Kinesthetic Aftereffect Scores Are Reliable

Brian L. Mishara
University of Massachusetts at Boston

A. Harvey Baker
Queens College,
City University of New York

The validity of the Kinesthetic Aftereffect (KAE) as a measure of personality has been criticized because of KAE's poor test-retest reliability. However. systematic bias effects render KAE retest sessions invalid and make test-retest reliability an inappropriate measure of KAE's true reliability. Internal consistency calculations, a better estimate of KAE's true reliability, have been flawed in the past. New analyses of internal consistency data from 10 samples using 2 different KAE procedures are presented. Internal consistency of KAE scores was found to be high (median .89 for 5 samples with Petrie's procedure and median .59 for 5 samples with Weintraub's procedure). Some increment in reliability can apparently be obtained by extending the number of trials in the Weintraub procedure.

Are Kinesthetic Aftereffect (KAE) scores reliable? According to KAE literature, which has generally answered this question in terms of test-retest reliability, KAE is not reliable. In this paper it is argued that test-retest reliability is not an appropriate statistic for KAE. When internal consistency, a more appropriate estimate of KAE's true reliability, is considered, KAE does show substantial reliability. A review of KAE's status and history shows that a proper conception of reliability is presently critical. The positive findings reported here strongly buttress the construct validity evidence for KAE.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 2 Spring 1978 pp. 239-247
© Copyright 1978 West Publishing Co.

KAE as a Measure of Personality

The major thrust of research using KAE to assess individual differences in personality has come from the hypothesis offered by Petrie (1967; Petrie, Collins, & Solomon, 1958). KAE scores are postulated to reflect the status of a stimulus intensity modulator mechanism which functions like a volume control, damping down subjective stimulus intensity in some individuals (reduction) and increasing stimulus intensity in others (augmentation). For individuals classified as reducers, subjective stimulus intensity is presumably attenuated. Thus, reducers are generally stimulus deprived and seek more stimulation to compensate. Although reducers can handle high-intensity stimulation quite well, they are uncomfortable when environmental stimulation is low. Augmenters show the opposite pattern: They are generally overloaded with stimulation, show stimulation-avoidant behaviors, but cope very well with low levels of stimulation. This formulation is supported by findings regarding pain tolerance (Petrie et al., 1958; Poser, 1960; Ryan & Foster, 1967; Sweeney, 1966); reactivity to sensory deprivation (Petrie et al., 1958; Sales, 1971); and stimulation-seeking tendencies (Ryan & Foster, 1967; Sales, 1971, 1972).

There are two variants of KAE procedure. In the Petrie (1967) version of KAE, subjects rub an aftereffect induction block with one hand while resting the other. In another KAE variant, subjects simultaneously rub two aftereffect induction blocks—a larger one with one hand and a smaller one with the other (e.g., Spilker & Callaway, 1969). Treatment of KAE reliability in this paper is limited to the one-hand variant and does not consider issues arising from the two-hand variant.

Recent Criticisms of KAE

The literature was largely supportive of the KAE modulator formulation from about 1958 to 1970. However, a sharp turn has taken place since 1970. Based largely on the findings of poor retest reliability and buttressed by some evidence of KAE's apparently intermittent validity, a number of critics suggested that KAE should no longer be used as an index of a stable trait (e.g., McDonald, 1974; Morgan & Hilgard, 1972; Platt, Holzman, & Larson, 1971; Sales & Throop, 1972; Weintraub, Green, & Herzog, 1973).

A New View of KAE

A view has recently emerged reinstating KAE as a valid measure of indivudual differences. Central to this new conceptualization are systematic individual differences in practice effects first noted by Bakan and Thompson (1962) and recently confirmed (Baker, Mishara, Parker, & Kostin, in press; Mishara & Baker, 1978a). These bias effects render second and later session KAE scores invalid as a measure of the stimulus-intensity modulator hypothesis (Baker, Mishara, Kostin, & Parker, 1976), but do not affect the validity of scores from the first KAE session. When KAE personality findings are considered for one-session studies, the validity literature is clearly supportive (see Table 1 in Baker, Mishara, Kostin, & Parker, 1976); and new one-session validity findings continue to emerge (Gupta, 1974; Mishara & Baker, 1978b, in press).

