Minding our p's and q's:

On finding relationships between

learning and intelligence

David F. Lohman

The University of Iowa

Draft of November 20, 2002

Paper presented at the Minnesota Conference on "The future of learning and individual difference research", University of Minnesota, October, 1997.

My contributions to this conference will be more conceptual than empirical. I will discuss conditions under which we might observe relationships between the constructs of learning and intelligence and reasons why we generally should not expect strong relationships between them. The history of attempts to relate these two constructs is remarkable in several respects. If there is a theme, it is the triumph of belief over evidence: i.e., belief that learning and intelligence ought to be strongly related versus the evidence that they are only weakly related.

The expectation of strong relationships between learning and intelligence is based on several logical, statistical, and conceptual confusions. The primary logical confusion is between learning and attainment; the primary statistical confusion is between row and column deviation scores in data matrices (or, more subtlely, between the mean of a deviation score and its variance or covariance); and the primary conceptual confusion is among psychological constructs defined by different aspects of score variation.

Background

From the earliest days of testing, scores on intelligence tests have been interpreted primarily as measures of scholastic aptitude, i.e., as the measures of the ability to learn the knowledge and skills taught in schools in the manner in which such things have been taught in schools. When used in this way, there is a clear expectation of relationship between scholastic aptitude and scholastic attainment and, to the extent that rank orders of students on measures of scholastic attainment vary over time, between scholastic aptitude and individual differences in scholastic learning. Those who interpret measures of intelligence more broadly have expected intelligence tests to predict individual differences in learning in any context. Several contributors to the 1921 symposium on intelligence offered definitions of intelligence as simply the ability or capacity to learn (see, e.g., Thorndike, 1921). It is this broader definition that has guided most research. In other words, most systematic efforts to explore relationships between learning and intelligence have sprung not from the school but from the laboratories of experimental psychologists and thus have focused on the learning of laboratory tasks rather than the learning of academic tasks. Woodrow's (1946) studies exemplify this tradition. The results of this effort are well known. His conclusions not only summarized his work, but also set the stage for subsequent efforts to relate learning and intelligence. The conclusions were:

1. The ability to learn cannot be identified with the ability known as intelligence.
2. Individuals possess no such thing as a unitary general learning ability.
3. Improvement with practice correlates importantly with group factors, that is, relatively narrow abilities, and also with specific factors.
4. Even the group actors involved in learning are not unique to learning but consist of abilities which can be measured by tests given but once. (Woodrow, 1946, pp. 148-149)

Responses to Woodrow

Woodrow's first two conclusions ran so completely counter to beliefs about the relationship between learning and intelligence that they were simply ignored by most differential psychologists and psychometricians (Cronbach & Snow, 1977, p. 111). One of the few who seemed to notice was Gulliksen who, with his students and colleagues (Allison, 1960; Stake, 1961; Bunderson, 1967), doggedly pursued the problem in a series of studies. Others joined the effort, or initiated independent attacks on the problem (see Gulliksen, 1961, for an overview of the early work). Their responses to Woodrow's challenge mostly fell in one of six categories: (1) that intelligence is not unitary; (2) that learning is not unitary; (3) that Woodrow's learning tasks were inappropriate; (4) that gain scores are bad; (5) that learning should be represented by attainment rather than by gain, and (6) that our measurement scales are inadequate. I briefly summarize each of these suggested solutions to the problem, and then offer my own.

1. Intelligence is not unitary. To those trained in the shadow of Thurstone (such as Gulliksen), the most obvious limitation of some of the early investigations was the failure to decompose intelligence into the proper sort of primary mental abilities. Although Woodrow (1946) had included group ability factors in his work (see Simrall, 1946), others believed his

tests and methods of factor analysis were inadequate. Correlations between learning and ability were indeed higher when learning tasks and ability tests were more carefully matched, although it now appears that the Cattell-Horn ability theory is more useful than Thurstone's theory for teasing out ability-learning relations (Snow, Kyllonen, & Marshalek, 1984). Nonetheless, replacing G with three broad group factors or with seven or more primary ability factors was at best a partial solution.

2. Learning is not unitary. In one of the earliest discussions of relationships between intelligence and learning, Thorndike, Bregman, Cobb, and Woodyard (1926) noted that most theorists distinguished between lower-order association processes and higher-order generalization, abstraction, and reasoning processes. One might reasonably expect individual differences in these different cognitive functions to show different relationships with intelligence (although Thorndike actually argued otherwise). Jensen (1973) offered a similar distinction between Level I and Level II learning abilities and their expected ability correlates. An even more extreme splintering was suggested by experiments showing low correlations among learning rate measures on different tasks (Woodrow, 1946), or when learning rate was estimated from different dependent variables (Cronbach & Snow, 1977). Nevertheless, many investigators hoped that stronger, more consistent relations between learning and intelligence would be found if the proper measure (or measures) of learning were employed.

Most significant in this respect were the attempts to define learning not by a simple gain score computed from two waves of data, but by the parameters of a learning curve fitted to each subject's performance over many trials. The preferred measure of learning was typically the curvature of this function, or the point at which its first derivative was maximum (i.e., the slope was steepest), indicating maximal rate of learning (Woodrow, 1946). A variant of this theme was to compute the slope at two points, one estimating early learning and the other late learning (Allison, 1960).

