Animal Learning & Behavior
1985, 13 (4), 345-348

Higher reliability and closer relationship between open-field test measures on aggregation data

TOSHIAKI TACHIBANA
Institute for Developmental Research, Aichi, Japan

Eighty male rats were tested in an open field. Correlation coefficients between aggregated test days were larger than those between nonaggregated test days, indicating that aggregation across days can enhance the reliability of scores in the open-field test. Also, absolute values of correlation coefficients among the seven open-field test measures based on the aggregated data were generally larger than those based on nonaggregated data, indicating that the correlation among measures may be closer than previously assumed on the basis of nonaggregated data. Issues concerning appropriate aggregation and limitations of aggregation are discussed. The technique of aggregation is recommended as a routine procedure in the analysis of open-field test results, because of the enhanced reliability obtained.

The open-field test (OFT) has been widely used in the behavioral study of rodents, but an inability to replicate published findings has frequently been pointed out. Two widely cited and influential reviews (Archer, 1973; Walsh & Cummins, 1976) have emphasized the unreliability of OFT results and have led to a rather pessimistic view of the utility of the OFT.

Two forms of reliability are at issue in the use of the OFT: between-study reliability and within-study reliability. The skepticism about between-study reliability, which means replicability of findings in independent studies, may be in large part a consequence of our excessive faith in adequate power of small-sample studies (Tachibana, 1982a). There is a widespread belief in the replicability of statistically significant results. However, statistical power is rarely so large that we can expect consistent statistically significant results. This erroneous belief has been uncovered by the studies based on questionnaire surveys (Tachibana, 1984; Tversky & Kahneman, 1971). What should be realized is that the sample size necessary to produce adequate power is substantially greater than psychologists have typically assumed.

Within-study replicability means the reliability of individual subject scores within a study. Inadequate reliability in individual subject scores also raises a serious problem. A poor within-study replicability results in a large variance in data and thereby leads to insensitive research on a treatment effect. One way to reduce the influence of error associated with incidental factors and thereby enhance the reliability of individual subject scores is to aggregate scores across occasions.

Address correspondence to: Toshiaki Tachibana, Institute for Developmental Research, Aichi Prefectural Colony, Kasugai, Aichi 480-03, Japan.

Recently, Epstein (e.g., 1983) has provided impressive evidence on the utility of aggregation in the area of personality research. As Epstein showed, aggregation of data is a very useful technique for enhancing reliability. Aggregation is not a new technique in OFT study. Some studies have performed analyses on the basis of aggregated data across total test days. However, such an aggregation has not been performed for the purpose of enhancing reliability, but has been used only for the sake of simplicity in the interpretation of results, without the theoretical basis for the aggregation that Epstein, for example, has pointed out.

Therefore, this study seeks to assess more explicitly how reliability of OFT measures may be obtained by aggregation. It also examines whether a new type of relationship among OFT measures can be obtained by aggregation across test days.

METHOD

Subjects

The subjects were male Sprague-Dawley rats about 5 weeks of age at the beginning of the OFT. They were obtained from the Shizuoka Laboratory Animal Center in Japan. The rats were the offspring of 20 dams and were identified by their litter numbers. Four animals were selected randomly from each litter, resulting in a total of 80 subjects.

Apparatus

The open field was 60×60 cm; it was enclosed by a wall 60 cm high and was composed of a 3×3 matrix of 20-cm squares. The floor and walls were painted white. A ventilating fan provided background noise (about 60 phons).

Procedure

All rats were tested under an illumination of 850 lx on the floor of the field. The rats were tested in the open field for 3 min/day for 4 consecutive days. Seven measures were used in the test: ambulation (AM), penetration into the inner square (PIS), rearing (RE),

defecation (DE), latency of defecation (LD), urination (US), and latency of urination (LU). Defecation was scored by the number of fecal boluses. Urination was graded on a scale from 0 (no urination) to 2 (large amount). If no defecation or urination occurrred, a 181-sec latency was recorded for the latency of defecation or urination.

RESULTS

In order to assess the reliability of individual measures, correlation coefficients (r) among 4 days were calculated for the seven measures; these coefficients are presented in Table 1. Latency of defecation and latency of urination were calculated on the square root transformed data. There was no negative correlation among days, so the aggregation across days was a reasonable technique.

