Psychol. Belg.
1988, XXVIII-1, 77-86.

A COMPARISON OF SIX STATISTICAL TESTS FOR EVALUATING PRESOLUTION STATIONARITY IN GRADUAL AND ALL-OR-NONE LEARNING

André VANDIERENDONCK

University of Ghent
Laboratory of Experimental Psychology

The hypothesis that a process is either all-or-none or gradual is usually tested with χ^2 tests based on a partitioning of the presolution trials or on the individual backward presolution curves. Both methods are to some extent biased towards null hypothesis acceptance. It is argued that techniques based on a correlation measure provide a better test of the two hypotheses. By means of Monte Carlo samples six tests for deciding whether a process is stationary are compared: the nonparametric correlation measures r_{Tukey} and $r_{Spearman}$ and the product-moment correlation coefficient tested against a bootstrap sampling distribution, the t test associated with the product-moment correlation coefficient and the χ^2 tests based on trial partitioning and on the backward learning curve. Within each cell of a completely factorial design based on sample size (5 vs. 20), sample homogeneity (homogeneous vs. heterogeneous), learning speed (slow or fast) and trial blocking, 40 Monte Carlo samples were used. In general, the test statistics were only moderately sensitive, except for the χ^2 trial partitioning technique in large samples. The χ^2 backward learning curve technique was very unsensitive under all conditions. In small samples, the test of the product-moment correlation coefficient against the bootstrap distribution, appears to be the most appropriate test.

In problem-solving and concept-learning experiments it is sometimes necessary to decide whether the subjects progress gradually to a solution or switch from a 'know-nothing' to a 'know-all' state. Group performance curves almost always reveal a gradual growth of performance. With a gradual solution process this is evident, since the group curve is obtained as the aggregation of a set of individual growth curves. If all subjects solve the problem in an all-or-none manner, the combined group curve also shows a gradual increment of performance, because the subjects switch from the nonsolution to the solution state at different moments in learning. This leads to an increasing number of subjects with errorless performance, and a gradual *average* curve.

To test whether learning progresses gradually or according to the all-

The author thanks Dr. Geert De Soete and an anonymous reviewer for their critical comments on an earlier draft of this paper.

or-none model, it is convenient to consider only the subjects' *presolution performance*, i.e., all trials *prior* to the last error. For each subject i, $1 \leqslant i \leqslant n$, we have then a score X_{ij} which expresses the performance of subject i on trial j, $1 \leqslant j \leqslant m_i$. In most applications this is a binary score (e.g., 0 = correct or 1 = incorrect) or a sum of such scores.

According to the all-or-none model the probability that a subject commits a presolution error is equal on all trials. An appropriate test should help to decide whether individual performance on the presolution trials is stationary. In practice, this is almost never possible, because the subjects do not contribute enough data for such a test.

Therefore, data of several subjects are pooled. However, the presolution sequences are of different length. One method, the *trial partitioning technique* (cf. Bower & Trabasso, 1964), partitions the subject's trials in a fixed number of blocks, and then pools the data per block, i.e.,

$$X_{.k} = \sum_{i=1}^{n} X_{ik} \tag{1}$$

$$= \sum_{i=1}^{n} \sum_{j \in J_k} X_{ij},$$

where k is the block index, and $J_k = \{j \mid j \text{ is a trial from block } k\}$. By means of the χ^2 statistic, the probability is estimated that the $X_{.k}$ are sampled from a uniform probability distribution.

The other method, the *backward learning curve technique*, first reverses all presolution runs (exclusive the last error) so that trial 1 is the last presolution trial, trial 2 the second but last, etc. Then the data are pooled per trial:

$$X_{.j} = \sum_{i \in I_j} X_{ij}/\text{card}(I_j), \tag{2}$$

where $I_j = \{i \mid 1 \leqslant i \leqslant n$, and i contributes data for trial $j\}$, and $\text{card}(I_j)$ is the cardinality of the set I_j. Again, a χ^2 statistic is used to estimate the probability that the $X_{.j}$ are sampled from a uniform probability distribution.

