# Individual Inconsistency: Implications for Test Reliability and Behavioral Predictability

**Susan E. Whitely**
**University of Kansas**

The nature of individual inconsistency in performance on trait measurements is an important topic in psychometrics because of its direct relevance to measurement reliability. Several studies have supported short-term inconsistency as a systematic source of variation among individuals by finding some evidence for generalizability and relationship to behavioral predictability. However, these findings are questionable, since these studies confounded change with short-term fluctuation in their response inconsistency measure. The current research separates these two sources of inconsistency in a reanalysis of the data from one major study on short-term consistency and finds little evidence for generalizability or a relationship to behavioral predictability. These results support the popular assumption that measurement error from short-term fluctuations is not due to systematic individual differences in response consistency, as well as supporting a more limited definition of the individual inconsistency construct.

Determining the nature of individual inconsistency in performance on trait measurements is an important topic in psychometrics because of its direct relevance to measurement reliability. Regardless of the theoretical conceptualization of reliability—classical test theory (Gulliksen, 1950), generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnum, 1972), or latent trait theory (Lord & Novick, 1968)—measurement error is estimated and interpreted without respect to possible individual differences in response consistency over the varying conditions of measurement. These approaches either explicitly or implicitly assume that measurement errors due to the individual, rather than the test items or testing conditions, are unsystematic and randomly distributed within the population. Research on individual inconsistency is directly related to the assumption of random individual error, as it has concerned the extent to which response inconsistency is itself a systematic source of variability among individuals.

Recently, Lumsden (1977) has distinguished conceptually between three types of individual inconsistencies—trends, swells, and tremors. These are defined as follows: Trends are changes in the trait; swells are short-term fluctuations over hours or days; tremors are momentary fluctuations that may be observed within a single testing session. Lumsden (1977) developed a normal ogive model for tremors and cited some supporting research. However, Lumsden deplored the lack of research on individual differences in swells as a systematic source of unreliability on cognitive tests. Contrary to Lumsden's summary on swell effects, a small set of studies has sought to examine short-term inconsistency for (1) generalizability over traits; (2) relationship to momentary incon-

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 2, No. 4 Fall 1978 pp. 571-579*

sistency; and (3) relationship to behavioral predictability. These will be described in turn.

Berdie (1969a) found modest support for the generalizability of individual inconsistency measures obtained from six cognitive traits. In this study, individual inconsistency was measured as score variability over 20 (daily) repetitions of the 6 trait measurements. Twenty alternate forms were available for each test, and test forms were carefully balanced over occasions so that the results could not be attributed to differences between forms. Berdie found moderate correlations between the inconsistency measures obtained from the six traits.

Hendel and Weiss (1970) and Weksel and Ware (1967) studied the relationship between short-term inconsistency and momentary inconsistency. In both studies, short-term inconsistency was measured by differences between scores obtained on two occasions, while momentary inconsistency was measured by preference ordering on a single occasion. A moderate correlation between the two kinds of inconsistency was found by both studies.

Berdie (1969a) found some support for short-term individual differences in response inconsistency to be related to the predictability of educational achievement by a standard ability test. Berdie hypothesized that if response inconsistency was a general personality attribute, it should be negatively correlated with the predictability of behavior. That is, persons with irregular habits and high inconsistency should be the least predictable. Berdie's (1969a) study used the same individual inconsistency measures as the Berdie (1969b) study, but it also had available both obtained college grade point averages and those predicted by a pre-entrance ability test. Some significant correlations were found, but the relationship complexly depended on both the specific index of individual inconsistency and behavioral predictability.

As a whole, these studies provide only modest support for short-term inconsistency as a systematic source of variation between individuals. However, a careful examination of the methods used to measure short-term inconsistency reveals a confounding which may have attenuated the results. In all these studies, within-individual variability over time was used to estimate short-term inconsistency. However, unless individuals show equivalent changes over time (an unlikely possibility with respect to prior literature on change, such as Stake, 1961), change is completely confounded with short-term inconsistency in the within-individual variability measures used in these studies. This confounding may not only have generally lowered the correlations found for generalizability and the external correlates of response inconsistency, but it is impossible to decide which type of inconsistency—change or short-term fluctuation—is responsible for the correlations that were found.