Two types of aggregation were applied to each measure: (1) aggregation over a period of 2 days (Day 1-Day 2, Day 3-Day 4), and (2) aggregation over a period of 3 days (Day 1-Day 3 or Day 2-Day 4). Correlation coefficients calculated on the basis of aggregated data are also presented in Table 1. It is clear that the correlation between 2 days' aggregation data (i.e., Day 1-Day 2 and Day 3-Day 4) was larger than that between nonaggregation data corresponding to the 2 days' aggregation data (see difference between aggregated and nonaggregated data in the same rows in Table 1). As to the correlation between Day 1 and 3 days' aggregation (i.e., Day 2-Day 4) or the correlation between another 3 days' aggregation (i.e., Day 1-Day 3) and Day 4, larger r scores than those from nonaggregated data were also obtained, although the magnitude of increment by aggregation is rather smaller than that by 2 days' aggregation

Correlation coefficients (rs) among the seven measures on each day were calculated, and the major results are presented in Table 2. Correlation coefficients among the seven measures were calculated on the basis of two types of aggregated data, and the major results of these calculations are also presented in Table 2. As can be seen, the absolute value of the correlation coefficient based on aggregated data tends generally to increase in magnitude in comparison with that of nonaggregation data.

A third type of aggregation (across subjects within a litter) is also possible. Table 2 also contains the results based on double aggregation, that is, first, aggregation across Day 2-Day 4 for each subject, and, second, aggregation across subjects within each litter. The result to be noted is the correlation between AM and DE. The coefficient increased to −.40, suggesting a rather strong relationship between two measures. The correlation between AM and LD also increased to a large magnitude (.63).

There was a considerable variation in the magnitude of scores among days. It seems unsuitable to weight equally all days on the basis of the raw score for the aggregation. Thus, a z transformation was performed within each test day. The results of the correlation based on the z scores, however, were quite similar to the results of the correla-

Table 1
Correlation Coefficients (r) Between Aggregated Data and Nonaggregated Data

	Nonaggregated				Aggregated
Measure	Day 1:Day 3	Day 1:Day 4	Day 2:Day 3	Day 2:Day 4	(Day 1+Day 2): (Day 3+Day 4)
AM	.14	.35	.45	.47	.51
RE	.34	.36	.45	.48	.56
PIS	.35	.37	.48	.42	.61
LD	.31	.27	.41	.31	.50
DE	.40	.27	.36	.25	.46
LU	.30	.35	.41	.39	.50
UR	.26	.23	.32	.25	.40

	Day 1:Day 2	Day 1:Day 3	Day 1:Day 4	Day 1: (Day 2+Day 3+Day 4)
AM	.30	.14	.35	.32
RE	.39	.34	.36	.50
PIS	.27	.35	.37	.42
LD	.16	.31	.27	.31
DE	.36	.40	.27	.46
LU	.33	.30	.35	.41
UR	.25	.26	.23	.34

	Day 1:Day 4	Day 2:Day 4	Day 3:Day 4	(Day 1+Day 2+Day 3): Day 4
AM	.35	.47	.57	.64
RE	.36	.48	.52	.59
PIS	.37	.42	.37	.51
LD	.27	.31	.49	.49
DE	.27	.25	.36	.38
LU	.35	.39	.53	.56
UR	.23	.25	.40	.42

Note − : *indicates correlation between two sets of data.* + *indicates aggregation. For example, (Day 1+Day 2) is the aggregation of data from Day 1 and Day 2.*

Table 2
Correlation Coefficients (r) Among Measures for Aggregation Data Across Days and for Nonaggregated Data

Aggregation	AM:RE	AM:PIS	AM:LD	AM:DE	LD:DE	LU:UR	DE:UR
Day 1	.67	.55	.17	.01	−.70	−.91	.24
Day 2	.79	.67	.07	−.04	−.67	−.87	.38
Day 1+Day 2	.74	.67	.15	−.11	−.73	−.89	.37
Day 3	.74	.70	.23	−.19	−.69	−.94	.43
Day 4	.77	.55	.26	−.15	−.75	−.82	.17
Day 3+Day 4	.77	.67	.30	−.29	−.78	−.91	.36
Day 1+Day 2+Day 3	.78	.74	.23	−.20	−.78	−.92	.44
Day 2+Day 3+Day 4	.81	.71	.28	−.22	−.81	−.91	.44
Litter (Day 2+Day 3+Day 4)	.67	.72	.60	−.40	−.87	−.92	.49

Note− : indicates correlation between measures. + indicates aggregation. For example, (Day 1+Day 2) is the aggregation of data from Day 1 and Day 2. Litter ( ) indicates double aggregation across days and across subjects within litter.

tion based on the original raw data, and the advantage of z transformation was negligible.