These methods have some severe drawbacks. Both are biased towards acceptance of the null hypothesis of presolution stationarity. An objection against the backward learning curve method relate to the differential weighting of slow and fast learning subjects in different phases of the solution process (cf. Atkinson, Bower & Crothers, 1965). If slow learners start with a low level of correct responding and improve slowly and fast learners start with high levels of performance and improve also over

trials, then the earlier trials of the pooled backward curve will be based essentially on data from slow learners, whereas the fast learners contribute data mainly to the later backward trials. In this way, the pooled averages can appear statistically invariant over trials and the null hypothesis is accepted when in fact it is false.

Although the trial-partitioning technique does not weight subjects differentially, it is also biased towards null hypothesis acceptance. First of all, the statistical test is not very sensitive as it uses only part of the data: in the process of partitioning, data from the central trials are excluded (e.g., when partitioning an odd number of trials into two blocks, the middle one belongs to neither block). Secondly, as most errors are committed by slow learners, these subjects are the main contributors to the data. If it can be assumed that their learning gradient is rather low, the test becomes biased in favor of the null hypothesis even if it is false.

Correlation measures

Within the context of the backward learning curve methodology another test statistic can be developed. There is a sample of subjects and each subject yields an ordered set of performance data $X_{i1}, \ldots X_{im}$, where X_{ij} is the performance of subject i on *backward* trial j. If the presolution stationarity model is correct, these values are sampled from a single distribution with mean μ and the null hypothesis states that

$$E(X_{i1}) = \ldots = E(X_{im}) = \mu. \tag{3}$$

On the other hand, if the gradual progress model is correct, error probability is expected to diminish as learning progresses:

$$E(X_{i1}) < E(X_{i2}) < \ldots < E(X_{im}), \tag{4}$$

so that in fact a *positive correlation* between performance and backward trial or block number is expected.

A stronger and more sensitive test statistic is therefore based on a correlation measure, ρ. The null hypothesis states that $\rho \leqslant 0$, whereas the alternative hypothesis is confirmed when $\rho > 0$. The measure can be said to be more sentitive because the order information in the sequence of trials or blocks is used and no data are excluded.

The assumption of bivariate normality is clearly not satisfied in the kind of data discussed here: one variable is uniformly distributed (block

numbers) and the performance variable is binomially distributed[1]. Therefore, the product-moment correlation coefficient is not very useful in this application. If the gradual progression hypothesis is correct, performance on each trial will be a monotonic function of trial number. Therefore a correlation index must be used which is sensitive to a monotonic relationship. It would appear that the Spearman rank correlation $r_{Spearman}$ is a good choice although the robust estimator of the product-moment correlation coefficient, r_{Tukey} (cf. Wainer & Thissen, 1976; Vandierendonck & De Soete, 1983) may be expected to do a good job also. The estimator r_{Tukey}, attributed to John W. Tukey, is defined by

$$r_{Tukey} = (1/4)\{[s(\tilde{x} + \tilde{y})]^2 - [s(\tilde{x} - \tilde{y})]^2\}, \quad (5)$$

where

$$s(x) = \frac{\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1)x_i$$

and

$$\tilde{x} = x/s(\tilde{x})$$

where $s(\tilde{x})$ is a robust estimator of scale, related to Gini's difference (cf. Vandierendonck & De Soete, 1983; Wainer & Thissen, 1976).

Constructing a confidence interval or testing the null hypothesis on the basis of a measure like r_{Tukey} by means of the classic test based on Student's t distribution, however, could invoke similar problems of bias as the χ^2 statistics discussed earlier.

The bootstrap technique

The bootstrap technique introduced by Efron (1979, 1981a, 1981b; Efron & Gong, 1983) offers a way out. In a backward learning curve each subject i contributes a series of performances X_{ij}, where j is the backward trial or trial block number. The complete sample is

$$X = (\{X_{ij}\}),\ i = 1, \ldots, n;\ j = m_i, m_i - 1, \ldots, 1.$$

By putting mass 1/n at each of the X_{ij} a sample probability distribution

[1] Strictly speaking, this statement only holds for the all-or-none hypothesis. For the sake of the argument, this is not important.

