# Person Reliability

**James Lumsden**
**The University of Western Australia**

Person changes can be of three kinds: developmental trends, swells, and tremors. Person unreliability in the tremor sense (momentary fluctuations) can be estimated from person characteristic curves. Average person reliability for groups can be compared from item characteristic curves.

Test theorists have long realized that person instabilities affect estimates of reliability. They have attempted to eliminate the effects of person changes by reducing the time interval between testings. These attempts have failed, however, because person changes during testing are uncontrolled even when internal consistency procedures are used. There have been attempts made to assess average person reliability for groups by estimating function fluctuation from gross test score relationships (Thouless, 1936; Garside, 1958; Cureton, 1971). These attempts have not led to fruitful applications; and, in any case, they provide no information about the reliability (stability) of individual subjects. It is the purpose of this paper to present an attribute-based model of test performance and to derive from it a different, and potentially fruitful, approach to person reliability.

Three kinds of change can be distinguished that can take place in a person's ability over time: developmental trend, swells, and tremors. Individual curves of mental growth and decline are attempts to represent developmental trends. Through maturation and experience a person's average ability level will exhibit a long-term trend over time. The rate of change will usually be gradual though sometimes (e.g., after brain damage) it can be steep. Swells are relatively short-term (days or hours) fluctuations of the average level. They are temporary elevations or depressions of ability arising from short-term conditions or states. Thus, a person may be in relatively good or poor form for short periods because of his/her emotional state. Quotidian variations would be classed as swells. Tremors are rapid momentary fluctuations in the level of ability. Tremors are considered either to be a random variable added to trend and swell level or to be a cyclic variable which oscillates so rapidly relative to irregularly occurring items that at the moment of encountering any item the tremor level is effectively random. These distinctions may be further illustrated by analogy with the depth of water in a mountain lake. Developmental trend refers to the effects on depth of long-term advances or retreats of the ice cap and progressive siltation; swells refer to effects from wet or dry spells and to tidal effects; tremors refer to ripples on the lake. It should be noted that the tremor effects are similar to the oscillation effects discussed by Spearman (1927) and Hull (1952).

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 1, No. 4 Fall 1977 pp. 477–482*

Although the different kinds of change can be differentiated conceptually, it is not easy to distinguish them operationally. It is assumed throughout this paper that during a single test administration trend and swell effects will be relatively small compared to the tremor effects. However, this may not always be the case. There appear to have been no empirical studies of intra-test practice effects. An interesting experiment would be to sample items from two or three equivalent forms and then to administer them as a single test. Comparisons of form results would enable estimates to be made of differential intra-test practice effects.

Where do the kinds of person change stand in relation to Thorndike's (1949) classification of sources of variation in test scores? It would seem that Thorndike would classify differential trend and swell effects between testings as temporary characteristics of the individual, and would consider them as error variance (i.e., as contributing to test score unreliability) in test-retest and parallel-form-with-interval estimates of reliability. Trend effects should not, however, be considered as an aspect of person or test score unreliability any more than different rates of growth in height should be considered as a reflection of unreliability. The extent and frequency of swell effects are probably important for human performance; these are properly considered as contributing to person unreliability. Little recent attention has been paid to swell effects in the ability domain, but considerable work has been done in the personality domain. Swell effects will not, however, be further considered in this paper; rather, it will concentrate on tremor effects. Thorndike assumes that no differential person changes take place during a single testing; therefore, he excludes person tremor from consideration.

The basic model for this paper is illustrated in Figure 1, which shows the attribute continuum and the locations of a set of items and persons. Each item is shown as having a point location on the attribute continuum. Each person has a distribution of attribute locations resulting from moment to moment changes, i.e., tremor. It is assumed that the distributions are normal. Note that the dispersions of attribute locations for the persons are not all the same. The model is essentially Thurstonian and is identical with that described by Torgerson (1958) for the law of categorical judgment. However, the discriminal dispersions of the items are all zero, and the score categories have been replaced by person locations. A person passes an item if, when attempting it, his/her momentary location is higher than the point location of the item; otherwise, he/she fails. It is, therefore, possible for a person to pass an item at a given location and fail another item at the same location, or an even lower location on another occasion.

