Psychological Methods
2004, Vol. 9, No. 3, 386–396

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.3.386

Properties of the Hubert–Arabie Adjusted Rand Index

Douglas Steinley
University of Illinois at Urbana–Champaign and University of Missouri—Columbia

This article provides an investigation of cluster validation indices that relates 4 of the indices to the L. Hubert and P. Arabie (1985) adjusted Rand index—the cluster validation measure of choice (G. W. Milligan & M. C. Cooper, 1986). It is shown how these other indices can be “roughly” transformed into the same scale as the adjusted Rand index. Furthermore, in-depth explanations are given of why classification rates should not be used in cluster validation research. The article concludes by summarizing several properties of the adjusted Rand index across many conditions and provides a method for testing the significance of observed adjusted Rand indices.

Since the inception of modern computing techniques, there has been a near-exponential increase in the number and variety of clustering procedures, making comparisons between the procedures' performances crucial for determining which clustering method is the most appropriate. Often these comparisons are conducted under the guise of Monte Carlo studies that compare a recovered structure to a generated structure with desired properties and summarize the similarity between the two by a cluster recovery index (e.g., see Brusco & Cradit, 2001; Milligan, 1980; Milligan & Cooper, 1985). Milligan and Cooper (1986) indicated the five cluster recovery indices that are most heavily used in the literature to be the Rand index (Rand, 1971), the Morey and Agresti (1984) adjusted Rand index, the Hubert and Arabie (1985) adjusted Rand index, the Jaccard index (Downton & Brennan, 1980), and the Fowlkes and Mallows (1983) index. Another measure that is often used (although advised against; see Steinley, 2003) relies on the computation of classification rates (e.g., see Balakrishnan, Cooper, Jacob, & Lewis, 1994; Beauchaine & Beauchaine, 2002; Vichi & Kiers, 2001; Waller, Kaiser, Illian, & Manry, 1998; Waller, Underhill, & Kaiser, 1999).

This article (a) relates the Hubert and Arabie (1985) adjusted Rand index to the other measures mentioned above and facilitates the translation of results from several research reports and articles into a common language to aid in comparison of techniques and analysis, (b) shows the use of classification rates is misguided and ill-advised in cluster validation research because of loss of information when compared with the Hubert and Arabie adjusted Rand index, (c) provides a set of general properties of the Hubert and Arabie adjusted Rand index as an aid to interpretation of results, and (d) develops a general procedure to test the significance of observed Hubert and Arabie adjusted Rand indices.

The Setting

According to Milligan (1996), the validation of a clustering technique requires the generation of artificial data sets and testing via Monte Carlo simulation so the researcher has prior knowledge of the exact structure of the data. After generation and the subsequent application of a clustering technique, the resulting clusters are compared with the known structure. Given the data matrix $\mathbf{X} = \{x_{ij}\}_{N \times V}$, where N is the number of objects and V is the number of variables, a partition of the N objects with R subsets (or groups) can be formed, $\mathcal{P} = \{p_1, \ldots, p_R\}$, such that the union of all the subsets is equal to the entire object set and the intersection of any two subsets in $\mathcal{P}$ is empty. Given two partitions, $\mathcal{P}$ and $\mathcal{Q}$, with R and C subsets, respectively, the table $\mathcal{T}$ (see Table 1) can be formed to indicate group overlap between $\mathcal{P}$ and $\mathcal{Q}$. In $\mathcal{T}$, a generic entry, t_{rc}, represents the number of objects that were classified in the rth subset of partition R and the cth subset of partition C. The indices cited in the introduction are measures of correspondence between $\mathcal{P}$ and $\mathcal{Q}$ based on how pairs of objects are classified in $\mathcal{T}$. Letting $\binom{N}{2}$ represent the total number of pairs results in four different types of pairs: (a) Objects in a pair are placed in the same group in $\mathcal{P}$ and the same group

Douglas Steinley, Department of Psychology, University of Illinois at Urbana–Champaign, and Department of Psychology, University of Missouri—Columbia.

This research was partially supported by Office of Naval Research Grant N000014-02-1-0877 to the University of Illinois. Most of this research was conducted while I was at the University of Illinois. I thank Lawrence Hubert for comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Douglas Steinley, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820. E-mail: steinley@cyrus.psych.uiuc.edu

Table 1
Co-Occurrence Between Two Partitions, 𝒫 and 𝒬

		𝒬				
Partition	Group	q_1	q_2	$\cdots$	q_C	Total
	p_1	t_{11}	t_{12}	$\cdots$	t_{1C}	t_{1+}
	p_2	t_{21}	t_{22}	$\cdots$	t_{2C}	t_{2+}
𝒫	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	p_R	t_{R1}	t_{R2}	$\cdots$	t_{RC}	t_{R+}
Total		t_{+1}	t_{+2}	$\cdots$	t_{+C}	$t_{++} = N$

Note. 𝒫 and 𝒬 are partitions, whereas *R* and *C* are subsets of the respective partitions. t_{rc} indicates the total number of objects that simultaneously belong to the *r*th and *c*th subset.

in 𝒬, (b) objects in a pair are placed in the same group in 𝒫 and in different groups in 𝒬, (c) objects in a pair are placed in the same group in 𝒬 and in different groups in 𝒫, and (d) objects in a pair are placed in different groups in 𝒫 and different groups in 𝒬. This leads to alternatively representing 𝒯 as a 2 × 2 contingency table (see Table 2) based on (a), (b), (c), and (d). If 𝒯 is represented by an $R \times C$ matrix, $\mathbf{T} = \{t_{rc}\}$, the four cells in Table 2 are calculated by

$$a = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2 - N}{2}, \tag{1}$$

$$b = \frac{\sum_{r=1}^{R} t_{r+}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2}{2}, \tag{2}$$

$$c = \frac{\sum_{c=1}^{C} t_{+c}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2}{2}, \tag{3}$$

and

$$d = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2 + N^2 - \sum_{r=1}^{R} t_{r+}^2 - \sum_{c=1}^{C} t_{+c}^2}{2}. \tag{4}$$

The formulation in Table 2 leads to the simple computation of several of the indices previously discussed.

