PSYCHOPHYSIOLOGY
Vol. 27, No. 5

Printed in U.S.A.

Methodology

An Approach to Artifact Identification: Application to Heart Period Data

GARY G. BERNTSON, KAREN S. QUIGLEY, JAYE F. JANG, AND SARAH T. BOYSEN
Ohio State University and Yerkes Regional Primate Research Center

ABSTRACT

A rational strategy for the automated detection of artifacts in heart period data is outlined and evaluated. The specific implementation of this approach for heart period data is based on the distribution characteristics of successive heart period differences. Because beat-to-beat differences generated by artifacts are large, relative to normal heart period variability, extreme differences between successive heart periods serve to identify potential artifacts. Critical to this approach are: 1) the derivation of the artifact criterion from the distribution of beat differences of the individual subject, and 2) the use of percentile-based distribution indexes, which are less sensitive to corruption by the presence of artifactual values than are least-squares estimates. The artifact algorithms were able to effectively identify artifactual beats embedded in heart period records, flagging each of the 1494 simulated and actual artifacts in data sets derived from both humans and chimpanzees. At the same time, the artifact algorithms yielded a false alarm rate of less than 0.3%. Although the present implementation was restricted to heart period data, the outlined approach to artifact detection may also be applicable to other biological signals.

DESCRIPTORS: Heart period, Artifacts, Artifact detection, Heart period artifacts.

Artifacts in recordings of bioelectric signals constitute an inevitable plague on psychophysiological studies. Clearly, the most effective approach to the management of artifacts is prevention. Post hoc solutions may require rejection of important segments of data, or may employ correction procedures that yield only approximations to the veridical values. Even with the most rigorous of recording techniques, however, some degree of artifact is typically encountered in psychophysiological recordings, and some means of artifact processing is generally necessary. The most critical aspect of this processing is artifact identification, because undetected artifacts can often compromise psychophysiological data far more than even the crudest means of artifact correction.

Artifact detection may be straightforward if the source of the artifact is known and can be independently monitored. In some cases, however, the psychophysiological data itself must be used for artifact detection. Fortunately, many types of artifacts have characteristics that differ from the target biological signals within which they are embedded, and such differences provide a potential means of artifact identification. Indeed, these differences may render many artifacts apparent on visual examination, and visual screening is often employed for artifact identification. This approach, however, may be difficult to quantify, can be time-consuming, may vary with the skills of the examiner, and can present significant vigilance problems for large data sets. Consequently, some researchers have adopted machine-based approaches to automatically flag potential artifacts that exceed a criterion level on some dimension (Berntson & Boysen, 1989; Cheung, 1981; Linden & Estrin, 1988). Al-

The present work was supported in part by NIH grant RR-0165 from the Division of Research Resources to the Yerkes Regional Primate Research Center. The Yerkes Center is fully accredited by the American Association for Laboratory Animal Care. We thank S. Cassini and S. Shell for technical assistance. We also thank Dr. J. T. Cacioppo for his invaluable comments.

Address requests for reprints to: G. G. Berntson, Ohio State University, 1885 Neil Avenue, Columbus, OH 43210.

though this approach is potentially preferable to visual scanning, its ultimate value depends on: 1) the selection of a rational criterion (see Rompelman, 1986), and 2) a quantitative evaluation of its performance, relative to the rate of hits, misses, and false alarms.

The present paper addresses these issues in the context of heart period data. Although algorithms for handling artifacts in different biological signals will be necessarily disparate, common problems faced by such algorithms suggest some general strategies for artifact detection that may transcend specific instances. Automated artifact detection would be possible if a critical dimension could be identified, along which artifacts and veridical signals differ. A questionable signal could then be compared against a template of the target signal or dimension. If the questionable signal deviated from the template, it could be flagged as a potential artifact. Even if the distributions of artifacts and veridical signals were partially overlapping, the target dimension would still allow the derivation of an expected probability of an artifact, which could serve to flag suspect values for further evaluation. Clearly, the initial phase in such an effort would be the identification of the features of biological signals that share the least overlap with artifacts.

A second phase would entail the derivation of a model or a set of descriptors that characterize artifacts and veridical signals on the criterion dimension. The establishment of such a model is complicated by the intrinsic variability of biological signals, and the descriptors will generally reflect a stochastic, rather than an absolute model of the biological signal. Hence, distribution indexes may be critical features of the model, likely including estimates of central tendency and variability (or higher order moments, Cacioppo & Dorfman, 1987). For a normal distribution of signals, the mean and variance may provide optimal indexes. Unfortunately, distributions of biological signals are often non-normal (e.g., Jennings, Stringfellow, & Graham, 1974), and the mean and variance may not provide an adequate characterization of the distribution. Moreover, because between-subject variability in psychophysiological systems is often considerably larger than within-subject variability, indexes derived from the broader population would likely overestimate the variance of signals for a given subject. Consequently, the performance of artifact algorithms may be superior if the signal template is based on distribution indexes derived from the individual subject. This imposes a further limitation on mean and variance measures. Given that they may be derived from the subject's own experimental data, these measures could be seriously corrupted by the very artifacts one wishes to identify.

In view of these considerations, alternative distribution indexes may be desirable. Percentile-based estimates of central tendency and variability, such as the median and interquartile range, are recognized to be less sensitive to deviations from the normal, and to extreme values introduced by artifacts. For non-normal distributions or artifact-laden records, percentile-based measures may provide more viable distribution indexes than least squares procedures.

