable until they display similar patterns in their correlational profiles. That is, two independent measures of the same ability cannot be considered interchangeable unless they display similar correlational patterns of corresponding convergent and discriminant correlations as well as other correlations with external criteria in the intermediate range.

TABLE 2

Upper and Lower Limits of the Possible Range of Criterion Validity (r_{cm_2}) for Test M_2 When Criterion Validity of Test M_1 Is r_{cm_1} and the Correlation Between the Two Tests Is $r_{m_1m_2}$

$r_{m_1m_2}$ (or r_{cm_1})	r_{cm_1} (or $r_{m_1m_2}$)								
	.10	.20	.30	.40	.50	.60	.70	.80	.90
.95	.41 −.21	.49 −.11	.58 −.01	.67 .09	.75 .20	.82 .32	.89 .44	.95 .57	.99 .72
.90	.52 −.34	.61 −.25	.69 −.15	.76 −.04	.83 .07	.89 .19	.94 .32	.98 .46	1.00 .62
.85	.61 −.44	.69 −.35	.76 −.25	.82 −.14	.88 −.03	.93 .09	.97 .22	1.00 .36	.99 .53
.80	.68 −.52	.75 −.43	.81 −.33	.87 −.23	.92 −.12	.96 .00	.99 .13	1.00 .28	.98 .46
.75	.73 −.58	.80 −.50	.85 −.41	.91 −.31	.95 −.20	.98 −.08	1.00 .05	1.00 .20	.96 .39
.70	.78 −.64	.84 −.56	.89 −.47	.93 −.37	.97 −.27	.99 −.15	1.00 −.02	1.00 .28	.94 .32
.65	.82 −.69	.87 −.61	.92 −.53	.96 −.44	.98 −.33	1.00 −.22	1.00 −.09	.97 .06	.92 .25
.60	.85 −.73	.90 −.66	.94 −.58	.97 −.49	.99 −.39	1.00 −.28	.99 −.15	.96 .00	.89 .19

To illustrate this point, Table 3 contains three extrinsic convergent validation profiles. These data are from the 12th-grade cohort of Project TALENT.[4] The first profile is for an information test of world literature (24 items, including prose and poetry, and several questions that are based on required reading in many high schools); the second is for vocabulary (30 items, designed to assess nontechnical words); the final profile is for a reading comprehension test (48 items, designed to assess comprehension of written material across a wide range of topics). The last two measures are excellent markers of verbal ability as well as general intelligence, whereas the first measure is typically considered farther from the centroid of the radex, but in the same slice of content. (Using conventional nomenclature, it would be considered more of an *achievement* measure.) The profiles consist of

TABLE 3

Extrinsic Convergent Validation Profiles Across Three Measures Having Verbal Content

	Literature	*Vocabulary*	*Reading Comprehension*
Aptitude Tests			
Mechanical reasoning	.43	.52	.54
2–D visualization	.25	.32	.35
3–D visualization	.35	.43	.47
Abstract reasoning	.45	.53	.61
Arithmetic reasoning	.54	.63	.63
High-school math	.57	.59	.57
Advanced math	.42	.43	.39
Information Tests			
Music	.67	.68	.62
Social studies	.74	.74	.71
Mathematics	.62	.63	.57
Physical science	.64	.67	.60
Biological science	.57	.61	.56
Interest			
Physical sciences	.24	.25	.22
Biological sciences	.26	.25	.22
Public service	.16	.12	.12
Literary-linguistic	.37	.32	.32
Social service	.07	.06	.07
Art	.32	.30	.29
Music	.23	.20	.20
Sports	.12	.12	.13
Office work	−.35	−.29	−.27
Labor	−.08	−.06	−.06

Note: These correlations were based only on female subjects (male profiles are parallel). N = 39,695. Intercorrelations for the three measures were the following: literature/vocabulary = .74, literature/reading comprehension = .71, and vocabulary/reading comprehension = .77.

correlations with an array of Project TALENT measures, with different locations in the radex as well as outside of it.

Vocabulary and reading comprehension display the highest degree of correlation, r_{xy} = .77. Recall from formula 1 that this degree of correlation can result in external correlations that can range over 1.25 correlational value points for r_{mzc} when r_{mic} = .30! However, our analysis reveals that both markers of verbal ability map essentially the same nomothetic span (network of correlates), at least as regards these criteria. These correlational profiles can be different when new variables are used, but in the context of the present variables, vocabulary and reading comprehension can be considered empirically interchangeable. In fact, the information test maps essentially the same space, which illustrates how fundamentally similar "achievement" and "aptitude" tests can be.

Building Measures of Constructs

In the early days of factor analytic research, most American investigators adopted the Thurstonian (1938) approach to guide test construction. Test batteries were constructed to map the "underlying" natural cleavages of the ability domain at uniform levels of molarity. The conviction that there was more to cognitive functioning than one dominant dimension was buttressed by the belief that all meaningful criteria were complex and multidimensional. The agendum was to build a finite set of measures that were relatively independent and specifically designed to account for different sectors of the complex criterion variable. What was needed to account for such criteria as school performance and job performance were multifactor test batteries with components that would map onto unique slices of the criterion pie. But as the evidence accumulated, it became apparent that the communality on both sides of the prediction equation was not fully appreciated.[5]

The Thurstonian approach makes some fairly strong assumptions about the nature and organization of cognitive functioning. It assumes that little is measured in common by the different tests, for example, of verbal, quantitative, spatial, pictorial, perceptual, and problem-solving abilities. Such multifactor batteries are predicated on the *specificity* of the measures rather than on the communality they share. So, if you have a battery of, say, 10 tests whose intercorrelations are all around .10, you have very little of the *general factor*—unless one test is a measure of general intelligence and the others have little to do with cognitive functioning. (Recall that Figures 1 and 2 show how it is possible for ability tests to covary on two dimensions, content *and* complexity.)

An alternative approach assumes that there are no "factor pure" tests—pure, that is, as conventionally conceived: a test that measures one, and only one, dimension. This notion of purity has tended to conflate statistical purity with psychological purity (Hulin & Humphreys, 1980). We maintain that, within the full range of human capability, cognitive functioning covaries with many measures having different content (e.g., verbal, spatial, quantitative, pictorial). Even Guilford (Guilford & Michael, 1948) seemed aware of this. When discussing the measurement of a person's pure factor score, he observed that it was usually necessary to add *suppressor variables* to partial out g or other unwanted content. As Vernon (1961) pointed out in response, why not just admit that all cognitive ability tests *do* involve g, instead of artificially removing it by rotation (pp. 133–134). If Guilford's suggestion were followed, the use of suppressors might actually result in the *loss* of scientifically significant variance when systematic bias in the measure is maximized.

The following methodology, developed by Humphreys (1952, 1962, 1984), can be used to build measures of human behavior capability, whether cognitive, motor, or perceptual, regardless of level of abstraction. (Later, we will show that this methodology also has important ramifications for criterion development.)

First, it is assumed that the variance of observed scores, X, can be analyzed into three components:

$$X^2\text{(total variance)} = A^2\text{(attribute variance)} + B^2\text{(bias variance)} + E^2\text{(error variance)}$$

Attribute variance comes from systematic sources of individual differences that constitute the attribute of interest. Ideally, attribute variance should approach unity. Bias variance

is also systematic (i.e., reliable), but reflects one or more consistent sources of variation distinct from the attribute of interest (e.g., a systematic response style or one of the several forms of method variance). Collectively, attribute and bias variance constitute true score variance in classical test theory (bias being reliable variance that is not associated with the attribute of interest). Finally, the third component of variance is error variance, random variation, or error in the classical sense.

In the context of factor analysis, it is easy to see that common variance cuts across both attribute and bias components to varying degrees as a function of the nature of the variables. Variance overlap among the several variables will involve both components, hence both will contribute to the communality in the matrix. Thus, factors can arise from either form of reliable variance or, more typically, from both.

The goal of measurement under these circumstances is to maximize attribute variance while minimizing both bias and error variance. To achieve this goal, the item pool has to be made more heterogeneous, the only requirement being that all items index the targeted construct to some degree. Item heterogeneity will serve to offset the relative prominence that any one component of bias can attain on the measure, at the same time that the items are adding, little by little, to the proportion of common attribute variance captured by the measuring instrument. This is the psychometric property of *aggregation* discussed by Green (1978) in his classic paper, "In Defense of Measurement." The idea is to have as many different types of items as possible, with each type possibly carrying with it a different kind of bias, so that, individually, each source of bias will tend to contribute minimally to the overall variance of the aggregate. This methodology is therefore designed to build up the communality of attribute variance running through the several items, while simultaneously minimizing any specific bias associated with any given type of item content or item format. The point is, *aggregation depends on covariation*. Without covariation, the benefits of aggregation will not be realized (see Roznowski, 1987; Roznowski & Hanisch, 1990; Rushton, Brainerd, & Pressley, 1983).

This approach to measurement recognizes that all forms of multi-item assessment will carry components of method variance bias, inextricably combined with the construct-valid variance of the attribute of interest. The objective is to minimize this unwanted bias variance. To use a psychophysical analogy, the goal is to amplify signal (attribute variance) and attenuate noise (bias variance). What we are discussing here at the item level is completely equivalent to our earlier discussion of aggregation at the group factor level illustrated in Figures 6 and 7.

Carroll (1985) has recently developed a similar procedure for building measures of factors:

> How can one develop or select two or more variables to reflect a factor without making them identical? It would be undesirable to make them identical or equivalent because the underlying common factor might then reflect also a specific factor in the variables, and the factor would be overdetermined. The solution is to include task variation, over the several variables, that is expected to be irrelevant to the definition of the ability.... Irrelevant task variation can be introduced by varying required knowledge bases. (p. 30)

For our earlier example (see Table 3), and given these considerations, it follows that a more optimal measure of verbal ability (Vernon's minor group factor *v*) would be obtained by aggregating *reading comprehension* and *vocabulary* than by using either measure alone, and an even more optimal measure would be obtained if the information test *literature* were combined with these two.

If the general factor captures much of the common variance and absorbs the variance due to major or minor group factors (constructed using the foregoing methodology), that should be no cause for concern. If the specific variance of a lower order factor is meaningful (i.e., has external correlates beyond the general factor), incremental validity beyond the general factor will reveal that fact. This methodology brings out the content-specific variance of lower order group factors without attempting to control for complexity.

There is another reason for considering major and minor group factors even when most of the common variance is captured by the general factor in samples with the full range of talent. In more select samples, the range of individual differences due to the general factor decreases and the demand on more concentrated or specialized cognitive abilities increases, thereby increasing the likelihood that measures with more focused content will account for more of the criterion variance. For example, philosophers and engineers may possess comparable levels of the general factor (i.e., their level of complexity on the radex is roughly equivalent), but the density of verbal skills relative to spatial skills will be greater for philosophers, and the reverse would be true for the engineers (i.e., their level of sophistication with contrasting symbolic content differs). If measures of verbal and spatial ability were correlated for a sample combining philosophers and engineers, the correlation will most likely be negative. Job performance for either group would not be predicted too well from a general factor measure because the level of complexity required is about the same for both occupations. Rather, because each occupation requires facility with different content, assessment of the cognitive ability repertoire beyond the general factor will be required for differential prediction. The major group factors *v:ed* and *k:m* are likely to be better predictors than the general factor, and perhaps even the minor group factors (e.g., *v*, *n*, and *s*) would contribute. These more circumscribed dimensions cannot compete with the general factor in normative samples, but in more select samples they may account for *more* criterion variance than the general factor. Herein may lie the solution to longstanding concerns about the significance of intraindividual differences in conventional multiability profiles. Profile scatter (dispersion) may be highly significant in certain contexts. However, these intraindividual differences, reflecting contrasting degrees of sophistication within different regions of the radex, can only be evaluated with precision after removing the communality these dimensions share, that is, the general factor.

Although the general factor is the most prominent dimension in the radex (absorbing about half the common variance of cognitive ability tests), there are other important lower order factors as well. Just as validity coefficients of minuscule size can be useful when the selection ratio is stringent and the base rate is low (Taylor & Russell, 1939), there are circumstances when lower order factors can be quite useful. Measures that are highly correlated in samples with the full range of talent will often pull apart (*dissociate*) in samples that are more select. In this circumstance, there is opportunity for lower order factors to become effective predictors, hence the scientific significance of such factors can be thought to be conditional, or population specific. In the psychological study of occupations, especially the higher level professions, an important first step might be to ascertain deviant attributes. If restriction of range is observed on any of these deviant attributes, a more focused assessment of lower order factors would be warranted.

The Criterion Problem

Assessing the construct validity of ability measures is inextricably intertwined with assessing the construct validity of relevant

criteria (Dunnette & Borman, 1979; James, 1973; Kavanagh, MacKinney, & Wolins, 1971; Smith, 1976; Tenopyr, 1977). Given this, before we can discuss the *prediction problem*—how to optimize prediction from abilities, some discussion of the criterion problem seems necessary. Indeed, few problems are more significant to industrial and organizational psychologists (Dunnette, 1963; Smith, 1976).

Ratings are the most used form of performance evaluation; hence, methods that enhance their quality are of great import. An important theory of ratings has reemerged that shares many elements in common with our previous discussion about aggregation. Landy and Farr (1980, pp. 98–99) called attention to this theory in their *Psychological Bulletin* article. The formulation was introduced by Wherry (1952) when he was working on problems of performance assessment in the military. Until recently, his treatment remained in esoteric technical reports (although Wherry taught this formulation to many students in the 1950s and 1960s at Ohio State University). More recently, Landy and Farr (1983) published a small chapter by Wherry (1983) as an appendix in *The Measurement of Work Performance: Methods, Theory, and Application*. Readers are referred to this source for a more detailed exposition (see also Wherry & Bartlett, 1982).