$\hat{F}$ can be constructed, and a random sample from this distribution is called a bootstrap sample (cf. also De Soete & Vandierendonck, 1982).

The mean and the variance of the bootstrap distribution, i.e., the distribution of bootstrap samples, can be used to approximate the sampling distribution of the random variables X_{ij}. With the help of Monte Carlo methods such a distribution is easily obtained. In practice, a number of bootstrap samples are generated by sampling with replacement from $\hat{F}$. Each bootstrap sample is used to compute the test statistic(s). The outcomes are accumulated to obtain the sampling distribution of the statistic(s). This approach has already been successfully used in studies of concept identification (e.g., Vandierendonck, 1984a, 1984b, 1987).

A Monte Carlo study

The remainder of this note will be devoted to a comparative study of the test statistics discussed. For the purpose of this study, artificial data sets were generated on the basis of either a gradual learning or an all-or-none learning model. The study was further arranged to capture effects of learning rate, degree of sample homogeneity, size of the trial blocks and sample size. Within each cell of a 2 (learning rate: fast vs. slow), × 2 (homogeneity: homogeneous vs. heterogeneous) × 2 (blocksize: 1 or 5 trials per block) × 2 (sample size: 5 or 20 statistical subjects per sample) factorial design, 40 data sets were generated, with the help of Schrage's (1979) portable random number generator. All programs needed for this study were written in FORTRAN77 running within a Unix environment on an Altos ACS68000 computer.

All-or-none homogeneous data sets were generated from the equation

$$\text{Prob}(X_{ij} = \text{error}) = qc(1 - qc)^{j-1} \tag{6}$$

(cf. Atkinson et al., 1965), with $q = .5$ and $c = .05$ (fast learning) or $c = .025$ (slow learning) in the homogeneous data sets. Heterogeneous data sets were obtained from the same equation, where the data for each statistical subject were generated while c was sampled from N(.05, 0.001) and N(.025, 0.001) for respectively fast and slow learning.

Data sets corresponding to a gradual model were obtained from the equation

$$\text{Prob}(X_{ij} = \text{error}) = q_j = q_1\alpha^{j-1} \tag{7}$$

with $q_1 = .5$ and $\alpha = .95$ (fast learning) or $\alpha = .975$ (slow learning) in

the homogeneous sets and with the actual α value sampled from N(.95, 0.001) and N(.975, 0.001) in heterogeneous sets.

The values of c and α were chosen such that the expected number of errors was equal in the corresponding conditions. Each data set was subjected to a bootstrap run with 500 samples, in which bootstrap confidence intervals were constructed. In each sample, furthermore, confidence intervals were estimated for r with the t test, and for the χ^2 trial partitioning and χ^2 backward learning curve techniques.

Tab. 1. — Number of Hits and False Alarms as a Function of Homogeneity, Learning Speed and Trial Blocking of the Six Test Statistics, in Samples of Size 5

	Bootstrap			Classic		
	r	r_{Tukey}	$r_{Spearman}$	r	χ^2_{bl}	χ^2_{tp}
Hits (maximum = 40)						
HEF1	21	19	15	18	1	7
HEF5	18	18	15	17	3	9
HES1	20	18	10	15	0	6
HES5	20	19	14	14	0	7
HOF1	22	16	10	16	0	9
HOF5	18	18	17	15	0	15
HOS1	27	22	16	16	2	18
HOS5	23	25	24	18	2	22
False alarms (maximum = 40)						
HEF1	2	4	2	3	0	0
HEF5	3	3	3	5	0	0
HES1	5	4	5	5	0	0
HES5	7	6	5	5	0	0
HOF1	1	1	1	6	0	0
HOF5	2	1	1	6	0	0
HOS1	3	3	2	6	0	0
HOS5	4	4	3	4	0	0

Note. χ^2_{bl} stands for χ^2 test based on the backward learning curve technique; χ^2_{tp} refers to the χ^2 test based on trial partitioning. The first three columns contain the results for the statistics based on the bootstrap technique. Furthermore symbols such as HEF1 should be read as HEterogeneous, Fast, Blocksize 1. So the meanings are: HE = heterogeneous, HO = homogeneous, F = fast, S = slow, 1 = blocksize 1 and 5 = blocksize 5.

