Assessing Spurious "Moderator Effects": Illustrated Substantively With the Hypothesized ("Synergistic") Relation Between Spatial and Mathematical Ability

David Lubinski and Lloyd G. Humphreys
University of Illinois at Urbana-Champaign

The traditional methodology for assessing moderator variables (hierarchical multiple regression analysis) is examined. Possible drawbacks of this technique for corroborating psychological theories (cf. Busemeyer & Jones, 1983), are illustrated empirically on the basis of an analysis of 400,000 subjects. This article tested a well-known (and currently popular) substantive hypothesis: A synergistic relation exists between mathematical ability and spatial visualization in the prediction and development of sophisticated levels of advanced mathematics. Using the traditional methodology, this hypothesis was confirmed; however, on further analysis, using a more systematic approach, it was demonstrated that this finding was spurious. Suggestions are offered for modifying the traditional methodology used for assessing moderator effects (for both applied and theoretical purposes). These amount to ways for minimizing Type I and Type II errors.

In an article aimed at addressing several problems frequently encountered when assessing "moderator effects," Busemeyer and Jones (1983) discussed a number of complex quantitative issues that compromise statistical results when *hierarchical multiple regression analysis* (HMRA) is used for testing theoretical predictions (e.g., Cohen, 1968; Cohen & Cohen, 1975). The purpose of this article is to highlight certain points of Busemeyer and Jones, empirically, with a substantive hypothesis currently receiving considerable attention: A synergistic relation exists between spatial visualization and mathematical ability in the prediction and development of exceptional levels of advanced mathematics. The foregoing hypothesized relation is analyzed in detail, not only for an illustrative context to frame recent methodological refinements (Busemeyer & Jones, 1983), but for contemporary theoretical interest as well (cf. Benbow, 1988; Lubinski & Humphreys, in press; and references therein).

Usefulness and Scope of Moderator Variables

The "moderator idea" was initially conceived in applied areas by psychologists interested in identifying subgroups of individuals for whom predictor–criterion relations are more valid than for other subgroups. Moderator variables, as such, were not of central concern. The concept was motivated by the desire to enhance atheoretical, criterion-related validity. Moderators were construed as relatively independent of criterion behaviors of interest and were thought of as tertiary variables on which group membership or individual differences reflect the extent to which more focal predictor–criterion relations are valid. Moderator variables were said to subdivide heterogeneous aggregations of individuals into homogeneous groups either categorically (e.g., by gender) or continuously (e.g., by attitude or personality dimensions), for purposes of "differential validity" (Berdie, 1961; Frederiksen & Gilbert, 1960; Ghiselli, 1956, 1960, 1963; Saunders, 1956; Zedeck, 1971).

More recently, theoretically driven ideas about specific trait constellations having "surplus properties" from the mutual integration of their constituents have been scrutinized in terms of moderator effects (Lubinski, 1983; Lubinski, Tellegen, & Butcher, 1981, 1983). So moderator variables emanating from theoretical considerations, like those traditionally assessed in applied settings, can be *dichotomous* (e.g., Class Membership × Trait Interactions: race × ability = performance; Hunter & Schmidt, 1976, 1978; Schmidt, 1988; Schmidt & Hunter, 1974, 1977) or *continuous* (e.g., Trait × Trait interactions: masculinity × femininity = androgyny; Lubinski et al., 1981, 1983). That moderator effects are currently relevant to both applied and theoretical issues in psychology is undeniable (Arnold, 1982, 1984; Chaplin & Goldberg, 1984; Cronbach, 1987; Dawis & Lofquist, 1984; Paunonen & Jackson, 1985; Stone & Hollenbeck, 1989; Tellegen, 1988; Tellegen, Kamp, & Watson, 1982; Tellegen & Lubinski, 1983). The "classic" equation used to assess the incremental validity gleaned from a moderator operation follows:

This investigation was supported by the National Institute of Mental Health, National Research Service Award No. 14257, Lawrence E. Jones, Training Director. The study was conducted while David Lubinski was a postdoctoral trainee in the Quantitative Methods Program of the Department of Psychology, University of Illinois at Urbana-Champaign. The Project Talent Data Bank was purchased through a research grant awarded to us by the University of Illinois Research Board, Award No. 1-2-69344.

Correspondence concerning this article should be addressed to David Lubinski or Lloyd G. Humphreys, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

Psychological Bulletin, 1990, Vol. 107, No. 3, 385-393
 0033-2909/90/$00.75

$$C = B_0 + B_1(X) + B_2(Z) + \quad \text{[Step 1]}$$
$$B_3(XZ), \quad \text{[Step 2]} \qquad (1)$$

where C = the criterion variable, X = the predictor, Z = the hypothesized moderator, XZ denotes the Linear × Linear interaction between the two main effects, and B_0, B_1, B_2, and B_3 are the structural parameters. By itself, Step 1 is known as the *additive* model; collectively, Step 1 and Step 2 define a *multiplicative* model. The moderator effect is assessed by statistically comparing the R^2 change, following Step 2, for incremental validity over that achieved by Step 1; the Linear × Linear interaction can only be assessed *after* removing the variance associated with its constituents, hence the hierarchical methodology: Step 1 followed by Step 2. It is often desirable (but not essential) to enter the main effects in an incremental stepwise fashion in order to ascertain the relative contribution of each (see the following).