Wherry (1983) partitions the variance in ratings into systematic and random components, and then further partitions the systematic into *true* and *bias* aspect components. This is done in a way very much in line with our earlier discussion of the decomposition of scale variance (i.e., attribute variance, bias variance, and error):

$$v_{\text{rating}} = v_{\text{true aspect}} + v_{\text{bias aspect}} + v_{\text{random error}}$$

Wherry's (1983) analysis also partitions the components of bias further and includes ways to minimize the influence of distinct forms of bias, the various nuisance variables that attenuate the predictor-criterion correlation. This analysis underscores the need to be wary about the components of reliability (e.g., during reliability checks on the raters). Rater reliability, like test reliability, can be enhanced by increasing true variance *or* bias variance or both. If it is bias variance that enhances reliability, we have the same problem that is encountered in test construction, namely, that what we are measuring more reliably is actually what we are less interested in. This perspective differs somewhat from that expressed by Schmidt, Hunter, and Pearlman (1981), who maintain that, because correlations between criterion measures typically approach 1.00 when such correlations are corrected for attenuation, such measures are essentially equivalent at the true score level. They go on to say, "These considerations point to the conclusion that the only function of multiple criterion scales is to increase reliability of the composite (overall) criterion measure" (p. 175). But optimal aggregation increases both reliability and validity.

The link between Wherry's theory of ratings and the aggregation of indicators of indifference (on the predictor side) is clear. Wherry (1983) even states that this "parallels Spearman's two-factor theory of intelligence. Obviously, the true expression should probably contain group or common factors as well as a general one, but for convenience we should restrict our discussion to the simpler case" (p. 288). We include this discussion of Wherry's theory because investigators may wish to consider aggregating criterion ratings in this manner so that systematic bias can be minimized. Aggregation is a powerful technique that should be used on both sides of the predictor-criterion equation. (It may be more than coincidence that Wherry, 1959, also developed one of the early hierarchical solutions in factor analysis.)

Systematic bias contaminates the validity of all psychometric measures. It cannot be removed completely in any assessment, but it can be minimized. When constructing *cross-validation profiles* with multiple criteria (see below), some criterion measures may collectively converge more accurately on a criterion dimension when they are aggregated. Furthermore, correlations with aggregated criterion measures as well as with the individual constituents of these composities can be included in cross-validation profiles, providing the investigators with the opportunity to have their cake and eat it, too. Such cross-validation profiles would allow research consumers to decide for themselves which criteria are most important for their purposes, how these criteria should be combined, and what the optimal predictor set is for their chosen criteria (Schmidt & Kaplan, 1971).

Tenopyr (1977) has argued that the process of construct validation involves establishing structural relations between aggregates of heterogeneous predictors and criteria that share common variance:

> We know from elementary factor theory that the correlation between any two variables equals the sum of the cross-products of their common factor loadings. Putting it more simply, to have high predictive value, a test must essentially involve the same constructs to the same degree as a measure of the job behavior. It would seem, then, that any interpretation of a content-based employment test strictly in terms of tasks is inadequate. A content-based test or any other test used in prediction must share common constructs with job behavior. (p. 50)

To the extent that large communalities are shared by both the predictors and the criteria, or by subgroupings of each, global models of structural relations are likely to serve applied and theoretical purposes best. For example, for a wide range of occupations that demand a lot of information processing, the *overall level* of the cognitive repertoire is what is important. To some degree, different forms of symbolic content can "compensate" for each other. (Perhaps most cognitively demanding occupations are more like this than not.) In psychology, for instance, if GRE-V provides as much information about the likelihood of completing graduate training as GRE-Q does (across most areas of psychology), the two measures can be combined. In this situation, disparate patterns that aggregate to the same level will reflect near-equivalent readiness to acquire the required repertoires of psychological knowledge. However, in other domains, the density of specific kinds of symbolic-manipulation ability may be more central. Under these circumstances, it is important to assess not only the complexity (level) of the cognitive repertoire, but also the content (density) of its particulars. For a number of academic and occupational environments, the same level of the general factor emanating from different forms of symbolic content does not make for the same ability requirements. Thus, in English literature and law, facility with linguistic material is much more important than facility with numerical or figural symbols, whereas the converse is true in architecture, engineering, and the physical sciences generally. The implication here for the criterion problem is that criterion measures should assess one's facility with the actual specific content and products of the domain.

In an article on the *real test bias*, Frederiksen (1984) makes a compelling case that tests *are* biased if they do not assess important facets of criterion behavior necessary for the required performance. This would imply that for figurally or spatially demanding occupations (such as architecture, engineering, and many of the creative arts), the appropriate criteria should include created products. Below, we will suggest that measures of spatial ability (*S*) would undoubtedly contribute incremental validity to measures of *G* (*general factor*), *V*, and *N*. This hypothesis is best tested with criteria in which variance overlaps with the *specificity* of spatial ability measures (good criterion examples include building, drawing, and designing).

Perhaps expert raters could independently assess the products of these more hands-on, less verbal, disciplines. For example, if raters

displayed an average interrater correlation of .45, an aggregate based on 5, raters would generate an estimated reliability (via Spearman-Brown) of .80. For ability measures distinct from the general factor to have incremental validity, they must share variance with criteria that are not loaded with the general factor. Such ratings may provide a needed vehicle to pull contruct-valid variance from the paper-and-pencil-loaded general factor. This again reflects Tenopyr's idea of establishing construct validity by linking predictor and criterion communalities, but the communalities in this case must be independent of the general factor. Criterion *attribute variance* that is independent of the general factor is the portion that will give scientific significance to measures of cognitive abilities beyond *g*. Ackerman (1987) provides examples of how perceptual and motor abilities factor into performance and how these attributes fit into dynamic behavioral changes throughout the skill acquisition process. But these abilities and processes will go undetected if the criterion variables used are not sensitive to systematic sources of individual differences beyond the general factor.

The Prediction Problem and Determining a Predictor Set: Combining Cross-validation With Extrinsic Convergent Validation

The preceding discussion suggests that it might be useful, when comparing competing predictors, to construct *cross-validation profiles* much like the extrinsic convergent validation profiles shown in Table 3. This would allow investigators to examine the comparative validity of contrasting predictor sets across several criteria. It is possible for different predictors to produce different validities for different criteria; therefore, comparison across multiple criteria is of the essence when evaluating competing predictors.

The comparison of several predictor sets in terms of multiple criteria is illustrated in Table 4, which contains a synthetic example. The five predictor sets being compared on *n* criteria include a measure of *g* and four combinations of major (*v:ed, k:m*) and minor (*V, N* & *S*) group factors. The question is whether the multiple predictors are getting at no more than the criterion variance captured by the general factor, or whether there are profitable components of specific variance that warrant the use of group factors in the prediction equation. In the illustration, only the major group factor combination (*v:ed, k:m*) consistently improves on the predictive efficiency of *g*, although for specific criteria, other combinations may improve on *g*.

Consider also our earlier synthetic example in Figures 6 and 7. The first equation would be a regression equation with one predictor, the three-variable composite. The fifth equation would involve three predictors, the three variable variate, say, the group factors, verbal (*V*), numerical (*N*), and spatial (*S*) ability. If the valid portion of the predicted variance acounted for by the three predictors is due solely to their communality and not to their specificity, on cross-validation the R^2 would shrink more for the three-predictor variate than for the composite, because having more predictors increases the likelihood of capitalizing on sample-specific noise (which is what cross-validation corrects for).

Consider, on the other hand, the one-predictor equation. Most of the variable's reliable variance is *g* variance. Therefore, the beta weight will be based mostly on the common variance (the *g* variance, found in all cognitive ability tests). The shrinkage in R^2 will be much less because only one parameter is being estimated, and it contains all of the relevant attribute variance. Thus, to the extent that criteria are saturated with *g*, cross-validation coefficients should be essentially equivalent for the single-predictor equation and the multiple predictor equation if the latter is composed of the constituents of the former.

TABLE 4

A Synthetic Example of Five Cross-validation Profiles for Predicting Multiple Criteria With Contrasting Predictor Scales

	Predictor Sets				
Criterion Variables	*g* R^2_{cv}	*v:ed, k:m* R^2_{cv}	*k:m, N* R^2_{cv}	*V, N, k:m* R^2_{cv}	*V, N, S* R^2_{cv}
C_1	.39	.43	.38	.37	.35
C_2	.34	.36	.34	.33	.31
C_3	.31	.34	.30	.29	.29
C_4	.30	.32	.30	.31	.28
C_5	.26	.35	.27	.30	.38
C_n	.31	.34	.34	.39	.30

Note: For each criterion variable, the beta weights for the predictor variable(s) will vary as a function of the least squares maximization procedure on the screening sample.
g = general intelligence
v:ed = verbal/numerical/educational
k:m = practical/mechanical/spatial
N = numerical ability
S = spatial ability
V = verbal ability

This analysis shows why multiple aptitude test batteries should be evaluated not only in terms of their incremental validity (over that achieved by measures of general intelligence) but also in terms of their cross-validation. Predictive validities can be spuriously inflated simply because adding variables can increase the R^2 significantly in large samples. It also distorts the interpretations we place on predictor constructs. (Statistically significant findings should *always* be evaluated for substantive significance.)

However, as we noted earlier, there is good reason to believe that in many applied contexts the construct of general intelligence may be splintered profitably. Vernon's major group factors, *verbal-numerical-educational* (*v:ed*) and *practical-mechanical-spatial* (*k:m*), collectively provide incremental validity over general intelligence in a number of contexts, as do the minor group factors, verbal, numerical, and spatial ability. The likelihood of such differential validities occurring is moderated by the intellectual level required of the occupation. Range truncation on the general factor will increase the likelihood that the major and minor group factors will be useful for prediction.

We offer a simple methodology for ascertaining when further splintering of general intelligence is warranted. Start with an ordinary univariate regression equation involving general intelligence to serve as the baseline. Then proceed to more complex variates, first, by splintering general intelligence into *v:ed* and

k:n, and then by splintering these major group factors, following a logical hierarchical descent from larger to more circumscribed factors. For example, *v:ed* may be splintered into verbal (*V*) and numerical (*N*) ability; *k:m* could be evaluated as a third variable, or splintered, and the process can continue.

The criterion for splintering is twofold. First, more complex predictor variates (i.e., multiple variable variates) must display incremental validity over and above that shown by less complex predictor variates (viz., composites). Second, on cross-validation, the more complex variates must account for more criterion variance than the less complex variates.

For cognitively demanding occupations, as in our earlier example of philosophers and engineers, the general factor (general intelligence) is expected to "fall out" of the predictor variate while lower order factors (more content-focused abilities) surface and become more predictively central.

Investigators must also take into account sample statistics that might vary from study to study, such as reliabilities and standard deviations of both predictors and criteria, level of the general factor for screening (development) and calibration (cross-validation) samples, and sample size. With what we now know about the roles such basic statistics play in affecting validity coefficients (see below), *it behooves all investigators to report these statistics routinely.*

The above methodology also speaks to Cronbach's (1971) point that all tests have *multiple* validities. Validity is based on the accuracy of inferences. It is not tests that are validated; inferences are. As cross-validation profiles are developed across different populations and occupational contexts, the construct validity of individual predictors and predictor sets will be better understood.

The importance of cross-validating predictor composites of a few variables selected from a larger pool (such as cognitive abilities) has recently been restressed in *Standards for Educational and Psychological Measurement* (APA, AERA, & NCME, 1985, Standard 1.25):

> When a small number of predictors is selected from a large pool and weights are simultaneously determined...selecting variables and weights and for estimating validity coefficients should take into account the bias in the weights and validity coefficients; otherwise the weights and validity coefficients should be cross-validated. (p. 18)

It is hoped that the procedure described above will forestall the possibility of ostensibly different measures appearing to show that different constructs are related to performance when, in reality, it is the common variance of more global measures and not the specificity of their constituents that is responsible for the meaningful predictor/criterion covariation that holds up under cross-validation.

Ability Domains: Cognitive, Perceptual, and Psychomotor

Human abilities are typically grouped into three domains: cognitive, perceptual, and psychomotor. But, of course, these are not discrete categories. All three categories provide unique and valuable information for predicting criterion behaviors, and dimensions from all three categories have met the criteria for scientific significance. Recent reviews have summarized the current status of the three domains; see, for example, Fleishman and Quaintance (1984), Hartigan and Wigdor (1989), Gottfredson (1986), Landy (1990), and Wigdor and Garner (1982). An issue of *Personnel Psychology* (Sackett, 1990) is devoted to the Army's Selection and Classification Project (Project A) and includes discussions of predictor and criterion development for both ability and nonability domains.

The above sources provide a broad review of contemporary validation research with the use of ability predictors. Many of the studies discussed draw on existing, well-known multitest batteries such as the *General Aptitude Test Battery* (GATB), the *Differential Aptitude Tests* (DAT), and the *Armed Services Vocational Aptitude Battery* (ASVAB).[5] These are excellent instruments, and the data they provide are useful. We hope that future research will expand coverage to other measures as well as to ways of aggregating existing measures that will enhance the validities of predictor composites and simplify the predictor variates.