Fortunately, however, change and short-term sources may be unconfounded, given a sufficient number of repeated observations. Modern perspectives on change and growth have shown that several parameters may be estimated for an *individual* to separate the various sources of individual inconsistencies in test performance over time (a sophisticated procedure is given in Bock, 1976, for example). Of the several studies on short-term inconsistency, Berdie's (1969a, 1969b) studies not only represent the most extensive efforts to establish individual differences, but these data contain enough repetitions over time to separate change from short-term fluctuations. The current study is a reanalysis of the Berdie data to examine separately change and short-term fluctuation as systematic sources of individual differences in test performance. Both sources of inconsistency will be examined for generalizability and for relationship to behavioral predictability.

## Method

The subjects, materials, and procedures will be summarized briefly, as a more complete description is available in Berdie (1969a, 1969b). The statistical analyses, however, will be dis-

cussed in detail, as they differ substantially from the original study.

### Subjects

The subjects were 79 college freshmen enrolled in the Institute of Technology of the University of Minnesota.[1] These subjects were paid to participate in the experiment and were screened from a larger pool of volunteers on the basis of availability for the 20 scheduled testing sessions. For all students, the Institute had available a predicted freshman grade point average (GPA), which was a regression estimate based on outcomes from preceding years using scores on the Minnesota Mathematics Test as the predictor. Obtained GPAs were available from student transcripts.

### Materials

The measures for individual consistency were taken from six traits which are frequently described in the factor analytic literature—Aiming, Flexibility of Closure, Number Facility, Perceptual Speed, Speed of Closure, and Visualization. These highly speeded tests are scored as number correct within a 2.5- to 3-minute time limit. The specific tests of these traits were the 20 equivalent forms developed by Moran and Melford (1959).

### Procedure

The students were tested for 20 consecutive days. On each of the 20 occasions they took a different form of each test, but test forms were counterbalanced over students so that differences among occasions would not reflect minor variations among the test forms. Additionally, students were tested in five separate subgroups so that differences in local testing conditions would not be confounded with occasions.

### Data Analysis

The data consisted of 120 scores—6 trait scores on each of 20 occasions—for each of the 79 students. Consistent with Berdie (1969a, 1969b), the raw scores on each form were standardized within the sample to control for any minor differences of means or standard deviations among the test forms.

A major difference between the original Berdie (1969b) study and this reanalysis is the way in which the estimates of individual inconsistency and level of performance were obtained. Berdie first computed a grand mean of the trait scores, over occasions, to represent the level of performance on the trait, and then obtained a variance from this mean as an estimate of individual inconsistency.

