THE JOURNAL OF SPECIAL EDUCATION VOL. 25/NO. 4/1992/pp. 504–526

ILLUSIONS OF MEANING IN THE IPSATIVE ASSESSMENT OF CHILDREN'S ABILITY

Paul A. McDermott
John W. Fantuzzo
University of Pennsylvania
Joseph J. Glutting
University of Delaware
Marley W. Watkins
Deer Valley Unified Schools
Phoenix, Arizona
Andrew R. Baggaley
University of Pennsylvania

In this study, we analyze the relative efficacy of normative and ipsative measures for the study of intra- and interindividual differences in child ability. With the use of representative data sets, including the WISC-R national standardization sample, purely ipsatized (or deviational ipsative) subtest scores were contrasted with conventional norm-based scores in terms of the evidential and consequential bases for validity. Internal and external evidence for validity was assessed for relative convergence of ability attributes, short- and long-term stability, and predictive efficiency. Comparative utility of each type of measure was explored for theoretical relevance, applicability in measurement work, and assessment of individualized intervention outcomes. Ipsative ability measures were found to be uniformly inferior to their normative counterparts, with ipsative measures conveying no uniquely useful information and otherwise impeding the versatility of assessment.

The past decade has brought a resurgence of interest in the study of *intraindividual* differences, especially as they relate to variation in children's ability. The general idea is to de-emphasize concern over more global measures of ability such as the IQ and, for any child assessed, to focus on the pattern of relative strengths and weaknesses across a variety of more specific ability traits. Essentially, one begins with the view that performance measures for underlying subareas of ability are more valid and revealing than global measures, whereupon one interprets performance for each subarea in terms of how much it deviates from a child's own average performance across all ability areas. Thus, emphasis is on within- rather than between-child comparison, a matter of particular relevance to those planning individualized counseling or intervention programs.

Such studies of intraindividual differences effectively constitute applications of what Cattell (1944) termed *ipsative* measurement. Notwithstanding new popular-

Address: Paul A. McDermott, Graduate School of Education, University of Pennsylvania, 3700 Walnut St., Philadelphia, PA 19104.

ity, ipsative measures are often mistaken as comparable in the theoretical and statistical sense to ordinary ability measures. Indeed, certain properties of ipsative measures tend to weigh against their current applications, and other properties have, until now, remained unexplored. This article examines through discussion and empirical tests the relative efficacy of ipsative approaches to psychological assessment. Using examples and representative national data sets, we open the discussion with an explanation of the ipsative concept and its most popular applications, and thereafter compare psychometric integrity and decision-making utility of ipsative and traditional norm-based ability indices. Finally, we offer recommendations for psychometric development and professional practice.

RAW, NORMATIVE, AND IPSATIVE PERSPECTIVES ON ABILITY

Cattell (1944) proposed that psychological phenomena could be expressed through three units of measurement: raw, normative, and ipsative scores. *Raw* scores are the most basic. Each psychological attribute is represented by its own unique distribution of raw scores, with score ranges, means, and other psychometric features differing from one attribute to another. Considering two children, it is neither readily apparent whether their scores are above or below the population mean, nor is it clear how much of a difference is represented by the distance between their scores. Thus, the scaling of a child's raw score on a given attribute remains interpretively isolated from the scores of all other children and from the child's own scores on other attributes.

Alternatively, *normative* scores retain a common distributional form, mean, and standard deviation. The scaling of a child's normative score is dependent on the scores of other children in the population, making it possible to determine the relative standing of any score with respect to the population and to assess the magnitude of differences separating any two children's scores. As is the case with raw scores, the scaling of a child's normative score on one attribute remains independent of his or her scores on other attributes. Normative scores are regarded as *population-relative* metrics, because any given value is an expression of performance relative to others in the population. It is reasonable to assume that individuals with high ability scores manifest more ability than those having low scores.

When ability is measured in *ipsative* score units, each child's set of ability attributes is treated separately so that concern is no longer with one's ability compared to others', but instead, with one's pattern of ability attributes relative to one's own average performance. Hence, we may regard these as *person-relative* metrics. Typically, the child's average normative score across the various attributes is subtracted from each component attribute score, producing a profile of positive and negative deviations from the child's average performance. These ipsatized scores hold a certain intuitive appeal because, by removing the general ability component as reflected in one's average performance level, the consequent score profile appears to isolate and amplify the pattern of abilities peculiar to the child. In contrast to both raw and normative scores, it can be seen that the scaling of a child's ipsative score for any ability attribute is dependent on that for every other attribute.

Perhaps most popular among contemporary practices is the method of ipsative ability assessment advocated by Kaufman (1979) for interpretation of the Wechsler Intelligence Scale for Children–Revised (WISC-R) (Wechsler, 1974). Kaufman, recognizing the practical and political limitations of general IQ measures, urged that psychologists focus "on ipsative (*intra*individual) more than normative measurement" (1979, p. 18). He cautioned practitioners not to overvalue IQ scores and advanced the notion that the WISC-R's 12 underlying subtest scales are unique or conjoint reflections of distinct human abilities. A major aspect of this interpretation process is the discovery of children's intellectual strengths and weaknesses by studying the magnitude and direction of each subtest score's deviation from a child's average subtest score. The Kaufman method is presently a common element in university curricula for preparing professional psychologists, with the ipsative procedure now generalized to many other ability tests (e.g., see Delaney & Hopkins, 1987, p. 83; Elliot, 1990, pp. 94-95; Kaufman, 1990, pp. 425–468; Kaufman & Kaufman, 1983a; Sattler, 1988) and guiding the development of popular computer software (Ingram, 1985; Psychological Corporation, 1986a, 1986b).

Table 1 illustrates corresponding normative and ipsative ability scores for a sample child. In this example and hereafter, we work only with normative and ipsative (and not raw) scores, because they are the scores generally used in the interpretation with published ability tests. We also work only with 11 of the WISC-R's 12 subtests, including the "mandatory" five Verbal and five Performance subtests (Wechsler, 1974, p. 8) and the supplementary Digit Span subtest (see Note 1).

The normative values shown in Table 1 are standard scaled scores ($M = 10$, $SD = 3$) as derived from a child's raw scores corrected for population variation within 4-month age intervals. Considering the sample profile, it will be noted that ipsative scores are reported as based both on all 11 subtests *and* on the six Verbal and five Performance subtests, respectively. This is because ipsatization can take different forms. Its most common form bases ipsative values on all attributes taken simultaneously (as in the example using all 11 subtests), but Kaufman (1979) recommended an alternative for the WISC-R. As determined through exploratory factor analyses of subtests within the standardization sample, Kaufman (1975) concluded that the various Verbal subtests tend to relate strongly to an overall verbal dimension (as represented by WISC-R's Verbal IQ scale) and the Performance subtests to a performance dimension (associated with the Performance IQ scale). Therefore, he recommended that more meaningful comparisons would be wrought by exploring deviations *within* the respective Verbal and Performance subtest groups. The example and subsequent illustrations in this article consider both approaches to ipsatization, thus permitting an empirical test of the claim that Verbal/Performance ipsatives provide more meaningful information than ipsatives based on all subtests (see Note 2).