Meaningful analyses can also be achieved with existing large-scale data sets. For example, Table 5 contains data from Hunter's (1983b) analysis of GATB data for 12,000 jobs. The GATB has eight scales and one general-factor composite scale. Hunter aggregated these scales to form three composites: cognitive, perceptual, and psychomotor, assembled from the GATB scales as follows: cognitive composite = *G* (general ability) + *V* (verbal ability) + *N* (numerical ability); perceptual composite = *S* (spatial ability) + *P* (form perception) + *Q* (clerical perception); and psychomotor composite = *K* (motor coordination) + *F* (finger dexterity) + *M* (manual dexterity). These composites were then validated as predictors of job performance for five job families (groupings of jobs based on their level of complexity). Table 5 summarizes their findings. Table 5 shows that, as job complexity increases, the regression weights for the cognitive composite increase, whereas those for the psychomotor composite decrease.

TABLE 5

Predicting Job Performance

Job Family	Complexity Levels	Regression Equation		Multiple Correlation	Number of Jobs
I	1 Set-up	JP = .40 GVN + .19 SPQ	+ .07 KFM	.59	21
III	2 Synthesize/coordinate	JP = .58 GVN		.58	60
IV	3 Analyze/compile/compute	JP = .45 GVN	+ .16 KFM	.53	205
V	4 Copy/compare	JP = .28 GVN	+ .33 KFM	.50	209
II	5 Feeding/offbearing	JP = .07 GVN	+ .46 KFM	.49	20

Note: These job families were based on the data-people-things classification system; however, numerical order does not reflect complexity.
Family I: Set-up precision work (e.g., machinist, cabinet maker, metal fabricator)
Family II: Feeding, offbearing (e.g., shrimp picker, corn husking machine operator, cannery worker, spot welder)
Family III: Synthesize, coordinate (e.g., retail food manager, fish and game warden, biologist, city circulation manager)
Family IV: Analyze, compile, compute (e.g., automotive mechanic, radiologic technician, automotive parts counterman, high school teacher)
Family V: Copy, compare (e.g., assembler, insulating machine operator, forklift truck operator)
GVN = Cognitive ability
SPQ = Perceptual ability
KFM = Psychomotor ability

From "Test Validation for 12,000 Jobs: An Application of Job Classification and Validity Generalization Analysis to the GATB" (Test Research Rep. No. 45) by J. E. Hunter, 1983, Washington, DC: U.S. Department of Labor, U.S. Employment Service.

Note that the perceptual composite contributed to the prediction in only one equation. Because the perceptual composite could be predicted almost perfectly from the other two composites, some authors have considered using only two composites, cognitive and psychomotor, to predict job performance. This may not be wise in certain circumstances. Our concern is this: There is reason to believe that spatial ability involves more than what the GATB's spatial measure, *S*, captures. The GATB does not measure all regions of the radex uniformly. Spatial ability also includes mechanical reasoning, which is not assessed by the GATB. There are many tests of spatial ability: Eliot and Smith (1983) have compiled a directory of almost 400 spatial tests. Lohman (1989) has stated that spatial abilities can be organized hierarchically, but so far, results of factor analytic studies focusing exclusively on this domain are less than clearcut. Nevertheless, if these factor analytic results were used to select tests that were then aggregated, we suspect that a spatial ability aggregate would contribute incremental validity in certain prediction contexts (e.g., scholastic prediction in engineering, architecture, physical sciences, and the creative arts). Although Hunter's (1983b) five-family categorization has its uses, narrower groupings might prove even better for prediction purposes.

Other studies have shown that spatial ability is a useful predictor. In Project A, for example, McHenry, Hough, Toquam, Hanson, and Ashworth (1990) found that a spatial test provided more incremental validity than any other cognitive ability beyond the general factor in the prediction of general soldiering proficiency and core technical proficiency. For these occupations, spatial ability added two to three percent to the predicted variance. Humphreys, Lubinski, and Yao (under review), using Project TALENT data, assembled two composites from three group factors: mathematical (*M*), spatial (*S*), and verbal (*V*) abilities. One composite (*V*-Math) consisted of *V* + *M* (essentially Vernon's *v:ed*), whereas the other composite (*S*-Math) was formed by combining *S* + *M* (so *M* was a component in both composites). Three groups of subjects were then selected using these composites: *high intelligence* (top 20% on both composites), *high spatial* (top 20% on the *S*-Math only), and *high verbal* (top 20% on *V*-Math only). Using longitudinal data obtained 11 years after high school graduation, these investigators found that for both genders, twice as many physical scientists were found in the high spatial group as in the high verbal group. A disproportionate number of creative artists came from the high spatial group as well. The authors suggest that a number of students who are academically able with respect to disciplines like architecture, engineering, and many of the physical sciences might not pass conventional screening for academic admission because such screening tends to be limited to the use of verbal and numerical/quantitative (i.e., *v:ed*) measures. Other studies have shown that spatial ability contributes incremental validity over a wide range of jobs (Humphreys, 1986; Smith, 1964; Vernon, 1961).

We suggest that investigators pursue spatial ability as a separate construct. We suspect that spatial ability draws on several underlying psychological systems. For instance, based on the correlates of many spatial ability measures, it is reasonable to hypothesize that spatial ability draws on both the cognitive and perceptual systems. This would mean that aggregating GATB spatial (*S*), form perception (*P*), and clerical perception (*Q*) may act to "pull" the construct validity of the spatial ability measure away from its cognitive components. Hunter (1983b) notes that spatial ability is more valid than the perceptual composite (*S*, *P*, *Q*) for artistic and scientific jobs (jobs classified at a more specific level than the five more general job families shown in Table 5), but concludes that this advantage is trivial. We think that this minuscule advantage can become more substantive if more broadly defined spatial measures were used.

Table 5 also depicts a well-established finding: As the information processing demands of jobs decrease, the validity of psychomotor composites increases (Hartigan & Wigdor, 1989; Hunter, 1983a, 1983b; Thorndike, 1985). But nonetheless, psychomotor abilities appear together with cognitive abilities as significant predictors for conceptually demanding jobs. Guion (1965) distinguished between *psychomotor* and *sensorimotor* abilities. The former stresses kinetic movement and control, whereas the latter requires some form of sensory discrimination to structure subsequent muscular movement. This distinction reflects again how abilities and skills resist classification into discrete categories. Readers are referred to Fleishman's work on motor ability dimensions (e.g., Fleishman, 1966; Fleishman & Quaintance, 1984). Dunnette's 1976 chapter in this *Handbook*'s first edition provides brief descriptions of these dimensions as well as dimensions of physical fitness. (See also Hogan's [1991] chapter on physical abilities in volume 2 of this *Handbook*.)

Psychomotor and sensorimotor abilities will continue to receive justifiable attention in future research on human abilities. Thorndike (1985), for example, aggregated the psychomotor abilities of the GATB, examined it as a predictor across a heterogeneous collection of job categories, and concluded that

> the motor dimension appears to be significant for job performance, and since it is relatively independent of the cognitive dimension ... it deserves to be given weight in any prediction.... It is likely that 10 or 15 percent increase in predicted criterion variance can be achieved. (p. 253)

Perhaps "any prediction" might be too sweeping, but these remarks do stress the importance of this class of abilities.

With respect to cognitive abilities, a number of contemporary investigators have commented on the central importance of general intelligence (i.e., the general cognitive factor) in the prediction of important criteria (Brand, 1987; Carroll, 1982; Gottfredson, 1986; Humphreys, 1979; Hunter, 1986; Jensen, 1980a; Schmidt, 1988; Thorndike, 1985). The *Journal of Vocational Behavior* devoted an entire issue to this topic in 1986 (Gottfredson, vol. 29, pp. 293–450). To be sure, the primacy of general intelligence in the prediction of academic and vocational criteria has always had its proponents (Berdie, Layton, Hagenah, & Swanson, 1962; Humphreys, 1962; Jenkins & Paterson, 1961; McNemar, 1964; Vernon, 1947, 1950). However, it is also well documented that in predicting these criteria it is possible to account for more criterion variance by using components (major and minor factors) of general intelligence instead of the usual full-scale score. In addition, perceptual and motor abilities can contribute to the prediction as well.

Given these studies of long standing, the current stress on the importance of the general factor would appear to be retrogressive. Why did Ackerman (1988b) remark, in his review of Gottfredson's (1986) "The *g* Factor in Employment Testing," that the special issue might have well been labeled, "Found: Our Intelligence. Why?"

The return of interest to general intelligence (or more precisely, the general factor in cognitive abilities) appears to be due to how well the construct has done by itself, so as to warrant a special status among predictors of criteria in the academic world and the world of work. Carroll (1982), for example, has noted that "a large part of whatever predictive validity the DAT and other multiple aptitude batteries have is attributable to an underlying general factor that enters into the various subtests" (pp. 83–84). Although incremental validity can be achieved by fractionating general intelligence, the amount of increase (even if valid and worthwhile) is minuscule compared with

what general intelligence does by itself (Gottfredson, 1986, vol. 29, no. 3, whole issue) *for the criterion variables it has been used to predict*. Table 6 gives recent data from Project A that show once again how significant a predictor the general cognitive factor is. Table 6 also shows the incremental validities for other predictors, some of these being nonability variables. However, the general factor accounts for the lion's share of criterion variance of all kinds. Other research programs have observed similar results.

Reviewing 20 years of research on aptitude-treatment interaction (ATI), Cronbach and Snow (1977) concluded:

> While we see merit in a hierarchical conception of abilities, with abilities differentiated at coarse and fine levels, we have not found Guilford's subdivision a powerful hypothesis.... Instead of finding general abilities irrelevant to school learning, we found nearly ubiquitous evidence that general measures predict amount learned or rate of learning or both. And, whereas we had expected specialized abilities rather than general abilities to account for interactions, the abilities that most frequently enter into interactions are general. Even in those programs of research that started with specialized ability measures and found interactions with treatment, the data seem to warrant attributing most effects to general ability. (pp. 496–497)

More recently, Snow (1989) made even stronger statements:

> In contrast to earlier predictions of Cronbach and Gleser, measures of general ability (G) enter interactions more frequently than other indices of aptitude, despite the fact that measures of G also typically show strong aptitude main effects. Many different measures have been used to reach this conclusion.... Given new evidence and reconsideration of old evidence, G can indeed be interpreted as 'ability to learn' as long as it is clear that these terms refer to complex processes and skills and that a somewhat different mix of these constituents may be required in different learning tasks and settings. The old view that mental tests and learning tasks measure distinctly different abilities should be discarded, even though we still lack a theory for integrating the two. (p. 22)

General ability, the general cognitive factor (or the complexity dimension of the radex), is indeed one of psychology's most powerful variables. The significance of this psychological parameter may stem in part from many of its correlates observed outside of academic or vocational settings, correlates that clearly contribute to instrumental effectiveness in these settings. Brand (1987) and Jensen (1980a, chap. 8) provide one with a good feel for these correlates. The nomothetic span of general intelligence is indeed among the broadest and deepest of any of psychology's constructs. The importance of this network for industrial and organizational psychologists extends beyond performance and training criteria. Many of the positive correlates have implications (corollary as well as direct) for academic or workplace behaviors—for example, breadth and depth of interests, physical health, responsiveness to psychotherapy, moral reasoning capacity, sense of humor, social skills, achievement motivation, and practical knowledge. Many of the negative correlates also have implications—accident-proneness, alcoholism, crime and delinquency, racial prejudice, smoking, and obesity, for example. (For additional correlates, the interested reader is referred to the table assembled by Brand, 1987, pp. 254–255.)[9] It would appear that a large number of socially valued attributes are "pulled along" by the general factor to such an extent that when we analyze the determinants of, say, supervisor or peer ratings or work performance in general, we have to be quite aware of the scope of this network.[10]

TABLE 6

MeanValidity and Incremental Validity for Multiple Sets of Predictors and Criteria

	Predictor Domain					
Job Performance Factor	*General Cognitive Ability (K = 4)*	*General Cognitive Ability Plus Spatial Ability (K = 5)*	*General Cognitive Ability Plus Perceptual Psychomotor Ability (K = 10)*	*General Cognitive Ability Plus Temperament/ Personality (K = 8)*	*General Cognitive Ability Plus Vocational Interest (K = 10)*	*General Cognitive Ability Plus Job Reward Preference (K = 7)*
Core technical proficiency	.63	.65	.64	.63	.64	.63
General soldiering	.65	.68	.67	.66	.66	.66
Effort and leadership	.31	.32	.32	.42	.35	.33
Personal discipline	.16	.17	.17	.35	.19	.19
Physical fitness and military bearing	.20	.22	.22	.41	.24	.22

Notes: Validity coefficients were corrected for range restriction and adjusted for shrinkage.
Incremental validity refers to the increase in *R* afforded by the new predictors above and beyond the *R* for the Army's current predictor battery, the ASVAB.
K is the number of predictor scores.

From "Project A Validation Results: The Relationship Between Predictor and Criterion Domains" by J. J. McHenry, L. M. Hough, J. L. Toquam, M. A. Hanson, and S. A. Ashworth, 1990, *Personnel Psychology, 43*, p. 346. Copyright 1990 by *Personnel Psychology*. Reprinted by permission.

Validity Generalization

Validity generalization is an important concept, not only for industrial and organizational psychology (where it was developed), but for psychology in general. Even sympathetic observers have remarked that psychology as a science lacks the systematic cumulation of knowledge so characteristic of the physical sciences (Meehl, 1978). Validity generalization (VG) is an exception. VG studies employ meta-analytic techniques to ascertain the generalizability and stability of forecasting equations based on individual differences in

abilities (e.g., those found in Table 5). These meta-analytic reviews have shown that validities are transportable across job situations and within job families.