Factor analysis by direct Varimax solution was performed on both the aggregated data and the nonaggregated data, using squared multiple correlations as the initial estimates of communality. The main results of the factor analysis are presented in Table 3. The results for both aggregated data and nonaggregated data showed essentially the same pattern in factor loading. AM, PIS, and RE showed a large loading on Factor I. LU and UR had a large loading on Factor II. LD and DE had a large loading on Factor III. Taking into consideration the percent of total cumulative contribution, one may consider that two factors were obtained. In the double aggregated data across days and across subjects within each litter, three factors are considered to have been obtained.

Table 3
Factor Loadings and Cumulative Factor Contribution for Aggregated and Nonaggregated Data

Aggregation	Measures	Factor I	Factor II	Factor III
Day 2	AM	.86	.10	−.03
	RE	.85	−.03	−.01
	PIS	.74	.16	.05
	LD	.02	.47	−.63
	DE	.02	−.37	.65
	LU	.09	.90	−.06
	UR	−.11	−.89	.06
	(%C)	(47)	(92)	(111)
Day 2+Day 3+Day 4	AM	.86	.13	−.15
	RE	.89	.07	.05
	PIS	.80	.06	−.08
	LD	.12	.46	−.75
	DE	−.08	−.38	.77
	LU	.17	.92	−.07
	UR	−.05	−.93	.09
	(%C)	(43)	(83)	(106)
Litter (Day 2+Day 3+Day 4)	AM	.79	.21	−.26
	RE	.88	−.05	.12
	PIS	.85	.08	−.26
	LD	.31	.50	−.74
	DE	−.19	−.38	.82
	LU	.14	.94	−.04
	UR	−.02	−.95	.15
	(%C)	(31)	(78)	(102)

Note− %C indicates percent of total cumulative contribution. + indicates aggregation. Day 2 + Day 3 + Day 4 is the aggregation of data from Day 2, Day 3, and Day 4. Litter ( ) indicates double aggregation across days and across subjects within litter.

DISCUSSION

Correlation between 2 days' aggregation data (i.e., aggregation of Day 1-Day 2 and Day 3-Day 4) gives evidence that reliability is increased by the use of aggregated data. This is also borne out by the correlation between Day 1 and Day 2-Day 4, as well as that between Day 1-Day 3 and Day 4, although the increment is rather small in comparison with the 2 days' aggregation. This rather small increment may be attributed mainly to the fact that Day 1 scores (or Day 4 scores) are not aggregated data. Despite the difference in magnitude of increment, these results indicate clearly that suitable aggregation across days can enhance the reliability of individual subjects' scores.

When the correlation among OFT measures is analyzed on the basis of aggregated data which are also associated with increased reliability, the results show clearly that almost all correlations among measures tend to increase in magnitude. This means that the correlation among measures may be closer than has been assumed on the basis of nonaggregated data.

The most extensively studied relationship has been between defecation and ambulation. The consensus seems to be that the correlation between ambulation and defecation is probably not very large in magnitude. Surely, considering each of 4 days separately, the r is small (for Day 1 and Day 2, for example, r = .01 and −.04, respectively). However, if one aggregates scores across the test days, r = −.11. This is also true for Day 2-Day 4 aggregation data. If the score for Day 2-Day 4 is aggregated across subjects within each litter, the r value comes to −.40. Such an aggregation across subjects within each litter is reasonable for the following reasons. Statistical inferences require the assumption of independence in

samples. Another study by the author (in preparation) has revealed a considerable correlation between individual subjects within a litter in a few OFT measures (such as DE and LD). Therefore, individual scores are inappropriate as the unit of analysis, and the litter should be considered as the basic unit, at least in DE and LD.

Present results show that the correlation between ambulation and defecation was −.40 on the basis of the double aggregation across Day 2-Day 4 and across subjects within each litter. Is the magnitude of −.40 sufficiently large to justify the relationship? Although there is no objective criterion for the judgment, most researchers would probably consider that it is.