Busemeyer and Jones (1983) provided several cogent quantitative arguments regarding problems inherent in this model. Two of the more central arguments address amplifications of Type I and Type II errors for the product term. Using Equation 1, enhanced Type I error for the product term can ensue when a function form similar to a Linear × Linear interaction—for example, a quadratic trend—better characterizes the structural relation between the predictor set and the criterion. Given this, a statistically significant Linear × Linear trend may result when, actually, a different higher-order trend better describes the covariation between the predictor set and criterion. A second problem with Equation 1 involves the *statistical power* of the test for the product term. This methodology may inordinately reduce the statistical power of the Linear × Linear trend, because when main effects are multiplied to generate terms for assessing Linear × Linear interactions, errors in measurement can be multiplied as well. So difficulties of Type II error can also occur with the present methodology, namely, statistical rejection of a Linear × Linear trend when one is actually present. The following exposition further explicates the importance of these two forms of error.

Type II error. To the extent that main effects are unreliable (or contain measurement error), the probability of Type II error is enhanced. Bohrnstedt and Marwell (1978) developed the reliability of a product of deviation scores as a joint function of the reliabilities of the components and the correlation between the components. For deviation scores only, the reliability of the product term is equal to the product of the reliabilities of *independent* main effects. As the correlations between main effects increase from 0 to 1.00, the reliability of the product of deviation scores approaches the reliabilities of the components. For example, given two relatively independent predictors, if one has low reliability, or if both have moderate reliabilities, the reliability of the product term (henceforth denoted $r_{(XZ)(XZ)}$) is severely attenuated (e.g., if $r_{XZ} = 0$ and $r_{XX} = r_{ZZ} = .60$, then $r_{(XZ)(XZ)} = .60 \times .60 = .36$).

Bohrnstedt and Marwell (1978) did not provide an analytic solution to the increases in reliability of products when the components are scaled about means other than zero, but a solution is not necessary in HMRA research. It seems entirely plausible that an origin other than zero has no effect as long as the evaluation is accomplished hierarchically. This inference about reliabilities follows from the fact that the contribution of XZ to the prediction of C is independent of the origins of X and Z in a hierarchical analysis in which X and Z are removed before XZ is evaluated. Thus, the reliability of the product becomes, in effect, the Bohrnstedt–Marwell expectation insofar as its effects on regression weights in a hierarchical analysis are concerned. (See McNemar, 1969, for further discussions on the effects of scale on correlations involving quotients or products.)

Consequently, it is critical, if one is interested in HMRA for testing theoretical predictions involving moderator effects (as opposed to simply a data analytic technique to account for additional criterion variance for applied purposes), to attend to the preceding considerations and especially to use highly reliable predictors. Perhaps this is one of the reasons that moderator variables have been particularly hard to tie down (Tellegen et al., 1982; Wiggins, 1973). In an article on statistical tests for moderator variables, Cronbach (1987, p. 417) has suggested, "[f]urther investigation of statistical power in studies of interactions and invention of more sensitive research strategies are much to be desired." Decreasing measurement error is one way of achieving this goal.

Type I error. In HMRA, Type I error for the product term is exacerbated by predictor–criterion functional relations involving nonlinear monotonic trends (i.e., positively or negatively accelerated function forms, for example, X^2 or $X^{1/2}$, respectively). This can result in false theoretical interpretations based on statistically significant, but spurious, moderator effects. Especially given that as $r_{XZ} \to 1.00$, $r_{(XZ)(X^2)} \to 1.00$, and, therefore, for all levels of r_{XZ} between .00 and 1.00 there is shared variance of XZ and X^2 (and this communality is typically appreciable). The significance of Type I errors when predictors are *not* linearly related to the criterion is illustrated later. We also discuss (with empirical examples) how erroneous theoretical interpretations can result if data analysis is terminated before scrutinizing limitations of the traditional methodology (i.e., Equation 1).

An empirical example. To underscore the importance of the aforementioned psychometric issues, we addressed a meaningful psychological hypothesis whose verisimilitude is still questionable: the hypothesized synergistic relation between quantitative and spatial ability in relation to the prediction and development of exceptional levels of mathematical sophistication. We tested the respective *null* (H_o) and *alternative* (H_a) hypotheses, respectively: H_o, is the genesis of brilliant mathematical accomplishment simply an extraordinary level of quantitative ability (a straightforward linear effect, the nature and strength of which can be empirically assessed and structurally characterized by the additive model)? Or H_a, is exceptional sophistication in mathematics the product of a synergistic relation between quantitative ability and spatial visualization (a straightforward Linear × Linear interaction indicative of a moderator variable as traditionally conceived, calling for a multiplicative model, viz., a Trait × Trait interaction)?