Partition Correspondence Indices

Rand index. The Rand index (Rand, 1971) has been used in several cluster validation studies (Dreger, Fuller, & Lemine, 1988; Dubes & Jain, 1976; Milligan, 1980; Milligan & Isaac, 1980; Milligan, Soon, & Sokol, 1983) and is calculated by

$$\text{Rand} = \frac{a + d}{a + b + c + d}. \tag{5}$$

Equation 5 gives weight to those objects that were concurrently classified together and apart in both 𝒫 and 𝒬. Unfortunately, Rand noted that the Rand statistic approaches its upper limit of unity as the number of clusters increases. Several measures have been created in an attempt to overcome such limitations.

Jaccard index. Although not as popular as the Rand (1971) index, the Jaccard index (Downton & Brennan, 1980) has also been used in cluster validation studies (Dubes, 1987; Milligan et al., 1983) and is calculated by

$$\text{Jaccard} = \frac{a}{a + b + c}. \tag{6}$$

Equation 6 differs from Equation 5 by removing *d* from both the numerator and denominator, placing the importance on *a*. However, because *b* and/or *c* must increase if *d* decreases, *d* is implicitly included in Equation 6.

Fowlkes and Mallows index. The Fowlkes and Mallows index (Fowlkes & Mallows, 1983) is calculated by

$$\text{Fowlkes–Mallows} = \frac{a}{\sqrt{(a + b)(a + c)}}. \tag{7}$$

The Fowlkes and Mallows index also removes *d* from the equation but, as in Equation 6, it is still implicitly included in the denominator.

Morey and Agresti adjusted Rand index. Morey and Agresti (1984) realized (5) may be overinflated due to chance assignment and corrected the Rand (1971) index for chance with the adjusted Rand index (ARI_{MA}). Unfortunately, ARI_{MA} cannot be computed by the notation in Table 2, but rather uses the notation of Table 1,

$$\text{ARI}_{\text{MA}} = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2 - \left(\sum_{r=1}^{R} \sum_{c=1}^{C} t_{r+}^2 t_{+c}^2 \right) \Big/ N^2}{\sum_{r=1}^{R} t_{r+}^2 \Big/ 2 + \sum_{c=1}^{R} t_{+c}^2 \Big/ 2 - \left(\sum_{r=1}^{R} \sum_{c=1}^{C} t_{r+}^2 t_{+c}^2 \right) \Big/ N^2}. \tag{8}$$

Upon its inception, Equation 8 was immediately adopted by

Table 2
2 × 2 Contingency Table Representation of the Partition Co-Occurrence Table

	𝒬	
𝒫	Pair in same group	Pair in different groups
Pair in same group	*a*	*b*
Pair in different groups	*c*	*d*

Note. 𝒫 and 𝒬 are partitions.

researchers (Milligan et al., 1983; Morey, Blashfield, & Skinner, 1983) to phrase the recovery capabilities of clustering algorithms in terms of an index corrected by chance.

Hubert and Arabie adjusted Rand index. Hubert and Arabie (1985) found something amiss with the unwieldily measure in Equation 8; specifically, Morey and Agresti (1984) incorrectly assumed that the expectation of a squared random variable is the square of the expectation. Hubert and Arabie corrected Equation 5 with the proper adjustment, creating the Hubert and Arabie adjusted Rand index ($\mathrm{ARI}_{\mathrm{HA}}$):

$$\mathrm{ARI}_{\mathrm{HA}} = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}. \quad (9)$$

Hubert and Arabie noted that the results provided by several Monte Carlo studies using the incorrect formulation provided by Morey and Agresti may provide misleading conclusions. $\mathrm{ARI}_{\mathrm{HA}}$ has been shown to be the most desirable index for measuring cluster recovery (Milligan & Cooper, 1986; Saltstone & Stange, 1996) and has been used in several cluster validation studies (Belbin, Faith, & Milligan, 1992; Donoghue, 1995; Dubes, 1987; Fisher & Hoffman, 1988; Klein & Dubes, 1989; Milligan, 1989a, 1989b; Milligan & Cooper, 1988). This widespread endorsement of $\mathrm{ARI}_{\mathrm{HA}}$ has lead to its inclusion in several cluster analysis applications beyond comparing two partitions. For example, $\mathrm{ARI}_{\mathrm{HA}}$ has been used to (a) aid in determining which variables to include in a cluster analysis (Brusco & Cradit, 2001; Carmone, Kara, & Maxwell, 1999), (b) develop a clustering procedure based on replicated results (Helsen & Green, 1991), (c) detect the influence of individual data points on a clustering procedure (Cheng & Milligan, 1996), and (d) obtain a consensus partition based on maximizing a weighted function of a set of adjusted Rand indices (Krieger & Green, 1999).

Classification Rates

Researchers often use classification rates in cluster analytic situations much as they would in a discriminant analytic situation (Balakrishnan et al., 1994; Beauchaine & Beauchaine, 2002; Breckenridge, 1989; Gierl & Schwanenberg, 1998; Vichi & Kiers, 2001; Waller et al., 1998, 1999) where a classification rate not equal to unity indicates an overlapping group structure (Scheiber & Schneider, 1985). Because the true cluster structure is known, one might expect that classification rates could be calculated for the simulated data sets by merely comparing the number of objects correctly allocated to each cluster. However, the concerns outlined below suggest that classification rates defined in this manner are somewhat problematic. *Classification rates* defined as the percentage of cases classified correctly by various methods unfortunately require an initial decision by the researcher of forcing the arbitrary assignment of cluster "labels." As an illustration, consider the example of Table 3 where it is ambiguous as to which sets in the two partitions, $\mathcal{P}$ and $\mathcal{Q}$, should be given the same "labels"—for example, which set $\mathcal{Q}$ should be "matched" with the first set of $\mathcal{P}$ ($\{a, b, c\}$)? Arguments could be made that either the first, third, or fourth groups of $\mathcal{Q}$ should be matched with the first group of $\mathcal{P}$. This arbitrariness of matching results in several problems: It implies the classification rate itself is arbitrary; the group assignment can be manipulated to generate either more or less favorable classification rates; and, finally, partitions can be compared only if they have same number of clusters.