The final steps in the development of an artifact-detection algorithm entail a determination of the optimal criterion placement along the target dimension, and an evaluation of its performance. The criterion placement is not a trivial consideration, because the expected values of different types of erroneous classifications may be widely divergent. The cost of a missed artifact is often considerably greater than that of a false alarm. A number of statistical and theoretical models, some perhaps employing simulated artifacts, could be used to derive the expected operating characteristics of an algorithm. Although such approaches are often desirable, they may not fully capture the diversity of actual artifacts occurring within biological signals. Consequently, evaluation of a potential artifact algorithm should probably include actual psychophysiological data, representative of those to which it will ultimately be applied. If the performance characteristics of the algorithm were expressed in terms of a signal detection/payoff matrix model, the relative discriminability of artifacts and veridical signals, as well as the consequences of varied criterion placements, could be quantitatively specified. In the present paper, we follow this general plan of development and evaluation of an algorithm for the detection of artifacts in heart period records.

Heart Period Data Bases

Subjects

Development and evaluation of the present algorithms was based on a data set of heart period records from 60 subjects, obtained during performance on cognitive tasks in two separate studies. All subjects were male undergraduate students (18–42 yrs), with no reported history of cardiovascular disorders. The present algorithms were also applied to heart period data of 6 infant chimpanzees (2–5 mo), derived from a study of the ontogeny of vocal perception (Berntson, Boysen, Bauer, & Torello, 1989).

ECG Data

The ECG was recorded using silver/silver chloride electrodes secured at thoracic monitor sites and connected to an Amerec ERM 101 cardiotachom-

eter. The raw ECG signal was recorded on a Grass Model 7 polygraph (60 mm/s), and the pulse output of the cardiotachometer was coupled to a microcomputer interface for on-line determination of heart periods (±1 ms).

The human heart period data were derived from two separate studies of cognitive performance, under typical laboratory conditions (Ns = 20 & 40 for Experiments 1 & 2, and baseline heart periods = 810.2 ms ± 69.4 SD and 888.5 ms ± 67.1 SD, respectively). To equate each subject's contribution to the overall data set, and avoid biases related to individual differences in baseline heart periods, analyses were based on the first 256 beats from each subject (yielding a total of 15,360 beats).

In addition to the raw heart period data, heart period records were edited to provide a corresponding artifact-free data set. We employed a highly conservative approach in this editing to ensure that the resulting data were accurate and artifact-free. First, the raw polygraph recordings were closely examined for the presence of electrical or movement-related noise, which could yield spurious R-wave detections. In addition, beat-by-beat heart periods were graphically displayed, and deviant values were checked and corrected, if necessary, by measurements from the polygraph record. As a final screening, potential artifacts in the heart period records were flagged off-line by the computer system, as beats that deviated from either of the surrounding periods by more than 30%. Again, all suspect values were verified or corrected. This redundant combination of manual and automated screening served to increase our confidence in the integrity of the resulting edited data. These methods identified a total of 33 (0.2%) artifacts across all subjects (4 due to failures to detect actual beats, and 29 due to spurious triggering on an ECG artifact).

In addition to the raw and the edited heart period data, one further transformation of this data set was employed, which entailed the systematic introduction of simulated artifacts. This served to increase the total number of artifacts in the records, and ensure that a balanced sampling of potential artifacts could be examined. For each subject, 10 simulated missed beats and 10 simulated artifactual beat detections were introduced into the edited data. Missed beats were simulated by summing adjacent heart periods, and extra beats were introduced by splitting given heart periods into two spurious interbeat intervals. The temporal position of each of these extra beats within the actual heart period was randomly determined to yield a ratio of the two resulting beats ranging from 1:1 to 1:9. The location of the artifacts within the serial stream of heart periods was determined by a constrained random procedure which, for the primary analysis, precluded immediately adjacent artifacts[1]. Overall, this approach yielded a total of 1200 simulated artifacts, which corrupted approximately 7.8% of the heart periods. We also examined an equivalent set of simulated artifacts, in which 50% of the artifactual values immediately followed another artifact.

Performance of the artifact algorithms was also evaluated by application to heart period data of infant chimpanzees. These data were included to broaden the range of species and experimental test conditions. The data were derived from a study of reactive heart rate changes to discrete acoustic stimuli. Evoked cardiac responses to the stimuli entailed both acceleratory and deceleratory changes, and notable somatic movements were elicited in some cases. Baseline heart periods of the chimpanzees were considerably shorter than those of the human subjects (mean heart period = 353.5 ms ± 37.8 SD). ECG and heart period data were processed as outlined above, except that simulated artifacts were not employed. In view of the smaller number of animal subjects tested, the present evaluation was based on all available data (17,361 heart periods over the 6 animals). The editing process outlined above identified 261 artifacts (1.5%) in the raw data files.

Artifact Identification

The initial step in the development of an artifact-detection algorithm is the identification of a critical dimension along which artifacts and veridical signals differ. Given that the cardiac beat is a quantal event, only two types of artifacts arise in heart period data. The monitoring system may either fail to detect an actual beat, or may spuriously report a nonexistent beat. A QRS complex that is not detected will result in a measured interbeat interval that is comprised of two actual heart periods, and the resulting spurious period will be exactly twice the mean of the two actual beats (or more if consecutive beats fail to be detected). In contrast, if an artifact results in the false detection of an R-wave within two actual QRS complexes, at

[1]These constraints for the simulated artifacts were conservative. One of the spurious beats resulting from the division of a normal heart period becomes progressively shorter, and thus more easy to detect, as the ratio of the resulting beats deviates from 1:1. The constraint against two adjacent simulated artifacts was to permit beat-to-beat deviations resulting from artifacts to be tested against the normal heart period variability of the surrounding beats. Sequentially contiguous artifacts were present in the unedited data files, and were also tested in an adaptation of the simulated artifact files, which permitted explicit evaluation of various sequential combinations of artifacts.