Annual Review chapters on personnel selection routinely include a section on this topic (Dunnette & Borman, 1979; Guion & Gibson, 1988; Hakel, 1986; Schmitt & Robertson, 1990; Tenopyr & Oeltjen, 1982; Zedeck & Cascio, 1984). Validity generalization is receiving as much attention as any topic in contemporary industrial and organizational psychology. A panel debate, "Forty Questions About Validity Generalization and Meta-analysis" (Schmidt, Pearlman, Hunter, & Hirsh, 1985) and "Commentary" (Sackett, Tenopyr, Schmitt, & Kahn, 1985, vol. 38), covered over 100 pages in *Personnel Psychology*, an exchange characterized by Cronbach (1988, p. 12) as "exhaustive." A more recent review is given in Hartigan and Wigdor (1989, especially chap. 7).

The general finding was that although the magnitude of the validity coefficients did vary across jobs that differed in complexity, cognitive abilities have shown at least some validity for all jobs. Monotonic increases in job complexity accompany increases in validity coefficients for broad groups of cognitive abilities and, conversely, decreases in validity coefficients for psychomotor abilities. Validity coefficients were often appropriately transportable across broad groupings of jobs and different organizational situations.[11] We know this now—these conclusions can be found in major textbooks in industrial and organizational psychology (e.g., Cascio, 1991; Landy, 1989; Muchinsky, 1989)—but we did not always know it (cf. Albright, Glennon, & Smith, 1963, p.18; Ghiselli, 1966, p. 28).

One of the most misleading myths in the history of industrial psychology was that validity coefficients and prediction equations for job performance were situation specific. This sentiment was captured by Guion (1965):

> The first and most persuasive generalization that can be made is that jobs within...various organizational groupings, as well as the organizational climates in which they may be found, will demonstrate extensive variability. A test or procedure that may be found highly predictive in one situation may, therefore, prove to be of no value at all in another apparently similar one. (p. 415)

Contrast the above with the following recommendations from the latest edition of *Standards for Educational and Psychological Measurement* (APA, AERA, & NCME, 1985):

> When adequate local validation evidence is not available, criterion-related evidence of validity for a specified test use may be based on validity generalization from a set of prior studies, provided that the specified test-use situation can be considered to have been drawn from the same population of situations on which validity generalization was conducted. (Primary)...Comment: Several methods of validity generalization and simultaneous estimation have proven useful. In all methods, the integrity of the inference depends on the degree of similarity between the local situation and the prior set of situations. Present and prior situations can be judged to be similar, for example, according to factors such as the characteristics of the people and job functions involved. (Standard 1.16, pp. 16–17)

and later:

> Employers should not be precluded from using a test if it can be demonstrated that the test has generated a significant record of validity in similar job settings and for highly similar people or that it is otherwise appropriate to generalize from other applications. (p. 59)

Finally, a recent publication by the National Academy of Sciences includes an entire chapter on validity generalization and supports its assumptions and methods (Hartigan & Wigdor, 1989).

Quite a change! Why?

A number of factors may have contributed to the myth of *situational specificity*, among them, small sample sizes characteristic of most validation studies, range restriction in the applicant and employee samples, and unreliability in the predictors and criteria. As Pearlman et al. (1980) insightfully observe: "The visions of complexity entertained by many psychologists in the field may stem largely from a tendency to interpret variance created by statistical artifacts as real" (p. 400).

Validity generalization research involves the estimation of two parameters: *true validity* and the *standard deviation of true validity*. Search for the best statistical estimation procedures has resulted in an extensive literature. Linn and Dunbar (1986) provide an excellent review of these procedures, finding that all of them produce estimates that are very close. If anything, the procedures tend to overestimate the true standard deviation. Schmidt and Hunter (1981, p. 1132) have listed a number of factors that contribute to variability in validity coefficients across situations:

- Sampling error
- Criterion unreliability
- Test (predictor) unreliability
- Range restriction
- Criterion contamination or deficiency (Brogden & Taylor, 1950)
- Computational and typographical errors (Wolins, 1962)
- Variation in the equivalence of measures employed

According to Schmidt and Hunter (1981), approximately 72 percent of the variance in observed validity coefficients is traceable to the first four sources of disturbance, and 85 percent of this 72 percent is due to the first source, sampling error.

To be sure, that validity generalization occurs across jobs and organizations does not preclude situational specificity under certain circumstances. Also, that validities are comparable does not imply that tests are unbiased in the sense of generating equally accurate inferences for different demographic groups. Systematic trends of over- or underprediction can occur. For example, underprediction for women in academia (Linn, 1973) and in the military (Linn, 1984) is discussed by Linn and Dunbar (1986).

Actually, validity generalization is only one manifestation of construct validity (Hunter, 1980). As Loevinger (1957) has perceptively noted some time ago: "Since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636). That is, construct validity is what enables scientists to extrapolate or infer from an organized body of data—knowledge—in a manner that increases the likelihood their inferences will be confirmed. Construct validity also embraces the concept of generalizability that Cronbach and his colleagues (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963) have studied extensively. As comparable results accrue with the use of different predictor or criterion measures targeted at the same constructs, the generalizability of measures and boundary conditions of constructs become more clearly defined. To know how a construct operates (which is the goal of construct validation research), we must first know the domains in which it operates.

This discussion may also be linked to Lykken's (1968) three-tiered concept of replication—constructive, operational, and literal (in order of greater to lesser degree of risk and scientific soundness). The many replications

of the general factor and the finding of similar predictor-criterion structural relations with the use of different measures constitute examples of constructive replication. This also captures the spirit of synthetic validity.

The concept of validity generalization continues to undergo refinement of the "normal science" kind (Kuhn, 1970). Both theoretical and applied psychologists will profit from assimilating the findings of validity generalization. The distributions of validity coefficients assembled in the validity generalization studies are rather impressive and may cause some to wonder if the fluctuations in the phenomena they are trying to understand (in other contexts) are not simply due to sampling error. For the applied areas, validity generalization may soon acquire the ascendancy that actuarial prediction has over clinical prediction (Meehl, 1954). Perhaps validity generalization will be assimilated by the field more quickly (Meehl, 1986b).

Predictor-Criterion Structural Relations

In applied psychology, the linear model is ubiquitous (Dawes, 1979). The preceding discussion was based entirely on linear relationships between predictor and criterion. People have tried repeatedly to transform existing predictors or look for systematic trait × trait (Ghiselli, 1972; Zedeck, 1971) or trait × treatment (Cronbach, 1957) interactions, but these more complex second-order relationships tended to fail to replicate (Cronbach, 1975a), even in fields outside of industrial and organizational psychology (Tellegen, 1988; Tellegen, Kamp, & Watson, 1982; Wiggins, 1973). In a special issue on moderator variables in the *Journal of Applied Psychology*, published more than 15 years after Saunders' (1956) "citation classic" on this concept, Ghiselli (1972) wrote, "It is possible that moderators are as fragile and elusive as that other will-o-the-wisp, the suppressor variable" (p. 270). The status of moderator variables has not changed, but they continue to be discussed at length in classrooms and seminars without any convincing examples. We continue to tell our students, "Keep looking. It's an attractive concept."

Perhaps it is not outrageous to suggest that to the extent that ability predictor-criterion variables covary, they "simply" do so linearly. At the level of analysis of conventional psychometric abilities, it may be that monotonic transformations (e.g., X^2 or $X^{1.2}$) as well as predictor-predictor interactions (X_1X_2) do not covary with meaningful criteria more precisely than linear trends (particularly in ways that replicate and cross-validate). This was essentially the conclusion that was reached about configural scoring at the item level (Lykken, 1956; Meehl, 1950), namely, it was an intriguing idea that unfortunately did not pan out empirically.

We are not suggesting that programs designed to uncover higher order trends and interactions be discontinued, but it is a reasonable hypothesis that linear relationships will continue to dominate forecasting equations (i.e., those that cross-validate). Some of the most solid work in psychological prediction is confined to the linear model (Dawes, 1979; Dawes & Corrigan, 1974; Dawes, Faust, & Meehl, 1989). Although there is much work yet to be done in validity generalization (including predictor and criterion development) given its current success using the linear model, substantial advances are unlikely to emanate from deviating too much from this framework. The amount of variance accounted for with the five equations in Table 5 is impressively high, high enough to make one wonder how much additional variance is left to account for, *using only ability dimensions as predictors, however transformed* (see below).

Prediction: The r_{xy} = .50 Barrier

How high should validity coefficients be? The applied psychological literature is full of examples of criterion-prediction validities that cluster around .20 to .30, with almost none of them exceeding .50. The notion that there is a *validity ceiling* at r = .50 (also known as the *.50 barrier*) can be traced back to Hull. Hull (1928, p. 193) believed that abilities contribute about 50 percent to criterion performance (other sources of influence include "industry or willingness," 35%, and "chance or accident," 15%), so his utopian expectation for this domain's predictive potential was about r_{xy} = 0.71. However, because of measurement error and the nature of psychological phenomena, he seriously doubted that validities above .50 would ever be obtained.

In recent articles, Meehl (1990a, 1990b) suggests that theoretical predictions in psychology, and indeed in all sciences, should be accompanied by ranges of tolerance. When faced with an estimation or prediction problem, such as accounting for the variance in work performance, psychologists should at least have *some feel* for how much to expect from their predictors. Given that all psychologically significant criterion behaviors are multiply determined and that many of these determining factors are unknown, estimates will always be much less than 1.0. But how much less than unity is tolerable? What fraction of the total criterion variance can we reasonably expect to account for with ability? This analysis is important because it adjusts our expectations and suggests when we should look beyond ability to other predictors.

Because we have had difficulty accounting for more than 25 percent of the criterion variance, several writers have expressed disenchantment with conventional ability tests as predictors and have argued that perhaps a different conceptualization of the problem is called for, stressing *process* factors, for example, as opposed to *products* (such as ability). Consider the following nonability factors that undoubtedly account for variance in work performance:

- *Sheer energy level.* This attribute might profitably be fractionated into cognitive versus psychomotor energy.
- *Interests and needs.* How satisfying does the individual find the occupation in general, and the job and work setting in particular?
- *Personality style.* To what extent does the individual's personality style correspond to the temporal characteristics of the job, such as the pace and cycles of the work (Dawis & Lofquist, 1984)? Style can be defined to include a variety of expressive and response tempo attributes, even such characteristics as speech (slow vs. rapid), certain mannerisms (some endearing, others aversive), and certain types of interpersonal presence (e.g., some extreme forms would be the animalistic presence of the socially poised psychopath or the debilitating preoccupation of the manic manifested in irritated impatience: "Why does everyone have to move *so* slowly?")
- *Nonbehavioral personal attributes.* These would include such attributes as body build, gender, height, race, weight, physical attractiveness, and health (which, of course, is relevant to behavior).
- *The short half-life of generalizations in the social sciences.* As pointed out by Cronbach (1975a), this is due to a multitude of factors including economic and political influences that moderate the significance of other personal attributes.
- *Chance or simply luck.* Chance is one of the most underestimated of factors. In psychology, for example, the demonstration that chance plays a critical role in

close interpersonal relationships (Berscheid & Walster, 1969) would suggest that psychologists should lower their expectations about criterion prediction.

The long-standing discussion of the .50 barrier always seems to see the glass half empty and then seeks to remedy the deficit by adding more abilities (or by improving the assessment of existing abilities), as opposed to trying other classes of attributes such as those enumerated above. This barrier may be merely the asymptotic limit of what the ability domain has to offer, and accounting for 25 percent of the variance in any socially important area with *one* class of attributes is impressive (contemporary findings are approaching 50%; see Tables 5 and 6). It is curious why some psychologists keep demanding more incremental validity from *this* domain instead of trying other classes of attributes.

Much contemporary discussion in cognitive experimental psychology expresses dissatisfaction with traditional psychometrics, arguing that no new knowledge is coming from this area. But then it proceeds to reconceptualize abilities as if the remaining unpredicted variance had to be accounted for by ability factors. Most of this variance is probably accounted for by other factors. If those with contrary views would assign percentages to these other factors, such as the variables listed above (we are sure there are others we have missed), we can then see what is left.

The extent to which these factors add incremental validity to ability variables will better enable us to determine the ceiling for psychological variables. Furthermore, there is good reason to suspect that these valued nonability factors have more differential validity for the less intellectually demanding occupations. Nonability attributes can be compensated for by other personal qualities. Someone who is physically attractive can "get away with" being a little aloof, whereas someone not as attractive can nevertheless be equally as effective with a charming personality. Other nonability attributes can be treated in similar fashion, but not ability. No matter how attractive or charming, a person requires a certain degree of intellectual ability to run a large corporation or to teach in law school. Given an intellectually demanding occupation such as lawyer, physician, military officer, business executive, or engineer, and assuming we can already account for 30 percent of the criterion variance (which seems a lot to us), how much more should we expect to glean from ability *if* our knowledge of this domain were comprehensive?

Other nonability attributes: Full explication of the following idea is beyond the compass of this chapter, but it might be useful to reflect more comprehensively on the concept of *response capability*, if only briefly. Throughout our treatment, we have focused on conventional abilities assessed psychometrically using T-data (i.e., tests having right and wrong answers). One can, however, think of abilities more broadly, say, as *response capabilities*. Wallace (1965) made this point years ago in a thought-provoking paper entitled, "An Abilities Conception of Personality: Some Implications for Personality Assessment," published in the *American Psychologist*. (Unfortunately, the article culminated with an unnecessary extreme environmentalistic position that undoubtedly put off otherwise sympathetic readers.) The idea is that the concept of *response capability* goes beyond conventional ability concepts.