As shown, ipsative scores based on all 11 subtests differ from those produced within respective Verbal and Performance scales. Evident also is the fact that, when the ipsative scores (for any individual) are themselves summed, the result is always zero. This property leads to the formal statistical definition of ipsative scores (Clemans, 1965) as a set of scores for which the sum of scores across the attributes

TABLE 1
NORMATIVE AND IPSATIVE SCORE VALUES FOR A GIVEN WISC-R SUBTEST PATTERN

Score form	Verbal IQ scale subtests							Performance IQ scale subtests						Sum of 11 subtests
	Inf	Sim	Arh	Voc	Com	DSp	Sum	PCp	PAr	BDn	OAs	Cod	Sum	
Normative	5	10	11	8	7	7	48	6	7	14	13	11	51	99
Ipsative														
Via 11 subtests[a]	−4.0	1.0	2.0	−1.0	−2.0	−2.0	—	−3.0	−2.0	5.0	4.0	2.0	—	0.0
Via IQ scales[b]	−3.0	2.0	3.0	0.0	−1.0	−1.0	0.0	−4.2	−3.2	3.8	2.8	0.8	0.0	0.0

Note. Inf = Information, Sim = Similarities, Arh = Arithmetic, Voc = Vocabulary, Com = Comprehension, DSp = Digit Span, PCp = Picture Completion, PAr = Picture Arrangement, BDn = Block Design, OAs = Object Assembly, Cod = Coding.

[a]Values are based on all subtests ipsatized simultaneously, where the normative score $M = 9.0$.

[b]Values are based on six Verbal and five Performance IQ scale subtests ipsatized separately, where the normative scores Verbal $M = 8.0$ and Performance $M = 10.2$

for each subject is a constant. These were more precisely termed "purely ipsatized scores" by Hicks (1970, p. 169) to differentiate them from the partially ipsatized scores sometimes produced by personality, attitude, or interest inventories (see Note 3). When, as in the present instance, the summative constant for every subject is zero, the scores are properly referred to as *deviational* ipsatives (Radcliffe, 1963), thus addressing the positive and negative values making up each child's profile. Since deviational ipsatives are computed directly from normative scores, where the normative mean for each attribute is identical, the mean of ipsative scores on any attribute for all children in the normative population will also be zero.

In subsequent sections of this article, we consider the value of ipsative approaches to ability assessment. Our strategy conforms to the unified validity model introduced by Messick (1989). Thus, we address the extent to which empirical fact and theoretical rationale support contemporary applications of ipsative ability measures. The judgment of validity emerges first from internal and external evidence on whether ipsatives appropriately represent intended ability constructs (or as Messick, 1989, p. 20, terms it, the "evidential bases" for validity claims) and ultimately from the determination whether consequent uses lead to intended outcomes in research, instruction, intervention, and so forth (the "consequential bases" for validity). To obtain a comparative perspective of ipsative measures, their relative efficacy is examined in the light of their natural counterparts—normative ability scores. The discussion turns first to test-internal evidence for the relative validity of normative and ipsative measures in reflecting the construct of intelligence.

INTERNAL EVIDENCE FOR VALID REPRESENTATION OF THE INTELLECTIVE CONSTRUCT

Convergence of Measures

Spearman (1927), Matarazzo (1972), and Brody (1985) each emphasized the fact that the attributes that compose omnibus ability (or intelligence) tests theoretically must be substantially and positively correlated. Thus, although certain tests may offer more or less distinguishable subsets of attributes, all ability measures are expected to carry a certain amount of common variance reflecting Spearman's *g* factor. As noted by Brody (1985) and Matarazzo (1972), all major published tests have satisfied the expectation.

Attempts to produce tests that measure intellective traits going appreciably beyond *g* have been largely unsuccessful, beginning with Vernon (1961) who, after extracting *g* as a general factor, tried to identify portions of residual test variance that might indicate meaningful areas of subability. The major contemporary proponents of the existence of specific rather than general ability factors have been Guilford (Guilford, 1964, 1967; Guilford & Hoepfner, 1971) and Cattell (1971). Guilford proposed some 120 independent dimensions of intelligence, and Cattell argued for subdomains of crystallized and fluid intelligence, each applying factor analysis to confirm theoretical posits. Nonetheless, such efforts have not been

regarded as fruitful and have been effectively criticized by Cronbach and Snow (1977), Eysenck (1979), and Brody and Brody (1976). Their general conclusion has been that, without evidence for some viable alternative, the positive intercorrelation of ability attributes and emergence of common *g* variance are requisites to the construct validity of omnibus ability devices.

The WISC-R manual (Wechsler, 1974, Table 15, p. 47) presents the correlations among the 11 subtests averaged across ages 6½ to 16½ for the 2,200 subjects forming the standardization sample. As anticipated, all correlations are positive and the grand mean intercorrelation (as we compute based on Fisher's *z* transformation) is .42, suggesting substantial common variation among the subtests. In addition, the degree to which the various subtests carry the common variance associated with *g* can be estimated by the magnitude of their correlations with the WISC-R Full Scale IQ (FSIQ). The grand mean correlation of subtests with FSIQ for the standardization sample is a sizable .69.

Table 2 displays comparable subtest intercorrelations, correlations with FSIQ, and grand means for the same 2,200 children after scores are rescaled to deviational ipsative form. Values appearing in the upper half of the table are based on ipsatization using all 11 subtests, and values in the lower half on ipsatization of subtests within respective Verbal and Performance scales. Several dramatic contrasts are evident between these values and those obtained from normative scores. First, whereas all normative intercorrelations are positive, the majority of ipsative intercorrelations are negative. Second, whereas the attributes in their normative form correlate on the average at .42, their average intercorrelation drops to near zero (−.09) after either form of ipsatization. Third, the strong average relationship seen between the normative ability attributes and general intelligence (.69) reduces to .02 for both ipsative score forms.

Through factor-analytic experiments with ability measures, Guilford (1952) surmised that about two-thirds of the correlations among ipsatized attributes would be found negative. As Clemans (1965) demonstrated, however, the number of negative values in an ipsative correlation matrix depends more precisely on the number of attributes involved and the distribution of correlations among the normative attributes. At least half of the ipsative intercorrelations will always be negative, with the proportion tending to increase as the number of attributes decreases or as the distribution of normative intercorrelations deviates from normality. Upon ipsatization of WISC-R scores using all 11 subtests, 42 of the possible 55 correlations (76.4%) are found negative, and using the Verbal and Performance scales, 38 (69.1%) are negative.

A key property of ipsative correlation matrices is the tendency for the average correlation, as in the WISC-R examples, to be negative or near zero. According to Clemans (1965) and Radcliffe (1963), when ipsative variances are equal, the average correlation will approximate $-1/(m-1)$ where m is the number of attributes. Effectively, this figure is zero, but it will deviate farther from zero as the number of attributes decreases. In contrast, Smith (1965) pointed out that the average intercorrelation of normative attributes is determined mainly by the relationships among the contents of attribute scales and the psychological characteristics of the respondent population.