Method

Simulation

The factors altered in simulation are reflective of several factors found in previous Monte Carlo studies on cluster validation: the number of clusters (K), the size of the kth cluster (n_k), and the distribution of n_k across the clusters.

The number of clusters. K was allowed to assume the values 2 to 8, a range that effectively covers the results reported in several studies (e.g., see Brusco & Cradit, 2001; Milligan, 1980; Milligan & Cooper, 1985).

Cluster sizes. The total number of objects considered, N, had four values (i.e., 50, 100, 200, 300). The size of each cluster can be represented in the $1 \times K$ vector, $\mathbf{n} = [n_1, \ldots, n_K]$, where n_k ($k = 1, \ldots, K$) represents the number of objects in the kth cluster. The distribution of N across K assumed three values: (a) equal cluster sizes for all K clusters, (b) one cluster must contain 10% of N and the remainder of N is distributed equally across the remaining clusters, and (c) one cluster must contain 60% of N and the remainder of N is distributed equally across the remaining clusters.

Simulating $\mathcal{T}$. Because Equations 5 to 9 can be calculated from $\mathcal{T}$, it is sufficient to simulate different structures of $\mathcal{T}$ by manipulating its matrix counterpart, $\mathbf{T}$. Specifically, $\mathbf{T}$ represents the body of the table $\mathcal{T}$. Furthermore, for the purposes of this study, the number of rows (R) and columns

Table 3
Arbitrariness of Group Assignment

Partition	Object assignment
$\mathcal{P}$	$\{a, b, c\}, \{d, e\}, \{f, g, h\}, \{i, j\}, \{k, l, m\}$
$\mathcal{Q}$	$\{a, m, f\}, \{d, i\}, \{b, g, l\}, \{j, c\}, \{e, h, k\}$

Note. $\mathcal{P}$ and $\mathcal{Q}$ are partitions. Objects together within braces are considered to be in the same group within a partition.

(C) will each be fixed to equal the number of clusters (i.e., $K = R = C$), whereas in the general situation it is not necessary for R and C to be equal. If the classes of $\mathcal{P}$ are considered the rows of **T** and $\mathcal{Q}$ the columns, perfect agreement can be represented by the diagonal matrix

$$\mathbf{T} = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & n_K \end{bmatrix},$$

and each index will obtain a value of unity. To facilitate overlapping partitions, we take objects from the diagonal of **T** and randomly place them in an off-diagonal cell. However, because the purpose is to emulate cluster validation studies, it is assumed that one of the partitions in **T** is fixed. For example, letting $\mathcal{P}$ be considered fixed results in randomly placing objects in the off-diagonal and maintaining the marginal totals of the rows.

The number of objects placed in the off-diagonal cells ranged from .05N to N in increments of .05. Combining these factors (number of clusters, number of objects, relative cluster size [density], and degree of partition overlap) results in 7 (number of clusters) $\times$ 4 (values of N) $\times$ 3 (relative density) $\times$ 20 (degree of overlap) = 1,680 conditions. One hundred random **T** matrices were generated for each condition, resulting in a total of 168,000 random matrices.

Calculating correct classification rates. In the discussion above, it was seen that classification rates cannot be calculated without an arbitrary decision concerning group membership. As an illustration of the shortcomings of classification rates, an arbitrary decision that results in optimal correct classification rates was made. The optimal correct classification rates can be arrived at by fixing the columns and permuting the rows of **T** (or fixing the rows and permuting the columns) to maximize the trace of **T**, tr(**T**), and correct classification rates can be calculated by

$$\frac{\mathrm{tr}(\mathbf{T})}{N}. \tag{10}$$

Thus, by calculating Equation 10, the benefit of doubt is given to the correct classification rates.

Results

ARI$_{HA}$ Related to Other Indices

Milligan and Cooper (1987) noted "it is important to note that criterion values used in the experiments may differ from study to study. Thus, the direct numerical comparisons of criterion values can be made only within a study and not across reports" (p. 336). Previously, it has been noted that the ARI$_{\mathrm{HA}}$ is the measure of choice when considering those mentioned above (Milligan, 1996; Milligan & Cooper, 1986; Steinley, 2003). Because of the previous support of ARI$_{\mathrm{HA}}$ as an index for cluster validation, Equations 5 to 8 were related to it through descriptive statistics (see Table 4) and simple linear regressions (see Table 5).

The descriptive statistics are calculated by collapsing over all 1,680 conditions. Given these conditions provide a broad range of settings to examine the indices, distribution-like statistics may be calculated. Table 4 indicates a close parallel between the two forms of the adjusted Rand index, which are quite different from the other measures. The small standard deviation and the smaller range of the Rand index makes it the least sensitive of all the measures; the ARI$_{\mathrm{HA}}$ has the largest standard deviation, implying it is more sensitive to different patterns in **T**.

All measures were related to ARI$_{\mathrm{HA}}$ through simple linear regressions (chosen because the information reported in several Monte Carlo studies collapses over several conditions to provide "overall" values of performance). It is shown in Table 5 that a large amount of variance is accounted for in the translation of measures to ARI$_{\mathrm{HA}}$; the least is associated with the unadjusted Rand index and is attributable to attempting to map a smaller range (1.00 − 0.47 = 0.53) to a larger range (1.00 + 0.11 = 1.11). Nonetheless, 66% variance accounted for is enough to give researchers a "rough" feeling for what the corresponding ARI$_{\mathrm{HA}}$ would be. For example, Milligan (1980) reported a Rand index of 0.974 for the single-link method in terms of cluster recovery. Translating 0.974 to the ARI$_{\mathrm{HA}}$ via the formula provided in Table 5 yields a value of 0.80, indicating a clear overestimation on the part of the unadjusted Rand index. Although not formally corrected for chance, the Fowlkes and Mallows (1983) and Jaccard (Downton & Brennan, 1980) indices are more sensitive measures than the unadjusted Rand index. In addition, it appears that the miscalculation by Morey and Agresti (1984) in computing the adjusted Rand index was not too grievous, resulting in a 0.02 average positive bias with respect to the correct calculation in Hubert and Arabie (1985).