Consider one of Wallace's examples, *social skill*, "the efficiency with which a person can elicit positive statements from others" (p. 132). Such skills clearly have implications for one's "net" interpersonal, and hence organizational, effectiveness. Should high levels of gregariousness be thought of as an ability? Perhaps, if one's occupation involves intense levels of social contact. But for a research scientist, this tendency might be construed in negative terms, say, *inability to withstand social isolation* (Wallace, 1965, p. 133). Several *nonability* attributes surely have implications for performance in industrial settings in both debilitating and facilitating ways. Extreme levels of *anxiety* or *negative affect* (Watson & Clark, 1984) and *anhedonia* (Meehl, 1975) are most likely detrimental across most contexts, whereas exceptional levels of *hedonic capacity* (Meehl, 1975) or *positive affect* (Tellegen, 1985) and *low anxiety* (Lykken, 1968) probably function more often than not in performance-enhancing ways.

In one of her early books on counseling psychology, Tyler (1953) noted:

> Many educators, counselors included, act as though they assumed that while ability is a fixed quantity not subject to change, effort and motivation can be manipulated at will. Nothing could be farther from the truth. One of the most difficult things to cope with in counseling is a state of chronic underachievement. The student with an IQ of 160 who has just slid by in high school and is making marginal grades in college is operating under a crippling set of habits which are probably as deep-seated as anything in his personality. (p. 120)

Perhaps the inverse of such debilitating attributes would be useful to assess in conjunction with conventional abilities to paint a more comprehensive picture of individuals' actual assets, capabilities, and tendencies. Some of this is currently being done by a Project A team (McHenry et al., 1990), and the results have been combined with ability assessment to enhance prediction. Perhaps underachievers (both academic and occupational) have particularly low status on some of these more central attributes (e.g., intellectual efficiency). This status could mark low levels of dimensions of great catalytic significance when measured in reverse.

The Prediction Problem: Is There a Threshold?

It is often remarked that having ability is all well and good, but there is a point of diminishing returns beyond which more ability has little to add (Wallach, 1976). People (psychologists as well as laypersons) have been heard to say this, and there are many reasons for this belief—not the least of which is selective recall of exceptions (Dawes, 1979). When based on data, this statement almost always involves a highly skewed distribution within a highly select range of ability (Linn, 1982b). Sometimes correlations are computed using dichotomous or trichotomous criterion variables. For example, the correlation of GRE scores with graduate GPAs (a criterion variable typically having only two or three values, with disparate frequencies) is usually in the .20 to .30 range (Linn, 1982b). The comment is often made that because GRE only accounts for less than 10 percent of the variance in graduate school grades, the test cannot be worth much.

The problem with this conclusion is that it is based on a severely restricted range of ability—students who have been admitted to graduate programs. This was the burden of McNemar's (1964) criticism of some early research on creativity. Several investigators had commented on the minuscule relationship between indices of creativity and general intelligence when, in fact, the samples only comprised individuals in the top five to six percent in ability. When ability distributions are restricted (e.g., a first-year graduate class, or a department's faculty), most individuals fall within a very small ability range. The people performing at exceptional levels (such as straight-A students or highly rated faculty) are so few that their bivariate points contribute little to the overall predictor-criterion correlation. The confounding that compensatory attributes create, discussed earlier, further confuses the picture.

The following example gives a more even-handed treatment to ability parameters within the upper ranges of talent. The data were collected on eighth-grade students who were identified by the Study of Mathematically Precocious Youth (SMPY) as being in the top 1 percent of ability. These students were initially given the *Scholastic Aptitude Test* (SAT) and subsequently tracked longitudinally. Predictive validities were obtained by correlating students' eighth-grade SAT-M with their scores on the *College Board Achievement Tests* (Math 1, Math 2, Biology, Chemistry, and Physics). The latter tests were given just before the students entered college (i.e., after a four-year interval). The predictive validities ranged from .16 to .57, with a mean of .40 for the females ($N = 95$), and from .39 to .52, with a mean of .45 for the males ($N = 223$). (Sample sizes given are average, inasmuch as not all students took all five tests.) On two Advanced Placement calculus tests, the females had validities of .42 and .44, and the males, .38 for both (these are raw correlations; they are not corrected for attenuation!). These correlations demonstrate that individual differences assessed among the top 1 percent of ability in the population can have meaningful predictive validities across a four-year temporal gap.

Furthermore, in a subsequent analysis using *nontest* criteria, Benbow (1992) compared the achievement profiles of the highest and lowest fourths of a larger group of gifted students defined as before (that included the previously mentioned group). Sample sizes averaged 100 females and 367 males for the top 25 percent, and 282 females and 248 males for the bottom 25 percent. Data on a variety of criteria—earning a college degree, intellectual level of college attended, honors won, college GPA, attending graduate school, and intensity of involvement with math and science—all favored the top fourth significantly and substantively, regardless of gender. Individual differences even in the top 1 percent of ability *do* have important psychological implications. However, it takes out-of-level testing to detect these differences (Keating & Stanley, 1972; APA, AERA, & NCME, 1985). We do not generally observe such differences because it is so seldom that we encounter a large enough sample of subjects at this level who span a substantial range of ability. But when representative samples within these ranges are assessed and criteria with sufficiently high ceilings are used, profound psychological differences emerge. This should not be surprising because this "tip" of the distribution does include about one-third of the ability range. Galton (1869) discussed this phenomenon in some detail, and the work of Julian Stanley (1983) is spiced with several fascinating examples documenting the particulars of such differences.

The most important attribute for successful performance in any highly select domain often has the least variation among the factors that contribute to achievement in that domain. This applies to all types of abilities: cognitive, perceptual, psychomotor, and even nonability attributes. (This would extend, for example, to running speed among National Football League [NFL] wide receivers, as well as height among players in the National Basketball Association [NBA]; Lubinski & Humphreys, 1990b, p. 390.) For further confirmation of this perspective, see convergent data bearing on this issue from an older age group in the latest volume of Lindzey's (1989) *A History of Psychology in Autobiography* (Vol. 8). The question to be answered is, What behavioral attributes distinguish this elite group of psychologists from psychologists in general?

Test Bias and Group Differences

Since the publication of the last *Handbook*, a voluminous literature has emerged on *test bias*. The scope of the scientific concern with this topic is reflected in the attention attracted by Jensen's book (1980a), *Bias in Mental Testing* (e.g., receiving "target article" treatment in *Behavioral and Brain Sciences;* Jensen, 1980b). Several scientific panels, both within psychology (Cleary et al., 1975) and in the larger scientific community (Hartigan & Wigdor, 1989; Wigdor & Garner, 1982), have been convened to examine the differential accuracy of statistical predictions of behavior made on the basis of test scores for different demographic groups. The *Journal of Vocational Behavior* (Gottfredson, 1988, vol. 33, no. 3) devoted a whole issue to "Fairness in Employment Testing" (edited by L. Gottfredson). Ackerman and Humphreys (1991) provide an excellent discussion of this topic in the first volume of this *Handbook*. Because we wholly concur with their treatment of the many issues they address, we will refer readers to their chapter and focus our discussion on extensions, implications, and new methodologies within this topic area. In addition to the Ackerman and Humphreys chapter, Cronbach (1975b), Humphreys (1988a, in press), and Linn (1982a, 1982b) will provide readers with other excellent reviews of the central issues and empirical findings that this literature has to offer. Also, a number of test validity issues are treated in more than just their psychometric aspects by Linn (1989) and Wainer and Braun (1988).

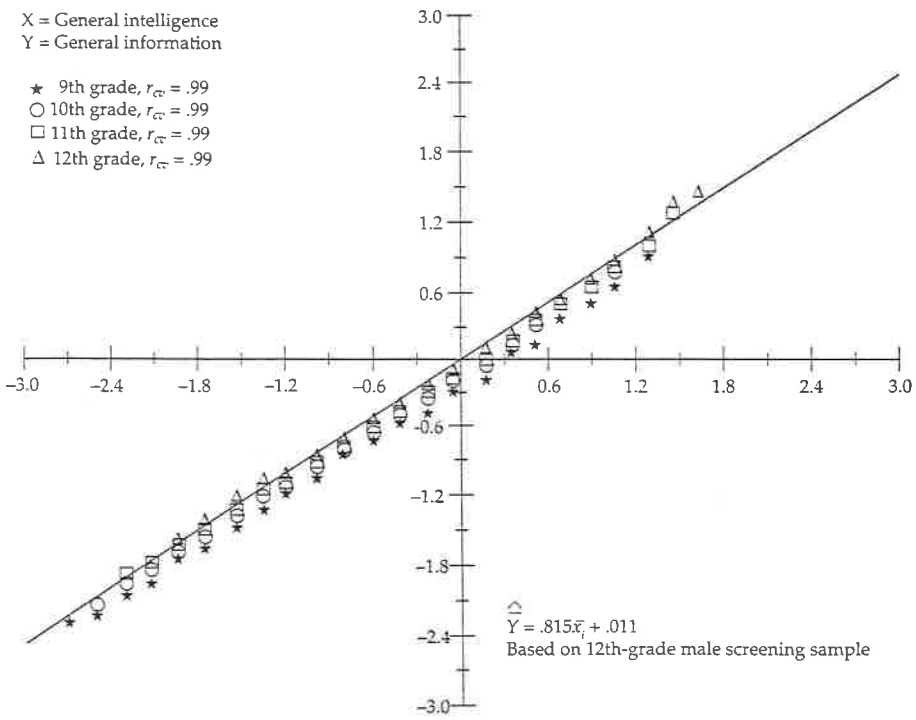
To add to this literature, we offer a recently developed methodology for the analysis of group differences in performance. Lubinski and Humphreys (manuscript under review) have shown that cross-validation coefficients, obtained with conventional predictor and criterion variables, can approach unity if the correlations are computed from the means for segmented intervals of the predictor and criterion variables (instead of using individual scores). This methodology is particularly noteworthy because when analyzing *group* differences in behavior, the appropriate unit of analysis should be an aggregate. This may seem obvious after the fact, but it has not been discussed in any great detail in the aforementioned literature. Typical test bias studies have employed conventional regression techniques applied to individual scores when comparing the accuracy of prediction for two or more groups, as discussed in detail in the references mentioned above.

Consider an alternative approach: Divide the predictor variable into approximately equal class intervals, say, by .20 standard deviations. Then compute bivariate means for individuals within each segment on both the predictor and the criterion variable. Then correlate these values. It is not atypical to observe correlations in the high .90s for conventional individual differences measures using this procedure. With the correlation this high, a regression equation can predict the mean level of performance in subsequent groups with great precision. Even group performance at different ages and different stages of development can be predicted with near certainty. Figure 8 presents data from Lubinski and Humphreys (manuscript under review). These data are from Project TALENT (Flanagan et al., 1962), and are based on male subjects from each of four classes, grades 9 through 12—over 180,000 high school students (about 45,000 from each class). The predictor variable for this analysis was Project TALENT's Intelligence composite and the criterion variable was a (nonacademic) General Information composite (consisting of 143 information items about, e.g., fishing, food, esoteric colors, architecture, law, the Bible, photography, games, ballet, accounting, business, and sales).

Figure 8 shows a regression equation based on a randomly selected screening sample of half of the male 12th-grade students (about 19,000 subjects). First, the Intelligence composite scale was divided into intervals of approximately .20 standard deviations. Means for the predictor and criterion variables were then computed for individuals within each predictor segment. These segment means were then standardized, based on the respective (predictor and criterion) means and standard deviations of the entire sample. A regression equation was then computed, based on 20 standardized data points (the predictor-criterion bivariate means; see Figure 8).

FIGURE 8

Cross-validation Analysis of the Regression of Group Means of General Information on Group Means of General Intelligence



Note: Data points represent bivariate means for General Information and TALENT's Intelligence composite.

Following this, the screening sample's raw-score interval values and statistics were used to divide and standardize the predictor scales for the remaining four groups: the other half of 12th-grade males (about 19,000 subjects) and the other three grade cohorts, grades 9 through 11 (about 45,000 subjects each). These groups served as the cross-validation *calibration* samples. The observed criterion values were then plotted (see Figure 8), along with the regression line based on the grade 12 screening sample. The cross-validated coefficients, computed by correlating these observed values with the predicted values, are also given in Figure 8. As Figure 8 reveals, the cross-validation coefficients all approached unity, and as grade level decreased (from 12th to 9th), systematic trends of *over*prediction were observed.

The overprediction trends reported here have also been observed in conventional regression analyses of contrasted groups (viz., groups having the lower mean on the predictor tend to be systematically overpredicted; Cleary et al., 1975; Linn, 1982b; Stanley, 1971). However, the present analysis offers a degree of precision heretofore unattainable. It would be informative to employ the foregoing methodology to other applied problems in industrial psychology as well as in psychology generally. For example, a pressing problem currently facing this country is the critical need for more engineers and physical scientists. Much research has focused on the black-white mean difference in cognitive abilities, but both these groups tend to be excelled by Asians in quantitative ability. Moreover, more than any other group, Asians tend to be limited by a GPA ceiling effect, inasmuch as Asians produce a disproportionate number of 4.0 (straight A) GPAs (Humphreys, 1988b). Hence, systematic trends of *under*prediction might be expected for this demographic group. The present methodology can be used to highlight the magnitude of such trends in a standard-score metric. (Interestingly, Asians are over-represented in the National Academy of Sciences by a ratio of 10 to 1; Havender, 1980; Vernon, 1982. Does this group differ from other groups on important work-related nonability attributes?)[12]

The foregoing methodology assumes a homoscedastic predictor-criterion relation. The precision offered by this analysis applies to *predicting group behavior, not the behavior of individuals*. Basic sample statistics should be reported so that the prediction error for individuals can be estimated. The same precision illustrated in Figure 5 may be shown by using conventional regression techniques and entering a group mean for X (in contrast to the raw score for an individual) to predict the average group performance (in contrast to an individual's criterion performance). The enhanced precision can be seen in the modification of the individual standard error of estimate (SEE) for group estimation:

$$\text{Individual SEE} = S_y(1 - r^2_{xy})^{1/2}$$

$$\text{Group SEE} = S_y(1 - r^2_{xy})^{1/2} / (n)^{1/2}$$

For the group SEE, because we are predicting a mean with a mean, the SEE is reduced by the conventional standard error of the mean: the square root of the sample size, n.