TABLE 2
AVERAGE INTERCORRELATION PATTERNS FOR WISC-R SUBTESTS AND FULL SCALE IQ BASED ON IPSATIVE SCORES IN THE STANDARDIZATION SAMPLE

Subtest	Inf	Sim	Arh	Voc	Com	DSp	PCp	PAr	BDn	OAs	Cod
Ipsatized using 11 subtests											
Sim	08					Mean subtest intercorrelation = −09					
Arh	04	−13				Mean subtest correlation with FSIQ = 02					
Voc	21	17	−07								
Com	00	07	−10	20							
DSp	−11	−16	13	−13	−21						
PCp	−19	−10	−17	−20	−10	−21					
PAr	−15	−15	−22	−16	−10	−18	−03				
BDn	−15	−17	−13	−26	−21	−18	07	−04			
OAs	−21	−18	−27	−28	−19	−21	10	03	24		
Cod	−21	−21	−04	−19	−16	05	−20	−09	−10	−14	
FSIQ	14	23	−04	25	05	−30	−02	−03	14	01	−20
Ipsatized using Verbal and Performance IQ scales											
Sim	−12					Mean subtest intercorrelation = −09					
Arh	−16	−33				Mean subtest correlation with FSIQ = 02					
Voc	−02	−04	−34								
Com	−20	−07	−27	02							
DSp	−31	−33	−01	−36	−37						
PCp	−01	07	−01	02	05	−09					
PAr	04	01	−08	07	05	−06	−23				
BDn	01	01	07	−03	−05	−02	−16	−29			
OAs	06	06	−07	00	02	−04	−18	−26	−04		
Cod	−09	−12	08	−05	−06	15	−37	−24	−29	−39	
FSIQ	12	21	−09	25	01	−34	01	00	20	05	−18

Note. Entries are average coefficients of correlation across 11 age levels computed through Fisher's *z* transformation. $N = 2{,}200$. Decimal points are omitted for convenient presentation. Inf = Information, Sim = Similarities, Arh = Arithmetic, Voc = Vocabulary, Com = Comprehension, DSp = Digit Span, PCp = Picture Completion, PAr = Picture Arrangement, BDn = Block Design, OAs = Object Assembly, Cod = Coding. The data in this table are from *Research Report No. 90-1* by P.A. McDermott, 1990, San Antonio, TX: The Psychological Corporation.

The discovery of predominantly negative bivariate relationships and near-zero average relationships among ability attributes runs contrary to the theoretical expectation for constructs that might reflect some meaningful aspect of Spearman's *g*. Of course, negative relationships by themselves do not preclude the possibility that ipsatized attributes are somehow strongly, albeit bidirectionally, related. To assess the possibility, the grand mean intercorrelations were recomputed using unsigned bivariate values. The average unsigned intercorrelation for ipsative attributes using all 11 subtests is .15, and that using Verbal/Performance scales is .13, indicating that ipsatization alters not only the direction of attribute relationships, but also substantially reduces the overall strength of those relationships.

A further consideration for the construct validity of ipsative ability measures is their relative insensitivity to the general ability factor permeating normative measures. Although the average unsigned correlation of the normative subtests with FSIQ remains .69, the corresponding value for both forms of ipsative subtests is only .13. The fundamental question is: What happens to Spearman's *g* upon ipsatization? The answer most likely rests in evidence from earlier empirical investigations of the influences of ipsatization on absolute measurement units. Radcliffe (1963) concluded that when normative scores are transformed into deviational ipsatives, all first-centroid variance is removed from resultant scores. The first centroid approximates Hotelling's (1933) first principal component, and removal of the latter, as shown by Clemans (1965), results in the loss of the greatest integral portion of common variance possible from a set of scores. For the case of normative ability scores, this means the loss of the general ability factor.

Clemans (1965) stressed the fact that ipsatization inherently diminishes the amount of useful information carried by test scores. Based on the first unrotated principal factor extracted for the WISC-R normative sample (drawn from data given by Kaufman, 1979, p. 110, Table 4-1) and the procedure advised by Wrigley (1958) and Silverstein (1976) for isolating variance sources, we estimate that ipsatization removes about 92% of the common variance from the 11 subtests. This constitutes a loss of about 55% of all reliable variance for the total test.

In summary, it is clear that ipsatized ability scores do not interrelate in a fashion similar to normative scores, do not maintain the strength of relationship characterizing normative scores, and are not sensitive to the general ability factor carried by normative scores. At this point in the analysis, one could not dismiss the possibility that ipsative scores might reflect some useful psychological qualities. Nevertheless, the relevant evidence indicates that ipsative scores do not measure the same constructs conveyed by conventional ability scores, and it is not known what constructs they do measure.

Reliability of Measures

Although certain properties of ipsative scores have been investigated previously by Harris (1953), Smith (1965), Clemans (1965), and Guilford (1952), research has not examined the reliability question. It seems logical to expect that any transformation procedure that would remove the more robust contribution of common variance is also likely to affect score reliability. The absence of inquiry probably is due in part to the fact that internal consistency estimates for ipsative measures cannot be produced because the necessary item components relate properly only to normative scores. But temporal stability approaches to reliability remain feasible, and, especially as pertains to ability measures intended for periodic reassessment, stability is an essential quality.

To this end, we examined both the short- and long-term stability for normative and ipsative scores. Short-term analyses were conducted with the WISC-R test–retest sample ($N = 303$) drawn representatively from the normative sample at the time of standardization (Wechsler, 1974, pp. 29–31). The sample included 97 children ages 6 and 7, 102 children ages 10 and 11, and 104 children ages 14 and

15, with the distributions of child gender, race, and parental occupation proportionate to the U.S. Census data. Each child was retested on the WISC-R approximately 1 month after original testing. Long-term analyses employed a cohort of special education children ($N = 189$) attending public schools in the Phoenix, Arizona, area. Ages ranged from 6 through 16 years, with 77.6% of the children having been classified with learning disabilities, 20% with emotional impairment, and 2.4% with mental handicaps. Each child was retested on the WISC-R approximately 3 years after initial testing, in accordance with school district policy affecting exceptional children.

Results are presented in Table 3. The average short-term reliability across normative ability attributes is .78, dropping to .63 upon ipsatization using all subtests and .62 using separate IQ scales. Both decrements are significant statistically at the .001 level. Similarly, the average long-term reliability for normative scores ($\bar{r} = .50$) was reduced significantly after ipsatization via all subtests ($\bar{r} = .37$, $p < .01$) and via Verbal/Performance scales ($\bar{r} = .28$, $p < .001$). For both short- and long-term analyses, ipsativity derived through separate IQ scales versus all ability attributes produced the least satisfactory results.

It was hypothesized that reliability would suffer by the loss of common variance associated with ipsative measures. Inspection of the coefficients posted in Table 3 supports this contention. Kaufman (1979, pp. 110–113) reported the proportions of common versus specific variance associated with each WISC-R subtest in the standardization sample. The subtests carrying the most common variance and least specific variance are Vocabulary and Similarities, and those carrying the most specific and least common variance are Coding and Digit Span. It follows, therefore, that if reliability decrements attendant upon ipsatization are due largely to losses of g variance, attributes that in the normative form retain the most common variance should be most affected, and conversely, those originally associated with specific variance should be the least affected. As expected, Vocabulary and Similarities evince the greatest reliability drops with ipsatization and Coding and Digit Span the least, and with Coding's ipsative reliability not differing significantly from its normative reliability.

Whereas examination of score stability provides important perspective for the ipsative question, it falls short in one respect. We have said that ipsative assessment is now commonplace in psychological practice. Proponents see value in the discovery of a child's unique pattern of stronger and weaker abilities, with the intent or implication that strengths might be fostered and weaknesses remedied. Within this context it can be seen that the reliability question pertains more properly to the stability of a detected strength or weakness, per se, than to stability of underlying scores. To provide the needed perspective, we applied one of the most popular methods by which practitioners are encouraged to distinguish ability strengths and weaknesses (Ingram, 1985; Kaufman, 1979). Specifically, for both test–retest samples employed above, any ipsative score at test or retest equal to or greater than +3 points was regarded as a strength, and any equal to or less than −3 points a weakness. Classificatory stability for observed strengths and weaknesses was assessed using Fleiss's (1971) extension of coefficient kappa to partial kappa via computer program CONGRU (Watkins & McDermott, 1979).