It is submitted that the model is plausible. The notion of intrinsic fluctuation of item attribute values makes little sense in psychological measurement. It is expected that the procedures of test construction will exclude ambiguous items, especially since the model is conceived as applying completely only to unidimensional tests constructed according to strict specifications (Lumsden, 1961, 1976).

If there are many items distributed evenly across the range of ability to be considered, then a person characteristic curve can be constructed for each subject. The person characteristic curve is the plot for a single subject of the proportion of items passed at different difficulty levels. It is perfectly analogous to the item characteristic curve. It should be noted that the idea of a plot of a person's trace line was first proposed by Weiss (1973) and later called a "subject characteristic curve" (Vale and Weiss, 1975).

Person characteristic curves enable comparisons to be made of person reliabilities. Figure 2 shows hypothetical person characteristic curves for three subjects. Persons *A* and *B* have the same average ability and pass the same number of items; however, the amplitude of *A*'s tremor is less than *B*'s so that he/she neither passes any difficult items nor fails any easy ones. Person reliabilities can be compared by inspection of the curves. If the person characteristic curves turn out to all have the same form (the model implies

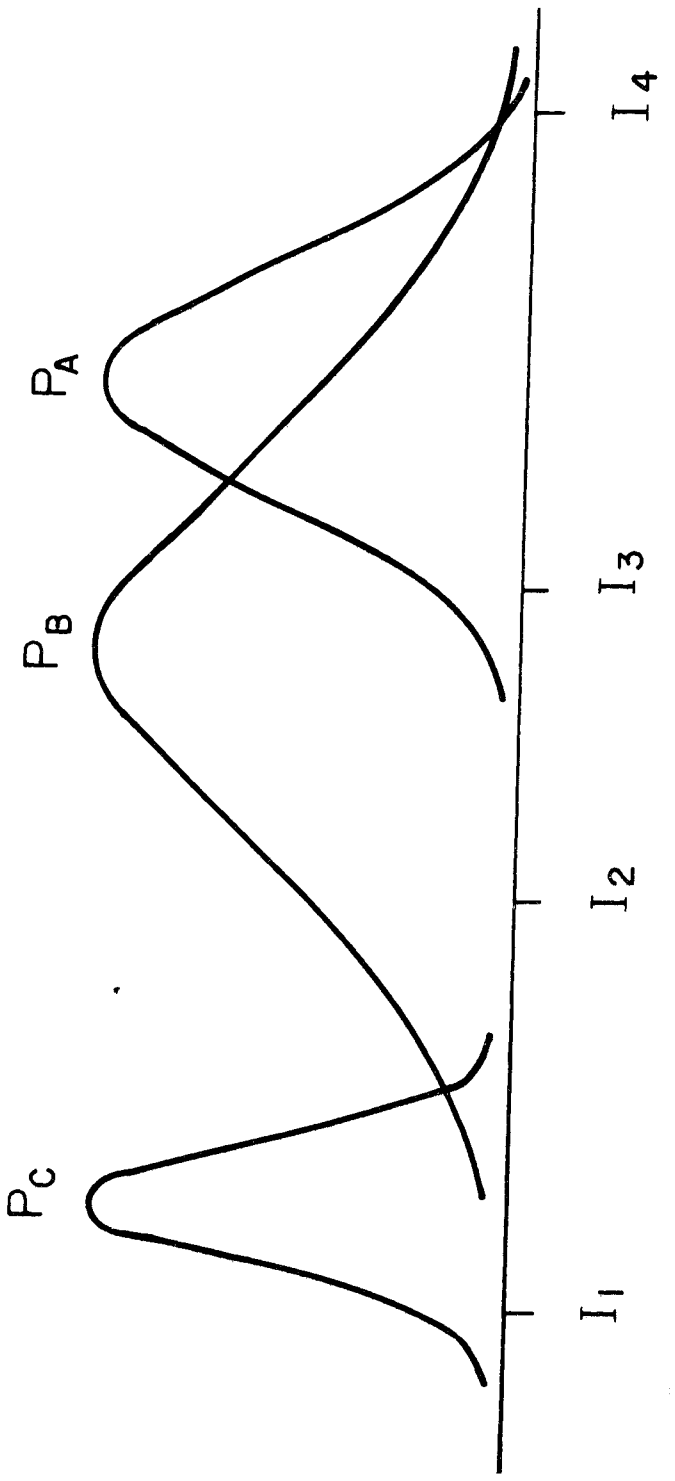


Figure 1: Attribute Continuum Showing Locations of Items (I) and Persons (P)