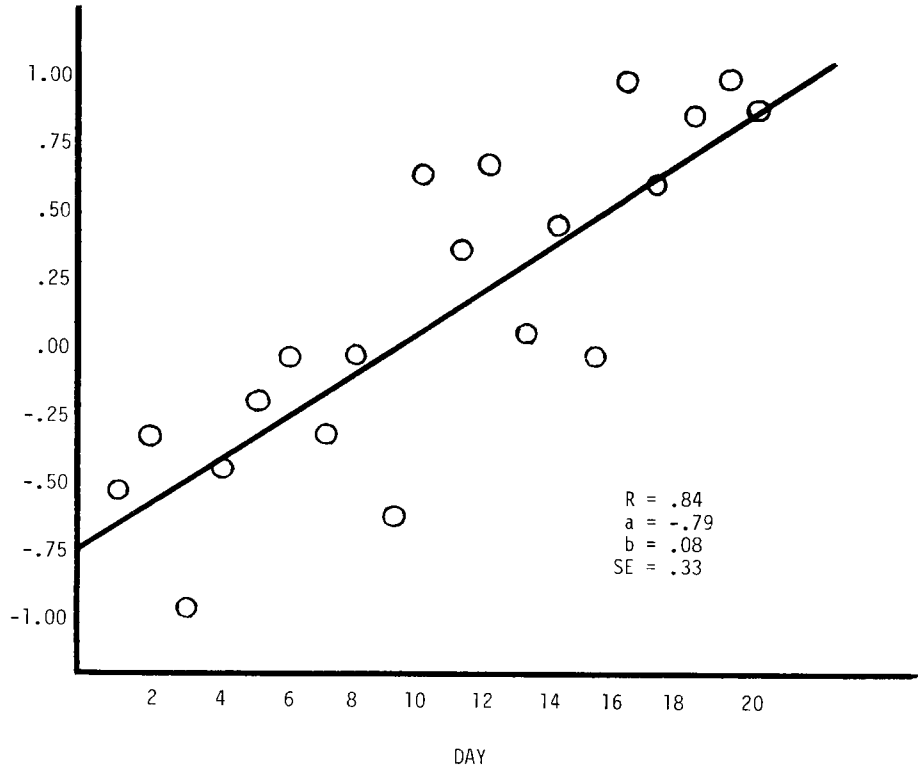
The way in which these computations confound change with both individual inconsistency and level of performance may be shown in Figure 1. Here, scores on the Aiming test are plotted for one student over days. It can be seen that scores are generally increasing over time for this student, although the relationship is by no means perfect. A regression line could be fitted to these data; and an intercept, slope, and standard error of −.79, .08, and .33, respectively, would be observed. Since this regression was significant ($p < .01$), the grand mean of test scores was influenced by both initial score level and rate of change. Furthermore, response variability from this mean (Berdie's index of individual inconsistency) can be seen to be the same as the variance of the dependent variable in the scattergram.

With the significant relationship between test scores and occasions that is depicted here, the test score variance represents the combined effect of change over time and inconsistency. That is,

$$\sigma_y^2 = \sigma_{y'}^2 + \sigma_{y-y'}^2 \qquad [1]$$

[1] The original Berdie study had 100 subjects. A comparison between means on the major variables showed that the loss of subjects was random. To compensate for the slightly lessened power to find significant differences, significance level was set at .10, rather than .05 as in Berdie's studies.

**Figure 1**
**The regression of test score on day of measurement**




where $y$ = test score, and
$y'$ = test score predicted by occasion.

Hence, since Berdie's measure of individual inconsistency is this test score variance, his measure represents the confounded influence of change and short-term inconsistency.

To eliminate the confounding of level and individual inconsistency with change, the following three parameters were estimated for each subject by regressing trait scores on occasions: (1) the intercept, for an estimate of initial score level; (2) the slope, for an estimate of change as a result of practice; and (3) the standard error of estimate, as an estimate of short-term fluctuation over trials. In this design, change is operationally defined as a systematic linear relationship between test scores and time, while short-term inconsistency is the nonlinear residual. To the extent that test scores are monotonically, but nonlinearly, related to time, change is underestimated and short-term inconsistency is overestimated. However, since the number of daily observations was too small for an adequate polynomial regression and estimation procedures for specified curves are not widely available, only linear trends were examined.

The 3 regression scores for each trait—a total of 18 scores—constituted the independent variables for the reanalysis of the Berdie (1969b) data. These scores were used to test hypotheses about both the generalizability of individual inconsistency and its relationship to predictability. In the Berdie (1969b) study, educational predictability was scored as the difference between obtained and predicted GPA. Both the absolute

and the signed difference were scored and analyzed separately, as in the original study.

## Results

The extent to which short-term inconsistency and change were confounded in Berdie's original variability indices was examined by calculating the percentage of subjects showing significant linear change over occasions. The following percentages showed significant correlations between test score level and day of measurement, by each trait: Aiming, 55.7%; Flexibility of Closure, 72.2%; Number Facility, 89.9%; Perceptual Speed, 63.3%; Speed of Closure, 88.6%; and Visualization, 72.2%. These data show that short-term fluctuation and change were substantially confounded in the variability indices for all 6 traits.

The generalizability of individual inconsistency measures over the six traits was tested separately for change and short-term inconsistency by maximum likelihood factor analysis. Prior to performing the factor analyses, the correlations of these measures between traits were tested for significance. Both the change and short-term fluctuation intercorrelation matrices were significant ($\chi^2_{15}$=42.77, $p < .001$ and $\chi^2_{15}$=38.54, $p < .001$, respectively). Thus, the data were suitable for factor analysis.