Curve fitting to individual data is a good idea, for both theoretical and methodological reasons (Bock, 1991; Estes, 1956; Rogosa, Brandt, & Zimowski, 1982; Woodrow, 1946). It is also useful to distinguish among various aspects of learning (such as initial level of performance, rate of improvement, and final level of performance) and to do so for different aspects of learning, retention, and transfer (Ferguson, 1956). But such methods are not without their problems. For example, in any curve-fitting exercise, subjects vary in the extent to which their data is well-described by functions fitted to the data. Should parameters of the learning curve be used only for those subjects well-fit by the model? If so, how should this be defined? More importantly, learning rate measures not only have the same sort of reliability, ceiling, and floor problems as gain scores, but also show extremely skewed distributions. Interpretation of correlations between rate measures and other variables is thus fraught with difficulties. Nonetheless, rate measures have sometimes shown interesting and interpretable patterns of relationships with ability variables (Snow, Kyllonen, & Marshalek, 1984).

3. Learning tasks are inappropriate. A common critique of the Woodrow (1946) studies is that the learning tasks he used were too simple or too short (Humphreys, 1979). Indeed, learning on more complex tasks (such as concept attainment tasks) often shows stronger correlations with ability. Taxonomies of learning tasks (e.g., Kyllonen & Shute, 1989; Gagne, 1965, Bloom, 1956) can help guide these efforts to discover ways in which task characteristics moderate ability learning relationships. Nevertheless, even on complex tasks, correlations between learning and intelligence are usually modest unless scores on the pretest are uniformly low or distributed randomly. Then, correlations between the learning gain score and other variables will be attenuated versions of the correlations between post (or final attainment) scores and these variables. Which brings us to the troublesome and confusing problem of gain scores.

4. Gain scores are bad. Probably the most common critique of Woodrow's (1946) studies--and many that followed in his wake--is that gain scores are unreliable, thereby severely attenuating relationships with other variables. For some, this means that the unreliable gain score should somehow be made more reliable, usually through disattenuation of its correlations with other variables. For example, Cronbach and Snow (1977) showed how

disattenuation could transform negative correlations between MA and yearly gains in MA to moderately positive correlations. Yet in a summary discussion, they eschew gain altogether and advise that

> Outcomes in learning...ought to be expressed in terms of level (i.e., attainment) scores collected at some terminal point (and perhaps at intermediate points also) (Cronbach & Snow, 1977, p. 116).

But in the very next sentence, the same authors advise treating initial status as a covariate, thereby bringing back a modified gain score. Clearly, there is some confusion about what to do. On the one hand, individual differences in gain scores are unreliable. On the other hand, learning is defined by a change in performance with practice. It is thus naturally operationalized by a gain score or, better, by a curve fitted to multiple waves of data, each representing performance at a particular point in time. In the simplest case in which the investigator has only two waves of data, the observed gain is an unbiased estimator of true gain. From a statistical point of view, then, raw gain scores are preferred over residualized gains as estimators of true gain (Rogosa et al., 1982). Further, the precision of estimated gain must not be confused with the reliability of individual differences in gain. For example, in the case in which all subjects show the same gain, we can say precisely what this gain is even though the reliability of individual differences in gain will be zero. The reliability coefficient indicates only the accuracy with which individuals can be ranked on the basis of the score. If there is little variability in gain, the reliability of individual differences in gains will be low even though the precision of each estimated gain is high. Conversely, if the variability in true gain is substantial, then reliability of gain scores can also be substantial.

The conditions under which this is likely occur can be deduced from an examination of the formula for the reliability of gain scores, which is shown as Equation 1 below.

$$\rho GG' = \frac{\rho_{11'}\sigma_1^2 + \rho_{22'}\sigma_2^2 - 2\,\rho_{12}\,\sigma_1}{\sigma_1^2 + \sigma_2^2 - 2\,\rho_{12}\,\sigma_1\sigma_2} \qquad (1)$$

The reliability of gains ($\rho_{GG'}$) is a complex function of the reliability of the pretest ($\rho_{11'}$), the reliability of the posttest ($\rho_{22'}$), the variance of the pretest (σ_1^2) and of the posttest (σ_2^2), and the correlation between pretest and posttest (ρ_{12}). One way to understand complex equations is to make some simplifying assumptions. The most common simplifying assumptions here are (a) the pretest and posttest have approximately the same reliabilities, and (b) that the variance of scores is uniform across time. Under these assumptions, Equation 1 reduces to Equation 2:

$$\rho GG' = \frac{\rho_{xx'} - \rho_{12}}{1 - \rho_{12}} \qquad (2)$$

where $\rho_{xx'}$ is the common pretest, posttest reliability. This formula shows clearly that the reliability of the gains decreases as the correlation between the two scores increases. In fact, $\rho GG'$ drops to zero when ρ_{12} equals $\rho_{xx'}$. Thus, it would seem desirable to reduce the pretest-posttest correlation in order to improve the reliability of the observed gains. However, some have argued that a high correlation between pretest and posttest (here, ρ_{12}) is a necessary condition for test validity. The assumption seems to be that a test that does not show good stability may not be measuring the same trait on both occasions. This is the so-called reliability-validity paradox of gain scores. It is a short step from Equation 2 to a general indictment of gain scores. But one gets to this unhappy place in part by the assumption that score variances should be equal and that stability of individual differences over time should be high. Neither assumption is particularly reasonable when significant learning is interposed between pretest and posttest. Indeed, the variability of scores often changes dramatically over time. For achievement test scores, variances tend to increase systematically over the grade school years. Figure 1 shows this graphically for the Reading Vocabulary subtest of the Iowa Tests of Basic Skills.

