Evaluation of unconstrained sorting data

CHRIS T. BERSTED, BILL R. BROWN, *and* SELBY H. EVANS, *TEXAS CHRISTIAN UNIVERSITY, Fort Worth, Texas 76129*

Two statistics are offered for evaluating unconstrained sorting performance in a specific task when categories are E-defined. One statistic is based upon empirically estimated sampling distributions and can be used for determining sorting performance significantly deviant from chance for any number of S-defined categories; the second statistic can be used to evaluate consistencies between S- and E-defined categories, regardless of the number of categories used by S. The present procedure provides a statistical basis for evaluation of performance where no adequate evaluation procedures are now available.

Sorting tasks have been used extensively (Haufmann & Kasanin, 1937; Evans & Arnoult, 1967; Rosser, 1967; Imai & Garner, 1968) for comparing consistencies between E's and S's classifications and to study strategy preferences (Imai & Garner, 1968; Bruner, Goodnow, & Austin, 1956). Unconstrained sorting tasks (i.e., number of categories unspecified to S) intended to evaluate consistencies between E's and S's categories, on the other hand, have rarely been utilized because of difficulties in evaluation. The present report provides a preliminary and limited statistical basis to support evaluation with such tasks. In addition, the report will also illustrate how Monte Carlo methods can be used to obtain a similar basis for other special cases Es might be interested in.

Two requirements are necessary if one wishes to evaluate performance in an unconstrained sorting task with E-defined categories. First, one must have an objective means of determining if a S's categories are consistent with E's categories. Second, one must have a means for the evaluation of changes in consistency across trials.

Several technical problems exist in assessing performance in a free sorting task. First, although descriptive statistics have been previously suggested (e.g., Shipstone, 1960), their sampling distributions have been neither theoretically nor empirically determined. The use of such statistics does not permit one to evaluate performance in terms of deviation from chance. In addition, and possibly of greater importance, the sampling distribution of any statistic for the evaluation of free sorting would be expected to change as a function of the number of categories used. If so, performance across trials could not be unambiguously evaluated if different numbers of categories were used on successive trials.

INDEX OF CATEGORICAL RESPONDING

One solution to the first problem mentioned above would be to devise a suitable descriptive statistic and then to use Monte Carlo methods to approximate its distribution. A reasonable basis for such a statistic is the following assumptions: The greater the number of stimuli from an E-defined category placed together in a category, the better or more consistent the sort; conversely, the greater the number of stimuli from different E-defined categories placed together in a category, the poorer the sort. A statistic embodying both of these assumptions is given in Eq. 1:

$$\mathrm{ICR} = \sum_{i=1}^{N} (F_i - D_i) + C \qquad (1)$$

where N is the number of categories used by S, and the quantities in parentheses are based on a partition of the set of all distinct unordered pairs of stimuli in the i^{th} category as follows: F_i is the number of such pairs in which both stimuli are from the same E-defined category; D_i is the number of such stimulus pairs in which the stimuli are from different E-defined categories. The additive constant (C) is chosen to avoid negative numbers. The statistic is referred to as the Index of Categorical Responding (ICR).

The null hypothesis against which obtained ICR values could be compared states that instances are randomly sorted into some chosen number of categories (i.e., equal probability of sorting each instance into each category). This is equivalent to saying that S chooses some number of categories and distributes stimuli into these categories in a fashion independent of E-defined categories. In the present case, a trial is defined as distributing 10 instances from each of three E-defined categories into some chosen number of categories.

MONTE CARLO SIMULATION

The sampling distribution of the value of ICR under the null hypothesis would give us the probability of any ICR value occurring by chance. Empirical sampling distributions of ICR were therefore obtained, by a Monte Carlo procedure, for each number of categories in the range 3 through 15. The value of the constant C in these runs was 150. Each distribution was based on 3,000 trials, as defined above. Sorting was based on numbers provided by a pseudorandom number generator using the power residue method as described by Van Gelder (1967). As noted by Van Gelder and others, this method has repeatedly been found to give good results with respect to uniformity of distribution, means, variances, and serial correlation. The method tends to fail on some more sophisticated tests of serial independence (MacLaren & Marsaglia, 1965). Some suggestions offered by Van Gelder for avoiding these defects were incorporated into the present generator, but major reliance was placed on the simple technique of running 17 different generating sequences in parallel and choosing at random among them on each call. The numbers were only tested for uniformity of distribution; they were satisfactory in that respect.

An IBM 1800 computer system was used to generate the empirical ICR sampling distributions. The time necessary to compute each distribution varied from 15 to 35 min, with longer times necessary as the number of categories increased. Computing time, of course, would be drastically reduced on faster machines.