In recent commentary in *Behavioral and Brain Sciences* on the relation between spatial ability and mathematical functioning at exceptional levels, Burnett (1988) noted a number of limitations in published research, for example, inadequate measures of spatial ability, only high school students in the average

ability range, or dependent measures having multifaceted complexity (e.g., grades) as opposed to dependent measures of ability with high ceilings. To alleviate these shortcomings, we analyzed data on 400,000 subjects from the entire Project Talent Data Bank (Flanagan et al., 1962).

This data bank consists of a stratified random sample of high schools collected in 1960. It contains four cohorts of students, Grades 9–12, with approximately 100,000 subjects per cohort. The data bank contains information on a number of distinct classes of psychological traits (abilities, interests, and personality), as well as autobiographical data. We selected from this huge fund of data a number of quantitative and spatial reasoning tests for the construction of two aptitude predictors, a *mathematical composite* and a *spatial composite* (hereafter labeled *M* and *S*, respectively). These larger aptitude composites were composed of a number of shorter tests to augment their psychometric properties (reliability and construct validity) through aggregation.[1]

Elsewhere, using cross-twin data (Humphreys, in press), the estimated reliability of both composites is $r_{mm} \approx r_{ss} \approx .90$, for both genders and all grade levels. If these two measures were independent, the reliability of the product term would approach the product of the two reliabilities (viz., $r_{(ms)(ms)} = .90 \times .90 = .81$). However, because *M* and *S* are appreciably correlated ($.61 \leq r_{ms} \leq .63$, for both genders, across all 4 cohorts), we know the reliabilities of all eight product terms, namely, 2 (genders) × 4 (cohorts), are within the 8-point range of $.81 < r_{(ms)(ms)} < .90$ (cf. Busemeyer & Jones, 1983, p. 557, Table 2). These reliabilities are more than adequate for research purposes. Also, the relative size of the regression weights of *M*, *S*, and their product are not compromised by appreciable differences in the reliabilities of *M* and *S*.

For our criterion variable we chose an Advanced Mathematics measure (*C*), the content of which included introductory calculus, solid and plane geometry, trigonometry, logarithms, probability logic, scientific notation, higher algebra, and elements of analytic geometry and clearly indexes higher levels of quantitative ability (with 14 items). This measure was designed specifically to assess students' understanding of advanced concepts, rather than rote memory. A test having the content described is obviously not a "fair" test of mathematical ability for even the average American 12th-grade student, let alone for Grades 9–12. It is, however, a valid criterion measure of how much mathematics has been acquired by students having high levels of talent. Just as the mathematics section of the Scholastic Aptitude Test is not a *fair* test for the average 7th-grade student, it can nevertheless be used as a valid tool for diagnosing mathematical giftedness in intellectually exceptional 12- and 13-year-olds (Benbow, 1988).

Level of talent is defined independently on a subject matter test appropriate to the typical level of formal preparation. The implicit assumption in moderator variable research is that the criterion variable performance is enhanced (or retarded) for some subset of the population sampled. Viewed in this way, it was anticipated that the subset with enhanced scores in the Project Talent data would increase from Grade 9 to Grade 12. In addition, because of the sophisticated ability level that this measure taps, its reliability increases markedly with grade and with ability level; we elaborate on this property in subsequent discussion.

Table 1
Moderator Effects Assessed by the Traditional Methodology

Step	Male subjects: Variable entered	Male subjects: Hierarchical stepwise R^2	Female subjects: Variable entered	Female subjects: Hierarchical stepwise R^2
		Grade 9		
Step 1	*M*	.062	*M*	.030
	S	.063	*S*	.030
Step 2	*MS*	.077	*MS*	.035
		Grade 10		
Step 1	*M*	.266	*M*	.188
	S	.266	*S*	.189
Step 2	*MS*	.289	*MS*	.209
		Grade 11		
Step 1	*M*	.515	*M*	.392
	S	.515	*S*	.392
Step 2	*MS*	.545	*MS*	.424
		Grade 12		
Step 1	*M*	.590	*M*	.429
	S	.590	*S*	.429
Step 2	*MS*	.623	*MS*	.467

Note. Sample *N*s are provided in Table 3. All R^2 increments for Step 2 are statistically significant at the $p < .001$ level.