Insert Figure 1 and Table 1 here

Kenny (1974) dubbed this pattern of increasing score variance the fan spread effect. It occurs in large measure because true gains are weakly but significantly correlated with initial status.

What is the effect of unequal variances on reliability of gains? Table 1 shows how the reliability of gain scores increases as the ratio of posttest to pretest standard deviation increases, and as the correlation between pretest and posttest declines from the average reliability of the pre- and posttests. The good news, then, is that gains can be quite reliable if there is significant variability in true gain. The bad news is that this does not happen quickly, at least for the sorts of broad abilities students acquire through formal schooling. Table 2 and Figure 2 show some of the correlations Martin (1985) obtained in his investigation of gains on the Iowa Tests of Basic Skills (ITBS) for a sample of 6,321 Iowa students retested every year from third through eighth grade. This particular table shows correlations between gains in different aspects of school achievement and general ability. One-year gains showed small correlations with ability, especially for individual subtests of the ITBS; but five-year gains showed substantial correlations with ability. [1]

Insert Table 2 and Figure 2 here

Thus, there is some confusion about gain scores. Sometimes we are admonished to avoid them altogether--or better--to reformulate our questions in ways that do not require that we use such measures (Cronbach & Furby, 1970); at other times we are admonished to disattenuate, or to examine relationships among latent rather than observed variables. Implicit in the first suggestion is the idea that learning could somehow be operationalized without contrasting initial and final performance; implicit in the second is the idea that the only thing separating constructs of learning and intelligence is error of measurement. Both suggestions mislead. Although there are times when attainment scores are to be preferred over learning scores, there are also times when the learning score is needed. Similarly, although errors of measurement are a problem, they are not the only reason, or even the main reason, for the relative independence of the constructs of learning and intelligence. About which more later.

5. Learning should be represented by attainment scores. One response to the measurement difficulties posed by gain scores is to avoid them altogether. There are indeed may times when questions may properly be formulated in terms of attainment rather than learning. A common example are questions about individual differences in what individuals can do at a particular point in time. Although one may infer that doing has something to do with learning, individual differences in attainment scores may reflect primarily individual differences in initial status. Indeed, individual differences in learning may be small even though learning is substantial. In other words, although attainment invariably reflects learning, individual differences in attainment may better reflect individual differences in initial status than individual differences in learning.

A variant on this scheme is to record attainment scores at various stages of practice or learning, and then to examine both the intercorrelations of the attainment scores and their correlations with other variables. This is precisely the methodology adapted by many who have investigated changes in ability correlations during skill acquisition (Fleishman, 1972; Ackerman, 1989). Fleishman kept all variables in one matrix; Ackerman follows Humphreys' (1960) recommendation to analyze attainment scores from the learning task separately. Attainment scores do not directly compare final status with initial status. Therefore, such analyses reflect individual differences in learning only to the extent that they reveal changes over time in correlations between attainment scores and ability factors. This happens only if the matrix of learning scores is not of unit rank, which is generally the case. Analyses of the external correlates of attainment scores are not only informative in their own right, but also avoid many of the difficult measurement problems that attend the use of learning scores. However, such analyses often show that the assumption that we are measuring the same thing early and late in learning may be unwarranted. Which brings us to the scaling problem.

6. Our scales are inadequate. One of the earliest criticisms of the Binet Scale was that it did not appear to be measuring the same thing for young children as for older children because there was little overlap in the tasks presented at different ages (Yerkes & Anderson, 1915). Yerkes thus proposed that intelligence tests contain subtests of relatively homogeneous items, such as series completion or vocabulary or analogies, that could be administered to all examinees. Yerkes' arguments and the tests he helped create (e.g., The Army Alpha and Beta, the National Intelligence Test) significantly altered testing in America. Later investigators often worried about the purity of the scales defined by ability tests, and some even proposed methods for constructing equal-interval scales. Concerns about the scale move center stage, however, when one seeks to estimate learning or growth. Does a 10-point gain mean the same thing at all points along the score scale? And if not, then how can gains be correlated meaningfully with other variables.

Recently, some have suggested that scales constructed by Item Response Theory (IRT) methods attenuate or even solve these problems (e.g., Embretson, 1991). Others, however, point to the fact that IRT-based ability estimates for high- or low-scoring examinees are quite unstable. For example, for the three-parameter logistic model, measurement error variance for extreme-scoring examinees can be 10 or even 100 times that for more typical examinees (Lord, 1980). Even more troubling is the fact that such methods can transform a raw score scale that shows a systematic increase in score variance with age into one that shows marked decrease in score variance with age (Hoover, 1984). Although score variance can reasonably remain steady or even decrease over time for closed-ended skills, there is little empirical or conceptual support for such effects in the complex, open-ended skills measured on school achievement tests. By the end of the grade-school years, some low-scoring children are still struggling with elementary addition and subtraction whereas their high-scoring peers are solving algebra problems. It makes little sense to say that somehow the variability in mathematics achievement or vocaubulary knowledge has declined over the grade school years. Yet this is precisely what happened when IRT methods were used to scale the California Achievement Tests (Hoover, 1984).[2]