TABLE 3
RELATIVE RELIABILITY OF NORMATIVE AND IPSATIVE WISC-R SUBTEST SCORES ACROSS 1 MONTH AND 3 YEARS

Score form	Inf	Sim	Arh	Voc	Com	DSp	PCp	PAr	BDn	OAs	Cod	Average
One-month standardization test–retest sample (N = 303)[a]												
Normative (r_N)	86	80	78	82	81	79	79	72	81	72	68	78
Ipsative (r_I)												
Via 11 subtests	68***	59***	60***	61***	57***	73**	65***	64***	64***	55***	66	63***
Via IQ scales	69***	57***	57***	55***	64***	71***	67***	56***	61***	49***	67	62***
Three-year special education test–retest sample (N = 189)												
Normative (r_N)	56	60	46	69	47	52	40	35	60	48	42	50
Ipsative (r_I)												
Via 11 subtests	36***	42***	23***	54***	22***	53	30*	18***	44***	38**	43	37**
Via IQ scales	29***	33***	21***	33***	10***	55	19***	18***	30***	18***	38	28***

Note. Statistical tests assess the significance of the difference between r_N and r_I, where r_N is the reliability coefficient for normative scores and r_I is the reliability coefficient for ipsative scores based on either all 11 subtests or on the Verbal and Performance scale subtests, respectively. Each test is a one-tailed application of the Pearson-Filon standard error of the difference between two correlation coefficients in the test–retest correlated series (Peters & VanVoorhis, 1940, p. 185). Decimal points are omitted from coefficients for convenience of presentation. INF = Information, Sim = Similarities, Arh = Arithmetic, Voc = Vocabulary, Com = Comprehension, DSp = Digit Span, PCp = Picture Completion, PAr = Picture Arrangement, BDn = Block Design, OAs = Object Assembly, Cod = Coding.

[a]The data pertaining to the WISC-R test–retest sample are from *Research Report No. 90-1* by P.A. McDermott, 1990, San Antonio, TX: The Psychological Corporation.

*$p < .05$. **$p < .01$. ***$p < .001$.

Considering the approach to ipsatization based on all 11 subtests, results showed that, given the presence of an ipsatively determined strength, the simple probability that it would have existed a month earlier or would continue to exist a month later is 42.6%, which is 16.0% above chance and not statistically significant. The general stability of ipsative weaknesses over a month is 34.5%, which is only 5.9% above chance expectancy (and also nonsignificant). Turning to the special education sample to examine stability over a 3-year interval, stability for ipsative strengths drops to 19.4% (which is 13.2% *below* chance expectation, indicating gross disparity), and stability for weaknesses is 21.7% (14.9% below chance). Also revealing was the observation that ipsative strengths and weaknesses would tend to remain unchanged about 30% to 32% of the time over a 1-month period and 29% to 32% over a 3-year period as dictated merely by *chance* factors alone (see Note 4).

Analyses based on ipsatives derived within separate Verbal and Performance scales indicated stability levels consistently below (by several percentage points) those using all 11 subtests. In sum, ipsative strengths and weaknesses reflected no statistically significant or practically consequential stability over 1 month and showed clear instability over 3 years.

Considering the reliability of normative and ipsative measures over both the short and long term, the evidence shows that ipsative ability measures are significantly less reliable than conventional norm-based ability measures.

EXTERNAL EVIDENCE FOR VALID REPRESENTATION OF THE INTELLECTIVE CONSTRUCT

Among the most critical tests of the comparative efficacy of normative and ipsative measures is their ability to inform an accurate prediction of some meaningful outcome. Observing the growing popularity among test developers, theoreticians, and practitioners for devices that promise understanding of specific rather than general human abilities, Brody (1985) commented:

> Any theory of the structure of the intellect that proposes to replace *g* with a more differentiated concept ought to have predictive validity that exceeds that obtained from a general factor. (p. 357)

Within this framework, we undertook a series of analyses to assess two aspects of predictive capacity; first, the relative strength of bivariate relationships for normative and ipsative scores against standardized achievement criteria and, second, the relative power of multivariate predictions for each score type against those same criteria.

The sample data were obtained from cooperating school systems in Arizona, Delaware, and Indiana ($N = 1{,}418$); they involve all of the 6½- to 16½-year-old children examined for possible classification as special education students during the 1986–88 school years. Approximately 49.3% of the children were ultimately classified with learning disabilities, 10.1% with mental retardation, and 9.2% with emotional disturbance. The remaining 31.4% of the sample had no classifiable handicaps. Each child was administered the 11 WISC-R subtests and relevant

achievement scales of at least one of the following individually administered achievement tests: Woodcock-Johnson Psycho-Educational Battery (W-J) (Woodcock & Johnson, 1977), Peabody Individual Achievement Test (PIAT) (Dunn & Markwardt, 1970), Wide Range Achievement Tests–Revised (WRAT-R) (Jastak & Wilkinson, 1984), and Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983b) (see Note 5).

Bivariate Prediction

Table 4 presents mean validity coefficients for normative and both forms of ipsative scores. The grand mean coefficient for normative scores is .24, whereas ipsatives based on all 11 subtests yield a value of .01 and ipsatives based on Verbal/Performance IQ scales a value of .00. These results conform to the ipsative property noted by Hicks (1970) and Clemans (1965) that, when ipsative variances are equal (as they tend to be for standardized ability tests), their average validity coefficient is zero.

The strength of relationships between the various ability predictors and achievement criteria is best estimated through unsigned validity coefficients. Viewed from this perspective, the grand mean coefficient for normative scores remains unchanged at .24, with coefficients for ipsative score forms .09 and .06, respectively. This suggests, on the one hand, that ipsative ability scores have a comparatively weak relationship to important validity criteria and, on the other hand, that there is no advantage to ipsatization by separate Verbal/Performance attributes as a means of mitigating such weakness.

Multivariate Prediction

Multiple R provides an excellent indication of predictive validity, as it is the correlation reflecting the best possible combination of the ability attributes for estimating criterion performance. Multiple Rs produced by simultaneous regression with WISC-R normative and ipsative subtest scores are listed in Table 4 (see Note 6). The values produced by ipsatized ability scores are markedly below those produced by normative scores. Overall, the normative predictors account for an average of 27.0% of criterion performance ($\overline{R} = .52$), whereas ipsatives derived from all 11 subtests account for 9.0% ($\overline{R} = .30$) and those derived from respective IQ scales only 6.1% ($\overline{R} = .24$). Thus, by ipsatization, about two-thirds to three-quarters of the predictive efficiency afforded by normative scores is lost (see Note 7). Also apparent is the fact that ipsatization based on separate IQ scales reaps poorer predictions than ipsatization based on all attributes taken simultaneously.