Equation 1 was applied to each gender within each cohort (2 × 4) to produce the eight regression analyses found in Table 1. In Step 1, the main effects were entered in an incremental stepwise manner; Step 2 followed. As shown in Table 1, the predictor *M* for all four regressions for both genders is significantly

[1] The scales chosen for *M* and *S* follow (with number of items and raw score weights, respectively, in parentheses): *M* = Mathematics Information (23 and .55) + Arithmetic Reasoning (16 and 1.0) + Introductory Mathematics (24 and .55); *S* = Visualization 3-d (16 and 3.0) + Visualization 2-d (24 and 1.0) + Mechanical Reasoning (20 and 1.5) + Abstract Reasoning (15 and 2.0). These composites were formed to represent the constructs of mathematical and spatial visualization abilities in accordance with the use of these terms in the literature concerning mathematical talent and the factor analytic findings for these tests (Humphreys, in press). Weights were assigned judgmentally by modifying raw score variances and covariances of the tests so that loadings on the common factors of mathematical and spatial ability, respectively, would be reflected in the composites. For example, use of raw score weights would have overweighted mathematics information and introductory mathematics in comparison to arithmetic reasoning in the mathematics composite. Selection of only one test for each construct would have been arbitrary and would have yielded less valid and less reliable measures of the constructs having larger nonerror-specific content as well. Alternatively, we could have used seven individual test scores, their powers, and their many cross-products in our regression equations, but the interpretation of the results would have presented an insurmountable problem.

and *substantially* related to C. (The term *substantially* is used here to describe proportion of variance accounted for, or R^2, because with such large samples statistical significance becomes practically meaningless; cf. Lykken, 1968; Meehl, 1967.) The second predictor S is entered, but with essentially nugatory effect on the R^2 increment.[2] It appears that whatever variance S shares with our criterion variable, this variance is also common to M. The inverse of this assertion is of course not true; M shares some common variance with C that is not common with S.

The findings of Step 2 indicate that although S does not add incremental validity to the prediction of C, following that accounted for by M, S does interact (synergistically) with M, and the prediction of C is enhanced by this Linear × Linear interaction, a "classic" moderator effect. If our analysis were to stop here, we would conclude that the writings of a number of theorists are empirically supported: There *is* a special synergistic relation between spatial ability and mathematical ability at high levels that enhances the prediction, and possibly the development, of sophisticated mathematical ability. But is this conclusion accurate?

As indicated earlier, product and quadratic terms may share substantial amounts of variance. If a quadratic M^2 or S^2 trend better characterizes the relation between our predictors and C, the significant Linear × Linear interaction could have resulted from simply being highly correlated with one of the quadratic trends. Motivation for entertaining this possibility is intensified by the realization that psychological predictors typically exhibit appreciable multicollinearity, or shared variance. To test for this possibility, Equation 1 must be expanded to at least include the squared constituent terms, and these components are now found in Equation 2 (with symbols representing our composites, M and S).

$$C = B_0 + B_1(M) + B_2(S) + \quad \text{[Step 1]}$$
$$B_3(MS) + B_4(M^2) + B_5(S^2) \quad \text{[Step 2]} \qquad (2)$$

Step 1 remains the same, but now three terms, MS, M^2, and S^2, are entered simultaneously in an incremental stepwise fashion (in competition with one another) to *empirically* assess which function form best characterizes the higher-order relation between the predictor set and C. The results are found in Table 2 (for additional descriptive data on these measures, see Appendixes A and B).

This analysis is illuminating; it reveals that for both genders in Grades 10–12, a quadratic trend (viz., M^2) best describes the function form between the predictor set and C. The quadratic term absorbs more criterion variance than MS in seven of the eight analyses; only for 9th-grade girls did MS emerge as the first entry in Step 2. In all analyses the two remaining terms in the equations contributed in combination less than .01 of the squared multiple correlation. Our findings for the 9th-grade girls are compromised, however, because predictability is so low for this cell. Even with our massive sample size, random error cannot be ruled out. To illustrate, we conducted an additional analysis on the 9th-grade girls to compare the results of the foregoing analysis with those obtained with the main effects *and* M^2 as the predictor set. Using M, S and MS as predictors, we obtained an R^2 of .0355, whereas with M, S, and M^2 as predictors, the R^2 decreased to .0347. As the results of the other analyses are more clear-cut, especially at higher grade levels, we attribute the findings obtained with the 9th-grade girls to sampling fluctuation compounded by especially low criterion reliability (r_{cc}) for this group (see the following paragraphs).

Table 2
Moderator Effects Assessed Simultaneously With Quadratic Trends

	Male subjects		Female subjects	
Step	Variable entered	Hierarchical stepwise R^2	Variable entered	Hierarchical stepwise R^2
		Grade 9		
Step 1	M	.062	M	.030
	S	.063	S	.030
Step 2	M^2	.081	MS	.035
		Grade 10		
Step 1	M	.266	M	.188
	S	.266	S	.189
Step 2	M^2	.301	M^2	.212
		Grade 11		
Step 1	M	.515	M	.392
	S	.515	S	.392
Step 2	M^2	.570	M^2	.450
		Grade 12		
Step 1	M	.590	M	.429
	S	.590	S	.429
Step 2	M^2	.656	M^2	.501

Note. All R^2 increments for Step 2 are statistically significant at the $p < .001$ level. All remaining trend components (for all eight analyses) accounted for less than 1% of additional variance following the first entry of Step 2, so consideration of these components was disregarded.