The problem is not that IRT methods are flawed, but rather that the model used may not really fit the data. Hoover's (1984) critique was aimed at attempts to fit a three-parameter logistic model to a heterogeneous achievement test. Furthermore, since no child takes the entire test, between-grade (or developmental) scales on such tests are often constructed by pasting together a chain of overlapping within-grade (or level) scales. Scaling problems at one level are compounded at the next. On the other hand, ability tests on which both person variance and item difficulty variance are unidimensional can be nicely scaled using the (much simpler) Rasch model, even when a multi-age developmental scale is constructed from a series of overalpping within-age scales (see, eg. Thorndike & Hagen, 1997, for one example). Athough we certainly need to devote more attention to the difficult issues of scale construction than we currently do, and even though IRT scales have many desirable properties and are known to work well when the underlying model fits the data, it is unlikely that IRT scales will resolve debates about the meaning of gain scores any more than improved methods of factor analysis have resolved debates about the number of ability factors and their organization (see also Cliff, 1991).

An even more difficult issue is the fact that scales are often bounded, and so gains are limited. For example, when latency is the dependent measure, those who respond faster in the initial task can improve less than those who respond more slowly. Transformations that linearize the scale provide some remedy, and probably should be employed more routinely than they are. At the very least, it needs to be demonstrated that effects attributed to differential gains are not as easily explained by differences in initial response latencies.

Summary. There was, and continues to be, an expectation that intelligence is related to learning. This assumption is particularly strong when intelligence is construed as scholastic aptitude and learning is assessed by complex educational tasks. Thus, claims by Woodrow (1946) that intelligence was unrelated to learning elicited either disbelief or a variety of repair strategies (cf. Brown & Burton, 1978). Some proposed that the problem lay in a conception of intelligence as unitary rather than as multiple. Others focused on the measures of learning and emphasized the construction of learning curves for extended, complex tasks. Others emphasized the unreliability of gain scores or the assumptions of interval scaling. Still others moved away from the measurement of learning per se and examined relations between ability tests and attainment scores at various stages of practice. And although each of these repair strategies has been shown to be effective to one degree or another, none has done so consistently and dramatically. As Cronbach and Snow (1977) put it "the results have been mixed rather than obviously coherent" (p. 133).

A Resolution?

Every now and then vigorous dispute about some experimental result can be quelled by the recognition of what (in retrospect, at least) is an obvious statistical necessity. For example, Bloom (1964) was puzzled by negative correlations between initial IQ and gain in IQ. However, correlated errors will produce such a result. More importantly, IQ scores have fixed SD's, and so highs cannot (on average) improve their scores since that would increase the SD. Rather, they must regress toward the mean, which contributes to the negative correlation. Thus, the negative correlation between gain in IQ and initial IQ is a statistical artifact, not a substantive finding.

Several years ago I experienced a similar dawning of the obvious as I puzzled over information processing analyses of ability tests (Lohman, 1994). The goals of this research were (a) to understand the mental processes subjects used when attempting to solve items on these tests, and (b) to explain overall individual differences on the tests in terms of individual differences on the component mental process. Although investigators were generally quite successful in meeting the first goal, they were less successful in achieving the second. In particular, although process-inspired models often showed good fits to the data, scores estimating the speed or efficiency with which subjects performed particular mental processes showed inconsistent and often low correlations with reference ability constructs (Carroll, 1980). The reason component scores fail to decompose individual differences on tasks is the same reason individual differences in learning are at best weakly-related to individual differences in intelligence.

Imagine a simple person-by-occasion data matrix whose entries X_{po} represent the scores of n_p persons on a task administered on n_o occasions. Figure 3 shows how the variability in scores may be partitioned into three sources: the person source, the occasion source, and the person x occasion/residual source. The person source represents variability in row means, that is, in the average attainment scores of persons on the tasks. The occasion source represents variability in column means, that is, in average differences in performance on the tasks across time. The person-by-occasion interaction captures systematic differences in the way subjects responded on each occasion that are not captured by the person or occasion sources. In the case of two occasions, it represents variability across persons in learning or gain scores. Changes in variability from pretest to posttest will be reflected in the person by occasion component as well. Note, however, that average learning is captured by the occasion source, and individual differences that are consistent across time are captured by the person source.

Insert Figure 3 here

Individual differences are represented by sources of variance that lie within the p circle in Figure 3. There are three claims here. First, if the goal is to understand individual differences on a task, then we are interested in <u>all</u> of the sources of variance in the p circle. Second, the p variance component is generally quite large when homogeneous ability tests or learning tasks are administered to samples that vary widely in ability. Third, learning or gain scores help explain variance in the i and pi variance components, not the p variance component. The mean of these learning scores is reflected in the o component; their variance in the p x o component. In other words, the primary contribution of individual learning scores is to help explain individual differences that are independent of both overall individual differences on the task and average learning on the task. Individual differences in learning scores help capture systematic variance from the p x o component that are ignored when performance is represented by average attainment. However, the learning or gain scores do not decompose and, therefore, cannot help explain the typically much larger p variance component. This is why attainment scores (initial, final, or average) show correlations with ability variables when gains or rate parameters often do not.

When are individual learning scores most useful? Generally, such scores are most useful when the p variance component is relatively small and the p by o variance component is relatively large. In the extreme, this would be on a task on which everyone received the same total score ($\sigma^2_p = 0$) and on which individual differences were entirely reflected in differential pre- to posttest gains. In the language of interclass correlations, this means that the correlation between pre- and posttest would be small and the variability in true gains large.