However, the most important point to note here is not the rather larger correlation coefficient of ambulation and defecation. In fact, other research has demonstrated larger rs than those found in the present research (see Table V of Archer, 1973). In such data, however, the question remains as to whether the results are adequately reliable. The most important point in the present results is that the r values were associated with increased reliability. The results of the present study make it possible to argue that a suitable aggregation not only enhances reliability, but also, in some cases, makes it clear that some relationships among OFT measures are closer than has been assumed.

Despite the fact that correlations between days for the same behaviors are usually quite small, the correlations between different behaviors for the same day are quite large and are consistent across days (Tables 1 and 2). This is not necessarily a curious phenomenon, for it suggests the consistency of the relationships between open-field behaviors despite fluctuation across days, and supports previous findings of consistency in the factor structure for open-field measures (Royce, 1977; Tachibana, 1982b).

Why were large correlations obtained in aggregation? Aggregation can cancel out incidental, uncontrolled factors and thereby enhance reliability. There is a common implicit assumption that open-field behavior on an individual test day is adequately consistent or reliable as the sample unit for statistical analysis. An ANOVA of repeated measures design, which is commonly employed in OFT studies, relies on such an assumption. However, open-field behavior on an individual test day tends to be dominated by incidental, uncontrolled effects. As a result, findings based on such data are unreliable. The larger correlation between measures gained by the aggregation of data is more clear, especially in the double aggregation by litter, in some cases (AM-DE and AM-LD) than in others. Why is there such a difference? The measures showing the larger gain show a relatively small correlation coefficient for the nonaggregated data; however, the reason for the difference in gain is not clear.

Aggregation, of course, does not always increase reliability, and the appropriate aggregation must be employed. In the present study, the 2 days' and 3 days' aggregation were employed instead of the total 4 days' aggregation of Day 1-Day 4. This is due to the fact that aggregation assured that the enhanced reliability should be employed, and assurance of the reliability of 4 days' aggregation data (Day 1-Day 4) is impossible without data to be compared other than the 4 days' aggregation data.

One shortcoming of aggregation is that it does not permit any assessment of the habituation process in open-field behavior in some cases. However, the possible enhancement of reliability by aggregation is so substantial that the technique should be recommended for routine use. Aggregation also overcomes to some degree the so-called unreliability of OFT, which is caused in part by an erroneous belief in the adequate reliability of individual test-day data. Needless to say, statistical inference, in this case, is restricted to the population constituted by such an aggregated unit.

The factor loading of aggregation data and nonaggregation data displayed essentially the same pattern. This pattern has been shown repeatedly in previous studies (e.g., Tachibana, 1982b). The defecation factor (Factor III) tends to appear as an independent factor in aggregation. In other words, the defecation factor might be regarded as an independent factor if OFT data were analyzed on the basis of more reliable data.

The arbitrary choice of value on the latency score in the case of no defecation or no urination may affect the correlation between amount and latency. However, in the present study, the concern is mainly with the magnitude of the correlation coefficient gained by aggregation. Therefore, the abstract magnitude for each measure, which may be flawed by the scoring method, is not a serious concern.

REFERENCES

ARCHER, J. (1973). Tests for emotionality in rats and mice: A review. *Animal Behavior*, **21**, 205-235.

EPSTEIN, S. (1983). Aggregation and beyond: Some basic issues in the prediction of behavior. *Journal of Personality*, **51**, 360-392.

ROYCE, J. T. (1977). On the construct validity of open-field measures. *Psychological Bulletin*, **84**, 1098-1106.

TACHIBANA, T. (1982a). A comment on confusion in open-field studies: Abuse of null-hypothesis significance test. *Physiology & Behavior*, **29**, 159-161.

TACHIBANA, T. (1982b). Open-field test for rats: Correlational analysis. *Psychological Reports*, **50**, 899-910.

TACHIBANA, T. (1984). A critical view of the utility of positive controls in a test battery. *Neurobehavioral Toxicology & Teratology*, **6**, 155-159.

TVERSKY, A., & KAHNEMAN, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.

WALSH, R. N., & CUMMINS, R. A. (1976). The open-field test: A critical review. *Psychological Bulletin*, **83**, 482-504.

(Manuscript received September 10, 1984;
revision accepted for publication February 22, 1985.)