Because of bias, second-session KAE scores measure something different from first-session scores. In this case, when *observed* retest reliability is low, test-retest reliability tells nothing about the *true* reliability of the first-session score and thus is no longer a criterion of construct validity (McNemar, 1969).

When test-retest reliability is an inappropriate statistic, as in this case, other estimates of true reliability, such as internal consistency, should be considered. Unfortunately, prior reports of high KAE internal consistency for first session administrations are flawed. Petrie (1967) reported .97 split-half reliability. The two terms which she correlated were (1) the mean of all judgments before aftereffect induction minus the mean of half the trials following aftereffect induction; and (2) the same mean of all judgments before aftereffect induction minus the mean for the other half of the trials following aftereffect induction. As Morgan and Hilgard (1972) pointed out, the same data appear in each term of this correlation. Such elements in common would tend to inflate artifactually the internal consistency correlation. Platt et al. (1971) also found high split-half reliabilities. However, since they failed to specify how each term in their correlation was defined, it remains unclear whether or not they avoided such elements in common.

In the absence of any unimpeachable published data, this paper explores the internal consistency for the first session administration of the KAE task. To provide a relatively comprehensive answer to the question, consistency values are reported on data from 10 samples run at 4 independent laboratories, encompassing 2 somewhat different KAE procedures. Positive consistency findings would strongly support the position that one-session KAE is a useful personality index.

Method

Table 1 presents the following information: (1) sample number; (2) type of subjects and source of sample; (3) number of subjects (and number for each sex separately); (4) size and type of aftereffect induction block (larger than, equal to, or smaller than the standard block whose size subject judges); (5) number of pretest trials; (6) number of test trials; (7) type of overall procedure [whether patterned after Petrie (1967 or Weintraub et al. (1973)]; and (8) specification of the exact computation of the two terms entered into the internal consistency computation.

Apparatus

For each sample, a 30-inch (76.2-cm) long, tapered, comparison wedge was used. For Samples 1–5, a 2.5-inch (6.35-cm) wide, aftereffect induction *(I)* block was employed, while for Samples 6–10, a 2-inch (5.08-cm) wide, *I* block was used. The test *(T)* block used varied in width according to the sample (see Table 1). A mounted ruler ran the length of the wedge, around which were a pair of finger guides, allowing the exact location of the subject's hand on the wedge to be recorded at the end of each trial. Similar finger guides were used while the subject was judging the *T* block or rubbing the *I* block. For the Petrie procedure, the subject was blindfolded and seated with the apparatus on a table in front of him/her. For the Weintraub procedure, the subject was standing and wore a cardboard collar which prevented sight of the apparatus.

Procedure

The procedure followed Petrie (1967, Appendix A) exactly for Samples 1–5. First, the subject rested his/her hands for 45 minutes so that nothing touched the thumbs and forefingers. During this time, the subject's attention was occupied with various orally administered questionnaires. The KAE task, involving 18 width judgments in all, was then given. On each trial, the subject held the *T* block with the thumb and forefinger of the *right* hand and indicated its width on the tapered block with the thumb and forefinger of the left hand. All trials were ascending, starting at the narrow edge of the wedge. The procedure consisted of

1. Two practice judgments;
2. Four pretest (prior to aftereffect induction) judgments;
3. A 90-second aftereffect induction period during which each subject rubbed the *I* block;
4. Four test (post-induction) judgments;
5. Two more periods of aftereffect induction (90 seconds; 120 seconds), each followed by a set of four test judgments.