Table 4
Hubert and Arabie Adjusted Rand Index (ARI$_{HA}$) as Predicted From Four Other Indices

	Descriptive statistics				
Index	N	M	SD	Minimum	Maximum
Rand	168,000	0.74	0.13	0.47	1.00
Jaccard	168,000	0.37	0.24	0.03	1.00
Fowlkes and Mallows	168,000	0.50	0.24	0.05	1.00
ARI$_{\mathrm{MA}}$	168,000	0.32	0.28	−0.10	1.00
ARI$_{\mathrm{HA}}$	168,000	0.30	0.29	−0.11	1.00

Note. ARI$_{\mathrm{MA}}$ = Morey and Agresti adjusted Rand index.

Table 5
Hubert and Arabie Adjusted Rand Index (ARI_{HA}) as Predicted From Four Other Indices

Response	Predictor	$\hat{\beta}$	Intercept	R^2
ARI_{HA}	Rand	1.82	−1.05	0.66
ARI_{HA}	Jaccard	1.10	−0.11	0.81
ARI_{HA}	Fowlkes and Mallows	1.06	−0.24	0.77
ARI_{HA}	ARI_{MA}	1.03	−0.03	0.99

Note. ARI_{MA} = Morey and Agresti adjusted Rand index.

ARI_{HA} Versus Classification Rates

On preliminary investigation of the descriptive statistics of correct classification rates ($M = 0.58$, $SD = 0.22$, minimum = 0.15, maximum = 1.00), the measure provided in Equation 10 does not seem unreasonable. To test whether the distributions of ARI_{HA} and correct classification rates were the same, the Wilcoxon rank sum test (Hollander & Wolfe, 1999, pp. 108–116) was used, indicating that the null assumption of distributional equality can be rejected at the $p < 10^{-7}$ level. When the distributions of correct classification rates (see Figure 1) and ARI_{HA} (see Figure 2) were examined, marked differences began to emerge. Figure 1 seems to indicate that the distribution of correct classification rates is somewhat uniform. Thus, if partition $\mathcal{P}$ is fixed and $\mathcal{Q}$ is generated randomly, correct classification rates between 30% and 90% are approximately equally likely to occur, indicating the mean of correct classifications is just the average of similarly probable events. Conversely, Figure 2 indicates the distribution of ARI_{HA} is highly skewed and the randomly generated partition would result in a value very near zero—a situation where the mean also denotes high levels of concentration of ARI_{HA}. Thus, if a high value is obtained by ARI_{HA}, it is more likely attributable to systematic properties of the clustering technique, whereas if a high value is obtained for correct classification, it is more likely to be caused by random variation.

Additional information concerning the discrepancies between the two measures can be discovered when examining their average values with respect to the percentage of overlap between the partitions (excluding the conditions containing only two clusters because a simple “flip” of **T** leads to symmetric **T** matrices with respect to percentage of overlap). Figure 3 indicates that the adjusted Rand index obtains values much closer to zero as percentage of partition overlap increases; however, the correct classification rate

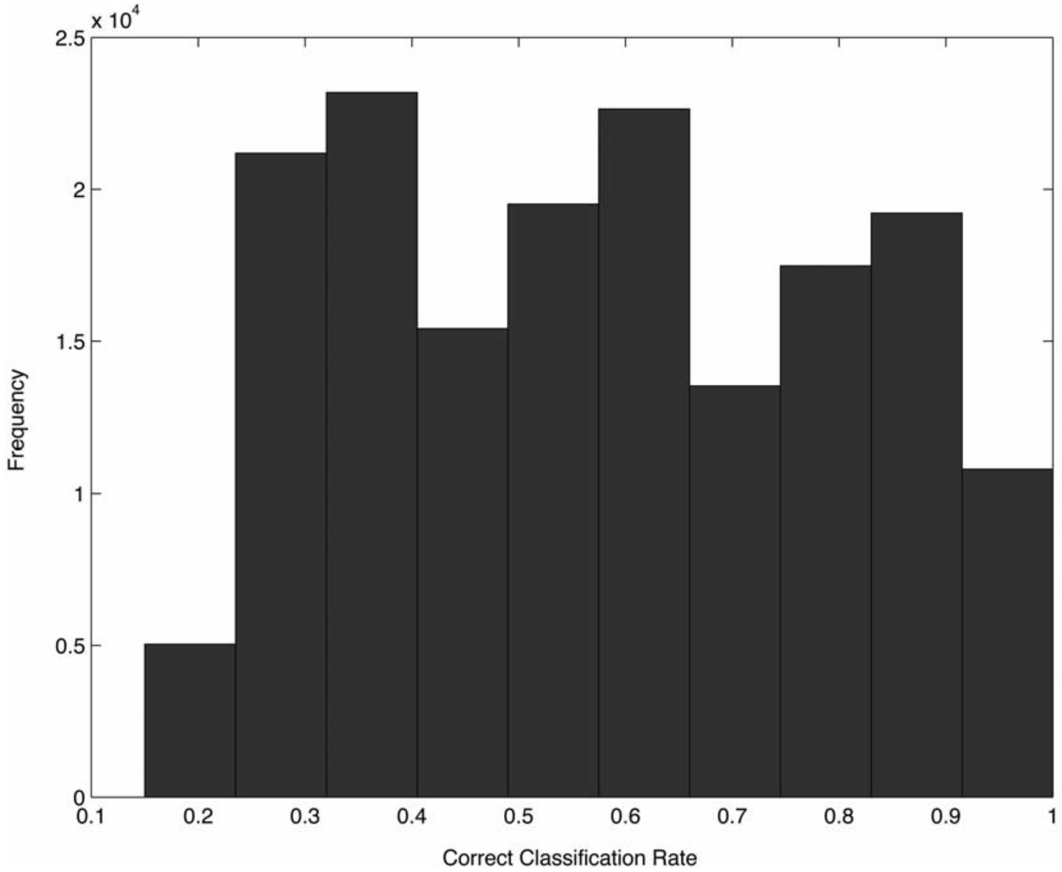

Figure 1. Distribution of correct classification rates.