A parallel scenario obtains when the goal is to estimate scores for mental processes. Such scores will be most useful when the person variance component is relatively small and the person by item variance component is relatively large. For example, Ippel (1986) found that the person by item (or person by task) variance component was large when different embedded figures tasks were administered simultaneously. For any task (or combination of tasks), measures of internal consistency (such as coefficient alpha) provide an estimate of the relative magnitudes of the person and person by facet interactions. When α is large, then p x i must be small. When α is small, then p x i may be large. Thus, if the goal is to measure individual differences in mental processes, then one should look for tasks that do not exhibit a high degree of internal consistency.

Like component scores for mental processes (e.g., Sternberg, 1977), learning is defined by some measure of within-person change. This can be a simple difference score or a more complexly computed difference score (such as a slope). (That slopes are difference scores is most readily seen in an ANOVA model in which a linear trend is estimated by multiplying cell means by coefficients such as -3, -1, 1, 3. One is simply computing a non-unit weighted difference score, but a difference score nonetheless)[3] . The main conceptual difference is that whereas component mental processes estimated on one task are thought to explain performance on that same task (e.g., Sternberg, 1977), learning scores on one task are typically related to status or attainment scores on another task -- at least when learning is correlated with intelligence. And although within-subject deviation scores may be independent of between-subject deviation scores on one task, these same within-subject deviation scores may indeed be related to between subject scores on another task. The key variable is the relationship between the between-subject scores on the two tasks. The relationship between learning scores and ability test scores will be limited to the extent that attainment scores on the learning task are highly correlated with status scores on the ability test. The limiting case is when the disattenuated correlation between attainment scores on the learning task and ability test is unity. In other words, from a statistical standpoint at least, learning scores on one task are more likely to relate to ability scores on another task when overall performance on the two tasks is not highly correlated. However, a psychological analysis would seem to lead to precisely the opposite conclusion. Perhaps this is the reliability-validity paradox writ in still another form, or perhaps it is the reason why (for example) Martin (1985) observed small correlations between ability and differential gains in achievement over an entire school year, whereas Woltz (this volume) finds relatively large correlations between repetition priming and academic achievement.

To recapitulate, then, the problem is not that measures of learning are unreliable (which they are), or that learning should be estimated from individual learning curves rather than simple gain scores (which is a good idea), or that learning needs to be measured on an interval scale (which is also a consummation devoutly to be wished). Nor is the problem that ability should be represented severally rather than singly. In short, my claim is that the primary problem lies not in the measurement of ability or learning, but in a more fundamental conceptual confusion about the meaning of constructs defined by different, often quite independent aspects of score variation. Further, the arguments apply with equal force to any measure of learning that is operationalized by subtracting one variable from another, whether the two variables are latencies, errors, or PET scan images; whether each is a simple sum of performance on several trials or the product of a more complex function; whether the theory that guides the process is behavioral, cognitive, or even socio-historical.

Critics of the arguments I have advanced here often point to cases in which measures of component mental processes or learning gain scores show moderate or even high correlations with other variables. (See, for example, the correlations between five-year gains in achievement and ability shown in Table 2). This can happen in a variety of ways, some of which are artifactual and some of which are not. First, it is important to emphasize that even relatively small amounts of variation in within-person scores can show substantial correlations with other variables, especially when correlations are disattenuated or reported as path coefficients among latent variables. Even in those cases where such relationships are replicable, they may not mean what they seem to mean. A common problem--especially when latency is the dependent measure--is that the variability of scores is much greater in one condition than in another condition, and the difference in variability is not due to learning but to differences in task demands or scale restrictions. For example, on mental rotation problems (Shepard & Metzler, 1971), variance in response latencies increases dramatically with amount of rotation required. A score estimating rate of rotation is typically computed by regressing response latency on angular separation between figures. This slope will show high correlations with time taken to solve problems requiring the most rotation. In such cases, correlations between individual slope scores and other measures may be little more than an attenuated version of the correlation between time required to solve problems in the condition with the greater variance. The limiting case here is when the variance is zero in one condition (i.e., all subjects have the same score). If scores in this condition are subtracted from scores in another condition, then the difference score will merely reflect the rank order of individuals in the condition with the non-zero variance. For example, if individuals are initially unable to perform a task but differ markedly in performance after practice, then gains will be highly correlated with final status. This is the model that seems to underlie many writer's expectations for strong relationships between learning and intelligence. What it fails to take into account is the fact that, on virtually any complex task, there will be substantial individual differences in initial performance and these differences will be correlated both with final performance and with academic aptitude (see Woodrow, 1946, for discussion and illustration).

A less extreme case occurs when the variance increases slightly across any two intervals, such as the year-to-year increases in score variance commonly observed on achievement tests (see Figure 1 and Table 1). It occurs because, although all children improve with education, high-scoring children tend to improve more. Over short intervals, however, differential gains are not large and are easily swamped by the doubly-error laden difference score. Over long periods, however, differential gains can be substantial. For example, Cronbach and Snow (1977) estimated that true mental age at age 6 correlated approximately r = .6 with true gain in MA between ages 6 and 7 in the Bayley (1949) data.

Construct Confusion

The difficulties in relating measures of learning and intelligence are but one example of a larger confusion in psychology. Many of us tend to think of all psychological constructs as belonging to the same conceptual category when in fact they may be conceptually and statistically independent. Although constructs in differential psychology are invariably defined by individual difference variance, constructs in other domains may be defined by changes in performance across conditions, by rules that map scores on to content domains or absolute scales, and in other ways that reflect individual-difference variance incidentally rather than directly (and may even obscure it altogether).