In our present example, the Information and Intelligence composites correlated .78. Using the standard deviation of the grade 12 males, which, for the Information composite was 20.41, the SEE for an individual would be 12.77. For a group of, say, 100 individuals, the group SEE would be (individual SEE/$100^{1/2}$) or 12.77/10 = 1.28. This amount of prediction error is quite small, considering that the standard deviation of the Information composite is over 15 times this value. Thus, we can achieve great precision in the prediction of group performance by using this relatively simple methodology. (For further treatment of this methodology, see Lubinski & Humphreys, manuscript under review.)

Gender Differences in Cognitive Abilities: Implications for Meta-analytic Reviews

Several writers discussing gender differences in cognitive abilities have concluded that females have excelled in the verbal domains and males in mathematics and spatial visualization (Maccoby & Jacklin, 1974; Tyler, 1965). To a degree, this generalization still holds today. But gender differences in means on these attributes appear to be decreasing. Several meta-analytic reviews (Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 1988; Linn & Hyde, 1989) and cross-sectional studies (Feingold, 1988) indicate that females and males are slowly

converging toward a common mean.[13] However, these reviews and studies of gender differences only discuss differences of the first moment, that is, differences in means. There are other moments about the mean on which differences between groups may be observed. The second, third, and fourth moments quantitatively characterize variance, skewness, and kurtosis, and there are data suggesting that these statistics deserve more attention (Lubinski & Humphreys, 1990a).

A long history of individual differences research finds that males are more variable than females on many abilities, even on abilities for which females have the higher mean (Lubinski & Benbow, 1992; Lubinski & Humphreys, 1990a). Perhaps one reason that the hypothesis of greater male variability has not received the attention it deserves is to be found in the unenlightened attitudes that developed about the subject, exemplified in these forgettable remarks made by E. L. Thorndike (1906):

> The restriction of women to the mediocre grades of ability and achievement should be reckoned with by our educational systems. The education of women for such professions as administration, statesmanship, philosophy, or scientific research...is far less needed than education for such professions...where the average is essential. (p. 213)

Fortunately, but rather belatedly, we have put such attitudes behind us (for the most part). Contemporary discussions of gender differences almost always contain the caveat that the overlap in distributions is far greater and much more significant than the difference between means.

The importance of the hypothesis of greater male variability (and its implications) has been revisited recently in *Behavioral and Brain Sciences* by Benbow (1988) and several commentators (Becker & Hedges, 1988; Humphreys, 1988b; Jensen, 1988). Benbow's work focused on mathematical talent, but, as several of the commentators pointed out, greater male variability has been observed on other cognitive abilities as well.

Discussing the hypothesis, Jensen (1980a) has placed the most stock on evidence based on "the largest, most representative sample ever tested in a single study on a group IQ test" (p. 628). This sample, although drawn back in 1932, consisted of all the children in Scotland between the ages of 10.5 and 11.5, except for the deaf and blind (N of about 87,000). The mean difference between the genders was trivial, but the male standard deviation was significantly larger than the female standard deviation (by one IQ point). Given a difference of this size, the overrepresentation of males at the extreme tails of the distribution (say, 3 or 4 standard deviations out) was striking. Gottfredson (1988) shows how, for groups manifesting a mean difference, group proportions change in different segments of the upper and lower tails of the distribution. The same phenomenon occurs for a variability difference.

Table 7 shows some Project TALENT data for over 360,000 students. Three ability composites were aggregated to measure English language, mathematical, and spatial abilities (see Lubinski & Humphreys, 1990a, for the psychometric details). For these three composites, plus the Intelligence composite previously discussed (Figure 8), males displayed greater variability, even for the English Language composite, on which the females consistently scored the higher mean.

Greater male variability was also observed in data for other nationally representative samples, collected for the National Growth Study of the Educational Testing Service (Hilton, Beaton, & Bower, 1971). This study ran from 1961 to 1967, with data collection every two years, to assess (among other things) representative statistics on the *Sequential Tests of Educational Progress* (STEP) and the *School and College Ability Test* (SCAT). These tests cover a range of academic topics including mathematics, science, social studies, reading, writing, and

TABLE 7

Means and Standard Deviations for Four Ability Composites From Project TALENT for Grades 9 Through 12 by Gender

	English Language		*Mathematics*		*Spatial*		*Intelligence*	
	M_x	S_x	M_x	S_x	M_x	S_x	M_x	S_x
Grade 9								
Females	87.65	17.29	15.35	6.43	60.15	20.07	145.36	52.58
Males	79.51	18.11	16.01	6.98	69.18	22.54	142.48	54.22
Grade 10								
Females	92.29	17.43	16.65	7.08	63.42	20.80	157.35	52.69
Males	84.37	18.12	18.05	7.65	74.45	22.77	156.40	54.27
Grade 11								
Females	96.63	16.91	17.77	8.02	66.22	20.94	169.97	52.13
Males	89.28	17.89	20.69	8.90	79.01	22.96	173.56	53.95
Grade 12								
Females	100.27	16.55	18.60	8.15	68.49	21.31	180.08	51.56
Males	92.73	17.56	22.46	9.32	82.35	23.31	184.98	53.82

Note: Sample sizes for each cohort by gender follow: grade 9, females = 49,393, males = 49,968; grade 10, females = 47,119, males = 48,543; grade 11, females = 45,428, males = 43,851; grade 12, females = 40,116, males = 38,392. Detailed descriptions of these ability composites may be found in Lubinski and Humphreys (1990a).

verbal and quantitative abilities. With few exceptions, males were found to display more variability on these measures than females. Stanley et al. (1992) recently analyzed female-versus-male effect sizes for 86 nationally standardized tests. This analysis included data collected in the 1980s. Tests studied included the *Advanced Placement* (AP) examinations, the advanced (subject) tests of the *Graduate Record Examinations* (GRE), and the *Differential Aptitude Tests* (DAT). Repeatedly, males displayed greater variability on these measures, even on scales for which the effect sizes favored the females.

The implication of greater male variability for gender representation when stringent selection criteria are employed is obvious. Furthermore, to the extent that these findings are valid and robust, an overrepresentation of males at the *lower* end of ability distributions should be anticipated as well. Because equal representation is a cherished value in our society, the facts about differences—in variability and not just in mean—should be scrupulously compiled. As a minimum, reporting descriptive statistics should be made a requirement for all published studies—to reverse a trend in research publication that has seen simple descriptive statistics sacrificed to accommodate the elliptical elaboration of abstruse analyses of the data. The development of methodologies for meta-analytic reviews has been fortuitous: We now appreciate the significance of standard deviations as well as means.

Whatever the causes of greater male variability may be, it is first important to establish the facts of the matter. The same principle applies to group differences in general.

Conclusion

We began our discussion with an important concept imported from experimental psychology: the criterion of scientific significance. We will close as well with another concept from experimental psychology: achievement versus topographical accounts of behavior.

A major emphasis throughout this chapter has been aggregation across diverse content, both within and between predictor and criterion domains. Psychologists have tended to label things *components* on the basis of surface features of the variables on both sides of our forecasting equations. These components typically share variance, and we have tended to underestimate their overlap. Just because entities or events share features that allow classification in a group, this does not constitute even a rudimentary basis for a scientific taxonomy. Cattell (1946) called such groupings without covariation *semantic traits*—classes found in the dictionary but having no scientific viability.

In contrast, the groupings we have recommended are predicated on parsimony emanating from covariation. At both item and scale level for predictors and criteria, indicators that have diverse content can be aggregated to converge collectively on the dominant dimensions defined by their communality. Aggregation also minimizes the bias inherent in all psychological measurement operations. The resulting structures will reflect broad dimensions that have meaningful correlates. Important systematic sources of individual differences tend to be associated with multiple indicators that have diverse content (Cattell's *source traits*, but not his *surface traits*). Indeed, given our discussion on how construct validity is maximized by aggregating content-heterogeneous items and scales, it is questionable whether *any* surface trait *could* possess scientific significance.

This discussion is linked to the distinction between an *achievement* versus a *topographical* account of behavior (MacCorquodale & Meehl, 1954). Topographical accounts focus on the surface features of behavior, its form and content, whereas achievement accounts focus on what the behaviors actually accomplish or achieve, regardless of their response topography (Lubinski & Thompson, 1986). In modern learning theory, achievement accounts are almost always preferred. Different behavior patterns (often *perceived* as distinct abilities or skills) can function equivalently in achieving the same goals. The question is, To what extent are these ostensibly different behaviors interchangeable? To what degree do different topographies reflect only a subtle psychological difference that does not prevent them from being psychologically equivalent? Part of being a good psychologist is knowing which classes of behavior are important (i.e., scientifically significant) and which are equivalent to others. Behavior topography is deceptive. We tend to perceive components in topography, components like those reified in faculty psychology or "vectors of mind." We tend to seek recurring features in recurring events. But to assess whether our perceptions are scientifically meaningful, we have to show covariation and forecasting efficiency. Disparate behavior topographies can be grouped if they share communality. Furthermore, to the extent this communality is shared with other broad classes of behavior (not only other classes of response capability but also criterion behavior classes of interest), it becomes possible to forecast the likelihood of developing subsequent behaviors. It is in this sense that the more behaviors one has, the greater the readiness one has for developing other behaviors.

Normal Science

If one thing is apparent from the foregoing treatment of human abilities, it is that we have stayed almost exclusively with well-known sources of individual differences and traditional methods of prediction (i.e., the linear model). This decision was intentional. These variables and methods have much to offer, and the scientific significance of contemporary and future innovations will have to be measured against their achievements. If reformulations framed as paradigmatic shifts are to achieve scientific significance, they must surpass what existing techniques offer. Too often, new paradigms are advanced as replacements for existing methods without any empirical basis to support their superiority.

For example, Pavlovian and instrumental conditioning procedures have been used, with remarkable results, in studies about the behavioral consequences of drugs. Commenting on the orderliness engendered by Pavlovian and instrumental conditioning procedures in the drug evaluation field, at a time when basic principles of learning theory were considered passé, MacCorquodale (1971) wrote:

> I suppose, however, that research workers in any specialized area would really prefer to get very durable, highly reproducible but wholly innovative and hopefully disconfirming outcomes. When this happens, one can get lots of extra mileage out of his results by brandishing a new paradigm at everyone else, or at least hinting at one, and proclaiming a scientific revolution. I have heard none of that sort here. We are all still in business so far as I can see, but we have a lot of new information about a new class of discriminative, reinforcing, and eliciting stimuli. That is news; it is useful and it is constructive. But it is not revolutionary, and I am delighted. (p. 217)

Today, some 20 years after those remarks, the animal model of abuse liability of drugs is arguably the most scientifically significant animal model in all of psychology

These remarks are offered to stress that like longstanding principles from learning theory, the available methodologies for assessing human abilities can generate new knowledge that will not have a "short half-life" (Cronbach, 1975a) so characteristic of the many fads and fashions in psychology's history (Dunnette, 1966). The field of human abilities in industrial and organizational psychology has much going for it. All indications are that more useful and enduring knowledge will be generated by existing techniques, if used appropriately. Many times, investigators all too lightly dismiss well-established principles and do not assimilate existing knowledge and techniques for acquiring further knowledge, in the hope of hitting on something wholly innovative and totally creative. In fact, it is our view that creative achievements typically involve the rearrangement of existing well-established facts in novel ways.

In this chapter, we have tried to indicate throughout which articles we think are essential reading to enhance the likelihood of instrumentally effective rearrangements of existing knowledge. To this list, we add the methodological treatments by Lykken (1968, 1991) and Meehl (1990a, 1990b). These contributions also discuss some sociological influences that impede the conduct of meaningful research.

These closing remarks are offered not with the intent of deterring novel approaches, but rather because we perceive the need for researchers on human abilities to have more of a common methodological and substantive knowledge base. *There isn't all that much to assimilate*, but what we do know ought to be known by all. So that readers are aware that others have expressed similar concerns, it might be useful to quote a recent exchange by two senior investigators who have built their

reputations on sound logic and normal science:

> The paradigm shift has been used to justify doing anything and everything except science.
> (Lloyd G. Humphreys, 1990, p. 153).

> I, too, think Kuhn's impact on the soft areas of psychology is unhealthy. That Humphreys arrives at views similar to mine...is reassuring to me.
> (Paul E. Meehl, 1990b, p. 173)

After all this, we leave you with not a few facts and ideas, and a not-so-short list of recommended readings.

We have profited from correspondence and discussions with a number of colleagues and friends whose talents are conspicuously public to all who know them: Camilla P. Benbow, Kathy A. Hanisch, Lloyd G. Humphreys, Paul E. Meehl, and Julian C. Stanley.

Notes

1 Although it is true that many behaviorists reject the idea of a general factor (Schmidt, 1988), upon closer analysis, concepts and findings from behavior-analytic perspectives are not incompatible with findings from individual differences research (Hull, 1945; Lubinski & Thompson, 1986; Meehl, 1986a).