We were reminded by a reviewer of an earlier version of this article that, as ipsatives are sometimes applied, their contribution is not viewed in isolation but in addition to that afforded by global IQ measures (i.e., FSIQ or VIQ and PIQ). Such application begs the question as to how much information is actually gained *beyond* that supplied already by general IQ. To this point we constructed for each of the 12 achievement areas noted previously a series of stepwise, hierarchical, regression models. The models alternatively entered FSIQ or VIQ and PIQ at the

TABLE 4
RELATIVE VALIDITY OF NORMATIVE AND IPSATIVE WISC-R SUBTEST SCORES FOR PREDICTING ACADEMIC ACHIEVEMENT

Achievement criterion	*N*	Mean bivariate r[a]			Multiple R		
			Ipsative scores			Ipsative scores	
		Normative scores	Via 11 subtests	Via IQ scales	Normative scores	Via 11 subtests	Via IQ scales
W-J Reading	322	.21	.01	−.00	.39	.26	.21
W-J Mathematics	323	.27	.01	.01	.53	.34	.28
W-J Written Language	289	.20	.01	.00	.33	.26	.23
PIAT Reading Recognition	586	.22	.01	−.00	.53	.30	.22
PIAT Reading Comprehension	490	.18	.00	−.00	.45	.28	.22
PIAT Mathematics	571	.26	.02	.00	.65	.31	.23
PIAT Spelling	250	.22	.01	.00	.54	.36	.28
PIAT Information	215	.29	.02	.01	.75	.53	.49
WRAT-R Reading	1095	.23	.00	.00	.38	.21	.17
WRAT-R Arithmetic	1095	.09	.01	.00	.52	.20	.17
WRAT-R Spelling	847	.13	.00	−.00	.24	.13	.10
K-ABC Achievement Scales	258	.38	.01	.00	.71	.37	.32

Note. W-J = Woodcock-Johnson Psycho-Educational Battery; PIAT = Peabody Individual Achievement Test; WRAT-R = Wide Range Achievement Test–Revised; K-ABC = Kaufman Assessment Battery for Children.
[a]Mean correlation coefficients are computed using Fisher's z transformation.

first step in forward inclusion, followed alternatively by either ipsative or normative subtest scores. The outcomes are summarized across the 12 achievement criteria.

On the average, FSIQ alone accounted for 19.6% of criterion variance or, alternatively, VIQ and PIQ accounted for 22.1% of such variance. Adding ipsatized subtest scores to FSIQ increased prediction efficiency by 7.8% to account for 27.4% of achievement variation, whereas adding ipsatives to VIQ and PIQ increased efficiency by 5.5% to account for 27.6% of criteria. This indicates that, indeed, ipsatives carry some predictive information missing from global ability measures. However, when normative subtest scores are substituted for ipsatives as a complement to FSIQ and VIQ/PIQ, essentially equivalent amounts of achievement variance are explained (27.6% and 27.7%, respectively), demonstrating that ipsatives carry no information not provided already by normative counterparts.

Hierarchical regression models help also to illustrate the comparative predictive value of ipsative and normative subtest scores. Here we ask, Given the fact that normative scores predict about 27% of achievement criteria and ipsative scores about 9% (refer to Table 4), how much are normative scores assisted by ipsative scores and vice versa? We found ipsative scores uniformly incapable of improving upon normative score prediction. In fact, as a signal for ipsatives' inability to assist normatives, ipsative scores were universally precluded equation entry by any attainable regression tolerance criteria. On the other hand, the addition of normative to ipsative scores consistently incremented predictive efficiency to the level attainable by normative scores alone (27%) (see Note 8).

The broader conclusions that we may draw from the comparative criterion validity of normative and ipsative measures are that (a) ipsative ability scores generally and substantially underperform counterpart normative scores in their capacity to relate to or predict meaningful phenomena outside the ability test and that (b) ipsative ability scores contribute no information to the prediction process not already provided by counterpart normative ability scores.

VALIDITY AS JUDGED BY USEFUL APPLICATIONS IN RESEARCH AND PRACTICE

The application and interpretation of any measurement unit must be guided by its enabling and limiting properties. In this final section we discuss some of the most commonly misunderstood properties of ipsative measures as they influence current educational and psychological practice.

Limited Relevance

Ipsative scores are frequently treated as if they are isomorphic variations or transformations of normative scores that would permit comparable interpretations. Thus, for example, users may be inclined to view ipsatives as relevant not only to ability distinctions within a given child, but also to distinctions between children, as in the case of profile typing or differentiation among children in the light of apparent patterns of strength and weakness. Actually, ipsative measures

represent a true change of metric. They require that every person's normative scores be altered by a *different* value, namely, his or her own personal average performance. The metric is no longer parametric across individuals but one pertinent only to the solitary individual under study.

Appreciating this idiometric quality of ipsatives, Cattell (1944) pointed out that once a person's scores have been ipsatized, their entire import is relative *to that person alone.* Allport and Vernon (1931), although not using the term *ipsative* in *A Study of Values,* similarly cautioned that such measures could have no meaning beyond a particular individual's personality. Perhaps the limited relevance of ipsatives is best illustrated by Clemans (1965). We are reminded that an obvious purpose of ability tests is to discern how much or how little ability one possesses. The quest is readily resolved by an examination of normative ability measures, because with this information we can tell not only how able one is compared to others (population-relevant performance) but how much more or less able one is across a profile of his or her own normative ability scores (person-relevant performance). Yet, upon ipsatization of those abilities, the former information is forfeited. Given the ipsative scores of two children for the same ability area, one child having a rather high score value and the other a low value, it is impossible to tell which child has more ability and it is entirely possible that the child having the higher score possesses less ability than the other child.

One must also ponder the implications of the primary statistical function differentiating normative and ipsative scores: Namely, deviational ipsatives are nothing more than normative measures devoid of all information pertaining to first-centroid or common variance. Thus, ipsatives tautologically convey less information than do normative scores, with ipsatives conveying no information not alternatively offered by normative scores. In this sense, ipsative ability measures not only have virtually no nomologic relevance, they further have no more idiologic relevance than their normative alternatives.

Limited Applicability

Several writers have noted the common misapplication of ipsative solutions to normative problems (Anastasi, 1988; Hicks, 1970; Smith, 1965). The personal frame of reference characterizing ipsative measures effectively precludes group comparisons that would assume a common metric or the preservation of common variance.

Broverman (1961) encouraged the factor analysis of ipsative measures to disclose "personological" ability dimensions. He observed that the consequent dimensions retained a certain reciprocally exclusive character whereby, as an individual exhibited greater ability in one area, commensurate inability was apparent in another area. He regarded this as an interesting and useful perspective on human ability, theorizing that it represented evidence of choice-points for every developing individual. Thus, in the normal course of growth, points are reached at which an individual must elect to specialize in one area of performance at the expense of another area. But Guilford (1952), however dedicated to the search for specific intellectual abilities, advised against the use of ipsatives for that effort based mainly

on the correlational evidence suggesting that ipsative measures change the very construct nature of the target phenomena. Harris (1953), intending to compare latent normative and ipsative dimensions, tried unsuccessfully to factor-analyze ipsative score matrices, and Tucker (1956) and Clemans (1965) subsequently recommended against the application of any type of traditional factor analysis for ipsative data. Clemans (1965) went on to stress that, given the nature and extent of the information lost with an ipsative correlation matrix, this fact alone would make it nearly impossible to reach any psychologically meaningful conclusions through factor-analytic procedures.