A more systematic accounting for what sort of variability is represented by different constructs may help us keep track of our constructs and keep in line our expectations for relationships among them. Learning and ability scores are defined by different partitionings of a simple person x occasion data matrix. Personality, developmental, and style variables complicate the picture. Figure 4 shows a modified version of Cattell's (1966) covariation chart: persons x items (nested within tasks) x occasions (or situations). Differential psychologists typically worry about person main effects (or covariation of person main effects across several tasks). Experimental psychologists are less uniform. Those who follow an information-processing paradigm worry about variation over trials with a particular task. Situationalists, however, worry more about covariation of either task main effects (e.g., delay versus no delay of reinforcement) or person main effects across occasions; they typically emphasize the magnitude of the former relative to the magnitude of the latter. Developmentalists do the opposite. Then there are those who worry about interactions. The point is that psychological constructs may be arrayed on several, not just two dimensions, and certainly not just one dimension. Person x situation is not the same as person x items within task. When constructs are defined by individual difference variance, then experimental analyses of tests may not tell us much about the source of these individual differences unless subjects solve the tests in different ways and the experimental analyses can identify them. Similarly, when a construct is defined by condition or stimulus variance, then correlating individual scores on a task with other variables may not tell us much about it either.

Insert Figure 4 here

Investigations of the generalizability of constructs show this confusion most clearly. The differential psychologist knows how to estimate the generalizability of individual differences across tasks, but that is not the aspect of generalizability that should most interest the experimentalist. In addition to the generalizability of individual differences across tasks, one can examine the consistency across tasks of treatment effects or even of score profiles (see Cronbach, 1957, for an example; also Cattell, 1966). In other words, the experimentalist should be more interested in covariation of response patterns between rows of the data matrix; not between columns, like the differential psychologist. Unfortunately, since the psychometrician is usually more adept at multivariate statistics, efforts to link experimental and differential psychology usually end up playing by differential rules. Entire research programs attempting to link experimental and differential psychology have risen and then collapsed on the basis of a few between-person correlation coefficients. A better strategy would be to exploit the separate strengths of the experimental and differential traditions rather than trying to reduce them to a common enterprise. In other words, one might look to an experimental analysis to explain how subjects solved tasks or learned from their exposure to them. Although such analyses can usefully inform our understanding of individual differences when individuals differ in the parameters of a common model, they are most informative when qualitatively different models are needed to describe the performance of different individuals, and these differences show systematic relations with overall performance on the task or with other variables. Stage-theoretic models of cognitive development (such as the one advanced by Piaget) are one example. Baddeley's neuropsychological examples (this volume) are another. In either case, logic and argumentation would seem to offer less hazardous avenues of commerce between the disciplines of psychology and their constructs than do between-person correlation coefficients, no matter how carefully estimated and adjusted.

Conclusions

So where does this lead us?

1. The fact that individual differences in measures of learning are generally weakly but positively related to measures of intelligence is an interesting fact of life, not a problem to be explained. Indeed, were the two not related at all, or more strongly related, then we would have more to explain. If learning and intelligence were not related, then differential maturation rates would constitute the only explanation for imperfect relationships among true status scores on mental tests over time (assuming, of course, that we can trust our scales). If learning and intelligence were more strongly related, then the differences between high and low ability learners would show a positively accelerated growth over time rather than the gentle fan that they seem to show.
2. The simple gain score is often the best measure of learning. If the data permit, then more sophisticated functions can be fitted to each subject's learning data and parameters of this function used to summarize the course of learning for each individual. (Aggregation of individuals comes later.) Gain scores should not be eschewed because of their low reliability. Indeed, ceiling effects (especially for accuracy) and floor effects (especially for latency) probably present more serious problems. Analyses that allow estimation of relations among latent variables can be performed and will often show patterns quite unlike those observed among raw scores. When this happens, it would seem wise to report both.
3. Although gain scores are not to be eschewed, neither should they be used when questions more properly concern changes in relationships of attainment scores with other variables over time as, for example, in Ackerman's (1987, 1989) studies of skill acquisition.

4. The expectation of strong relationships between learning and intelligence is based on several logical, statistical, and conceptual confusions. The primary logical confusion is between learning and attainment. Indeed, although they are usually moderately correlated, individual differences in learning may be independent of individual differences in attainment. This illustrates the primary statistical confusion, which is between row and column deviation scores in data matrices, or between within-person and between-person variation. More subtlely, it is between the mean of a deviation score and its variance or covariance. Finally, the primary conceptual confusion is among psychological constructs defined by these (and other) different aspects of score variation.

* * * * * * *

In the old days, typesetters assembled words from individual letters, each stored alphabetically in small compartments. P's and q's were thus stored in adjacent bins, and would appear reversed when arranged on the composing stick. It was easy to confuse them. Apprentices who disassembled type and re-stored the letters were thus admonished to "mind their p's and q's." A more colorful tale is that tabs in pubs once consisted of lists of p's (for pints) and q's (for quarts). Both the inn keeper and the drinker would have a stake in making sure p's were not confused with q's when the tally was reckoned. Like apprentice typesetters or wary pub sitters, we must attend carefully to the orientation of our data matrices. Constructs that explain much of the within-person variation on a task need not explain much of the between-person variation either on that task or on other tasks.