Table 1

Effects of artifacts on estimates of central tendency, variance, and criterion indexes of successive heart period differences

Experiment	Mean	SD	Median	Quart Dev[a]	MAD[b]	MED[c]
Edited Data						
1 (N=20)	0.0	56.9	0.5	35.2	222.3	116.9
2 (N=40)	0.0	57.4	0.6	35.8	253.4	118.9
Unedited Data						
1 (N=20)	0.0	65.7	0.1	35.7	221.7	118.5
2 (N=40)	0.0	60.0	0.6	35.8	253.6	118.9
Simulated Artifacts						
1 (N=20)	0.0	280.2	−0.2	46.2	214.4	153.4
2 (N=40)	0.0	308.3	0.1	46.3	245.6	153.7

Note.—Data shown are the average values (in ms) for standard and percentile-based indexes of central tendency and variability of beat-to-beat heart period differences.

[a]Quartile Deviation = Interquartile range/2.

[b]Minimal expected beat difference associated with an artifact.

[c]Maximal expected beat difference between veridical beats.

least one of the two resulting spurious interbeat intervals would be equal to or less than one-half of the actual heart period value. Because these artifact-related changes in period are large relative to normal beat-to-beat heart period variability (Table 1), this may serve as a criterion dimension for artifact identification. The remaining problem is to derive a sufficiently accurate, and statistically defensible expected value for the target interbeat interval, against which the measured value could be compared. The subject's average heart period provides one such estimate. Heart periods, however, are often subject to notable tonic and phasic trends associated with drifting baselines, respiratory sinus arrhythmia, or event-related phasic heart period responses. Consequently, the averaged interbeat interval over a long epoch may be a poor predictor of an individual beat within that epoch.

Prediction of a given heart period could be improved by a knowledge of the interbeat intervals occurring close in time to the target beat. If a temporal epoch were short relative to the time course of tonic and phasic baseline changes, the mean heart period within this local epoch would provide an estimate of the individual beats that was less confounded by baseline changes. In this case, the predicted value would be limited primarily by beat-to-beat heart period variability, with minimal influence of slow baseline drifts or more rapid phasic responses. Indeed, physiological constraints on the rate of change of heart period may render the immediately prior and subsequent beats the best predictors of an intervening target beat. Based on this model, an index of the error of prediction could be obtained from the average difference between successive heart periods. Comparing beat-difference measures of heart period variability to conventional measures of variance, Heslegrave, Ogilvie, and Furedy (1979) reported that a beat-by-beat difference measure was less contaminated by linear trends in heart period data, and yielded the lowest overall variance.

The superior predictive ability of local heart periods is documented by analysis of the edited (artifact-free) data from the 60 human subjects of the present study. For each subject, the error of estimate for individual heart periods was obtained as the average absolute deviation of each beat from the mean for that subject, from the median, and from the immediately surrounding interbeat intervals[2]. Results indicated that the immediately surrounding interbeat intervals (IBIs) were a significantly better predictor of individual heart periods than was the overall mean or median (average error of prediction from surrounding IBIs = 31.5 (±0.33) ms; from the mean IBI = 54.1 (±0.38) ms; and from the median IBI = 53.6 (±0.37) ms; paired t's$(59)>7.0$, $p<.001$).

The surrounding beats may therefore offer the most viable expected value for a target beat, and the disparity between the actual and predicted beats

[2]A least squares procedure may have some statistical advantages over an absolute difference method. However, the effects of extreme heart period values generated by artifacts tend to dramatically corrupt the mean and least squares estimates of variance, rendering medians and absolute differences more appropriate in the present context.

could be expressed as the average difference score between successive heart periods. Consequently, a measure of the normal beat-to-beat differences, together with an estimate of the variance of this measure, may offer a means of detecting the presence of artifactual beats within a serial stream of heart period data. This approach would be viable to the extent to which the distribution of differences associated with veridical beats and that associated with artifacts are non-overlapping. Fortunately, beat differences associated with artifacts are large relative to normal heart period differences. Thus, average beat-to-beat heart period differences of the present 60 human subjects (edited data) had an overall mean of 0.0, and a standard deviation of 57.2. Given this standard deviation, and assuming a normal distribution of differences, over 99 percent of successive beat differences would be less than 133 ms. In contrast, expected differences resulting from artifacts are considerably larger. For a typical 800-ms period containing an artifactual beat, at least one of the resulting spurious beats would be 400 ms or shorter. This yields an expected beat-to-beat difference (from surrounding normal beats) of 400 ms or greater. Similarly, a missed beat would yield a spurious heart period of about 1600 ms, which would differ from normal surrounding beats by an average of 800 ms.

In summary, artifacts in ECG records yield highly deviant heart periods, and resulting beat-to-beat heart period differences are large relative to normal heart period variability. Beat-to-beat differences in heart period thus may provide a critical dimension for artifact identification.

Artifact-Detection Criteria

ECG artifacts could be reliably detected by obtaining accurate estimates of the largest expected beat difference among normal beats, and the smallest expected difference associated with an artifact, and determining whether the former is smaller than the latter. Obtaining an accurate estimate of normal beat-to-beat differences, however, can be problematic. In the absence of artifacts, arrhythmias, or linear trends in heart periods over the test interval, the distribution of beat-to-beat differences is approximately normal, with a mean of zero. Figure 1 illustrates this distribution for the 60 human subjects. Although slightly leptokurtic (kurtosis = 4.19), the distribution approximates the normal. Under these conditions, the mean and standard deviation represent viable distribution indexes. Unfortunately, linear trends in the heart period data, or exaggerated beat differences associated with artifacts, can render the distribution of differences nonnormal, and can grossly distort the variance and