Collectively, the preceding analyses demonstrate that our earlier conclusions (based on results of Equation 1) were spurious due to multicollinearity between MS and M^2. This, in turn, changes our theoretical interpretation offered earlier. On the basis of the present findings, the posited synergistic relation between spatial visualization and quantitative ability is rejected. Advanced mathematics (C) appears to be a function of the first- and second-order trends of mathematical ability (M); by and large, spatial ability (S) does not appear to explain any additional incremental variance, either additively or multiplicatively. From a verisimilitude framework of empirically based competitive support, the quadratic trend clearly wins out over the Linear × Linear interaction.

To illustrate these quadratic trends graphically, the quartiles

[2] Some investigators do not specifically delineate whether the contribution of spatial ability to mathematical excellence is additive or synergistic (cf. Sherman, 1967). To the extent that investigators posit the former, these findings negate that position.

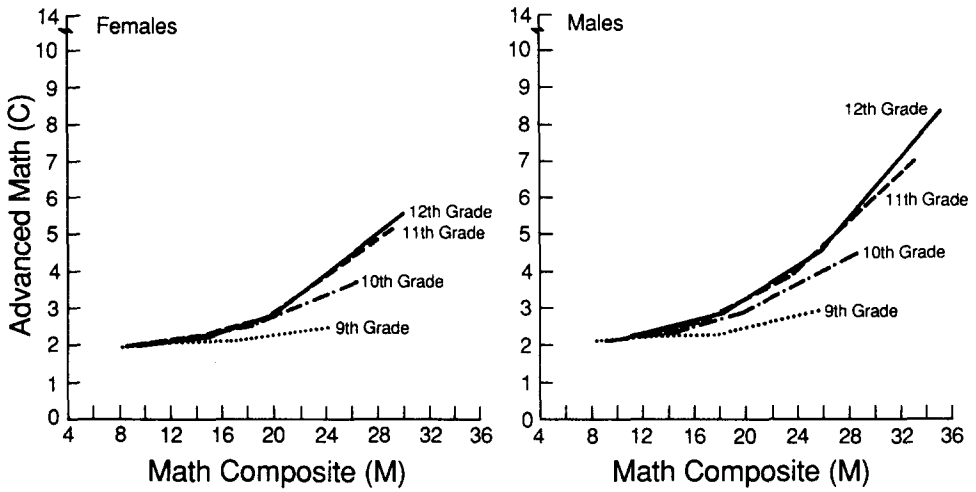

Figure 1. The regression of Advanced Mathematics on the quartiles of the Mathematics Composite. (For all four cohorts, by gender, quartiles were computed for the Mathematics Composite; the means for each of these four segments are plotted on the x axis and the corresponding means on Advanced Mathematics for these four segments are plotted on the y axis.)

on the Mathematics Composite were computed for all four cohorts, by gender. Second, means on the Mathematics Composite and Advanced Mathematics were computed for subjects within each quartile. Finally, these means were used to plot bivariate points; lines connecting these points bring out the positively accelerating function forms indicated by our regression analyses (see Figure 1). Because the Mathematics Composite was parsed in this somewhat arbitrary fashion, the full form of the quadratic trend is actually suppressed a little. Nevertheless, with only four points to graphically represent the curvilinearity of the quadratic function (M^2), this higher-order trend is revealed convincingly. Given these findings, one could venture the following conclusion: To the extent that students acquire exceptionally sophisticated mathematical skills, *whether developed through formal instruction or especially on their own,* these skills are likely to be better characterized by a quadratic transformation of their general mathematical ability rather than by a synergistic interaction between the latter and their level of spatial visualization.[3]

Moreover, a quadratic trend, like a Linear × Linear trend, *also has psychological meaningfulness for the present example.* One possibility is the following: If, for example, individual differences below the normative mean on M are *un*correlated with sophisticated levels of mathematical skill C, and, to the extent that individuals are located within higher ability ranges, the correlation between M and C increases. Psychologically, this could occur if individuals at lower levels of functioning are so far from the rudimentary prerequisites for acquiring skills at advanced mathematics that individual differences observed within this truncated segment of the ability distribution are essentially equivalent and of no consequence (even though they may be psychologically related to other important criteria).