Samples 6 and 7 were run by Weintraub et al. (1973); and Samples 8-10, run by the present authors, used their procedure. In Samples 6–10 there was no 45-minute rest period. The subject began the KAE task by feeling the entire length of the wedge; two practice judgments followed. The subject then made four pre-induction (pretest) judgments of the width of *T*, holding it in the *left* hand and simultaneously indicating the apparent width on the tapered wedge with the right hand. On half of the trials, ascending judgments were made and on the remaining trials, descending judgments. By mounting the wedge in a track that allowed its point of objective equality to be moved 2.5-inches (6.35-cm) forward and 2.5-inches (6.35-cm) back with respect to the subject's body, the wedge position was also alternated.

The same fixed order of trials was used for each subject. After the pretest, a 60-second induction period took place during which the subject rubbed the *I* block with the left hand. The *T* block was then replaced in the subject's left hand and four post-induction (test) judgments were made. For Samples 8–10 only, another 60-second aftereffect induction period took place and 4 further post-induction judgments were made (8 test judgments in all).

Scores

The usual KAE score consists of the difference between the mean of all pretest judgments and the mean of all test judgments. In computing internal consistency coefficients for each sample, scores were devised with half of the pretest and half of the test trials entering into each of the two terms to be correlated, so that there were no elements in common. Table 1 presents the precise way in which these scores were defined for each sample.

Petrie procedure. For Samples 1–5, there were four pretest trials and three sets of four test trials each. The two terms to be correlated were: (1) the mean of six test trials (the first and last test trials from each of the three sets) minus the mean of two pretest trials (the first and last pretest trials) and (2) the mean of the remaining six test trials minus the mean of the remaining two pretest trials. This method of combining scores resulted in two terms composed of measures administered at approximately the same time, minimizing any temporal difference between the two terms.

Weintraub procedure. For Samples 6 and 7, there were four pretest trials and four test trials. The two terms were: (1) the mean of the first two test trials minus the mean of the first two pretest trials; and (2) the mean of the last two test trials minus the mean of the last two pretest trials. For Samples 8–10, there were four pretest trials and two sets of four test trials each. The two terms were: (1) the mean of four test trials (the first two test trials from each of the two blocks of test trials) minus the mean of the first two pretest trials; and (2) the mean of the remaining four test trials minus the mean of the last two pretest trials.

This method of combining scores resulted in two terms with exactly equal representation of ascending and descending trials as well as of back and forward trials. However, these two terms differ temporally. Given the actual sequence of trials within each set of trials used in these studies (ascending-back; descending-forward; descending-back; ascending-forward), there was no way to derive two terms with equal representation of ascending-descending and back-forward trials which did not result in some temporal difference. It was felt to be more important to equalize ascending/descending and back and forth trials. Weintraub et al. (1973) reported significant effects for both these variables—the mean judgments for ascending trials differed from that for descending trials as did the mean judgments for back as compared to forward trials. Therefore, it was decided to make the scores comparable in ascending-descending and back-forward status rather than in terms of temporal proximity.

Results

The internal consistency coefficients are reported in Table 1, both uncorrected and corrected, for scores with twice the number of trials, using the Spearman-Brown formula.

Petrie procedure. The internal consistency coefficients for Samples 1–5 ranged from .86 to .95 with a median value of .89 (each $p < .001$). The Spearman-Brown corrected values ranged from .92 to .97 with a median of .94, indicating very high reliability.

Weintraub procedure. The internal consistency coefficients for Samples 6–10 ranged from .43 to .63 with a median value of .59 (each $p < .005$). The Spearman-Brown corrected values ranged from .60 to .77 with a median value of .74, indicating an acceptable level of reliability.

The two procedures differed in internal consistency: There was no overlap across procedures in observed internal consistency coefficients, and the between-procedure difference in median corrected values (.20) was greater than the range of corrected values observed within either procedure.