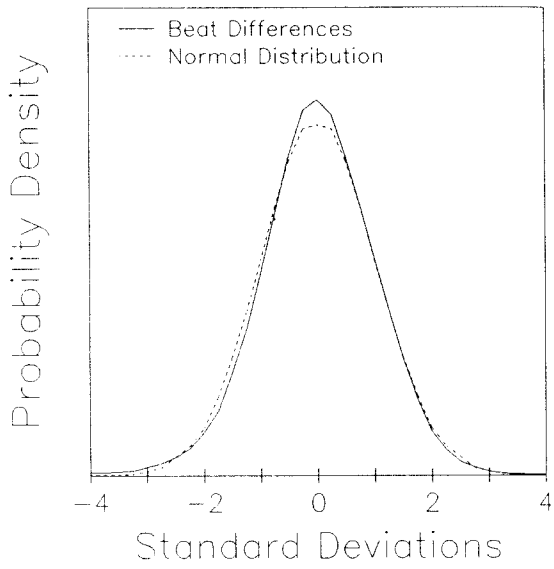

Figure 1. Distribution of beat differences over all subjects. To normalize differences in overall heart period variability, data for each subject were converted to *z*-score values prior to averaging. A corresponding normal distribution is shown for reference.

standard deviation. Consequently, alternative distribution indexes, which are less sensitive to the extreme values generated by artifacts, may be preferable. Candidate measures include the median, and a percentile-based index of variability, such as the interquartile range (IQ) or the quartile deviation (QD = IQ/2). Percentile-based distribution measures are characteristically less sensitive to extreme scores, and thus may provide superior estimates of the distribution of normal beat differences within artifact-laden records.

Although the mean beat-to-beat difference is minimally affected by the presence of artifacts, artifacts can grossly distort the standard deviation of heart period differences (Table 1). In contrast to marked effect on the standard deviation, however, a moderate level of artifacts yields minimal distortion of quartile-based estimates of variance (Table 1). Given that it is minimally corrupted by the presence of artifacts, the quartile deviation (QD) may serve as a viable index of variance of normal heart period differences, even within artifact-laden records.

The following identity holds for a normal distribution:

$$\text{Standard Deviation} = \frac{\text{Quartile Deviation}}{0.675} = 1.48 \cdot \text{Quartile Deviation}$$

where 0.675 is the *z*-score associated with the quartile. Thus, an estimate of the standard deviation

can be derived from the QD, and this quartile-based estimate agrees well with the standard deviation of beat differences for the artifact-free data set from the 60 human subjects (QD-based estimate of SD = 52.7, actual SD = 57.2). Unlike the standard deviation, which is seriously distorted by the presence of artifacts, this quartile-based estimate continues to provide a reasonable predictor of the standard deviation of normal beats in the artifact-laden data set (QD-based estimate = 68.5, SD = 298.9). Inasmuch as percentile-based indexes are considerably less sensitive to artifacts than are those derived from least-squares procedures, they offer more viable estimates of the distribution of the non-artifactual beats in heart period records.

For a normal distribution of beat differences, values within the range $\pm 3.32 \cdot QD$ (2.24 SD) would encompass approximately 97.5% of all beat-to-beat differences. Because non-artifactual heart period differences are approximately normally distributed, this value, the Maximum Expected Difference (MED) for non-artifactual beats, could serve as one index for artifact identification:

$$MED = 3.32 \cdot QD$$

The MED index is conservatively biased toward the detection of artifacts, and would flag 2.5% of veridical beats in a normal distribution of artifact-free heart periods (it flagged 3.8% of the beats in the present edited data set).

An additional index of the minimum expected difference score associated with artifacts would also be desirable. The difference scores associated with missed beats are considerably larger than those related to spurious extra beats. The artifactually long period generated by a missed beat is equal to the sum of the two constituent beats, and yields an expected beat difference (from normal surrounding beats) equal to one heart period. In contrast, a spurious extra beat yields two resultant periods, and the consequent beat differences may be only one third of the normal interbeat interval (see Figure 2). Thus, it is the spuriously detected extra beats that determine the lower limit for a minimal expected artifact-related difference score. Two beat-difference comparisons are relevant for each of the two resulting artifactual periods, these are between: 1) the preceding beat and the artifactual beat, and 2) the artifactual beat and the subsequent beat. Inasmuch as an extreme value obtained on either of these two comparisons could be used to identify the artifact, the largest of the two resulting differences is most relevant for artifact detection. Given this consideration, it can be shown (Figure 2) that the smallest beat difference generated by a spuriously detected R-wave complex would arise when the artifact divided the actual beat by a 1:2 ratio (i.e., fell at either one-third or two-thirds of the interbeat interval)[3]. In view of these considerations, an estimate of the Minimal Artifact Difference (MAD) would be:

$$MAD = \text{Shortest expected veridical beat}/3$$

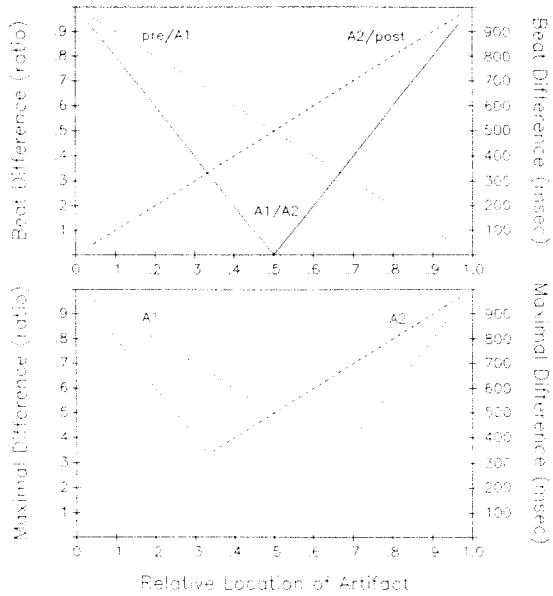

Figure 2. Expected beat-to-beat heart period differences generated by spuriously detected R-waves, as a function of the location of the artifact within the heart period. For illustration, beat differences are expressed as a ratio of the original heart period (left axis), and as a change (ms) from a standard 1000-ms period (right axis). The top graph illustrates the heart period differences associated with each of the three relevant beat difference comparisons resulting from the spurious detection of an R-wave. Two beat differences are associated with each of the two resultant spurious periods, and an extreme value on either of these comparisons would serve to flag the target beat. The largest of the two resultant difference scores, for each artifactual period, is plotted on the bottom graph. (A1 and A2 indicate the first and second resultant artifactual beats. PRE and POST designate the prior and subsequent normal beats).

[3]Actually, three relevant differences emerge, between: 1) the prior beat and the first artifactual beat, 2) the first artifactual beat and the second artifactual beat, and 3) the second artifactual beat and the subsequent beat. Given that an extreme difference on any of these three comparisons would serve to identify the artifact *complex*, a 50% division of the heart period could also be used as the basis for an artifact criterion. The 1:2 split represents a more conservative basis for a criterion, however, because it would serve to flag each of the resulting artifactual beats.

where SEB is the shortest expected veridical beat, estimated as follows:

$$SEB = \text{Median Beat} - 2.9 \cdot \text{Quartile Deviation}$$

For a normal distribution, SEB would be smaller than 97.5% of all heart periods. Thus:

$$MAD = (\text{Median Beat} - 2.9 \cdot QD)/3$$

Although the MAD index nominally covers only 97.5% of artifact-related beat differences, it constitutes a rather conservative index for two reasons. First, the QD is slightly inflated by the presence of artifacts (see Table 1), and thus would tend to underestimate the value of the shortest beat. Secondly, only a 1:2 split of a heart period yields a difference value as small as MAD, any other artifact location would result in a greater beat difference. Thus, there remains only a vanishingly small probability of an artifact generating a beat-difference less than MAD.

As long as the maximum expected difference score for veridical beats (MED) is smaller than the minimum expected difference score associated with an artifact (MAD), these two indexes should serve to differentiate artifacts from veridical beats. A reasonable placement of a criterion difference score might be midway between MED and MAD. The criterion difference score would then be:

$$\text{Criterion Beat Difference} = (MAD + MED)/2$$

Evaluation of the Criterion Dimension

Figure 3 illustrates the overall distribution of beat differences, relative to the criterion score, for veridical beats and for beat differences associated with artifact complexes in the present human data sets. As is apparent, the distributions are well separated, although slight overlap occurred in the tail regions. These data support the viability of the beat-difference measure in differentiating artifacts from veridical signals[4]. Moreover, the criterion difference score, described above, offers a rational means for selecting a criterion placement along this dimension. This criterion difference score proved highly effective in identifying artifactual heart periods in the present human data sets, with each of the 1200 simulated artifacts and the 33 actual artifacts yielding beat differences that exceeded the criterion. The criterion difference score was highly conservative, however, and although it identified every artifact complex in all records, it also yielded a low rate of false alarms. A total of 0.73% of the veridical beat differences exceeded criterion in the simulated artifact data, 0.95% in the unedited data, and 0.94% in the edited data. These false alarm rates were similar for the two experiments that comprised the present data sets (1.03% overall for Experiment 1, and 0.79% for Experiment 2).

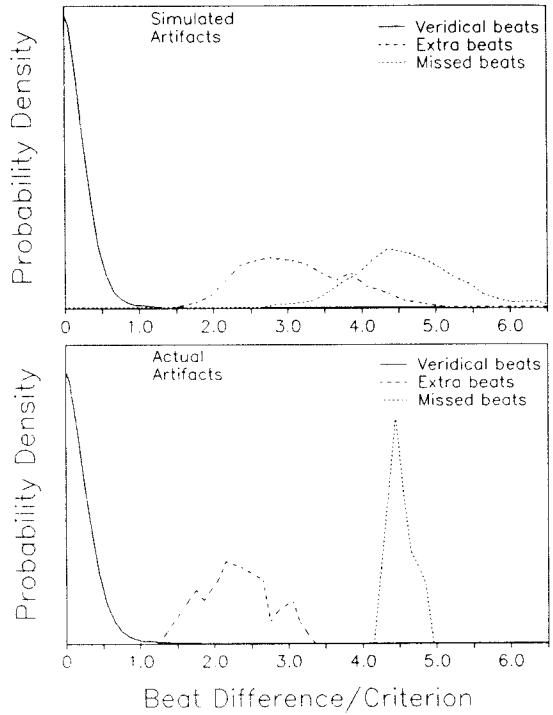

Figure 3. Distributions of absolute successive beat differences associated with veridical heart periods and artifacts. Data are expressed as a ratio of the beat difference to the criterion difference score. For artifactual beats, the graphs illustrate the largest difference scores generated by the spurious beats.

As discussed above, the MAD value (minimum artifact deviation) should ideally be larger than the MED index (maximum expected deviation of veridical beats) to maximize discrimination. This was the case for each of the 60 subjects' edited data. Inflation of the MED index by artifacts, however, resulted in the reversal of these values in one subject's unedited data, and in 4 of the 60 simulated

[4]An *a priori* basis for the selection of a criterion dimension was available in the present effort. For other psychophysiological signals, however, such may not be the case. In these instances signal detection analyses may be helpful in identifying candidate dimensions with high discriminability. The derivation of the parametric index d' assumes that the differences associated with veridical beats and artifacts are normally distributed, and have equal variance. However, this may not be the case, so a more appropriate estimate of discriminability might be the nonparametric measure A' (Norman, 1964; Craig, 1979). This analysis requires at least some degree of overlap between artifacts and signals (Caldeira, 1980). In the absence of such overlap, however, analysis would be unnecessary.