Given the foregoing state of affairs, a quadratic trend, namely M^2, would be expected to better characterize the structural relation between the predictor set and the criterion. Moreover, the obtained signs of the beta weights (for all eight regressions reported in Table 2) were in the proper direction for this interpretation: All were negative for M and positive for M^2. The negative M weight, for low scorers, adjusts for or cancels out the positive M^2 weight in an offsetting manner; however, as scores on M increase, the weight assigned to positive M^2 increases inordinately, in contrast to that subtracted by negative M, and the estimate of C increases in a positively accelerated manner characteristic of a quadratic trend (see Figure 1). Our interpretation of the positively accelerated trend corresponds to a similar curvilinear (quadratic) phenomenon observed within a variety of disparate behavioral domains; athletic ability, for example, is a case in point (albeit remote from the substantive issue currently under analysis).

[3] Smith (1964) has suggested that the special relation between mathematical ability and spatial visualization becomes more important at higher levels of intellectual functioning. To address this idea, we split all eight groups of subjects roughly in half, into those above the mean (high ability subjects) and those below the mean (low ability subjects), using Project Talent's IQ composite. This composite comes closest to matching the content found on the Stanford–Binet Intelligence Scale (Terman & Merrill, 1960) and the various Wechsler (1974) tests of general intelligence as could be achieved with Project Talent's group tests. We did not want to split the groups on the Mathematics Composite, because this would have produced a pronounced positively skewed distribution of subjects on the predictor M, and conducting regression analyses on such groups is undesirable from an interpretive point of view. With the high ability subjects only, we then conducted the same regression analyses as before, using Equation 2. These results are presented in Appendix C. For all eight groups, the quadratic trend was entered first in Step 2, and none of the remaining variables individually or collectively accounted for more than 1% of criterion variance. That this was true for the 9th-grade girls supports our early interpretation of why MS accounted for more variance and was entered before M^2 in Step 2, using the full range of talent for this group.

Table 3
Correlations Between the Mathematical Composite and Advanced Mathematics and Reliabilities for Low and High Ability Segments by Gender for All Four Cohorts

Grade level	Male subjects				Female subjects			
	Low ability		High ability		Low ability		High ability	
	r_{cm}	r_{cc}	r_{cm}	r_{cc}	r_{cm}	r_{cc}	r_{cm}	r_{cc}
Grade 9	.08	.17	.33	.28	.07	.18	.21	.21
Grade 10	.12	.18	.52	.36	.10	.18	.45	.25
Grade 11	.20	.26	.68	.56	.12	.21	.66	.52
Grade 12	.26	.35	.72	.66	.15	.21	.69	.60

Note. Reliability estimates for Advanced Mathematics were computed by Kuder–Richardson (1937) 21, because p values were not available for individual items. Sample sizes follow: Grade 9—boys, low ability (L) = 27,943, high ability (H) = 22,025; girls, L = 27,454; H = 21,939; Grade 10—boys, L = 26,878, H = 21,665; girls, L = 26,072, H = 21,047; Grade 11—boys, L = 23,636, H = 20,215; girls, L = 25,564, H = 19,864; Grade 12—boys, L = 20,250, H = 18,142; girls, L = 22,482, H = 17,634.

Consider the following hypothetical example. Using the same range metric, athletic ability in high school for the bottom half of the adolescent distribution (even if selection is based only on those who go out for extracurricular sports) probably correlates zero (concurrently and predictively) with instrumental effectiveness in professional careers in the National Football League (NFL) or the National Basketball Association (NBA). Most high school athletes simply do not have the necessary level of antecedent skills necessary to compete at this elite level (even following 4 years of extensive training in college), and ability deficits for such behavioral domains are even more pronounced for athletes whose ability level is below their peer group's norm. Although this example is far removed from intellectual functioning, it is conceptually and psychologically isomorphic with the earlier scenario involving low levels of mathematical ability (as illustrated in Figure 1). Other psychologically meaningful examples could be offered from the arts and the humanities, as well as other content domains that require an appreciable level of certain "standardized" antecedent skills on which profound individual differences are displayed, before they can be profitably entered.

The quadratic trend also indicates that the reliability of the criterion C increases for individuals at the upper half of the M distribution (or reliability increases as individual differences move across upper ranges of ability segments). To illustrate, we split all 8 groups of individuals into high and low ability subsets based on a mean split on the M predictor. We then computed correlations between M and C for all 16 groups (see Table 3). Reliability estimates for the criterion C (r_{cc}) are also provided for all 16 upper and lower (i.e., above the mean on M vs. below the mean on M) ability groups. Clearly, the reliability of r_{cc} and the validity of r_{mc} is greater for the more mathematically talented students.[4]

These correlations and reliability estimates reveal, in yet another way, the phenomenon discussed in our previous example: In the case of low ability subjects, their individual differences in quantitative ability are not much related to performance in advanced mathematics. Similar to the lower truncated segment of high school athletes, individual differences within low ability ranges are, for all practical purposes, essentially equivalent when it comes to providing the necessary antecedents for effective functioning at *exceptional* levels. Individuals within low mathematical ability segments are, in a very real sense, too psychologically removed from the necessary requisite skills to instrumentally enter *exceptionally sophisticated* quantitative domains. So although individual differences within the lower ability range are psychologically significant across a variety of meaningful behavioral criteria, they are essentially equivalent, *psychologically*, as antecedents to and predictors of highly complex quantitative skills.