Table 1 presents several common significance levels and associated values of the ICR statistic for 3 through 15 categories estimated by the Monte Carlo procedure. These significance levels represent one-tailed statistical tests of the hypothesis that a given ICR value associated with a S's sorting performance occurred by chance. It should be noted that larger values of ICR indicate greater agreement between E-defined and S-defined categories. Likewise, larger ICR values are in general necessary to reach any significance level as the number of S-defined categories increases. For example, an ICR value of 114 for a S's sorting performance is significantly different from chance at the .05 significance level when three categories are used, whereas an ICR value of 144 is necessary to meet the same significance level when 10 categories are used by an S.

In addition to the significance levels, Table 1 gives estimates of the expected value and mode; these were also derived from the empirical sampling distributions. The maximum values of ICR, also given in the table, were calculated by first assuming $D_i = 0$. Then a maximum for F_i was determined by starting with three categories corresponding perfectly to the E-defined categories and distributing one

Table 1
Significance Levels and Summary Statistics Associated with Various Value of ICR for Different Numbers of Categories

	Categories												
	3	4	5	6	7	8	9	10	11	12	13	14	15
p(ICR) <.50	93	108	116	122	126	129	131	133	136	138	140	142	143
p(ICR) <.20	102	116	123	129	132	135	137	138	141	143	145	146	146
p(ICR) <.10	108	121	127	132	136	138	140	141	143	145	147	148	148
p(ICR) <.05	114	124	132	136	139	141	143	144	146	148	149	150	150
p(ICR) <.01	126	136	141	143	146	147	150	150	151	152	153	153	153
p(ICR) <.001	141	150	154	154	156	158	160	157	158	159	163	158	161
EV(ICR)	95	109	117	123	126	129	131	133	136	138	140	141	142
Mode (ICR)	90	106	116	120	125	127	131	132	136	137	141	142	143
Maximum (ICR)	285	276	268	261	255	250	246	243	241	240	231	223	216
Minimum (ICR)	6	15	25	34	42	51	59	66	74	81	87	94	100

instance from one of the E-defined categories to each of the remaining categories. Minimum values of ICR were obtained by putting as many instances as possible from each E-defined category into one category, resulting in a maximum number of inconsistent pairs, with each of the remaining categories being filled with one instance, these instances being drawn as equally as possible from each E-defined category.

The ICR statistic serves as a descriptive statistic for comparing a S's performance across trials, if the number of categories by the S does not change, and it can also be utilized as a test of significance. The use of ICR must be tempered with the knowledge that, as shown in Fig. 1, values of the statistic at different numbers of categories are not directly comparable, especially with a small number of categories.

INDEX OF RELATIVE PERFORMANCE

In order to be able to compare scores across different numbers of categories, some other descriptive statistic is needed. A position that might be adopted is that ICR scores at different numbers of categories may be regarded as equal if they have equal probabilities of occurrence under the null hypothesis. One solution, then, would be to standardize the ICR scores at each number of categories in terms of the mean and standard deviation for that distribution. This procedure was rejected, however, since the sampling distributions of the ICR statistic at different numbers of categories all exhibited a slight positive skewness, although all were unimodal and subjectively highly similar. Usual standardization procedures would therefore tend to underestimate probability densities associated with ICR values on the skewed end of the distributions, and similarly, overestimate the cumulative probability associated with these ICR values.

In most cases, only those scores above chance, indicating degrees of correspondence between E- and S-defined categories, would be of interest to the E. It was therefore felt that a statistic that only standardized the ICR scores with respect to the tails of the sampling distributions above the expected value might suffice for equating ICR values at different numbers of categories. A modified standard deviation (i.e., one that only reflected deviations above the expected value) could be used to accomplish this type of standardization. Rather than employ this unusual statistic, however, it seemed better to use the range of each distribution above its expected value as the representative of its variability. Inspection of the distributions indicated that they were similar in form and differed primarily in range, so that a standardization for range would remove most of the difference. Moreover, the resulting score would be directly interpretable as a proportion of the maximum possible deviation from chance. This statistic, termed relative performance (RP) is given in Eq. 2:

$$\mathrm{RP} = \frac{\mathrm{ICR\ (obtained)} - \mathrm{EV\ (ICR)}}{\mathrm{MAX\ (ICR)} - \mathrm{EV\ (ICR)}} \quad (2)$$

Figure 1 presents the RP value associated with the .05 significance level at each of the different numbers of categories. As can be seen from the figure, the same RP value is associated with this significance level at each number of categories, with only minor variations. These relatively small variations could