Discussion

The internal consistencies of the KAE scores in 10 samples are impressive, most notably with those samples tested using the Petrie (1967) procedure. This is important evidence for the construct validation of KAE. These results do not refute findings of low KAE test-retest reliabilities. Retest reliabilities, as has been previously argued (Baker, Mishara, Parker, & Kostin, in press; Mishara & Baker, 1978a) are

TABLE 1

Internal Consistencies of KAE Scores in Ten Independent Samples

Sample Number	Nature and Source of Sample	Sample Size/ Sex	Type of After-effect Induction Block	Number of Trials		Proce-dure Used	Computation Formulae [a]	Internal Consistency	Spearman-Brown Correction
				Pre-test	Test				
1	College students (Mishara, Baker, Parker & Kostin, 1973, Sample 4)	79 (32 Male, 47 Female)	Large	4	12	Petrie	$\left[\frac{T_1+T_4+T_5+T_8+T_9+T_{12}}{6} - \frac{P_1+P_4}{2}\right]$ by $\left[\frac{T_2+T_3+T_6+T_7+T_{10}+T_{11}}{6} - \frac{P_2+P_3}{2}\right]$	.86	.92
2	Community Active Elderly (Mishara & Baker, 1974)	40 (20 Male, 20 Female)	Large	4	12	Petrie	Same as Sample 1	.88	.94
3	College Students (Mishara, Baker, Parker & Kostin, 1973, Sample 1)	24 (10 Male, 14 Female)	Large	4	12	Petrie	Same as Sample 1	.89	.94
4	Nursing School Students (Barcus, Note 3)	41 (41 Female)	Large	4	12	Petrie	Same as Sample 1	.93	.96
5	Normal Adults (Petrie et al., 1972; Petrie et al., 1973)[b]	32 (10 Male, 22 Female)	Large	4	12	Petrie	Same as Sample 1	.95	.97

(Continued on next page)

Table 1, continued

6	College Students (Weintraub, Green & Herzog, 1973)	35 (19 Male, 14 Female, 2 unknown)	Large	4	4	Wein-traub	$\left[\frac{T_1+T_2}{2} - \frac{P_1+P_2}{2}\right]$ by $\left[\frac{T_3+T_4}{2} - \frac{P_3+P_4}{2}\right]$	.49	.66
7	College Students (Weintraub, Green & Herzog, 1973)	34 (13 Male, 18 Female, 3 unknown)	Small	4	4	Wein-traub	Same as Sample 6	.59	.74
8	College Students (Baker & Mishara, un-published data)	36 (all female)	Small	4	8	Wein-traub	$\left[\frac{T_1+T_2+T_5+T_6}{4} - \frac{P_1+P_2}{2}\right]$ by $\left[\frac{T_3+T_4+T_7+T_8}{4} - \frac{P_3+P_4}{2}\right]$	.43	.60
9	College Students (Baker & Mishara, un-published data)	36 (all female)	Large	4	8	Wein-traub	Same as Sample 8	.60	.75
10	College Students (Baker & Mishara, un-published data)	36 (all female)	Same	4	8	Wein-traub	Same as Sample 8	.63	.77

[a] In these formulae for internal consistency, T refers to Test (Post-Induction) trials and P refers to Pretest (Pre-Induction) trials, with the subscripts indicating the consecutive order of Test and Pre-test trials.

[b] Data for this sample are also discussed in Mishara, Baker, Parker & Kostin, 1973 (Sample 2).

attenuated by individual differences in bias or practice effects and are inappropriate for the KAE. Low test-retest reliabilities only reflect the fact that a second KAE administration does not measure the same thing as the first (Baker, Mishara, Kostin, & Parker, 1976).

Since test-retest reliability is an inappropriate measure, alternative estimates of true reliability should be considered. The high internal consistency reported here should correct the false impression created by the test-retest findings that KAE has never demonstrated "true" reliability. The present findings converge with the strong evidence of KAE first-session validity (Baker, Mishara, Kostin, & Parker, 1976), indicating that a one-session KAE procedure remains a viable index of personality functioning.

Morgan and Hilgard (1972) and Sales and Throop (1973) have argued that mathematical relations among KAE scores preclude their showing adequate reliability. They applied the following formula for the reliability of a difference score (McNemar, 1969; Gulliksen, 1950):

$$r_{DD} = (r_{XX} + r_{YY} - 2r_{XY})/(2 - 2r_{XY}), \qquad [1]$$

where r_{DD} is the reliability of the KAE (test minus pre-test) difference score, r_{XX} is the reliability of the KAE pre-test scores; r_{YY} is the reliability of the KAE test score; and r_{XY} is the correlation between the pretest and test scores.