For the six change measures, the log likelihood chi-square test indicated that a single factor adequately reproduced the patterns of correlations ($\chi^2_9$=6.80, $p > .10$), but the average communality was only .21. Thus, substantial trait-specific variance remained. For the six short-term fluctuation indices, a single factor also adequately reproduced the pattern of correlations ($\chi^2_9$=16.53, $p > .10$), and the average communality was .17. Again, substantial trait-specific variance remained.

The relationship of change and short-term fluctuation to behavioral predictability was determined by multiple regression analysis. Two measures of predictability—the absolute difference and the signed difference between predicted and observed college GPA—were used as dependent variables. The independent variables were the three parameters estimated from the within-individual regression—initial score level, change, and short-term consistency. The variables were interpreted to have a significant relationship to behavioral predictability if they obtained significant beta weights in a regression with a significant multiple correlation. As indicated by Darlington (1968), this significance criterion controls for correlations among the predictors by requiring that an independent variable has a contribution beyond that which it shares with the other independent variables, as well as for overall significance.

To use multiple regression to test for the relationship between behavioral predictability and the generalizable component of the independent variables, factor scores were obtained separately for initial level, change, and short-term fluctuations from the maximum likelihood factor analyses described above. The behavioral predictability measures were then separately regressed on the three factor scores. Since the overall multiple correlation for neither the signed GPA discrepancy nor the absolute GPA discrepancy reached significance ($F$'s =.86, 1.20 respectively), the beta weights were not interpreted. Thus, the generalizable components of initial level, change, and short-term fluctuation were not significantly related to behavioral predictability.

The factor score analysis may underestimate the relationship between inconsistency and the behavioral predictability measures, however, since substantial trait-specific variation was found among both types of inconsistency measures. Therefore, the two GPA discrepancy indices were regressed on the initial level, short-term fluctuation, and change measures obtained for each trait. Table 1 presents the multiple correlations, overall $F$ values, and beta weights for the six regressions from each GPA discrepancy index.

With the signed differences as the dependent variable, significant multiple correlations with

Table 1
Multiple Correlations and Beta Weights for Prediction of GPA Discrepancies from Sample Standardization

| Variable | Multiple R | Beta Weights: Initial Level | Beta Weights: Long-Term Change | Beta Weights: Short Term Inconsistency |
|---|---|---|---|---|
| GPA discrepancy, signed | | | | |
| Aiming | .36* | .39** | .15 | -.09 |
| Flexibility of Closure | .13 | -.08 | .12 | -.03 |
| Number Facility | .28 | .23* | .19 | -.13 |
| Perceptual Speed | .31+ | .24 | .29* | -.11 |
| Speed of Closure | .05 | .04 | .01 | -.03 |
| Visualization | .33* | .24* | .12 | .27* |
| GPA Discrepancy, absolute | | | | |
| Aiming | .25 | .06 | -.22 | .07 |
| Flexibility of Closure | .22 | .19 | -.04 | -.06 |
| Number Facility | .11 | -.09 | -.09 | -.06 |
| Perceptual Speed | .22 | -.13 | -.25 | .00 |
| Speed of Closure | .21 | -.11 | -.16 | .00 |
| Visualization | .20 | -.09 | -.10 | .14 |

+p < .10, *p < .05, **p < .01

the independent variables were found for three traits. Aiming, Perceptual Speed, and Visualization significantly predicted the signed GPA difference; however, the magnitude of these multiple correlations was low ($R \leqslant .36$). For Aiming only initial level obtained a significant beta weight, while for Perceptual Speed only change was significant. Two predictors from Visualization were significant—initial level and short-term fluctuation. In contrast, with the absolute difference between predicted and obtained GPA discrepancy as a dependent measure, none of the multiple correlations reached significance. An inspection of the zero-order correlations of the predictors with two GPA measures revealed only a few significant, but small, correlations ($-.23 < r < .33$).