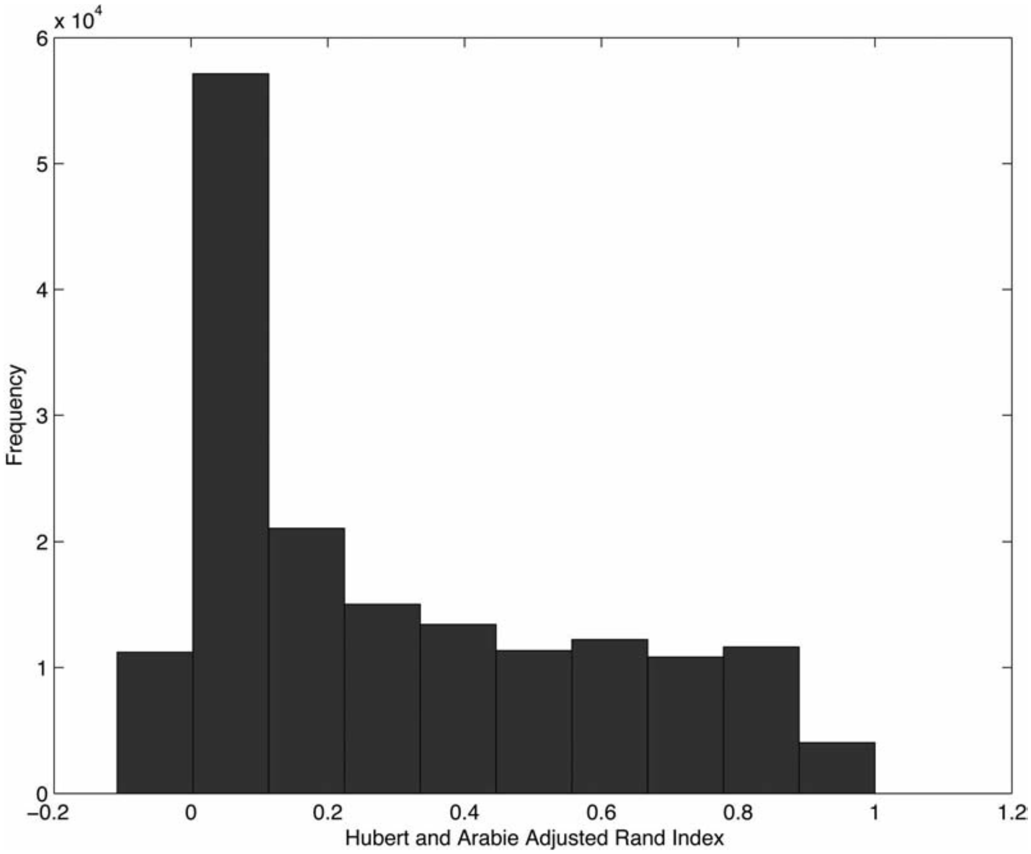

Figure 2. Distribution of Hubert and Arabie adjusted Rand index.

reaches an average lower value around 0.30 at approximately 80% overlap between the partitions. The slight increase at the end of each curve can be easily explained for each situation: (a) For correct classification rates an optimal rule is being used to calculate the rates, resulting in an increase when off-diagonal cells are filled “too full” by chance because of the number of objects allocated in the 0.85–1.00 percentage of overlap conditions; (b) for ARI_{HA}, the slight increase is attributed to the *d* term from Table 2 becoming disproportionately large because of the number of off-diagonal cell allocations in the 0.90–1.00 percentage of overlap conditions. Regardless, the steeper slope exhibited by ARI_{HA} indicates it penalizes more for increased overlap between partitions than correct classification rates, another indication that ARI_{HA} may be preferred over classification rates.

General Properties of the ARI_{HA}

Given that ARI_{HA} has been the index of choice and it has more desirable properties than classification rates, its properties are studied in more detail. Table 6 presents an analysis of variance (ANOVA; main effects only) for the manipulated variables (i.e., distribution of objects, cluster size, number of clusters, and percentage of overlap) in the simulation. Because of the large sample size, all of the effects are statistically significant at very small alpha levels. However, when we compare by effect size, the only practically significant effect is due to percentage of overlap between the two partitions. Table 7 provides means and standard deviations of ARI_{HA} across levels of the different conditions (i.e., the cell means of the ANOVA), indicating many desirable properties of ARI_{HA}.

First, ARI_{HA} seems to be fairly invariant in terms of means and variance across density, the number of objects, and the number of clusters (there was a minor increase in the mean value when one cluster had 60% of all observations), making ARI_{HA} a suitable measure irrespective of these conditions. In addition, ARI_{HA} does not exhibit constant variance across the fourth condition, percentage of overlap. Although this may seem undesirable, it is a beneficial property because it indicates that ARI_{HA} takes into account different patterns of partition overlap, displaying sensitivity not only to the degree of overlap between the partitions but also to the type of overlap present.

If all 168,000 observations of ARI_{HA} are ranked, a simple procedure based on cutpoints can be used to