Table 2

Performance of the criterion index as a function of MAD-MED differences

MAD-MED Difference	Number of Subjects	p(HIT)	p(FA)
Edited Data			
<0	0	–	–
0–49	5	–	.051
50–99	15	–	.010
100–149	21	–	.004
>150	19	–	.003
Overall	60	–	.009
Unedited Data			
<0	1	1.0	.094
0–49	4	1.0	.037
50–99	15	1.0	.011
100–149	22	1.0	.004
>150	18	1.0	.003
Overall	60	1.0	.009
Simulated Artifacts			
<0	4	1.0	.042
0–49	15	1.0	.008
50–99	17	1.0	.005
100–149	18	1.0	.003
>150	6	1.0	.003
Overall	60	1.0	.007

artifact files. These cases raise a question concerning the appropriate placement of the criterion, relative to the MED and MAD indexes. Even in instances where MED was larger than MAD, however, the midpoint criterion detected all artifacts, although at the expense of an increase in false alarms (Table 2). The rate of false alarms was thus dependent on the MAD-MED difference for that subject (see Table 2), with small MAD-MED differences, or a MED value larger than MAD, yielding much higher false alarm rates.

In view of the conservative nature of the artifact criterion, a reduced incidence of false alarms could be achieved by lowering the criterion. An additional consideration in setting the criterion level, however, is the payoff matrix related to the expected value of false alarms vs. artifacts (Green & Swets, 1966). The cost of a missed artifact is often substantially greater than that of a false alarm. This is especially true in the present application, where false alarms could be eliminated by post-hoc evaluation of flagged beats. These considerations dictate a downward shift of the criterion. To further examine the performance of the algorithm at various criterion placements, the simulated artifact data were reanalyzed at different criterion positions within the MAD to MED interval (at 0% (MAD), 25%, 50%, 75%, and 100% (MED)).

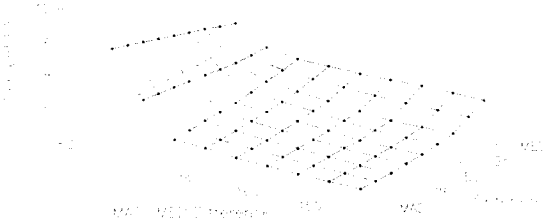

Figure 4. False alarm surface illustrating the false alarm rate as a function of the criterion setting (from MAD to MED), and the MAD-MED difference.

All artifacts were detected for all subjects at every criterion placement, although false alarm rates varied. As expected, the highest false alarm rates were seen with the lowest criterion placement, and for subjects with MAD indexes that were lower than MED values. Figure 4 illustrates the relationships among false alarm rate, MAD-MED differences, and criterion placement. Although the false alarm surface in this figure indicates that a variable criterion setting based on the MAD-MED difference could further minimize false alarms, the 50% criterion value provides a reasonably low false alarm rate at all MAD-MED differences.

The sensitivity of the criterion (50% of the MAD-MED interval) to successive artifacts was confirmed by application of the algorithm to an additional simulated artifact set in which half of the artifacts appeared on adjacent beats. Performance was comparable to that described above, and each artifact complex was identified (at a cost of 0.8% false alarms).

To provide an additional test of performance, the beat-difference criterion (50% of the MAD-MED interval) was also applied to a heart period data set derived from infant chimpanzees (17,361 heart periods from 6 animals). Based on the editing process described above, the raw records were known to contain 52 isolated artifacts (24 long periods due to undetected R-waves, and 28 short periods related to spurious R-wave detections). In addition to these isolated artifacts, the raw records also contained 75 artifact complexes, comprised of various combinations of two or more sequential artifacts (yielding a total of 209 artifactual periods). Collectively, these artifacts corrupted approximately 1.5% of the heart periods in the raw records. As with the human data, the beat-difference criterion was highly effective in identifying artifacts, flagging each of the 261 artifactual values. Some false alarms

were again evident, with 51 veridical heart periods (0.29%) exceeding criterion in the unedited data, and 54 (0.31%) emerging from the edited data files.

Given that slight overlap can exist in the tail-regions of the beat-difference distributions associated with artifacts and with veridical beats (Figure 2), some false alarms may be inevitable. This is especially true if a conservative criterion is employed in order to avoid missed detections of actual artifacts. Post-identification processing of flagged beats could reveal false alarms, whereas an actual artifact that is not detected could substantially distort the final data. Given the low rate of false alarms emerging from the criterion in the present data sets, only 3–10 beats out of 1000 would require further evaluation.

Identification of False Alarms

The artifact criterion yielded a low rate of false alarms, and these erroneously flagged beats could be so identified by direct measurements from polygraph records or digitized ECG data. An automated alternative could further reduce the required manual effort. Such an approach is feasible by virtue of an important property of artifacts. Artifacts can often be corrected, either precisely or to a close approximation, by rather simple procedures. Thus, a spuriously detected R-wave results in two artifactual beats, with no loss of information. If these resultant spurious beats are added together, the true heart period will necessarily be restored. Moreover, the restored heart period should yield beat differences that fall within the criterion outlined above. Similarly, if an R-wave was not detected, the resulting spurious heart period would be the sum of the periods of the constituent beats. Although, in this case, information is lost on the precise temporal location of the missed beat, dividing the artifactual beat in half would provide a close approximation to the original values. The resulting heart period differences should pass the beat-difference criterion. In contrast, false alarms do not generally evidence this property. Indeed, of the 143 false alarms flagged by the artifact-criterion in the edited human data-sets, none could be so "corrected," and this difference between artifacts and false alarms may serve as the basis for post hoc identification of false alarms. This distinction, however, does not necessarily hold for false alarms surrounded by artifacts.