2 Cattell (1950, pp. 23–27) provides an excellent introduction to the geometric interpretation of factors.

3 This idea is actually an extension of the multitrait-multimethod matrix (Campbell & Fiske, 1959). The complementarity of the two ideas is especially featured by organizing the matrix around traits, as opposed to methods, and placing the correlation profiles of different measures of the same trait directly beneath the resulting monotrait triangles (see Lubinski, Tellegen, & Butcher, 1983).

4 Throughout this chapter, data from Project TALENT (Flanagan et al., 1962) will be presented to illustrate principles as well as to shed light on some unresolved empirical issues. These data were obtained from a stratified random sample of the nation's high schools. The sample contains four cohorts, grades 9 through 12, with approximately 100,000 subjects per cohort. The information collected by Project TALENT includes scores for several dozen conventional ability, interest, and personality measures, as well as biographical information, with many of the same measures used across all four cohorts. These subjects were also followed up longitudinally at three time points: 1, 5, and 13 years after high school graduation. For more detail on this huge data set, readers are referred to Wise, McLaughlin, and Steel (1979).

5 Humphreys (1962) described his experience of attempting, for seven years during the 1950s, to assemble unique predictor equations for military assignments, using Thurstone's primaries. He was able to achieve differential validity only with two large composites corresponding to Vernon's two major group factors, *v:ed* and *k:m*. Vernon (1947, 1961) has reported on a similar experience. See also Thorndike 1985, 1986.

6 Guilford has been inconsistent in his discussions of the general factor, sometimes even in the same article. In discussing optimal conditions for investigating his SI model, Guilford (1961) noted that "the selection of the population of individuals is important. In the study of intellectual abilities, there should be relative homogeneity in age, education, sex, and *general intellectual level*" (p. 8, italics added). And later, "These and other factorial learning studies should nail in its coffin for all time the notion that there is a single, general learning ability" (p. 15).

7 Both applied and theoretical psychologists might employ this procedure profitably in other contexts. For example, construct-irrelevant specific variance, rather than construct-relevant variance, is a problem with many paradigms in experimental cognitive psychology. We have concluded that the concepts of true and error scores are meaningful only in the context of reliability. As soon as we move to validity, the reliable variance of a predictor variable invariably splits into common and specific components. These two components, in turn, always split into construct-relevant and construct-irrelevant (bias) variance components.

8 Detailed descriptions of these instruments are provided by Dunnette (1976) and in *Personnel Psychology* (1990, 43, whole issue).

9 For a discussion of the influence of socioeconomic status versus intelligence on socially valued criteria, see Lubinski and Humphreys (1992).

10 In a recent development, Ackerman (1987) has augmented the conceptual scheme of the radex to include perceptual and psychomotor abilities. The resulting conceptual structure, represented as a cylinder, is useful in conceptualizing the acquisition of skill and the role of different abilities in skill acquisition. Detailed development of this model, including its relation to information processing accounts of cognitive functioning and learning, are given in Ackerman (1987, 1988a).

11 Vernon (1961), while commenting on his work in the military, actually formulated the hypothesis for validity generalization with respect to the general factor:

> It often happened that recruits who failed in one Service job had to be reallocated, and it was usually found necessary to move them to a job requiring lower general ability and application. If they were transferred to another job at the same level, in which they claimed some interest or experience, only too often they failed again. The layman's notion that there exists a niche or special type of work ideally suited to the specialized aptitudes of each individual appeared to be much less true than the view that all types of work and all employees fall along a single high-grade to low-grade continuum. The success of women workers at skilled engineering jobs during the war further supports this view. (p. 122)

However, Vernon (1961) also appreciated the differential validity offered by more detailed ability assessment.

> Psychologists giving guidance are justified in making the fullest possible use of *g*, *v:ed*, and *k:m* tests, but thereafter their success is likely to depend chiefly on the extent to which they can gauge each candidate's previous relevant experience, ... motivation, ...[and] specific attitudes. (p. 128)

12 For a fascinating discussion of how a "minority" population, having slightly more talent than a "majority" population (as a group), may be overrepresented when extreme cutting scores are employed, see Page (1976, pp. 305–306).

13 Stanley et al. (1992) has suggested that these trends should be interpreted with caution. He writes:

> Are girls catching up with boys, and boys with girls, in cognitive respects where they once differed considerably? Feingold (1988) thought they were, based largely on his analysis of scores from two test batteries over a long period. This is hard to determine, partly because for the last 20 years or longer many test publishers have tried to minimize what some call gender "bias" by studying each test item carefully. Inevitably, a number of the "worst offenders" tend to be discarded from one revision of the test or test battery to the next. Some idea of how extensive and intensive this screening is could be obtained by studying the research and operational resources that Educational Testing Service and the College Board now devote to remedying item bias of many sorts, to what might be called "equity in testing." Their PSAT, SAT, GRE, high school achievement tests, Advanced Placement Program examinations, and others are combed over for gender, ethnic, and racial differences in various items, item types, subtests, and tests. It seems reasonable to conjecture that gender differences for its tests may decline, even without any alteration in the cognitive behavior of the examinees. (p. 43)

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3–27.

Ackerman, P. L. (1988a). Determinants of individual differences during skill acquisition: Cognitive processes and information processing. *Journal of Experimental Psychology: General, 117*, 299–329.

Ackerman, P. L. (1988b). A review of Linda S. Gottfredson (Ed.), The *g* factor in employment. *Journal of Vocational Behavior* [Special issue]. *Educational and Psychological Measurement, 48,* 553–558.

Ackerman, P. L. (1989). Individual differences and skill acquisition. In P. L. Ackerman, R. J. Sternberg, & G. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 164–217). New York: Freedman.

Ackerman, P. L., & Humphreys, L. G. (1991). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., vol. 1). Palo Alto, CA: Consulting Psychologists Press.

Albright, L. E., Glennon, J. R., & Smith, W. J. (1963). *The use of psychological tests in industry.* Cleveland, OH: Howard Allen.

American Psychological Association, American Educational Research Association, & National Counsel on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Becker, B. J., & Hedges, L. H. (1988). The effects of selection and variability in studies of gender differences. *Behavioral and Brain Sciences, 11,* 183–184.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences, 11,* 169–183, 217–232.

Benbow, C. P. (1992). Academic achievement in mathematics and science of students between ages 13 and 23: Are there differences among students in the top one percent of mathematical ability? *Journal of Educational Psychology, 84,* 51–61.

Berdie, R. F., Layton, W. L., Hagenah, T., & Swanson, E. O. (1962). *Who goes where to college?* Minneapolis: University of Minnesota Press.

Berscheid, E., & Walster, E. (1969). *Interpersonal attraction.* New York: Addison-Wesley.

Brand, C. (1987). The importance of general intelligence. In S. Magil & C. Magil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 251–265). New York: Falmer Press.

Brogden, H. E., & Taylor, E. K. (1950). A theory and classification of criterion bias. *Educational and Psychological Measurement, 10,* 159–186.

Brunswik, E. (1956). *Perception and the representative design of experiments.* Berkeley: University of California Press.

Burt, C. (1949). The structure of the mind: A review of the results of factor analysis. *British Journal of Educational Psychology, 19,* 100–114.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 93,* 81–105.

Carroll, J. B. (1982). The measurement of intelligence. In R. G. Sternberg (Ed.), *Handbook of human intelligence* (pp. 29–120). Cambridge: Cambridge University Press.

Carroll, J. B. (1985). Exploratory factor analysis: A tutorial. In D. K. Detterman (Ed.), *Current topics in human intelligence: Vol. 1. Research methodology* (pp. 25–58). Norwood, NJ: Ablex.

Carroll, J. B. (1989a). Factor analysis since Spearman: Where do we stand? What do we know? In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology* (pp. 43–67). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1989b). Intellectual abilities and aptitudes. In A. Lesgold & R. Glaser (Eds.), *Foundations for a psychology of education* (pp. 137–197). Hillsdale, NJ: Erlbaum.

Cascio, W. F. (1991). Applied psychology in personnel management (4th ed.). Englewood Cliffs, NJ: Prentice-Hall

Cattell, R. B. (1946). *Description and measurement of personality.* Yonkers-on-Hudson: World Book Company.

Cattell, R. B. (1950). *Personality: A systematic theoretical and factual study.* New York: McGraw-Hill.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15–41.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (1st ed., pp. 223–326). Chicago: Rand McNally.

Cronbach, L. J. (1957). Two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1971). Test validity. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443–507). Washington, DC: American Council of Education.

Cronbach, L. J. (1975a). Beyond the two disciplines of scientific psychology revisited. *American Psychologist, 30,* 116–127.

Cronbach, L. J. (1975b). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1–14

Cronbach, L. J. (1977). *Educational psychology* (3rd ed.). New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.) New York: Harper & Row

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–18). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validity after thrity years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

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

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment: An individual differences model and its applications.* Minneapolis: University of Minnesota Press.

Dunnette, M. D. (1963). A note on the criterion. *Journal of Applied Psychology, 47,* 251–254.

Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist, 21,* 343–352.

Dunnette, M. D. (1976). Aptitudes, abilities, and skills. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (1st ed., pp. 473–520). Chicago: Rand McNally.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. In M. R. Rosenzweg & L. W. Porter (Eds.), *Annual Review of Psychology* (Vol. 30, pp. 477–525). Palo Alto, CA: Annual Reviews.

Eliot, J. C. , & Smith, I. M. (1983). *An international dictionary of spatial tests.* Windsor, England: NFER-Nelson.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43,* 95–103.

Ferguson, G. A. (1954). On learning and human ability. *Canadian Journal of Psychology, 8,* 95–112.

Fiske, D. T. (1971). *Measuring the concepts of personality.* Chicago: Aldine-Atherton.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldberg, I. (1962). *Design for a study for American youth.* Boston: Houghton Miffin.

Fleishman, E. A. (1966). Human abilities and the acquisition of skill. In E. A. Bilodeau (Ed.), *Acquisition of skill.* New York: Academic Press.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks.* Orlando, FL: Academic Press.

Frederiksen, N. R. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193–202.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences.* London: Collins.

Ghiselli, E. E. (1966). *The validity of occupational tests.* New York: Wiley.

Ghiselli, E. E. (1972). Comment on the use of moderator variables. *Journal of Applied Psychology, 56,* 270.

Gottfredson, L. S. (Ed.). (1986). The *g* factor in employment [Special issue]. *Journal of Vocational Behavior, 29,* 3, 293–450.

Gottfredson, L. S. (Ed.). (1988). *Journal of Vocational Behavior, 32.*

Gottfredson, L. S. (Ed.). (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33*(3), 293–319.

Green, B. F. (1978). In defense of measurement. *American Psychologist, 33,* 664–670.

Guilford, J. P. (1961). Factorial angles to psychology. *Psychological Review, 68,* 1–20.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P. (1985). The structure-of-intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence theories, measurement, and applications* (pp. 225–266). New York: Wiley.

Guilford, J. P., & Michael, W. B. (1948). Approaches to univocal factor scores. *Psychometricka, 13,* 1–22.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw-Hill.

Guion, R. M., & Gibson, W. M. (1988). Personnel selection and placement. *Annual Review of Psychology, 39,* 349–374.

Gustafsson, J. E. (1984). A unifying model for structure of intellectual abilities. *Intelligence, 8,* 179–203.

Gustafsson, J. E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 35–71). Hillsdale, NJ: Erlbaum.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). Glencoe, IL: Free Press.

Hakel, M. D. (1986). Personnel selection and placment. *Annual Review of Psychology, 37,* 351–380.

Harman, H. H. (1976). *Modern factor analysis* (rev. 3rd ed.). Chicago: University of Chicago Press.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Havender, W. R. (1980). Individual versus collective social justice. *Behavioral and Brain Sciences, 3,* 345–346.

Hilton, T. L., Beaton, A. E., & Bower, C. P. (1971). *Stability and instability in academic growth: A compilation of longitudinal data.* Princeton, NJ: Educational Testing Service.

Hogan, J. C. (1991). Physical abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., vol. 2, pp. 753–831). Palo Alto, CA: Consulting Psychologists Press.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurement and application* (pp. 267–300). New York: Wiley.

Horn, J. (1986). Intellectual ability concepts. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp. 35–78). Hillsdale, NJ: Erlbaum.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253–270.

Hulin, C. L., & Humphreys, L. G. (1980). Foundations of test theory. In A. P. Maslow, R. H. McKillip, & M. Thatcher (Eds.), *Construct validity in psychological measurement* (pp. 5–12). Princeton, NJ: Educational Testing Service.

Hull, C. L. (1928). *Aptitude testing.* Yonkers, NY: World Book Co.

Hull, C. L. (1945). The place of innate individual differences in a natural science theory of behavior. *Psychological Review, 52,* 133–142.

Humphreys, L. G. (1952). Human abilities. *Annual Review of Psychology, 3,* 5–15.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475–483.

Humphreys, L. G. (1976). A factor model for research on intelligence and problem solving. In L. B. Resnick (Eds.), *The nature of intelligence* (pp. 329–339). Hillsdale, NJ: Erlbaum.

Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence, 3,* 105–120.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–120). Plenum Press.

Humphreys, L. G. (1982). The hierarchical factor model and general intelligence. In N. Hirschberg & L. G. Humphreys (Eds.), *Multivariate applications in the social sciences* (pp. 223–240). Hillsdale, NJ: Erlbaum.

Humphreys, L. G. (1984). General intelligence. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives in mental testing* (pp. 221–248). New York: Plenum.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurement and application* (pp. 201–224). New York: Wiley

Humphreys, L. G. (1988a). Trends in levels of blacks and other minorities. *Intelligence, 12,* 231–260.

Humphreys, L. G. (1988b). Sex differences in variability may be more important than sex differences in means. *Behavioral and Brain Sciences, 11,* 195–196.