that they will be normal ogives), then the slope will be an adequate summary statistic for comparing person reliabilities. The argument has assumed that items have no intrinsic difficulty fluctuation. In any case differences between the curves are clearly independent of item fluctuation in the large sample case assumed here.

Suppose that the curves of Figure 2 arise from some familiar ability test, for example, number ability. Persons *A* and *B* get the same total score, so that they will be estimated by the usual unit weight methods to have the same location parameter. They differ, however, in the pattern of their results; and it is this pattern which may be important. If testees are being selected for a job in which the number work is easy and the mistakes are costly, then *A* should be selected. If testees are being selected for a job in which the number work is difficult and triumphs highly rewarded, then *B* should be selected. Spearman (1927), in one of his rare flashes of humour, pointed out that extreme oscillation is desirable in an airplane designer but disastrous in an airplane pilot.

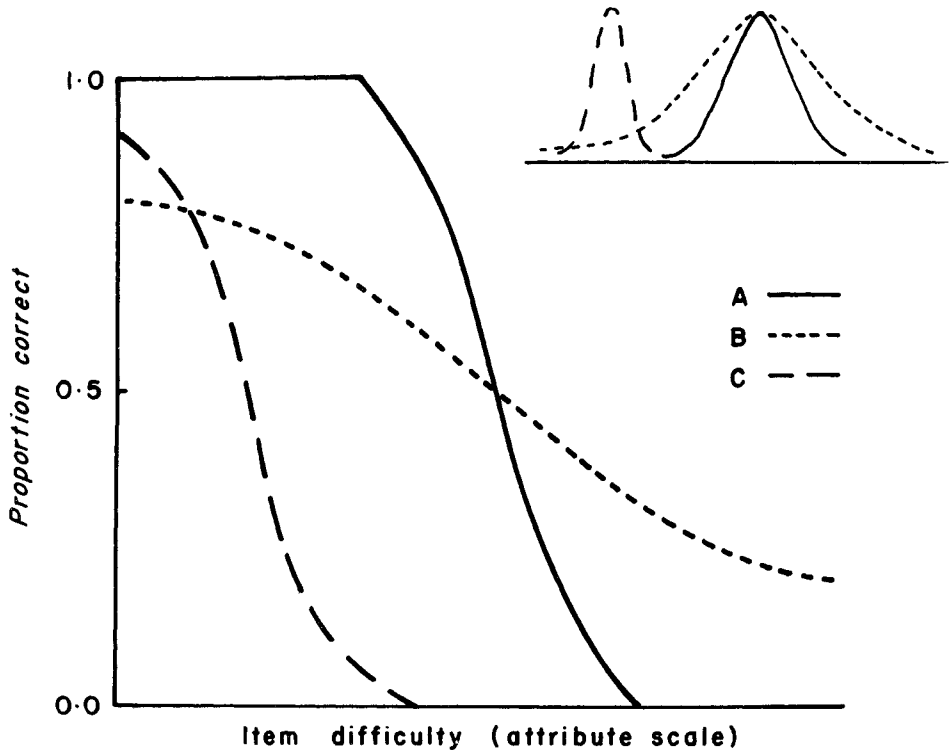


Figure 2: Hypothetical Person Characteristic Curves for Three Subjects. (Attribute locations for the three subjects at right top)

It is clear that if person characteristic curves intersect, then estimates of relative ability will be biased by the difficulty of the items. On an easy test *A* will be seen as more able than *B*, while on a difficult test *B* will be seen as more able than *A*. Note, too, that on very easy items, *C*, whose overall performance is much inferior, will be seen as more able than *B*. It is also obvious that no weighting procedure intended for use over all subjects can correct the bias. It follows, generally, that statements may be misleading to the effect that particular estimation procedures yield sufficient statistics for estimating average ability levels. Even when the sufficiency claim is true, there is another aspect of performance, the person reliability, which cannot be estimated by a single test score.