Use of deviational ipsative rather than normative measures frequently is recommended for typological studies such as clustering of score profiles (e.g., see Borgen & Barnett, 1987). The purpose is twofold: first, to accentuate the role of the presumably idiologic patterns or shapes defining profiles and, second, to preclude profile levels or average performance from dominating the typal solution. At first glance the strategy appears sensible, given the intention and the fact that ordinary *Q*-type analyses do not demand the use of parametric scores. But as we have gleaned from the psychometric properties of ipsatives, any consequent typology will almost inevitably be less reliable and valid than its normative counterpart. As observed in typological investigations with nationally standardized intelligence tests (McDermott, Glutting, Jones, & Noonan, 1989; McDermott, Glutting, Jones, Watkins, & Kush, 1989), profile cluster solutions based on ipsatized scores tend to be unreplicable, temporally unstable, unrelated to meaningful external criteria, and very difficult to interpret. Moreover, inasmuch as typologies are built upon nonparametric measures, they preclude application of the more useful statistics designed for group comparison and multivariate classification.

Diminished Gains

We noted previously the popular practice of identifying ability strengths and weaknesses using ipsative scores, the implication being that detected strengths should be maintained or exploited and that weaknesses should be avoided or remediated. Notwithstanding the dubious value of detected strengths or weaknesses given their temporal instability, the practice is logically faulted because it ignores another unfortunate aspect of ipsative concepts. That is, there can be no gains without corresponding losses—what we call the seesaw effect.

To demonstrate the phenomenon, Table 5 presents normative and ipsative ability scores for a hypothetical youngster. Pattern A shows the child's relative strengths and weaknesses at the time of initial assessment. To strengthen the example, notable score deviations are those attaining statistical significance (Sattler, 1988, Table C-3, p. 815), with significant positive deviations indicating strengths (appearing in italics) and negative deviations indicating weaknesses (boldface). (For the sake of brevity, the example considers only that form of ipsatization based on all 11 subtests.)

The ipsative pattern shows three areas of relative strength and two of weakness, but the area of greater weakness is Coding performance (ipsative score = −5.0). Now let us assume that Coding performance per se reflects some important

TABLE 5
NORMATIVE AND IPSATIVE SUBTEST PATTERNS FOR A GIVEN CHILD BEFORE AND AFTER IMPROVEMENT IN THE WEAKEST AREA OF PERFORMANCE

Score form	Inf	Sim	Arh	Voc	Com	DSp	PCp	PAr	BDn	OAs	Cod	Sum
Pattern A: Initial performance (FSIQ = 107)												
Normative	14	17	9	16	15	10	7	10	8	9	6	121
Ipsative	3.0	*6.0*	−2.0	*5.0*	*4.0*	−1.0	**−4.0**	−1.0	−3.0	−2.0	**−5.0**	0.0
Pattern B: Improved Coding, other abilities constant (FSIQ = 111)												
Normative	14	17	9	16	15	10	7	10	8	9	11	127
Ipsative	2.6	*5.6*	−2.5	*4.6*	3.6	−1.5	**−4.5**	−1.5	**−3.5**	−2.5	−0.5	0.0

Note. Ipsatization is based on all 11 subtests where the normative score $M = 11.00$ for Pattern A and 11.45 for Pattern B. Ipsative scores are shown in italics to indicate statistically significant areas of relative strengths and in boldface to indicate relative weakness, where statistical significance is determined through the standard error of the deviation of test scores from the mean of such test scores, with Type I error controlled by Bonferroni correction (see Sattler, 1988, Table C-3, p. 815). FSIQ is based on 10 subtests excluding Digit Span, which remains constant across subtest patterns. INF = Information, Sim = Similarities, Arh = Arithmetic, Voc = Vocabulary, Com = Comprehension, DSp = Digit Span, PCp = Picture Completion, PAr = Picture Arrangement, BDn = Block Design, OAs = Object Assembly, Cod = Coding.

intellective construct and that we have available an intervention process that will improve both Coding performance and the construct it measures. After many months of treatment, we reassess the child's abilities and discover that ipsative Coding performance has indeed improved (no longer as deviant as −5.0 points). The question becomes, Is all well, or is it just an illusion?

Pattern B offers one potential outcome. Considering its normative score version, the child's Coding proficiency has risen from 6 to 11 scaled score points, every other ability area has remained happily constant, and the Coding improvement is properly reflected in an FSIQ increment from 107 to 111. But the ipsative version of the pattern tells quite another tale. Here, the elimination of Coding as an area of weakness if offset by the loss of Comprehension as a strength and by the addition of Block Design as another weakness. From the ipsative perspective, the child is no better off after than before intervention. This points to the natural economy of ipsative assessment which dictates that, for every gain, there must be a reciprocal loss.

As we close the discussion on consequential validity of ipsative measures, we would hope that, by the foregoing hypothetical example on outcome validity, readers do not acquire the impression that there exists some body of knowledge effectively showing that ipsative abilities are somehow malleable to treatment. As we have emphasized with respect to internal and external evidence of validity, there is no clear understanding as to what constructs might be carried by ipsative ability measures nor as to what meaningful criteria those constructs might relate. So far, claims that ipsative abilities would respond to intervention or would inform

instruction, and whether observed changes would translate into meaningful benefits for children, are entirely speculative.

CONCLUSIONS AND FURTHER IMPLICATIONS

The evidence on the comparative efficacy of normative and ipsative ability assessment is consistent and compelling. Ipsative measures have insufficient reliability for individual educational decisions, are significantly less reliable than normative measures, and are relatively insensitive to sources of individual variation that characterize omnibus ability measures. Further, any argument in favor of ipsatized assessment certainly is vitiated by the fact that such approaches fail to predict outcomes as well as normative approaches. And, were all of this not the case, we would still be left with uncertainty about the meaning of ipsative constructs and their limited utility for either group or individual studies.

We must stress also that the limiting aspects of ipsative assessment are not mitigated by procedures that would have users concentrate on deviations within theoretically integral or factorially verified subsets of ability attributes. Thus, as is clear for the WISC-R case, there is no evidence to warrant comparison of Verbal subtest scores with the mean of Verbal subtests nor of Performance subtest scores with their mean. In fact, given the tenuous reliability of ipsative measures, any method that bases assessment on less than the full complement of meaningful attributes is likely to exacerbate rather than remedy the problems. This is particularly true when considering the relative predictive efficiency and temporal stability wrought by measures ipsatized with respect to a general ability factor versus some more elaborate ability dimensions subsumed within the general factor.

The most compelling evidence is that demonstrating a substantial loss of information by way of ipsatization. Normative ability measures carry all of the potential advantages afforded by ipsative measures. Ipsative approaches provide but a portion of the information already available through conventional norm-based ability measures and bring with them a great variety of disadvantages. Consequently, we must recommend against ipsative approaches for assessing human ability.

Can we then at least endorse comparisons among regular norm-based subtest scores? We cannot. That would be legitimate only were the individual subtests found to have unique construct identity and were knowledge of their variation or covariation found to be useful in classification or treatment.

Individual subtests do tend to retain some distinct variability, which we have referred to as specific variance. But specific variance by itself is not sufficient grounds for a claim that a subtest measures something useful. In spite of all the speculation about what subtests might uniquely measure, there is virtually no empirical support for the many claims. Nor, as Cahan (1986) and others have admonished, can we think that statistically significant comparisons among subtests somehow transform those subtests into meaningful traits. To accept this notion is to fall prey to a subtest-trait fallacy (see Tryon, 1979) holding that, just because a test author groups together items sharing similar format or materials, those item groups must necessarily tap some distinct characteristic. Indeed, with

respect to the Wechsler scales, such a premise is antithetical to the developer's basic intent (Zachary, 1990).