Implementation and False Alarm Detection

The present algorithms are designed for off-line processing, with the entire sample of heart periods

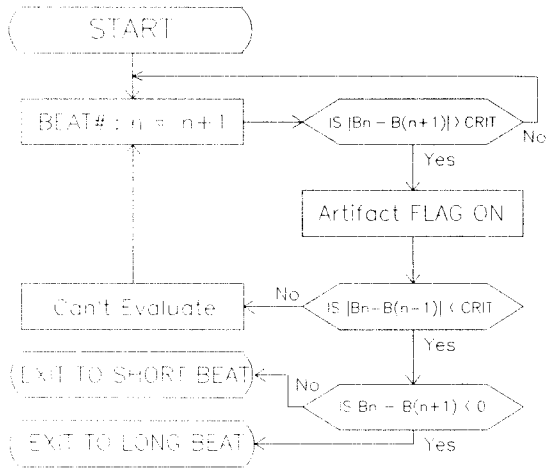

Figure 5. Flow diagram of the START program, which checks for criterion-level beat differences. If a criterion difference is found, an artifact flag is set, and the integrity of the prior beat difference is evaluated. If the prior beat difference is within criterion, the program then determines whether the criterion-level difference was due to the presence of a long beat or a short beat, and control is then passed to the appropriate subroutine.

available for derivation of the artifact criterion[5]. Flow charts illustrating the present implementation are shown in Figures 5–7. The first step of the START program (Figure 5) increments a beat pointer and then compares the absolute beat difference between that heart period and the next against the artifact criterion. If the beat difference is within criterion, the program increments the beat counter and continues evaluation of subsequent beats. If the beat difference exceeds the criterion, however, an artifact flag is set, and the program proceeds to evaluate the possibility of a false alarm.

A potential false alarm can best be evaluated if the surrounding beats are non-artifactual. The first step of false alarm processing is thus a test of whether the prior beat difference was within criterion. If not, the artifact flag remains set, and the program returns to the testing of subsequent beat differences. If the prior beat difference is within criterion, the program then determines whether the target beat difference was associated with a missed beat or the spurious detection of an extra beat. This is accomplished in the final portion of the START program. The occurrence of a missed beat is identified by the

[5]An on-line implementation is also possible. In this case the artifact criterion could be periodically updated by the accumulating beats, or calculated over a discrete time window. For most applications, however, there would probably be little advantage to such an approach.

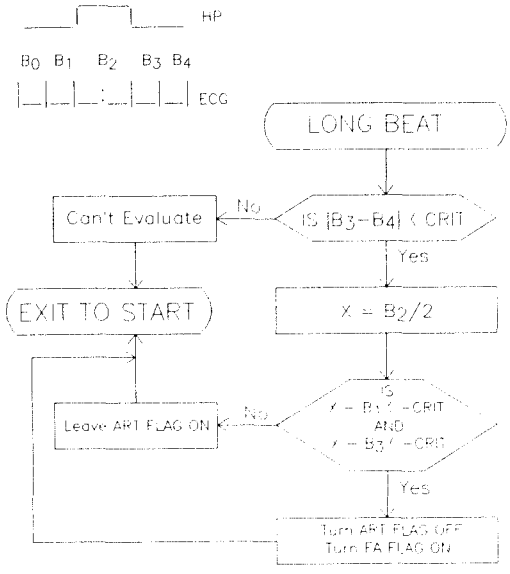

Figure 6. Flow diagram of the LONG BEAT routine, which eliminates false alarms associated with long, but non-artifactual periods. The initial step tests the integrity of the subsequent beats, by applying the beat difference criterion. If these beats fail to pass the criterion, the routine returns control to START. Otherwise, the target beat is divided in half, and the resulting periods are tested against the immediately surrounding beats. If the split beat is now too short, relative to each of the surrounding beats, the artifact flag is turned off, and a false alarm flag is set. Control is then returned to START.

direction of heart period change associated with the target beat difference. If the artifact was a missed beat, the target beat difference should reflect a criterion-level increase in heart period. If this condition is fulfilled, control passes to the LONG BEAT routine, otherwise it goes to SHORT BEAT.

The LONG BEAT routine is illustrated in Figure 6. The first step further evaluates the integrity of the surrounding beats, by comparing the difference of the two beats after the artifact against the criterion. If this test is not passed, control passes back to START, with the artifact flag remaining set. If passed, the routine then determines whether the beat can be "corrected." The long beat is divided in half, and resulting beat differences are tested against the criterion, relative to the surrounding beats. A long but non-artifactual beat (false alarm), when so divided, will frequently yield resultant beats too short to meet criterion. In this case, the program declares the long beat a false alarm, turns off the artifact flag, turns on a false alarm flag, and passes control to START. In contrast, a long artifactual beat comprised of two heart periods would yield resultant split beats that approximate the normal periods, and thus would meet criterion. In this case, control returns to START with the artifact flag remaining set.