Person characteristic curve differences will, in some cases, reveal obvious but important interactions between aptitudes and instructional methods. For example, a sensible intructional program for person *B* would be to give rewards for accuracy on easy tasks and some teaching on checking procedures. For *A*, help with easy items would be futile; what is required in this case is instruction or, perhaps, encouragement in tackling the more difficult items.

Whether there exist sizable numbers of people who are as different in reliability as *A* and *B* is an empirical matter, although teachers often report encountering them. Weiss (1973), using a stratified vocabulary test, found differences in the slopes of the person characteristic curves. Using the standard deviation of the difficulties of items attempted as an estimate of the slope, he found that the differences were moderately consistent in the test-retest sense. The result is encouraging, since Weiss' tailored test was quite short. Anderson (1958) attempted to estimate individual function fluctuation from total test scores by retesting with the same test and parallel forms. He found no evidence for a general factor either within or between different types of tests; and he concluded that the study of individual differences in function fluctuation was impossible. The Anderson procedure, however, confounds tremor, swell, and trend effects. His results cannot be regarded as cogent evidence vis-à-vis the Weiss result.

The reliabilities of groups could be compared by using some average of the slopes of the person characteristic curves, since there is a simple way of, in effect, experimentally averaging. Consider the item characteristic curves for a single item given to two groups, e.g., men and women. The slopes of the curves will be determined by the average person fluctuation and (if there is any) by item fluctuation. It is clear again that differences between the curves for the two groups will reflect differences in group reliability independent of item fluctuation. The suggested method is superior to the usual procedure of comparing standard deviations of total test score, since it is independent of the distributions of average attribute level. The method can also handle the case in which the differences in the reliabilities are not constant across attribute levels. It would be interesting if, for example, men were found to be more variable than women only at the extremes of the range.

The interpretation of differences between person characteristic curves for individuals, and between item characteristic curves for groups, as reflecting differences in reliability, depends critically on the unidimensionality assumption. Similar curves to those of Figure 2 could be obtained if the more difficult items measured some attribute different from that of the easier items. The differences between persons *A* and *B* would simply represent their relative standing on different attributes. Nevertheless, the person characteristic curve remains a useful device for drawing attention to these differences, which are likely to be obscured by standard methods of treating test performance.

## References

Anderson, C. C. Function fluctuation. *British Journal of Psychology, Monograph Supplements,* 1958, *30.*

Cureton, E. E. The stability coefficient. *Educational and Psychological Measurement,* 1971, *31,* 45–53.

Garside, R. F. The measurement of function fluctuation. *Psychometrika,* 1958, *23,* 75–83.

Hull, C. L. *A Behavior System.* New Haven: Yale University Press, 1952.

Lumsden, J. The construction of unidimensional tests. *Psychological Bulletin,* 1961, *58,* 122–131.

Lumsden, J. Test theory. *Annual Review of Psychology,* 1976, 251–280.

Spearman, C. *The Abilities of Man.* New York: MacMillan, 1927.

Thorndike, R. L. *Personnel Selection: Test and Measurement Techniques.* New York: Teachers College Press, 1949.

Thouless, R. H. Test unreliability and function fluctuation. *British Journal of Psychology,* 1936, *26,* 325–343.

Torgerson, W. S. *Theory and Methods of Scaling.* New York: John Wiley, 1958.

Vale, C. D. & Weiss, D. J. *A study of computer-administered stradaptive testing.* (Research Report 75–4) Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD-A018758).

Weiss, D. J. *The stratified adaptive computerized ability test.* (Research Report 73–3) Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD-768376).

## Author's Address

James Lumsden, Department of Psychology, The University of Western Australia, Nedlands, Western Australia 6008