In addition to the regression analysis, the individual inconsistency variables were also examined as moderators for the correlation between pre-entrance ability and GPA. Mirroring Berdie's analysis on the general variability indices, high- and low-inconsistency groups were formed by averaging the inconsistency measures on the six traits and then splitting on the mean of the average scores. This procedure was followed separately for the change and short-term inconsistency indices. The far right column of Table 2 presents the resulting within-group correlations. These correlations did not differ significantly for either index ($p > .10$).

Also following Berdie, an additional grouping factor was added to control for score level differences. In this case, the initial level scores were averaged over traits, and an additional mean split on level resulted in four groups for the change and short-term indices. Rao's (1970) $\chi^2$ test for the equality of correlations obtained from separate samples yielded no significant differences for either the long-term groups ($\chi^2$=4.47, $p > .10$) or the short-term groups ($\chi^2$=5.99, $p > .10$). Furthermore, none of the individual correlational comparisons between the high- and low-inconsistency groups would have reached significance, even if tested following the non-significant overall test.

Table 2
Correlations of Ability with GPA and Sample Sizes for Groups Varying in Initial Trait Level and Inconsistency

| | Initial Level | | | | | |
|---|---|---|---|---|---|---|
| | low | | high | | Summed over Level | |
| Inconsistency | r | n | r | n | r | n |
| Change | | | | | | |
| low | .31 | 22 | .37 | 13 | .53 | 35 |
| high | .35 | 18 | .62 | 26 | .23 | 44 |
| Short-term | | | | | | |
| low | .54 | 17 | .52 | 24 | .31 | 41 |
| high | .12 | 23 | .62 | 15 | .49 | 38 |

## Discussion

Contrary to expectation, rather than providing stronger support for short-term fluctuation as a systematic source of variation among individuals, the current study substantially weakens the construct and, additionally, finds little support for systematic individual differences in change. These findings will be discussed separately for generalizability and behavioral predictability and will be compared to Berdie's (1969a, 1969b) results. Furthermore, the implications of the results for test theory and the general construct of individual inconsistency will be explored.

The generalizability results are consistent with Berdie's findings: Positive correlations were found among the set of six traits for both the change and short-term inconsistency measures. Additionally, the correlations were significantly accounted for by a single common factor in both cases. However, the proportion of total variance accounted for by the factor was extremely small in both cases—.17 for short-term inconsistency and .21 for change.

Thus, although these data show some support for generalizability of individual inconsistency over separate traits, the importance of the generalizable component for contributing to instability on the trait measures is questionable. The trait-specific contribution to unstable performance is, on the average, four times as large as the generalizable component. These results are more supportive of the contention that unstable performance on psychometric instruments is due to random, situation-specific factors than they are supportive of the contention that unstable performance arises from systematic individual differences on a general inconsistency construct. With respect to contemporary reliability theories, some empirical support is given here to the common practice of calculating and interpreting a single standard error of temporal instability for individuals in similar testing conditions (or at similar score levels) without reference to individual differences in inconsistency.

For the relationship between individual inconsistency and behavioral predictability, separating short-term inconsistency from change did not lead to stronger results in any of the three analyses. Similar to Berdie's analysis with the confounded inconsistency measures, neither change nor short-term fluctuation was related to the absolute difference between obtained and predicted college grade-point averages. This same pattern was found for the generalizable component among the trait inconsistency scores, as well as for the inconsistency measures for each trait. In contrast, the analysis of the signed discrepancy between obtained and predicted college grade-point averages was *less* related to the inconsistency measures than in

Berdie's analysis. Berdie had found that the signed discrepancy was significantly related to a general inconsistency measure, averaged over traits. Unfortunately, as indicated by Berdie, this index was not easily interpretable.