References

Ackerman, P. L. (1989). Individual differences and skill acquisition. In P.L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences: Advances in theory and research (pp. 164-217). NY: W. H. Freeman & Co.

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

Allison, R. B. (1960). Learning parameters and human abilities. Unpublished report, Educational Testing Service. UM 60-4958.

Anderson, J. E. (1939). The limitation of infant and preschool tests in the measurement of intelligence. Journal of Psychology, 8, 351-379.

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. Journal of Genetic Psychology, 75, 165-196.

Bloom, B. S. (1964). Stability and change in human characteristics. New York: Wiley.

Bock, R. D. (1991). Prediction of growth. Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 126-136). Washington, DC: American Psychological Association.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Bunderson, C. V. (1967). Transfer of mental abilities at different stages of practice in the solution of concept problems. Research Bulletin 67-20, Educational Testing Service. UM 66-4986.

Carroll, J. B. (1980). Individual differences in psychometric and experimental cognitive tasks (NU 150-406 ONR Final Report). Chapel Hill, NC: University of North Carolina, L. L. Thurstone Psychometric Laboratory.

Cattell, R. B. (1966). Handbook of multivariate psychology. Chicago: Rand McNally.

Cliff, N. (1991). Comments on "Implications of a multidimensional latent trait model for measuring change" by S. E. Embretson. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 198-201). Washington, DC: American Psychological Association.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 761-684.

Cronbach, L. J. & Furby, L. (1970). How should we measure "change" -- or should we? Psychological Bulletin, 74, 68-80.

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

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 184-197). Washington, DC: American Psychological Association.

Estes, W. K. (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53, 134-140.

Ferguson, G. A. (1956). On transfer and the abilities of man. Canadian Journal of Psychology, 10, 121-131.

Fleishman, E. A. (1972). On the relation between abilities, learning, and human performance. American Psychologist, 11, 1017-1032.

Gagne, R. M. (1965). Problem solving. In A. W. Melton (Ed.). Categories of human learning. NY: Academic Press.

Gulliksen, H. (1961). Measurement of learning and general abilities. Psychometrika, 26, 93-107.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. Educational Measurement: Issues and Practices, 3, 8-14.

Hoover, H. D., Hieronymous, A. N., Frisbie, D.A., & Dunbar, S.B. (1993). Iowa Test of Basic Skills: Norms and scire conversions (Form K). Chicago: Riverside Publishing Company.

Humphreys, L. G. (1960). Investigation of the simplex. Psychometrika, 25, 313-323.

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

Hunt, E. G., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), The psychology of learning and motivation (Vol. 7). New York: Academic.

Ippel, M. J. (1986). Component-testing. Amsterdam: Free University Press.

Jensen, A. R. (1973). Educability and group differences. NY: Harper and Row.

Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), A model for intelligence (pp. 93-132). Prenger-Verlag.

Kenny, D. A. (1974). A quasi-experimental approach to assessing treatment effects in nonequivalent control group design. Psychological Bulletin, 82, 345-362.

Kyllonen, P. C., & Shute, V. J. (1989). A Taxonomy of learning skills. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences: Advances in theory and research (pp. 117-163). NY: W. H. Freeman & Co.

Lohman, D. F. (1994). Component scores as residual variation (or why the intercept correlates best.) Intelligence, 19, 1-12.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Martin, D. J. (1985). The measurement of growth in educational achievement. Unpublished doctoral dissertation. The University of Iowa, Iowa City, Iowa.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726-748.

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

Simrall, D. V. (1946). The effects of practice on the factorial equations for perceptual and visual-spatial tests. Unpublished doctoral dissertation, University of Illinois, Urbana.

Snow, R. E., Kyllonen, P. C., & Marshalek, E. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 2, pp. 47-103). Hillsdale, NJ: Erlbaum.

Stake, R. E. (1961). Learning parameters, aptitudes, and achievement. Psychometric Monographs, No. 9.

Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum.

Thorndike, E. L. (1921). Intelligence and its measurement: A symposium. Journal of Educational Psychology, 12, 124-127.

Thorndike, E. L. , Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). Measurement of intelligence. NY: Teachers College Press.

Thorndike, R. L. (1966). Intellectual status and intellectual growth. Journal of Educational Psychology, 57, 121-127.

Thorndike, R. L. & Hagen, E. P. (1997). Cognitive Abilities Test (Form 5) Research Handbook. Itasca, Il: The Riverside Publishing Company.

Woodrow, H. (1946). The ability to learn. Psychological Review, 53, 147-158.

Yerkes, R. M., & Anderson, H. M. (1915). The importance of social status as indicated by the results of the point-scale method of measuring mental capacity. Journal of Educational Psychology, 6, 137-150.

Footnotes

[1] The low correlations between one-year gains and ability mean that a model that claimed that status on year n was the sum of status on year n-1 and random growth (e.g., Anderson, 1939) would fit the year-to-year gains quite well. However, such a model would not predict that growth over a longer period would show higher correlations with ability. Rather, the model would need to incorporate a small correlation between true gain during any one year period and ability. These small advantages cumulate over the years.