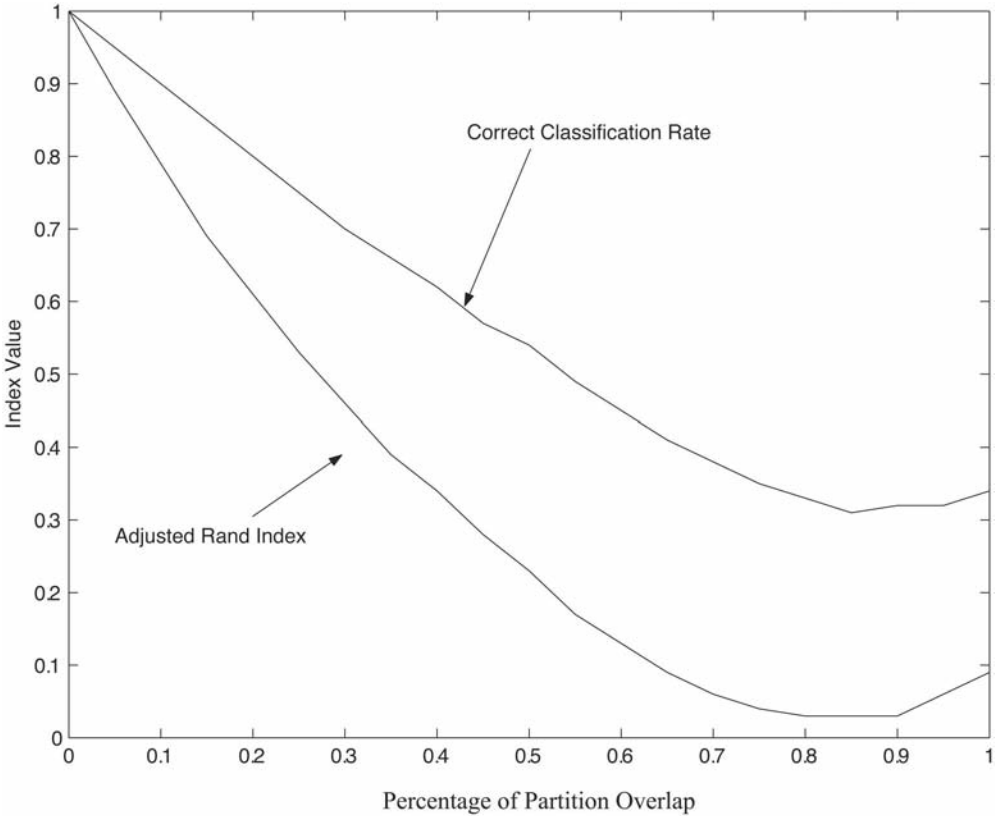

Figure 3. Comparison of classification rates and Hubert and Arabie adjusted Rand index.

obtain general guidelines to give values of ARI_{HA} interpretive credence. If the ranked observations are cut at the 95th, 90th, 85th, and 80th percentiles, the respective values of ARI_{HA} are 0.86, 0.77, 0.67, and 0.60. Thus, when one is validating clustering techniques, a set of heuristics for determining the quality of cluster recovery as related to ARI_{HA} could be (a) values greater than 0.90 can be viewed as excellent recovery, (b) values greater than 0.80 can be considered good recovery, (c) values greater than 0.65 can be considered moderate recovery, and (d) values less than 0.65 reflect poor recovery.

Significance Testing

The above procedure lends itself directly to significance testing of observed values of ARI_{HA}. A Monte Carlo sampling procedure can be used to generate a baseline probability distribution, and probability values can be derived for specific null-hypothesis tests of ARI_{HA} against the baseline model. This procedure is comparable to those discussed by Hubert (1987) and leads to easily interpretable hypothesis tests that control for cluster density, the number of objects, and the number of clusters—creating a test that focuses on

Table 6
Analysis of Variance for the Hubert and Arabie Adjusted Rand Index With Regard to Simulation Variables

Effect	*df*	*SS*	*MSE*	*F*	*p*	η^2
Object distribution	2	390.3	195.1	8,056.8	$< .0001$	.03
Cluster size	4	32.5	8.1	335.2	$< .0001$	.00
No. of clusters	6	105.9	17.6	728.5	$< .0001$	.01
Percentage of overlap	19	9,645.0	507.6	20,957.6	$< .0001$	.68
Error	167,968	4,068.5	0.02			
Total	167,999	14,242.1				

Table 7
Mean Hubert and Arabie Adjusted Rand Index (ARI_{HA}) by Clusters, Number of Objects, Density, and Percentage of Overlap

	ARI_{HA}			ARI_{HA}			ARI_{HA}			ARI_{HA}	
Density	*M*	*SD*	No. of objects	*M*	*SD*	No. of clusters	*M*	*SD*	Percentage of overlap	*M*	*SD*
Equal	.26	.27	50	.29	.29	2	.29	.31	.05	.89	.03
10%	.26	.28	100	.30	.29	3	.25	.25	.10	.79	.05
60%	.36	.31	200	.30	.29	4	.27	.27	.15	.69	.08
			300	.30	.29	5	.29	.29	.20	.61	.10
						6	.31	.30	.25	.53	.11
						7	.32	.30	.30	.46	.13
						8	.32	.31	.35	.39	.14
									.40	.34	.15
									.45	.28	.15
									.50	.23	.14
									.55	.17	.12
									.60	.13	.09
									.65	.09	.07
									.70	.06	.05
									.75	.04	.03
									.80	.03	.02
									.85	.03	.03
									.90	.03	.04
									.95	.06	.06
									1.00	.09	.08

Note. Each column is collapsed across the other three conditions.

partition overlap, the most influential factor of all those examined.

First, the baseline probability distribution must be established. When we are interpreting ARI_{HA}, the null model assumes there is nothing more than random agreement between the two partitions. Thus, we are immediately led to testing the following null hypothesis:

> $H_0(1)$: The observed ARI_{HA} value does not indicate significantly better recovery than can be expected by chance agreement between the two partitions given that the properties of the original data are known (i.e., cluster density, the number of objects, and the number of clusters).

To test $H_0(1)$, we implement the following algorithm.

1. Set $A = 0$, $B = 0$; define ARI^*_{HA} as the observed adjusted Rand index, α as the desired significance level, and B_s as the number of samples to be drawn from the baseline distribution.

2. Draw a random matrix, **T**, from the conditional distribution (conditioned on row marginal totals where all cell entries have a uniform, equal chance of assuming all acceptable values) of the baseline probability distribution.

3. Calculate ARI_{HA} for **T**.

4. If $\text{ARI}_{\text{HA}} \geq \text{ARI}^*_{\text{HA}}$, then $A = A + 1$.

5. If $B < B_s$, then $B = B + 1$, and return to Step 2.

6. If $A/B_s < \alpha$, reject $H_0(1)$.

Using the notation just introduced, we can rewrite the null hypothesis as

$$H_0(1)\text{: } \text{ARI}^*_{\text{HA}} = 0.$$

Because these significant tests are performed in the domain of cluster recovery, it only makes sense to perform one-sided hypothesis testing that the recovered structure is recovered better than chance would have allowed. This allows A/B_s to serve as a probability value that can be interpreted as (a) the probability of observing such a high value for ARI_{HA} if the true nature of the agreement between the partitions is merely chance agreement or (b) the probability of observing such a high value on a matrix, **T**, from a uniform sample of matrices conditioned on cluster density, the number of objects, and size.