In the current study, the factor scores for change and short-term inconsistency measures provided a meaningful general index; however, neither was related to the signed discrepancy criterion of behavioral predictability. Furthermore, if the relationship of signed discrepancy index to inconsistency is examined at the individual trait level, only short-term inconsistency of Visualization was significantly related to signed GPA discrepancy, while change was significantly related only to Perceptual Speed. Lastly, an analysis of change and short-term inconsistency as possible moderators of the correlation between preentrance ability and GPA, controlling for initial level, yielded no significant differences in the overall $\chi^2$ test or in the individual $t$-tests. In contrast, Berdie had found the high-inconsistency individuals to be less predictable than the low-inconsistency individuals, but this difference was significant only for those with high means on the traits. However, no overall significance test was given for differences in correlations between the various groups in Berdie (1969b) to control for experiment-wise error.

The best conclusion from the results is that change and short-term inconsistency on the six psychometric traits measured in these studies is not highly related to behavioral predictability in college achievement. The size and magnitude of the few significant correlations obtained here, and in Berdie (1969b), could be attributed to experiment-wise Type I error, due to the multiple comparisons involving the inconsistency measures.

The results have implications for the general nature of the individual inconsistency construct. It has been shown that neither short-term inconsistency nor change is highly generalizable over cognitive traits. Furthermore, inconsistency on these traits is not related to behavioral predictability in an important applied setting—achievement in higher education. Thus, either the inconsistency construct itself does not describe a meaningful dimension of individual differences, or the construct does not apply to performance inconsistency on cognitive tests. General personality theory that is relevant to individual consistency would probably favor the "limited construct" approach given by the latter possibility.

As Maddi (1976) points out, several personality theories are based on a consistency model of human behavior. Fiske and Maddi (1961) present a consistency model of personality, for example, in which impact-modifying behavior can vary between individuals. Impact modification refers to the individual changing the meaning, intensity, or variety of stimuli he or she encounters to maintain some "optimal" arousal level. This model, then, focuses on the choice of stimuli, rather than consistency of response to the same repeated stimuli. The kind of behavior implied in the measurement of this type of inconsistency would be inconsistency of choice and judgment tasks, rather than inconsistency of cognitive performance.

In summary, the current study supports the notion that measurement error from short-term fluctuations is not due to systematic individual differences in response consistency, as well as supporting a more limited definition of the individual inconsistency construct.

## References

Berdie, R. F. Consistency and generalizability of intra-individual variability. *Journal of Applied Psychology*, 1969, *53*, 35–41. (a)

Berdie, R. F. Intra-individual temporal variability and predictability. *Educational and Psychological Measurement*, 1969, *29*, 235–257. (b)

Bock, R. D. Basic issues in the measurement of change. In D. N. Degruijter & L. J. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: John Wiley & Sons, 1976.

Cronbach, L. J., Gleser, B. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements*. New York: John Wiley & Sons, 1972.

Darlington, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin*, 1968, *69*, 161–182.

Fiske, D. W., & Maddi, S. R. *Functions of varied experience.* Homewood, IL: The Dorsey Press, 1961.

Gulliksen, H. *Theory of mental tests.* New York: John Wiley & Sons, 1950.

Hendel, D. D., & Weiss, D. J. Individual inconsistency and reliability of measurement. *Educational and Psychological Measurement,* 1970, *30,* 579–593.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Lumsden, J. Person reliability. *Applied Psychological Measurement,* 1977, *1,* 477–482.

Maddi, S. R. *Personality theories: A comparative analysis.* Homewood, IL: The Dorsey Press, 1976.

Moran, L. J., & Melford, R. B. Repetitive psychometric measures. *Psychological Reports,* 1959, *5,* 269–275.

Rao, C. R. *Advanced statistical methods in biometric research.* Darien, CT: Hafner Publishing Co., 1970, pp. 234–235.

Stake, R. Learning parameters, aptitudes and achievements. *Psychometric Monographs,* 1961, *9,* 1–70.

Weksel, W., & Ware, E. E. The reliability and consistency of complex personality judgments. *Multivariate Behavioral Research,* 1967, *2,* 537–541.

## Acknowledgment

*The author is indebted to the late Ralph F. Berdie for generously providing his data for reanalysis from another point of view. Only in recent times have the conceptual and methodological perspectives become available that do justice to his well-designed and executed experiment.*

## Author's Address

Susan E. Whitely, 449 Fraser, Department of Psychology, University of Kansas, Lawrence, KS 66045.