As for the usefulness of normative subtest scores in classification, Hale (1979), Hale and Landino (1981), and Thompson (1980, 1981) have found subtest variation rather unhelpful in discriminating among groups. Hale and his associates (Hale & Raymond, 1981; Hale & Saxe, 1983) have further established that subtests add nothing to the classificatory proficiency of global IQ. McDermott, Fantuzzo, and Glutting (1990) noted also a rather troubled history and pervasive methodological problems in the search for diagnostic relevance of ability subtests.

We have discussed the popular impression that for treatment or for educational purposes it is better to have knowledge of an individual's many specific abilities than knowledge of general ability. It is this belief that drives the search for discovery of intellectual strengths and weaknesses, as illustrated for the case of ipsative assessment. But what of the evidence? One cannot help recall the landmark reviews by Cronbach and Snow (1977) of the many years of "aptitude–treatment" interaction research. They discovered that, across the vast literature and with few exceptions, more differentiated and *specific* views of intellectual abilities were *not* the most useful bases for individualizing instruction and curriculum. To the contrary, attention given to global measures of ability reaps noticeably better treatment and instruction outcomes. These observations do not constitute a tacit endorsement for *all* applications of global ability measures. Appropriate uses of global ability are controversial from the standpoint of educational utility (Heller, Holtzman, & Messick, 1982; Reschly, 1988). For example, homogeneous ability grouping for intact classes or consequent ethnic segregation are generally undesirable, both in the policy sense and in their failure to improve learning (Peterson, Wilkinson, & Williams, 1984; Snow, 1986).

Thus we cannot recommend either ipsative or normative approaches for subtest interpretation. Such approaches essentially violate primary principles guiding valid test interpretation as set forth in the American Psychological Association's (1985) *Standards for Educational and Psychological Testing*—specifically, Standard 1.8 concerning construct integrity for subtest interpretations, and Standards 1.2 and 1.3 regulating claims for proper interpretation of subtest variability and discrepancy. Perhaps more important is the fundamental threat that subtest interpretation poses for the basic tenet of scientific parsimony. Namely, subtest analysis is a practice in which "what can be explained by fewer principles is explained needlessly by more" (Occam's Razor) (Jones, 1952, p. 620).

Authors' Note

We gratefully acknowledge the advice and materials given by William V. Clemans, Lou E. Hicks, Paul Horst, and Julian C. Stanley in preparation for the research reported in this article.

Notes

1. Because the more popular Coding subtest is included as a mandatory part of the Performance scale, the alternate Mazes subtest is unused. Both Digit Span and Coding are regarded as primary compo-

nents in most subtest interpretation schemes (Kaufman, 1979, pp. 149–152, 170–171), whereas Mazes often is excluded (e.g., see Bannatyne, 1974; Guilford, 1967).

2. Kaufman (1990, p. 428) recently altered in part his original recommendation to suggest that ipsatization within respective Verbal and Performance scales be carried out only when Verbal and Performance IQs are significantly disparate. It is, nonetheless, the original procedure that has been popularized. Also, the revised recommendation, by application to some but not all children, preserves the original procedure for many children and introduces an aspect of nonparametric score transformation across case studies.

3. The term *ipsative,* as applied mainly in the field of ability assessment, pertains to rescaled normative scores, in which each original normative score represented overall performance on a specific set of items that are similar in format or task requirements (hence, an attribute). Cattell (1957) used the term *normative–ipsative* for such scores. Ipsative may refer also to measurement units applied with certain personality and interest tests to correct for response bias and faking. Here, respondents are forced to choose among items that are grouped into sets of two or more, with competing items typically matched in attractiveness but discrepant in validity against a single criterion, or of comparable validity, but against different criteria. If matched items are relegated to different criteria (attributes) and a final test score is computed for each attribute such that the score for one attribute decreases as that for another increases, the personality or interest scores are regarded as ipsative.

4. With respect to 1-month stability of ipsative strengths, the proportion of simple agreement is .426, simple chance agreement .317, partial kappa .160, and standard error of partial kappa .286. Thus, the percentage of simple agreement is .426 times 100, simple chance agreement .317 times 100, and agreement beyond chance = (.426 – .317/1 – .317)(100) = 16.0%. For 1-month stability of weaknesses, the proportion of simple agreement is .345, chance .304, partial kappa .059, and the standard error .287. Comparable figures for 3-year stability of strengths are simple agreement .194, chance .288, partial kappa –.132, standard error .407, and for stability of weaknesses, simple agreement .217, chance .318, partial kappa –.149, and standard error .394.

5. Because of the unconventional content of the K-ABC's Faces & Places and Riddles achievement subtests and limited age range of the Reading/Understanding subtest, the composite score across achievement subtests was used as the single most relevant scale reflecting K-ABC achievement performance. This decision comports with the discovery of a single latent dimension explaining variation among K-ABC achievement subtests (Wilson, Reynolds, Chatman, & Kaufman, 1985).

6. Computing multiple R for the normative score case is a straightforward application of the least-squares method, a procedure involving the inverse of the matrix of attribute intercorrelations. However, a regular inverse cannot be computed from a singular matrix, and ipsative correlation matrices are distinctly singular. Consequently, an alternative approach is needed if ipsative score sets are to be evaluated through multiple regression.

For those less familiar with matrix algebra, it may help to recall than an individual's ipsative scores are interdependent and always sum to zero. Returning to the WISC-R example, it would follow that if the sum across 11 subtests must be zero, it is also possible to tell an individual's ipsative score on any one subtest merely by knowing the scores on the other 10 subtests. Therefore, ipsative score matrices always contain some redundant information, inasmuch as knowledge of all but one of the attributes is sufficient and equivalent to knowledge of all attributes. This redundancy is, in a sense, the source of the problem preventing a natural solution to the inverse of a full ipsative correlation matrix. But as Clemans (1965) has proven, the problem implies its own solution, because an ipsative matrix can be made basic and nonsingular merely by deleting any one attribute from the matrix. That is, for ipsative measures, the least-squares multiple regression solution based on all but one of the ipsative attributes is identical to the solution that would result using all attributes, were that possible. In the present application, R values were computed by using 11 WISC-R subtests for normative scores and 10 subtests for ipsative scores. Prior to each analysis, it was confirmed empirically, through attempts to conduct regression with and without an 11th subtest, that the corresponding correlation matrices were singular prior to removal of one subtest and nonsingular subsequent to removal of one subtest.

7. These findings comport generally with the mathematical theorem derived by Clemans (1965) for estimating ipsative predictive efficiency from normative predictive efficiency. Essentially, R^2 for any ipsative set can be expressed as a function of the R^2 for the corresponding normative set *minus* the square of the sum of beta coefficients derived for the full normative set. Therefore, except in the rare

event that the beta coefficients for a normative set all equal zero (meaning that even the normative form of a set of test scores has no correlation with criterion), the multiple *R* rendered by ipsative scores will consistently be less than its normative counterpart.

8. W.V. Clemans (personal communication, December 31, 1990), using an unpublished derivation by Paul Horst, has proved with matrix algebra that "any transformation that operates only on the scores of individuals will not result in new measures that will add to the precision of prediction."

References

Allport, G.W., & Vernon, P.E. (1931). *A study of values: Manual of directions.* Boston: Houghton Mifflin.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bannatyne, A. (1974). Diagnosis: A note on recategorization of WISC scaled scores. *Journal of Learning Disabilities, 7,* 272–274.