Conclusions

Just as the concept of the moderator variable reveals that for predictive and theoretical purposes it is desirable to segregate individuals into homogeneous subsets as a function of predictor–criterion differential validities (moderated by discrete subgroup membership or a continuous trait level), the present analysis of squared components illustrates the importance of differential reliability and validity as a function of contrasting ability ranges (as a function of predictor level). We recommend inspection of squared terms concurrently with analytic treatments aimed at assessing moderator effects.[5] Ideally, this is best achieved by assessing the Linear × Linear interaction (XZ) with the squared components (X^2 and Z^2), simultaneously, in an incremental stepwise fashion in competition with one another. This proposed methodology will thereby let the data decide on the precise functional relation responsible for observed incremental validity. More accurate theoretical interpretations of data will necessarily follow.

Finally, although our discussion is framed in the context of ability assessment, the foregoing exposition is also relevant to several contemporary issues in personality assessment. Paunonen (1988) and Tellegen (1988) provided two particularly engaging contributions containing psychometric issues related to the present article. Among other things, these articles deal with theoretical ideas of trait level and trait relevance; both concepts are evaluated by multiple regression analysis, the former with linear components and the latter with curvilinear components (e.g., X^2).

---

[4] This phenomenon is similar to that characterized by Fisher's (1959) "twisted pear." Typically, range truncation is thought to attenuate predictor–criterion correlates (and it usually does), but it is not *necessarily* so. For certain segments of individual differences, restriction of range can enhance reliability and validity.

[5] Investigators should be alerted to the possibility of encountering more complex higher-order trends, for example XZ^2, X^2Z, X^2Z^2; these trends were inspected in the present study but all individual components accounted for less than 1% of criterion variance. For additional and more detailed treatments of nonlinear trends, beyond the scope of the present discussion, readers are referred to Busemeyer and Jones (1983).

References

Arnold, H. J. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance, 29,* 143–174.

Arnold, H. J. (1984). Testing moderator variable hypotheses: A reply to Stone and Hollenbeck. *Organizational Behavior and Human Performance, 34,* 214–224.

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.

Berdie, R. F. (1961). Intra-individual variability and predictability. *Educational and Psychological Measurement, 21,* 663–676.

Bohrnstedt, G. W., & Marwell, G. (1978). The reliability of products of two random variables. In K. F. Schuessler (Ed.), *Socological methodology* (pp. 254–273). San Francisco: Jossey-Bass.

Burnett, S. A. (1988). Spatial reasoning and mathematical reasoning abilities. *Behavioral and Brain Sciences, 11,* 187–188.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93,* 549–562.

Chaplin, W. F., & Goldberg, L. R. (1984). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology, 47,* 1074–1090.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin, 102,* 414–417.

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.

Fisher, J. (1959). The twisted pear and the prediction of behavior. *Journal of Consulting Psychology, 23,* 400–405.

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, MA: Houghton Mifflin.

Frederiksen, N., & Gilbert, A. (1960). Replication of a study of differential predictability. *Educational and Psychological Measurement, 20,* 759–767.

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology, 40,* 374–377.

Ghiselli, E. E. (1960). The prediction of predictability. *Educational and Psychological Measurement, 20,* 3–8.

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology, 47,* 81–86.

Humphreys, L. G. (in press). Some unconventional analyses of resemblance coefficients for male and female monozygotic and dizygotic twins. In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl.* Minneapolis: University of Minnesota Press.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of "test bias." *Psychological Bulletin, 83,* 1053–1071.

Hunter, J. E., & Schmidt, F. L. (1978). Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology, 63,* 1–11.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Lubinski, D. (1983). The androgyny dimension: A comment on Stokes, Childs, and Fuehrer. *Journal of Counseling Psychology, 30,* 130–133.

Lubinski, D., & Humphreys, L. G. (in press). A broadly based analysis of mathematical giftedness. *Intelligence.*

Lubinski, D., Tellegen, A., & Butcher, J. N. (1981). The relationship between androgyny and subjective indicators of emotional well-being. *Journal of Personality and Social Psychology, 40,* 722–730.

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.

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

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

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

Paunonen, S. V. (1988). Trait relevance and the differential predictability of behavior. *Journal of Personality, 56,* 599–619.

Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review, 92,* 486–511.

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

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

Schmidt, F. L., & Hunter, J. E. (1974). Racial and ethnic bias in psychological tests: Divergent implications for two definitions of test bias. *American Psychologist, 29,* 1–8.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Sherman, J. A. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review, 74,* 290–299.