These critics argued that this formula shows that when any two responses are highly correlated, a difference score based upon those two responses must necessarily show low reliability. For example, Sales and Throop (1973) commented that "because of these high correlations [that is, between pretest and test], KAE scores should be unreliable [in the psychometric sense] . . ." (p. 496). They concluded that KAE scores can only have a small percent of total variance attributable to true score variance.

These conclusions, dramatically at variance with the present paper, are in error. The critics failed to note that r_{XY} and r_{DD} could be quite high, as long as the average of r_{XX} and r_{YY} is suitably larger than r_{XY}. For example, median values for the Petrie-type samples were $r_{XY} = .80$; $r_{XX} = .97$; $r_{YY} = .99$. The high value of r_{XY} did not prevent a finding of even higher r_{DD} (median = .89) precisely because r_{XX} and r_{YY} were extremely high, a possibility overlooked by the critics.

Why does the Petrie-type procedure show higher reliability? Perhaps the Petrie procedure gives rise to scores with inflated observed reliability because only ascending trials are employed; this might produce correlated errors of anticipation and habituation. On the other hand, the Weintraub-type procedure may show lower observed reliability because it involves a smaller number of trials (8 in Samples 6 and 7; 12 in Samples 8 to 10) than the Petrie procedure (16 trials). When the Spearman-Brown formula was used to estimate the reliability of a 16-trial-long procedure for each of the Weintraub-type samples, the median of these 5 estimated values was .80. Although this more closely approximates the .94 value of the Petrie-type procedure, some additional data suggests that an equal number of trials may not result in equal reliability. Equation 1 indicates that, with other things being equal, higher r_{XX} results in higher r_{DD}. For Petrie-type samples, r_{XX} ranged from .94 to .98 with a median of .97; whereas, for Weintraub-type samples, r_{XX} ranged from .72 to .84 with a median of .75. (Applying Spearman-Brown, these medians become .98 and .86, respectively.) However, both procedures have the identical number of pretest trials.

This issue was further explored with some recently collected data. Twenty-six right-handed subjects were run through a variant of the Weintraub procedure involving 4 pretest and 16 test trials, making the procedure slightly *longer* than the Petrie-type studies. The procedure was identical to the prior Weintraub-type studies except that back and forward variation in wedge position was eliminated (the wedge was always in the middle position). The aftereffect induction block was equal in size to the standard block, both being 2.5 inches (6.5-cm) wide. The two terms entered into the correlation were: (1) the mean of the first (ascending) and third (descending) test trials minus the mean of the first and third pretest trials from each of the four sets of test trials; and (2) the mean of the remaining eight test trials minus the mean of the second and fourth pretest trials. The internal consistency coefficient was .72. The Spearman-Brown corrected value was .84, a value precisely halfway between the median corrected values for the Weintraub-type (.74) and the Petrie-type (.94) procedures.

It would appear that increasing the number of trials does increase the internal consistency of the Weintraub-type procedure, but not enough to match that of the Petrie-type procedure. However, caution is indicated before drawing final conclusions. As Lord and Novick (1968, p. 135) noted, the Spearman-Brown formula will underestimate reliability if the two terms correlated do not exactly parallel each other. The two terms entered into correlations for the Petrie-type samples were equated temporally, but this was not possible for the Weintraub-type procedure (see Method). Even in the case of the additional sample just reported, the temporal disparity between the two terms was only reduced (the trials show partial temporal overlap, which was not the case in any of the five earlier discussed Weintraub-type samples), but not eliminated. Such a temporal difference could artifactually make the Weintraub-type procedure appear less reliable than it actually is.

In conclusion, both procedures provide adequate reliability. Future users of the Weintraub-type procedure should be able to obtain an increment in reliability by extending the number of trials; however, the possibility exists that even with an equal number of trials, the Petrie-type procedure may have higher reliability. This issue requires further research.