Borgen, F.H., & Barnett, D.C. (1987). Applying cluster analysis in counseling psychology research. *Journal of Counseling Psychology, 34,* 456–468.

Brody, E.B., & Brody, N. (1976). *Nature, determinants, and consequences.* New York: Academic Press.

Brody, N. (1985). The validity of tests of intelligence. In B. Wolman (Ed.), *Handbook of intelligence* (pp. 353–389). New York: Wiley.

Broverman, D.M. (1961). Effects of score transformations in *Q* and *R* factor analysis techniques. *Psychological Review, 68,* 68–80.

Cahan, S. (1986). Significance testing of subtest score differences: The rules of the game. *Journal of Psychoeducational Assessment, 4,* 273–280.

Cattell, R.B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51,* 292–303.

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

Cattell, R.R. (1957). *Personality and motivation structure and measurement.* New York: World Book.

Clemans, W.V. (1965). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs* (No. 14).

Cronbach, L.J., & Snow, R.E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Delaney, E.A., & Hopkins, T.F. (1987). *The Stanford-Binet Intelligence Scale: Fourth edition examiner's handbook.* Chicago: Riverside.

Dunn, L.M., & Markwardt, F.C., Jr. (1970). *Peabody Individual Achievement Test manual.* Circle Pines, MN: American Guidance Service.

Elliot, C.D. (1990). *Differential Ability Scales introductory and technical handbook.* San Antonio, TX: Psychological Corp.

Eysenck, H.J. (1979). *The structure and measurement of intelligence.* Berlin: Springer-Verlag.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382.

Guilford, J.P. (1952). When not to factor analyze. *Psychological Bulletin, 49,* 31.

Guilford, J.P. (1964). Zero intercorrelations among tests of intellectual abilities. *Psychological Bulletin, 61,* 401–404.

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

Guilford, J.P., & Hoepfner, R. (1971). *The analysis of intelligence.* New York: McGraw-Hill.

Hale, R.L. (1979). The utility of WISC-R subtest scores in discriminating among adequate and underachieving children. *Multivariate Behavioral Research, 14,* 245–253.

Hale, R.L., & Landino, S.A. (1981). Utility of WISC-R subtest analysis in discriminating among groups of conduct problem, withdrawn, mixed, and non-problem boys. *Journal of Consulting and Clinical Psychology, 41,* 91–95.

Hale, R.L., & Raymond, M.R. (1981). Wechsler Intelligence Scale for Children-Revised (WISC-R) patterns of strengths and weaknesses as predictors of the intelligence-achievement relationship. *Diagnostique, 7,* 35–42.

Hale, R.L., & Saxe, J.E. (1983). Profile analysis of the Wechsler Intelligence Scale for

Children-Revised. *Journal of Psychoeducational Assessment, 1,* 155–162.

Harris, L.W. (1953). Relations among factors of raw, deviation, and double-centered score matrices. *Journal of Experimental Education, 22,* 53.

Heller, K.A., Holtzman, W.H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity.* Washington, DC: National Academy Press.

Hicks, L.E. (1970). Some properties of ipsative, normative, and forced normative measures. *Psychological Bulletin, 74,* 167–184.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441, 498–520.

Ingram, R. (1985). *The Kaufman method of WISC-R hypothesis generation* [Computer program]. New York: Wiley.

Jastak, S., & Wilkinson, G.S. (1984). *Wide range achievement tests–Revised.* Wilmington, DE: Jastak Associates.

Jones, W.T. (1952). *A history of Western philosophy.* New York: Harcourt, Brace.

Kaufman, A.S. (1975). Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology, 43,* 135–147.

Kaufman, A.S. (1979). *Intelligent testing with the WISC-R.* New York: Wiley.

Kaufman, A.S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kaufman, A.S., & Kaufman, N.L. (1983a). *Interpretive manual for the Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A.S., & Kaufman, N.L. (1983b). *Kaufman assessment battery for children.* Circle Pines, MN: American Guidance Service.

Matarazzo, J.D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins.

McDermott, P.A., Fantuzzo, J.W., & Glutting, J.J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290–302.

McDermott, P.A., Glutting, J.J., Jones, J.N., & Noonan, J.V. (1989). Typology and prevailing composition of core profiles in the WAIS-R standardization sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 118–125.

McDermott, P.A., Glutting, J.J., Jones, J.N., Watkins, M.W., & Kush, J. (1989). Core profile types in the WISC-R national sample: Structure, membership, and applications. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 292–299.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Peters, C.C., & VanVoorhis, W.R. (1940). *Statistical procedures and their mathematical bases.* New York: McGraw-Hill.

Peterson, P., Wilkinson, L.C., & Williams, M. (Eds.). (1984). *The social context of instruction: Group organization and group processes.* New York: Academic Press.

Psychological Corporation. (1986a). *WAIS-R microcomputer-assisted interpretive report* [Computer program]. New York: Author.

Psychological Corporation. (1986b). *WISC-R microcomputer-assisted interpretive report* [Computer program]. New York: Author.

Radcliffe, J.A. (1963). Some properties of ipsative score matrices and their relevance for some current interest tests. *Australian Journal of Psychology, 15,* 1–11.

Reschly, D.J. (1988). Obstacles, starting points, and doldrums notwithstanding: Reform/revolution from outcomes criteria. *School Psychology Review, 17,* 495–501.

Sattler, J.M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Silverstein, A.B. (1976). Variance components in the subtests of the WISC-R. *Psychological Reports, 39,* 1109–1110.

Smith, H.E. (1965). A critique of ipsative measures with special reference to the Navy Activities Preference Blank. In *U.S. Naval Personnel Activity* (Technical Bulletin STB 65-16). San Diego: Bureau of Naval Personnel.

Snow, R.E. (1986). Individual differences and the design of educational programs. *American Psychologist, 41,* 1029–1039.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Thompson, R.J. (1980). The diagnostic utility of WISC-R measures with children referred to a developmental deviation center. *Journal of Consulting and Clinical Psychology, 48,* 440–447.

Thompson, R.J. (1981). The diagnostic utility of Bannatyne's recategorized WISC-R scores with children referred to a develop-

mental evaluation center. *Psychology in the Schools, 18,* 43–47.

Tryon, W.W. (1979). The test-trait fallacy. *American Psychologist, 34,* 402–406.

Tucker, L.R. (1956). *Factor analysis of double-centered score matrices* (Research Memorandum No. 56-3). Princeton, NJ: Educational Testing Service.

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

Watkins, M.W., & McDermott, P.A. (1979). A computer program for measuring levels of overall and partial congruence among multiple observers on nominal scales. *Educational and Psychological Measurement, 39,* 235–239.

Wechsler, D. (1974). *Wechsler intelligence scale for children–Revised.* San Antonio, TX: Psychological Corp.

Wilson, V.L., Reynolds, C.R., Chatman, S.P., & Kaufman, A.S. (1985). Confirmatory analysis of simultaneous, sequential, and achievement factors on the K-ABC at 11 age levels ranging from 2½ to 12½ years. *Journal of School Psychology, 23,* 261–269.

Woodcock, R.W., & Johnson, M.B. (1977). *Woodcock-Johnson psycho-educational battery.* Hingham, MA: Teaching Resources.

Wrigley, C. (1958). Objectivity in factor analysis. *Educational and Psychological Measurement, 18,* 463–476.

Zachary, R.A. (1990). Wechsler's intelligence scales: Theoretical and practical considerations. *Journal of Psychoeducational Assessment, 8,* 276–289.