An interesting question concerns the sensitivity of the test statistics employed. How well do these statistics discriminate between the two models of presolution stationarity? To answer this question, for each statistic, each sample received a score of 1, if the null hypothesis was rejected and a score of 0 otherwise. In samples based on the gradual model, a score of 1 means a hit, because the null hypothesis is correctly rejected. In samples based on the all-or-none model, to the contrary, a

score of 1 means a false alarm, as the null hypothesis is rejected when this is not appropriate.

Table 1 displays the number of hits and the number of false alarms for each of the 6 statistics as a function of homogeneity, learning speed and blocksize for the samples of size 5. The data for the samples of size 20 are displayed in Table 2.

Tab. 2. — Number of Hits and False Alarms as a Function of Homogeneity, Learning Speed and Trial Blocking of the Six Test Statistics, in Samples of Size 20

	Bootstrap			Classic		
	r	r_{Tukey}	$r_{Spearman}$	r	χ^2_{bl}	χ^2_{tp}
Hits (maximum = 40)						
HEF1	23	22	8	19	1	31
HEF5	26	28	26	28	14	36
HES1	24	24	20	22	0	24
HES5	28	29	28	28	10	25
HOF1	25	22	13	18	0	36
HOF5	19	21	22	19	13	37
HOS1	32	27	10	25	1	39
HOS5	27	29	30	26	15	40
False alarms (maximum = 40)						
HEF1	3	3	3	4	0	0
HEF5	8	8	7	7	0	0
HES1	4	4	4	6	0	1
HES5	7	5	5	9	0	1
HOF1	2	2	1	6	0	0
HOF5	2	1	1	4	0	1
HOS1	2	2	1	8	0	0
HOS5	2	2	0	6	0	0

Note. The same comments as in Table 1 apply.

On the basis of the proportions of hits and false alarms, measures of sensitivity can be computed. To avoid discussions about the appropriateness about certain assumptions, a nonparametric measure of sensitivity was employed. It is defined as

$$d = \frac{p(\text{hits}) - p(\text{false alarm})}{1 - p(\text{false alarm})}, \tag{8}$$

where p(hit) is the proportion of hits and p(false alarm) is the proportion of false alarms. The measure can be considered as an estimate of the probability of detecting a real difference.

Table 3 contains the sensitivity values of each test statistic as a function of the factors of the experimental design. The most obvious observation is that the sensitivity is, in general, moderate to low, except

for the trial partitioning test in the sample size-20 case. Another very saliant finding is that the χ^2 backward learning test is quite insensitive in all conditions. Furthermore, the data in Table 3 show that each of the factors involved in the factorial design affects sensitivity.

Tab. 3. — Sensitivity of Six Test Statistics as a Function of Sample Homogeneity, Learning Speed, Trial Blocking and Sample Size

Group	Bootstrap			Classic		
	r	r_{Tukey}	$r_{Spearman}$	r	χ^2_{bl}	χ^2_{tp}
sample size = 5						
HEF1	.50	.42	.34	.41	.03	.18
HEF5	.41	.41	.32	.34	.08	.23
HES1	.43	.39	.14	.29	.00	.15
HES5	.39	.38	.26	.26	.00	.18
HOS1	.54	.39	.23	.29	.00	.23
HOF1	.42	.44	.41	.27	.00	.38
HOS5	.65	.51	.37	.29	.05	.45
HOF5	.53	.58	.57	.39	.05	.55
sample size = 20						
HEF1	.54	.51	.14	.42	.03	.78
HEF5	.56	.63	.58	.64	.35	.90
HES1	.56	.56	.44	.47	.00	.59
HES5	.64	.69	.66	.61	.25	.62
HOS1	.61	.53	.31	.35	.00	.90
HOF1	.45	.51	.54	.42	.33	.92
HOS5	.79	.66	.23	.53	.03	.98
HOF5	.66	.71	.75	.59	.38	1.00
sample size						
5	.48	.44	.33	.32	.05	.29
20	.60	.60	.46	.50	.17	.84
homogeneity						
HE	.50	.50	.36	.43	.09	.45
HO	.58	.54	.43	.39	.10	.68
speed						
Fast	.50	.48	.36	.39	.10	.56
Slow	.58	.56	.43	.43	.09	.56
block size						
1	.58	.50	.28	.38	.02	.53
5	.51	.54	.51	.44	.18	.60