References

Bakan, P., & Thompson, R. On the relation between induced and residual kinesthetic aftereffects. *Perceptual and Motor Skills,* 1962, *15,* 391–396.

Baker, A. H., Mishara, B. L., Kostin, I. W., & Parker, L. Kinesthetic aftereffect and personality: A case study of issues involved in construct validation. *Journal of Personality and Social Psychology,* 1976, *34,* 1–13.

Baker, A. H., Mishara, B. L., Parker, L., & Kostin, I. W. When "reliability" fails, must a measure be discarded?—The case of kinesthetic aftereffects. *Journal of Research in Personality,* in press.

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

Gupta, B. S. Stimulant and depressant drugs on kinesthetic figural after-effect. *Psychopharmacologia,* 1974, *36,* 275–280.

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

McDonald, A. The lack of effect of D-amphetamine of perceptual reactance and personality. *Journal of Abnormal Psychology,* 1974, *83,* 87–90.

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

Mishara, B. L., & Baker, A. H. *Bias in Petrie's alternate form kinesthetic aftereffect procedure.* Unpublished manuscript. University of Massachusetts at Boston, 1978. (a)

Mishara, B. L., & Baker, A. H. *Stimulus intensity modulation: A perceptual style approach to individual differences in old age.* Unpublished manuscript. University of Massachusetts at Boston, 1978. (b)

Mishara, B. L., & Baker, A. H. Individual differences in old age: The stimulus intensity modulation approach. In R. Kastenbaum (Ed.), *Old age on the new scene.* Springer, in press.

Morgan, A., & Hilgard, E. The lack of retest reliability for individual differences in the kinesthetic aftereffect. *Educational and Psychological Measurement,* 1972, *32,* 871–878.

Petrie, A. *Individuality in pain and suffering.* Chicago: University of Chicago Press, 1967.

Petrie, A., Collins, W., & Solomon, P. Pain sensitivity, sensory deprivation, and susceptibility to satiation. *Science,* 1958, *128,* 1431–1433.

Platt, D., Holtzman, P., & Larson, D. Individual consistencies in kinesthetic figural aftereffects. *Perceptual and Motor Skills,* 1971, *32,* 787–795.

Poser, E. Figural aftereffect as a personality correlate. *Proceedings of the XVIth International*

Congress of Psychology. Amsterdam: North Holland Publishing Company, 1960, 748–749.

Ryan, E. D., & Foster, R. Athletic participation and perceptual reduction and augmentation. *Journal of Personality and Social Psychology,* 1967, *6,* 472–476.

Sales, S. Need for stimulation as a factor in social behavior. *Journal of Personality and Social Psychology,* 1971, *19,* 124–134.

Sales, S. Need for stimulation as a factor in preferences for different stimuli. *Journal of Personality Assessment,* 1972, *36,* 55–61.

Sales, S., & Throop, W. Relationship between kinesthetic aftereffect and strength of the nervous system. *Psychophysiology,* 1972, *9,* 492–497.

Spilker, B., & Callaway, E. "Augmenting" and "reducing" in averaged visual evoked responses to sine wave light. *Psychophysiology,* 1969, *6,* 49–57.

Sweeney, D. R. Pain reactivity and kinesthetic aftereffect. *Perceptual and Motor Skills,* 1966, *22,* 763–769.

Weintraub, D., Green, G., & Herzog, T. Kinesthetic aftereffects day by day: Trends, task features, reliable individual differences. *American Journal of Psychology,* 1973, *86,* 827–844.

Acknowledgments

We thank Jean Barcus, Arlene Morgan, Asenath Petrie, and Daniel Weintraub for making their data available to us. Thanks are also extended to Scot Tempel, who served as experimenter for Samples 8, 9, and 10; to Anthony Risser, who was experimenter for the additional sample mentioned in the discussion; and most especially to our colleagues, Irene W. Kostin and Laurence Parker, whose comments were invaluable at all stages of this research.

Author's Address

Brian L. Mishara, Psychology Department, University of Massachusetts at Boston, Boston, MA 12125