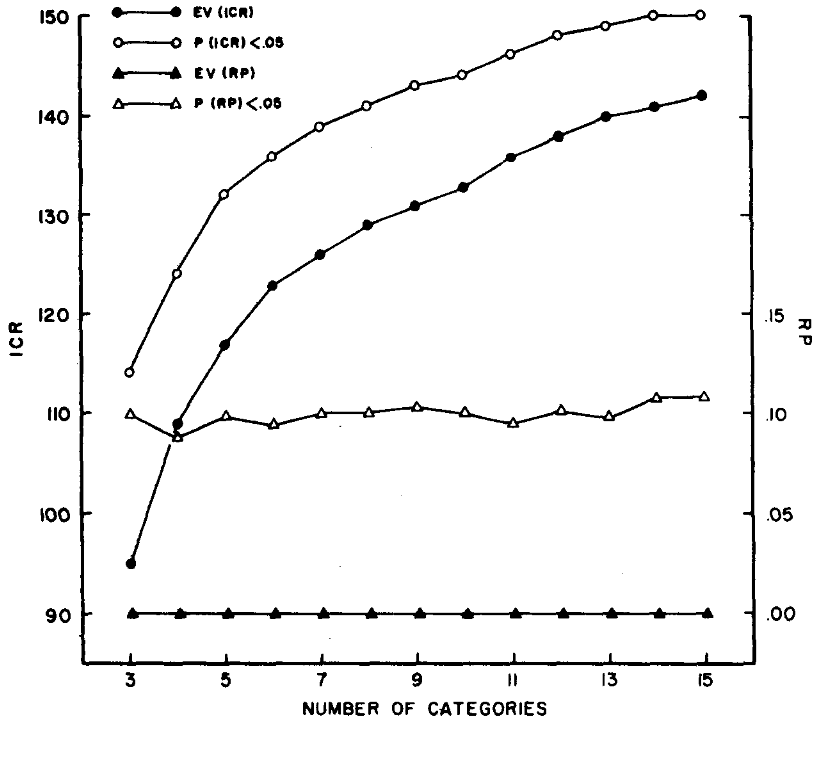

Fig. 1. Expected value and .05 significance levels of ICR and RP statistics at different numbers of categories.

easily result from sampling variability. Similarly, RP values calculated at other significance levels also remained constant at different numbers of categories, with only slight variations. Therefore, at least in the present case, RP seems to afford a rather efficient means for equating and evaluating performance at differing numbers of categories (e.g., provides an index that takes on the same value at any number of categories for a specified significance level). Whether or not the method used for the derivation of the RP statistic will prove useful in similar applications is at present uncertain. In any case, the RP statistic appears to exhibit the necessary invariance quality for at least the experimental situation used as a basis for this report.

It should be emphasized that, while the approach used in this paper is adequate for evaluation of performance in the specified sorting task and is likely generalizable to similar experimental situations, it is only a first approximation to more theoretical mathematical developments to be used in evaluation of sorting data. Several investigators (e.g., Johnson, 1969) are currently working on the derivation of theoretical statistics that can be used more generally for the evaluation of free sorting tasks. The present statistics and approach may be regarded as useful interim methods until more elegant procedures are available.

REFERENCES

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking.* New York: Wiley, 1956.

EVANS, S. H., & ARNOULT, M. D. Schematic concept formation: Demonstration in a free sorting task. Psychonomic Science, 1967, 9, 221-222.

HAUFMANN, D., & KASANIN, J. A method for the study of concept formation. Journal of Psychology, 1937, 3, 521-540.

IMAI, S., & GARNER, W. R. Structure in perceptual classification. Psychonomic Monograph Supplements, 1968, 2(9, Whole No. 25).

JOHNSON, S. C. Metric clustering. Paper presented at the Newport Beach Conference, Newport, Rhode Island, June 1969.

MacLAREN, M. D., & MARSAGLIA, G. Uniform random number generators. Journal of the Association for Computing Machinery, 1965, 12, 83-89.

ROSSER, E. M. Categorization and discrimination of tone sequences. Unpublished doctoral dissertation, Harvard University, 1967.

SHIPSTONE, E. I. Some variables affecting pattern conception. Psychological Monographs, 1960, 74(17, Whole No. 504).

VAN GELDER, A. Some new results in speudo-random number generation. Journal of the Association for Computing Machinery, October 1967, 14, 785-792.

NOTE

1. This research was supported by the Department of Defense, Project THEMIS Contract (DAAD05-68-C-0176), under the Department of the Army to the Institute for the Study of Cognitive Systems through the TCU Research Foundation. Further reproduction is authorized to satisfy needs of the U.S. Government.