Humphreys, L. G. (1990). View of a supportive empiricist. *Psychological Inquiry, 1,* 153–155.

Humphreys, L. G. (in press). Intelligence from the standpoint of a (pragmatic) behaviorist. In D. K. Detterman (Ed.), Current topics in human intelligence: Series of intelligence (Vol. 4). Norwood, NJ: Ablex.

Humphreys, L. G., Lubinski, D., & Yao, G. (manuscript under review). *Engineering schools can profitably broaden the base from which they recruit students.*

Hunter, J. E. (1980). Construct validity and validity generalization. In *Proceedings of a Conference on Construct Validity in Psychological Measurement* (pp. 199–129). Princeton, NJ: U.S. Office of Personnel Management and Educational Testing Service.

Hunter, J. E. (1983a). *The dimensionality of the General Aptitude Test Battery and the dominance of the general factor over specific factors in the prediction of job performance for USES* (Test Res. Rep. No. 44). Washington, DC: U.S. Department of Labor, U.S. Employment Services.

Hunter, J. E. (1983b). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery* (Test Res. Rep. No. 45). Washington, DC: U.S. Department of Labor, U.S. Employment Services.

Hunter, J. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340–362.

Hyde, J. S., & Linn, M. G. (1988). Gender differences in verbal ability: A meta analysis. *Psychological Bulletin, 104,* 139–155.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematical performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin, 80,* 75–83.

Jenkins, J. J. (1989). The more things change, the more they stay the same: Comments from an historical perspective. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivations, and methodology: The Minnesota symposium on learning and individual differences* (pp. 475–491). Hillsdale, NJ: Erlbaum.

Jenkins, J. J., & Paterson, D. G. (1961). *Studies in individual difference: The search for intelligence.* New York: Appleton-Century-Crofts.

Jensen, A. R. (1980a). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1980b). Précis of bias in mental testing. *Behavioral and Brain Sciences, 3,* 325–371.

Jensen, A. R. (1988). Sex differences in arithmetic computation and reasoning in prepubertal boys and girls. *Behavioral and Brain Science, 11,* 198–199.

Kanfer, R., Ackerman, P., & Cudeck, R. R. (1989). *Abilities, motivation and methodology: The Minnesota symposium on learning and individual differences.* Hillsdale, NJ: Erlbaum.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis. *Psychological Bulletin, 75,* 34–49.

Keating, D. P., & Stanley, J. C. (1972). Extreme measures for the exceptionally gifted in mathematics and science. *Educational Researcher, 1,* 3–7.

Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differential mental abilities.* Stanford, CA: Stanford University Press.

Kelley, T. L. (1939). Psychological factors of no importance. *Journal of Educational Psychology, 30,* 139–143.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Landy, F. J. (1990). *Psychology and work behavior* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.

Lindzey, G. (1989). *A history of psychology in autobiography: Volume 8*. Stanford, CA: Stanford University Press.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*, 17–19, 22–27.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–164.

Linn, R. L. (1982a). Admissions testing on trial. *American Psychologist, 37*, 279–291.

Linn, R. L. (1982b). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335–388). Washington, DC: National Academy Press.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21*, 33–47.

Linn, R. L. (1989). *Educational measurement* (3rd ed.). New York: Collier Macmillan.

Linn, R. L., & Dunbar, S. B. (1986). Validity generalization and predictive bias. In R. A. Berk (Ed.), *Performance assessment methods and applications* (pp. 203–236). Baltimore, MD: Johns Hopkins Press.

Loevinger, J. (1954). Effect of distortion on item selection. *Educational and Psychological Measurement, 14*, 441–448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph No. 9]. *Psychological Reports, 3*, 635–694.

Lohman, D. F. (1989). Human intelligence: An introduction to advances in theory and research. *Review of Educational Research, 59*, 333–373.

Lubinski, D., & Benbow, C. P. (1992). Gender differences in abilities and preferences: Implications for the math/science pipeline. *Current Directions in Psychological Science, 1*, 61–66.

Lubinski, D., & Humphreys, L. G. (1990a). A broadly based analysis of mathematical giftedness. *Intelligence, 14*, 327–355.

Lubinski, D., & Humphreys, L. G. (1990b). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin, 107*, 385–393.

Lubinski, D., & Humphreys, L. G. (1992). Some bodily and medical correlates of mathematical giftedness and commensurate levels of socioeconomic status. *Intelligence, 16*, 99–115.

Lubinski, D., & Humphreys, L. G. (manuscript under review). *Seeing the forest from the trees: When predicting the behavior or status of groups, correlate means.*

Lubinski, D., Tellegen, A., & Butcher, J. N. (1983). Masculinity, femininity, and androgyny: Viewed and assessed as distinct concepts. *Journal of Personality and Social Psychology, 44*, 428–439.

Lubinski, D., & Thompson, T. (1986). Functional units of human behavior: A dispositional analysis. In T. Thompson & M. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 315–334). Hillsdale, NJ: Erlbaum.

Lykken, D. T. (1956). Method of actuarial pattern analysis. *Psychological Bulletin, 53*, 102–107.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Chiccetti & W. Grove (Eds.), *Thinking clearly about psychology*. Minneapolis: University of Minnesota Press.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

MacCorquodale, K. (1971). Reinforcing stimulus functions of drugs: Interpretations II. In T. Thompson & R. Pickens (Eds.), *Stimulus properties of drugs* (pp. 215–217). New York: Appleton-Century-Crofts.

MacCorquodale, K., & Meehl, P. E. (1954). Edward C. Tolman. In W. K. Estes et al., *Modern learning theory* (pp. 177–266). New York: Appleton-Century-Crofts.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*, 107–127.

McCornack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology, 40*, 343–344.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. A. (1990). Project A validation results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335–353.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19*, 871–882.

Meehl, P. E. (1950). Configural scoring. *Journal of Consulting Psychology, 14*, 165–171.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

Meehl, P. E. (1975). Hedonic capacity: Some conjectures. *Bulletin of the Menninger Clinic, 39*, 295–307.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Meehl, P. E. (1986a). Trait language and behaviorese. In T. Thompson & M. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 335–354). Hillsdale, NJ: Erlbaum.

Meehl, P. E. (1986b). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370–375.

Meehl, P. E. (1990a). Why summaries in psychological research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.

Meehl, P. E. (1990b). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108–141, 173–180.

Muchinsky, P. M. (1990). *Psychology applied to work: An introduction to industrial and organizational psychology* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Page, E. B. (1976). A historical step beyond Terman. In D. P. Keating (Ed.), *Intellectual talent: Research and development* (pp. 295–307). Baltimore, MD: Johns Hopkins University Press.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin, 87*, 1–28.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373–406.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72*, 480–483.

Roznowski, M., & Hanisch, K. A. (1990). Building systematic heterogeneity into work attitudes and behavior measures. *Journal of Vocational Behavior, 36*, 361–375.

Rummel, J. R. (1970). *Applied factor analysis* (2nd ed.). Evanston, IL: Northwestern University Press.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94*, 18–38.

Sackett, P. R. (Ed.). (1990). *Personnel Psychology, 43*.

Sackett, P. R., Tenopyr, M. L., Schmitt, N., & Kahn, J. (1985). Commentary on forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697–798.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16*, 209–222.

Scarr, S., & Carter-Saltzman, L. (1980). Twin method: Defense of a critical assumption. *Behavior Genetics, 9*, 527–542.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometricka, 22*, 53–61.

Schmidt, F. L. (1988). The problem of group differences in ability scores in employment selection. *Journal of Vocational Behavior, 33*, 272–292.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36*, 1128–1137.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task difference and the validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66*, 166–185.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution to the controversy. *Personnel Psychology, 24*, 419–434.

Schmidt, F. L., Pearlman, K., Hunter, J. E., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697–789.

Schmitt, N., & Robertson, I. (1990). Personnel selection. *Annual Review of Psychology, 41*, 289–319.

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. New York: Appleton-Century-Crofts.

Smith, I. M. (1964). *Spatial ability: Its educational and social significance*. San Diego, CA: Knapp.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (1st ed., pp. 745–775). Chicago: Rand McNally.

Snow, R. E. (1980). Aptitude processes. In R. E. Snow, P. A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction: Vol.1. Cognitive analyses of aptitude*. Hillsdale, NJ: Erlbaum.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. G. Glasser, *Learning and individual differences: Advances in theory and research* (pp. 13–59). New York: Freedman.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–104). Hillsdale, NJ: Erlbaum.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Eds.), *Educational measurement* (3rd ed., pp. 263–331). New York: Collier Macmillan.

Snow, R. E., & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 493–585). Cambridge, England: Cambridge University Press.

Spearman, C. (1904). "General intelligence": Objectively determined and measured. *American Journal of Psychology, 15*, 201–292.

Spearman, C. (1914). Theory of two factors. *Psychological Review, 21*, 101–115.

Spearman, C. (1927). *Abilities of man: Their nature and measurement*. New York: Macmillan.

Spearman, C., & Jones, L. L. (1950). *Human ability*. London: Macmillan.

Spence, K. W. (1948). The postulates and methods of "behaviorism." *Psychological Review, 55*, 67–78.

Stanley, J. C. (1971). Predicting college success of the educationally disadvantaged. *Science, 171*, 640–647.

Stanley, J. C. (1983). Introduction. In C. P. Benbow & J. C. Stanley (Eds.), *Academic precocity: Aspects of its development* (pp. 1–8). Baltimore, MD: Johns Hopkins University Press.

Stanley, J. C. (1991). Personal communication from a panel discussion. In *The Henry B. and Jocelyn Wallace National Research Symposium on Talent Development*. Iowa City, IA.

Stanley, J. C., Benbow, C. P., Brody, L. E., Dauber, S., & Lupkowski, A. (1992). Gender differences on eighty-six nationally standardized achievement and aptitude tests. In N. Colangelo, S. G. Assouline, and D. L. Ambroson (Eds.), *Talent development: Proceedings from the 1991 Henry B. and Jocelyn Wallace national research symposium on talent development* (pp. 42–65). New York: Trillium Press.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychology, 36*, 1181–1189.

Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences, 7*, 269–287.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (in press). Theory-based testing of intellectual abilities: Rationale for the Sternberg triarchic abilities test. In H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement*. Hillsdale, NJ: Erlbaum.

Taylor, R. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23*, 565–578.

Tellegen, A. (1985). Structure of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681–706). Hillsdale: NJ: Erlbaum.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality Assessment, 56*, 621–663.

Tellegen, A., Kamp, J., & Watson, D. (1982). Recognizing individual differences in predictive structure. *Psychological Review, 89*, 95–105.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30*, 47–54.

Tenopyr, M. L., & Oeltjen, P. D. (1982). Personnel selection and classification. *Annual Review of Psychology, 33*, 581–618.

Thomson, G. (1951). *The factor analysis of human abilities* (5th ed.). New York: Houghton Mifflin.

Thorndike, E. L. (1906). Sex in education. *The Bookman, 23*, 211–214.

Thorndike, E. L. (1926). *The measurement of intelligence*. New York: Bureau of Publications, Teachers College, Columbia University.

Thorndike, R. L. (1985). The central role of general ability in prediction. *Multivariate Behavioral Research, 20*, 241–254.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29*, 332–339.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs* (No. 1).

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs* (No. 2).

Tryon, R. C. (1935). A theory of psychological components—an alternative to "mathematical factors." *Psychological Review, 42*, 425–454.

Tyler, L. E. (1953). *The work of the counselor*. New York: Appleton-Century-Crofts.

Tyler, L. E. (1965). *The psychology of human differences* (3rd ed.). New York: Appleton-Century Crofts.

Vernon, P. E. (1947). Research on personnel selection in the Royal Navy and British Army. *American Psychologist, 2*, 35–51.

Vernon, P. E. (1950). *The structure of human abilities*. Andover Hants, England: International Thompson Publishing Services.

Vernon, P. E. (1961). *The structure of human abilities* (2nd ed.). London: Methuen London Ltd.

Vernon, P. E. (1982). *The abilities and achievements of Orientals in North America*. New York: Academic Press.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Wallace, J. (1965). An abilities conception of personality: Some implications for personality measurement. *American Psychologist, 20*, 132–138.

Wallace, S. R. (1965). Criteria for what? *American Psychologist, 20*, 411–417.

Wallach, M. A. (1976). Tests tell us little about talent. *American Scientist, 64*, 57–63.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience negative emotional states. *Psychological Bulletin, 96*, 465–490.

Wherry, R. J. (1952). *The control of bias in ratings: A theory of rating* (Personnel Research Board Rep. No. 922). Washington, DC: Department of the Army, Personnel Research Section.

Wherry, R. J. (1959). Hierarchical factor solutions without rotations. *Psychometrika, 24*, 45–51.

Wherry, R. J. (1983). Wherry's theory of ratings. In F. J. Landy & F. L. Farr (Eds.), *The measurement of work performance: Methods, theory and applications* (pp. 283–303). New York: Academic Press.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*, 521–551.

Wigdor, A. K., & Garner, W. R. (1982). *Ability testing: Uses, consequences, and controversies*. Washington, DC: National Academy Press.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Wise, L. L., McLaughlin, D. H., & Steel, L. (1979). *The Project Talent Data Handbook*. American Institutes for Research, Palo Alto, CA.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist, 17*, 657–658.

Zedeck, S. (1971). Problems with the use of moderator variables. *Psychological Bulletin, 76*, 295–310.

Zedeck, S., & Cascio, W. F. (1984). Psychological issues in personnel decisions. *Annual Review of Psychology, 35*, 461–518.