[2] Thorndike (1966) leveled the same criticism at the use of IQ scores to study the relationship between intellectual status and intellectual growth: "By eliminating from the score scale the differences in standard deviation at different ages, that which is the essence of growth is eliminated -- the greater variability of specimens as they mature. Imagine a group of adults whose heights and weights showed no greater standard deviations than those of newborn babies! ... A statistical treatment that ... excludes greater variability in intellect as we go from birth to maturity is equally absurd." (p.126)

[3] Kyllonen pointed out that an alternative way to think of this is that the slope is just the average of the difference scores between adjacent levels (e.g., the average of 5-4, 4-3, 3-2, and 2-1).

Table 1

Reliability of Raw Gain Scores as a Function of the Ratio of the

Pretest and Posttest True Score Standard Deviations

	σ_1 / σ_2			
ρ_{12}	.40	.60	.80	1.00
.50	.70	.65	.61	.60
.60	.66	.58	.52	.50
.70	.61	.48	.37	.33
.80	.55	.33	.09	.00

Note: Assuming $\rho_{11'} = \rho_{22'} = .80$

Table 2

Correlations Between Average One- to Five-Year Gains in Achievement and General Ability for 6321 Iowa Students (after Martin, 1985)

	Years between pretest and posttest				
Test	1	2	3	4	5
VOCABULARY	0.167	0.255	0.334	0.398	0.437
READING COMP	0.141	0.270	0.357	0.441	0.486
Spelling	0.139	0.281	0.379	0.452	0.470
Capitalization	0.126	0.243	0.342	0.400	0.453
Punctuation	0.112	0.216	0.299	0.361	0.412
Language Usage	0.093	0.179	0.240	0.314	0.354
LANGUAGE TOTAL	0.183	0.337	0.441	0.514	0.554
Visual Materials	0.109	0.218	0.282	0.378	0.421
References	0.134	0.262	0.354	0.433	0.480
WORK STUDY TOTAL	0.154	0.298	0.386	0.477	0.519
Math Concepts	0.140	0.260	0.362	0.440	0.503
Math Problems	0.127	0.238	0.327	0.393	0.440
Math Computation	0.141	0.244	0.318	0.392	0.462
MATH TOTAL	0.193	0.335	0.435	0.510	0.567
COMPOSITE	0.262	0.445	0.548	0.625	0.664

Note. Entries in Column 1 are average correlations with ability for the five 1-year gains; in Column 2 for the four 2-year gains; in Column 3 for the three 3-year gains; in Column 4 for the two 4-year gains; in Column 5 for the one 5-year gain.

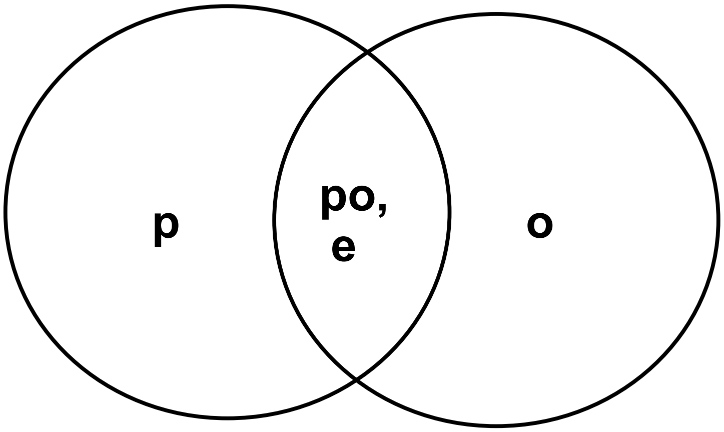
Figure Captions

Figure 1. Developmental standard scores on the Reading Vocabulary subtest of the Iowa Tests of Basic Skills from Kindergarten through Grade 12, for students scoring at the 1st, 20th, 50th, 80th, and 99th percentiles within each grade (spring norms) (After Hoover, Hieronymous, Frisbie, & Dunbar, 1993).

Figure 2. Average correlation between gains over periods of one to five years on subtests of the ITBS and an estimate of general ability for 6,321 Iowa students retested every year from third through eight grade (after Martin, 1985).

Figure 3. Venn diagram showing the partitioning of sources of variation in a person by occasion data matrix. Variation in row means ($\bar{X}_{p.}$) captures differences among individuals in overall performance whereas variation in column means ($\bar{X}_{.o}$) captures differences across occasions. Individual differences in learning scores capture neither of these sources of variation, but instead reflect some portion of the p x o interaction.

Figure 4. A person by task (with items nested within tasks) by occasion data matrix. Note that constructs in psychology are often defined by quite different, often independent aspects of score variation.


p
po,
e
o

[1] The low correlations between one-year gains and ability mean that a model that claimed that status on year n was the sum of status on year n-1 and random growth (e.g., Anderson, 1939) would fit the year-to-year gains quite well. However, such a model would not predict that growth over a longer period would show higher correlations with ability. Rather, the model would need to incorporate a small correlation between true gain during any one year period and ability. These small advantages cumulate over the years.

[2] Thorndike (1966) leveled the same criticism at the use of IQ scores to study the relationship between intellectual status and intellectual growth: "By eliminating from the score scale the differences in standard deviation at different ages, that which is the essence of growth is eliminated -- the greater variability of specimens as they mature. Imagine a group of adults whose heights and weights showed no greater standard deviations than those of newborn babies! ... A statistical treatment that ... excludes greater variability in intellect as we go from birth to maturity is equally absurd." (p.126)

[3] Kyllonen pointed out that an alternative way to think of this is that the slope is just the average of the difference scores between adjacent levels (e.g., the average of 5-4, 4-3, 3-2, and 2-1.