Smith, I. M. (1964). *Spatial ability.* London: The University of London Press.

Stone, E. F., & Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology, 74,* 3–10.

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

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

Tellegen, A., & Lubinski, D. (1983). Some methodological comments on labels, traits, interaction, and types in the study of "femininity" and "masculinity": Reply to Spence. *Journal of Personality and Social Psychology, 44,* 447–455.

Terman, L., & Merrill, M. (1960). *Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M.* Boston: Houghton Mifflin.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for children.* New York: The Psychological Corporation.

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

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

(Appendixes follow on next page)

Appendix A

Means and Standard Deviations for Advanced Mathematics and the Mathematics and Spatial Composites by Gender for All Four Cohorts

Variable	Males		Females		Variable	Males		Females	
	M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*
	Grade 9					Grade 11			
Mathematics Composite	16.18	6.91	15.47	6.37	Mathematics Composite	20.90	8.83	17.89	7.93
Spatial Composite	69.51	22.33	60.40	19.93	Spatial Composite	79.25	22.76	66.45	20.81
Advanced Mathematics	2.39	1.61	2.16	1.50	Advanced Mathematics	3.91	2.65	3.05	2.19
	Grade 10					Grade 12			
Mathematics Composite	18.07	7.65	16.66	7.08	Mathematics Composite	22.65	9.26	18.70	8.11
Spatial Composite	74.55	22.77	63.53	20.77	Spatial Composite	82.54	23.11	68.65	21.19
Advanced Mathematics	2.98	1.93	2.59	1.72	Advanced Mathematics	4.49	3.12	3.19	2.38

Appendix B

Intercorrelations Between Advanced Mathematics, the Mathematics and Spatial Composites, and Their Transformations M^2, S^2 and MS, by Gender for All Four Cohorts

Variable	1	2	3	4	5	6	Variable	1	2	3	4	5	6
			Grade 9							Grade 11			
1. *C*	—	.172	.108	.173	.184	.118	1. *C*	—	.626	.392	.617	.663	.416
2. *M*	.249	—	.635	.918	.974	.638	2. *M*	.717	—	.627	.930	.976	.634
3. *S*	.137	.623	—	.848	.607	.978	3. *S*	.451	.622	—	.826	.595	.978
4. *MS*	.247	.928	.828	—	.920	.868	4. *MS*	.704	.934	.820	—	.929	.849
5. M^2	.274	.974	.590	.926	—	.625	5. M^2	.749	.979	.593	.933	—	.616
6. S^2	.154	.637	.979	.854	.618	—	6. S^2	.478	.635	.981	.845	.619	—
			Grade 10							Grade 12			
1. *C*	—	.434	.297	.437	.456	.314	1. *C*	—	.655	.396	.642	.697	.422
2. *M*	.516	—	.635	.921	.975	.637	2. *M*	.768	—	.609	.924	.977	.621
3. *S*	.332	.624	—	.844	.605	.978	3. *S*	.474	.615	—	.823	.578	.979
4. *MS*	.513	.931	.824	—	.922	.863	4. *MS*	.749	.931	.822	—	.923	.846
5. M^2	.544	.976	.590	.929	—	.622	5. M^2	.803	.981	.590	.931	—	.603
6. S^2	.355	.637	.980	.850	.617	—	6. S^2	.505	.632	.980	.846	.618	—

Note. For all four cohorts, intercorrelations for female subjects are located above the diagonal; intercorrelations for male subjects are found below the diagonal.

Appendix C

Moderator Effects Assessed Simultaneously With Quadratic Trends for High I.Q. Students

	Male subjects		Female subjects			Male subjects		Female subjects	
Step	Variable entered	Hierarchical stepwise R^2	Variable entered	Hierarchical stepwise R^2	Step	Variable entered	Hierarchical stepwise R^2	Variable entered	Hierarchical stepwise R^2
		Grade 9					Grade 11		
Step 1	M	.101	M	.044	Step 1	M	.524	M	.448
	S	.106	S	.048		S	.527	S	.451
Step 2	M^2	.137	M^2	.059	Step 2	M^2	.553	M^2	.483
		Grade 10					Grade 12		
Step 1	M	.297	M	.221	Step 1	M	.592	M	.448
	S	.302	S	.229		S	.595	S	.491
Step 2	M^2	.323	M^2	.244	Step 2	M^2	.627	M^2	.535

Note. All R^2 increments for Step 2 are statistically significant at the $p < .001$ level. All remaining trend components (for all eight analyses) accounted for less than 1% of additional variance following the first entry of Step 2, so consideration of these components was disregarded. Sample sizes follow: Grade 9—males = 24,253, females = 25,426; Grade 10—males = 24,650, females = 24,430; Grade 11—males = 23,768, females = 23,636; and Grade 12—males = 21,728, females = 21,323.

Received November 17, 1988
Revision received March 31, 1989
Accepted June 15, 1989 ■