For example, consider a situation with 120 objects partitioned into 4 clusters where the "true" number of objects in each cluster is 20, 30, 30, and 40. Imagine using two different clustering techniques, C_1 and C_2, to obtain two clusterings of the data set, which can be compared to the original, "true" structure of the data in two matrices, $\mathbf{T}_1$ and $\mathbf{T}_2$, respectively. Letting the row totals be fixed as the aforementioned distribution (20, 30, 30, 40), $\mathbf{T}_1$ and $\mathbf{T}_2$ could be represented as

$$\mathbf{T_1} = \begin{bmatrix} 15 & 5 & 0 & 0 \\ 10 & 10 & 5 & 5 \\ 0 & 12 & 18 & 0 \\ 1 & 2 & 14 & 23 \end{bmatrix},$$

and

$$\mathbf{T_2} = \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 25 & 0 & 5 \\ 0 & 0 & 25 & 5 \\ 0 & 0 & 1 & 39 \end{bmatrix}.$$

The ARI^*_{HA} for $\mathbf{T}_1$ is $ARI^*_{T1} = .2456$; similarly, for $\mathbf{T}_2$, $ARI^*_{T2} = .7401$. If we set $B_s = 10{,}000$ and $\alpha = .05$ and proceed as described above, the probability values for $\mathbf{T}_1$ and $\mathbf{T}_2$, represented as p_{t1} and p_{t2}, respectively, are $p_{t1} =$ .0001 and $p_{t2} = .0001$ (note that each of the matrices is tested separately). Thus, it is clear that these two matrices are significantly different than one would expect if the only driving force was chance agreement. This result isn't particularly surprising because one would expect even subpar cluster recovery methods to perform better than chance.

However, this formulation leads nicely into a more interesting hypothesis testing situation. Although it is clear that we can test the uninteresting (and almost always significant) case of $H_0(1)$: $ARI^*_{\mathrm{HA}} = 0$, it is not as clear how to test the general hypothesis

> $H_0(2)$: $ARI^*_{\mathrm{HA}} = r$, where r is a constant value greater than zero.

For example, this type of hypothesis would be of interest in testing whether ARI^*_{HA} is different from the heuristic values provided at the end of the previous section. Unfortunately, the difficulty in formulating such a hypothesis and testing it in the manner used above lies in the general combinatorial intractability of enumerating all possible partitions of a set of objects specified by the row totals of the general matrix **T**. However, it is possible to skirt this problem by altering Step 2 of the algorithm—the type of baseline distribution **T** is drawn from. Although it is impossible to fix r to an exact value, r can be estimated using the external information available provided in Table 7—namely, percentage of partition overlap. This changes $H_0(2)$, which can now be represented as

> $H_0(2)^*$: $ARI^*_{\mathrm{HA}} \approx r$.

Thus, Step 2 in the above algorithm can be altered as follows:

> 2*. Draw a random matrix, **T**, from the conditional (conditioned on row marginal totals and percentage of overlap between the two partitions) distribution of the baseline probability distribution (see the Simulating $\mathcal{T}$ section).

For example, if we test the hypotheses $ARI^*_{\mathrm{HA}} \approx .90$, $ARI^*_{\mathrm{HA}} \approx .80$, and $ARI^*_{\mathrm{HA}} \approx .65$, the appropriate amount of percentage of overlap to approximate these values of r would be .05, .10, and .17, respectively. Given $ARI^*_{T1} =$ .2456, it is uninteresting to compare this value to any of the heuristic cutoff values provided; however, because $ARI^*_{T2} =$ .7401, it may be of interest to test whether ARI^*_{T2} is statistically different than 0.80. This can be tested by generating 10,000 separate 4×4 matrices with an overlap of 0.10, yielding a probability value of 0.003—indicating the observed matrix does not fall into the "good" recovery category but is most assuredly in the "moderate" recovery category. Furthermore, it is important to note that this simulated procedure only approximates the heuristic value (in this instance, 0.80 is only approximated). Thus, to find the true value used to test the hypothesis, the mean ARI_{HA} of the simulated matrices is calculated and found to equal 0.78, providing more credence to the "moderate" recovery category.

Conclusion

This article fills several gaps in the literature. First, it provides a template for translating measures previously used in cluster validation research into the common framework of the $\mathrm{ARI}_{\mathrm{HA}}$. Although the variance accounted for is not perfect, rough approximations can be obtained to give an overall sense of the relationship. This is particularly useful in making clustering decisions for complex data sets because several characteristics of varying algorithms were discovered and expressed under the guise of different recovery indices over the span of several years.

This article expands on Steinley's (2003) assertion that classification rates should not be used in cluster validation research for a variety of reasons. When the true structure is known, the $\mathrm{ARI}_{\mathrm{HA}}$ corrects the classification rates for chance. Thus, it removes the bias for chance correct classifications (introduced by the arbitrary nature mentioned above), making it a more precise measure of the true recovery abilities. Second, classification rates and $\mathrm{ARI}_{\mathrm{HA}}$ are not one-to-one functions. The implications of this are that two different $\mathrm{ARI}_{\mathrm{HA}}$ values can have the same classification rates. This is because the $\mathrm{ARI}_{\mathrm{HA}}$ considers the pattern of classified observations. In addition, classification rates may provide seemingly "good" solutions that may be attributable to random chance, and classification rates do not provide an adequate penalty for increases in partition overlap. Along with the fact that $\mathrm{ARI}_{\mathrm{HA}}$ is a steeper, more discriminating function, these properties indicate it is the more informative of the two measures.