In general, tests based on larger samples (size 20) are more sensitive than those based on smaller samples (size 5): .318 vs .527 on the average. The effect is especially strong with the χ^2 trial partitioning test.

Tests based on heterogeneous data sets tend to be less sensitive than tests based on homogeneous data sets: on average .389 vs. .453. However, the effect is in the opposite direction for the t test associated with the product-moment correlation coefficient.

The effect of learning speed is difficult to predict. Faster learning tends to lead to steeper learning curves. Hence, nonstationarity should be easier to detect. At the same time, the number of data points is drastically diminished, so that the test becomes less sensitive. It turns out that, in general, the statistical tests are more sensitive when learning is slow: .399 vs. .442. However, for the techniques based on the χ^2 statistic, speed of learning seems to be of no importance.

Finally, it can be observed that tests based on an analysis with trial blocking (size 5) are more sensitive than an analysis based on individual trials (blocksize 1): .379 vs. .453 on the average. Again, there is an exception: the bootstrap r test is more sensitive with individual trials.

Conclusions and discussion

On the basis of the sensitivity data shown in Table 3, the following conclusions may be proposed:

1. If possible, perform a test on rather large samples: a sample size of 20 still leads to moderately sensitive tests.
2. If the sample size is about 20 or more, use the χ^2 trial partitioning technique with trial blocking (e.g., blocks of 5). This is even a good choice in absolute terms.
3. If the sample size is small, use the bootstrap methodology on r, but do not block trials.

In any case, unless the sample size is large, the bias will be toward null hypothesis acceptance. With small samples the test statistics compared in the present study are only sensitive enough to detect a graded process less than half of the time. The tests based on the χ^2 statistic perform even worse.

There are, of course, some limitations to the present study, in that each factor included in the design was represented by two levels only. Nevertheless, the results are quite clear: in general, sensitivity of the tests is moderate, and the alleged drawbacks of the tests based on the χ^2 statistic are confirmed, except for the trial partitioning technique with large samples. The present study could be extended to other sample sizes, other heterogeneity levels, etc. It does not seem likely that such a study would lead to different conclusions, unless a more efficient test statistic than the ones currently in use, is employed.

REFERENCES

Atkinson, R.C., Bower, G.H., & Crothers, E.J. (1965). *An introduction to mathematical learning theory*. New York: Wiley.

Bower, G.H., & Trabasso, T. (1964). Concept identification. In R.C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 32-96). Stanford: Stanford University Press.

De Soete, G., & Vandierendonck, A. (1982). On the use of the jackknife and the bootstrap for estimating a confidence interval for the product-moment correlation coefficient. *Psychologica Belgica, 22*, 87-97.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Efron, B. (1981a). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68*, 589-599.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics, 9*, 139-172.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *American Statistician, 37*, 36-48.

Schrage, L. (1979). A more portable FORTRAN random number generator. *ACM Transactions on Mathematical Software, 5*, 132-138.

Vandierendonck, A. (1984a). Evidence for two levels of processing in concept learning and abstraction? *Cahiers de Psychologie Cognitive, 4*, 217-244.

Vandierendonck, A. (1984b). Concept learning, memory and transfer of dot pattern categories. *Acta Psychologica, 55*, 71-88.

Vandierendonck, A. (1987). *Typicality gradient in well-defined categories*. Unpublished manuscript.

Vandierendonck, A., & De Soete, G. (1983). Some robust statistics for psychologists. *Psychologica Belgica, 23*, 73-83.

Wainer, H., & Thissen, D. (1976). Three steps towards robust regression. *Psychometrika, 41*, 9-34.

H. Dunantlaan 2
9000 Gent

Received October 1987