In addition to showing $\mathrm{ARI}_{\mathrm{HA}}$'s superior performance, several other "nice" properties were illustrated. $\mathrm{ARI}_{\mathrm{HA}}$ is practically invariant to changes in the number of clusters, objects, and relative cluster size, tending to focus all of its discriminability on the presence and type of partition overlap. The sensitivity for the type of overlap is especially attractive because it indicates $\mathrm{ARI}_{\mathrm{HA}}$ is taking into account several higher level interactions not noticed by other recovery functions, such as classification rates. Furthermore, a set of heuristic values was determined by ordering the 168,000 simulated observations. The values indicate various degrees of cluster recovery. Taking advantage of numerous randomization procedures makes it possible to test the observed values of cluster recovery indices, helping researchers give qualitative values to numbers that were previously uninterpretable. These results suggest further research in extending the nonparametric testing techniques to parametric testing techniques, requiring the derivation of distributional families for $\mathrm{ARI}_{\mathrm{HA}}$.

References

Balakrishnan, P. V., Cooper, M. C., Jacob, V. S., & Lewis, P. A. (1994). A study of the classification capabilities of neural networks using unsupervised learning: A comparison with *K*-means clustering. *Psychometrika, 59,* 509–525.

Beauchaine, T. P., & Beauchaine, R. J., III. (2002). A comparison of maximum covariance and *K*-means cluster analysis in classifying cases into known taxon groups. *Psychological Methods, 7,* 245–261.

Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research, 27,* 417–433.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research, 24,* 147–161.

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for *K*-means clustering. *Psychometrika, 66,* 249–270.

Carmone, F. J., Kara, A., & Maxwell, S. (1999). HINoV: A new model to improve market segment definition by identifying noisy variables. *Journal of Marketing Research, 36,* 501–509.

Cheng, R., & Milligan, G. W. (1996). Measuring the influence of individual data points in cluster analysis. *Journal of Classification, 13,* 315–335.

Donoghue, J. R. (1995). Univariate screening measures for cluster analysis. *Multivariate Behavioral Research, 30,* 385–427.

Downton, M., & Brennan, T. (1980, June). *Comparing classifications: An evaluation of several coefficients of partition agreement.* Paper presented at the meeting of the Classification Society, Boulder, CO.

Dreger, R. M., Fuller, J., & Lemine, R. L. (1988). Clustering seven data sets by means of some or all of seven clustering methods. *Multivariate Behavioral Research, 23,* 203–230.

Dubes, R. C. (1987). How many clusters are best?—An experiment. *Pattern Recognition, 20,* 645–663.

Dubes, R., & Jain, A. K. (1976). Clustering techniques: The user's dilemma. *Pattern Recognition, 8,* 247–260.

Fisher, D. G., & Hoffman, P. (1988). The adjusted Rand statistic: A SAS macro. *Psychometrika, 53,* 417–423.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78,* 553–569.

Gierl, H., & Schwanenberg, S. (1998). A comparison of traditional segmentation methods with segmentation based upon artifical neural networks by means of conjoint data from a Monte Carlo simulation. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 386–392). Berlin, Germany: Springer.

Helsen, K., & Green, P. E. (1991). A computational study of replicated clustering with an application to market segmentation. *Decision Sciences, 22,* 1124–1141.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods.* (2nd ed.). New York: Wiley.

Hubert, L. (1987). *Assignment methods in combinatorial analysis.* New York: Marcel Dekker.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Klein, R. W., & Dubes, R. C. (1989). Experiments in projection and clustering by simulated annealing. *Pattern Recognition, 22,* 213–220.

Krieger, A. M., & Green, P. E. (1999). A generalized Rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification, 16,* 63–89.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45,* 325–342.

Milligan, G. W. (1989a). A study of the beta-flexible clustering method. *Multivariate Behavioral Research, 24,* 163–176.

Milligan, G. W. (1989b). A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification, 6,* 53–71.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 341–375). River Edge, NJ: World Scientific.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50,* 159–179.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21,* 441–458.

Milligan, G. W., & Cooper, M. C. (1987). Methodological review: Clustering methods. *Applied Psychological Measurement, 11,* 329–354.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5,* 181–204.

Milligan, G. W., & Isaac, P. D. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition, 12,* 41–50.

Milligan, G. W., Soon, S. C., & Sokal, L. M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5,* 40–47.

Morey, L., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement, 44,* 33–37.

Morey, L. C., Blashfield, R. K., & Skinner, H. A. (1983). A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research, 18,* 309–329.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66,* 846–850.

Saltstone, R., & Stange, K. (1996). A computer program to calculate Hubert and Arabie's adjusted Rand index. *Journal of Classification, 13,* 169–172.

Scheiber, D., & Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms: A comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research, 20,* 283–304.

Steinley, D. (2003). Local optima in *K*-means clustering: What you don't know may hurt you. *Psychological Methods, 8,* 294–304.

Vichi, M., & Kiers, H. A. L. (2001). Factorial *K*-means analysis for two-way data. *Computational Statistics and Data Analysis, 37,* 49–64.

Waller, N. G., Kaiser, H. A., Illian, J. B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika, 63,* 5–22.

Waller, N., Underhill, J., & Kaiser, H. (1999). A method for generating simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation. *Multivariate Behavioral Research, 34,* 123–142.

Received October 9, 2003
Accepted February 3, 2004 ■

New Editor Appointed for *Journal of Occupational Health Psychology*

The American Psychological Association announces the appointment of Lois E. Tetrick, PhD, as editor of *Journal of Occupational Health Psychology* for a 5-year term (2006–2010).

As of January 1, 2005, manuscripts should be submitted electronically via the journal's Manuscript Submission Portal (www.apa.org/journals/ocp.html). Authors who are unable to do so should correspond with the editor's office about alternatives:

Lois E. Tetrick, PhD
Incoming Editor, *JOHP*
George Mason University
Department of Psychology, MSN, 3F5
4400 University Drive, Fairfax, VA 22030

Manuscript submission patterns make the precise date of completion of the 2005 volume uncertain. The current editor, Julian Barling, PhD, will receive and consider manuscripts through December 31, 2004. Should the 2005 volume be completed before that date, manuscripts will be redirected to the new editor for consideration in the 2006 volume.