Although beat difference criterion tests are generally based on absolute differences, a directional test is employed in LONG BEAT. That is, the routine asks whether the difference between one half of the long beat (X in Figure 6) and the prior beat (B_1) exceeds the criterion and that the value ($X - B_1$) is negative. This captures the false alarm condition outlined above, should the split beat now be too short. A similar test is applied to the subsequent beat ($X - B_3$). If both of the split beats are too short, the routine declares the beat a false alarm, turns off the artifact flag, and sets the false alarm flag. Otherwise, the artifact flag remains set. Control then returns to START. The directional criterion test is important to deal with a failure of the acquisition system to detect two or more consecutive R-waves. In that event, the resultant artifactual beat could be of sufficient duration to yield split beats that are still too long to meet criterion. Thus, the program adopts the conservative approach of declaring a false alarm only if: 1) the target beat is too long to meet criterion, and 2) the resultant split beats are too short to meet criterion for both the prior and subsequent beats. The latter also precludes confound from the failure of the acquisition system to detect two, or more, alternate R-waves.

The SHORT BEAT routine, illustrated in Figure 7, follows a similar conceptual approach to LONG BEAT. As with LONG BEAT, the SHORT BEAT routine further tests the integrity of the surrounding beats by applying the difference criterion to the two beats after the artifact. If this test is not passed, the artifact flag remains set, and control returns to START. Otherwise, processing continues. In fact, this initial step serves to exclude many extra-beat artifacts from further consideration, because spuriously detected R-waves result in two artifactual beats, which frequently yield three successive criterion-level beat differences. A false alarm surrounded by non-artifactual beats could thus pass this initial screening, as could one class of artifacts—those in which the spuriously detected R-wave fell near the beginning or end of the heart period. In such cases, one of the resultant artifactual periods could be extremely short, and therefore exceed criterion, whereas the second remains close to the true period. Subsequent steps of SHORT BEAT differentiate false alarms from this type of artifact.

If an artifact arose from the spurious division of an actual heart period, the true heart period could be restored by summing the two artifactual beats, and the restored beat should pass relevant beat dif-

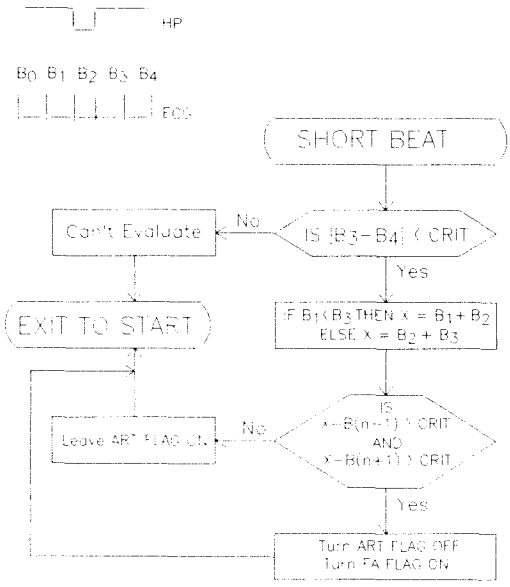

Figure 7. Flow diagram of the SHORT BEAT routine. This routine eliminates false alarms associated with short, but non-artifactual periods. The initial step tests the integrity of the subsequent beats, by applying the beat difference criterion. If these beats fail to pass the criterion, the routine returns control to START. Otherwise, the target beat is added to the shortest surrounding heart period, and the resulting period is tested against the immediately surrounding beats. If the summed period is now too long, relative to each of the surrounding beats, the artifact flag is turned off, and a false alarm flag is set. Control is then returned to START.

ference criterion tests. In contrast, a false alarm often cannot be added to either of the surrounding beats without generating a resultant beat that is now too long. Therefore, SHORT BEAT adds the target beat to the shortest of the two immediately surrounding heart periods. If the resulting beat is now too long to meet the beat difference criterion, relative to its surrounding beats, SHORT BEAT declares it a false alarm, turns the artifact flag off, sets the false alarm flag, and returns control back to START. For reasons paralleling those outlined above, a directional criterion test is also used here; the resultant summed beat must fail to meet criterion because it is too long. Thus, if multiple extra beats were detected within a single heart period, the sum of two of the resulting artifactual periods may still be unable to pass criterion. In this case, however, the summed beat will fail because it is too short. Hence, it would not be declared a false alarm. SHORT BEAT further requires that a resultant summed beat be too long for both the prior and the subsequent beats. This precludes complications from spurious R-wave detections that occur over successive beats, and yield a series of artifactually short beats, all of which are within criterion difference from each other. Summing the first two such artifacts would then restore a normal beat that is now too long for the subsequent artifactual beat. It would not be too long for the prior beat, however, and thus would not be declared a false alarm.

The performance of these false alarm algorithms was tested on the present data sets. As indicated above, the beat-difference criterion set 143 artifact flags for the edited human data. Of these, 121 (85%) were correctly eliminated by the false alarm algorithms, whereas the remaining escaped detection and thus continued to be flagged as potential artifacts. Thus, the final false alarm rate in the edited data set was reduced from 0.94% to 0.14% (0.29% for Experiment 1 and 0.06% for Experiment 2). The presence of artifacts would be expected to have two effects on false alarm detection. First, they would raise the criterion for false alarm identification (by increasing the beat-difference criterion). Secondly, if an artifact is proximate to a target beat, it may preclude application of the false alarm algorithms to this beat by violating the requirements for prior and subsequent beat differences. These expectations were born out by the lower proportion of false alarms that were able to be eliminated (59%) in the simulated artifact files. Consequently, a slightly higher residual false alarm rate (0.24%) emerged from the simulated artifact files (0.35% for Experiment 1 and 0.19% for Experiment 2). Because of the lower incidence of artifacts in the unedited data files, false alarm identification in these data was comparable to that in the artifact-free files (85%), yielding a residual false alarm rate of 0.14% (0.24% for Experiment 1 and 0.08% for Experiment 2). Importantly, none of the 1200 simulated artifacts or the 33 actual artifacts were misclassified as false alarms.

A comparable level of performance was obtained for the chimpanzee data. Of the 54 false alarms (0.31%) in the edited data, 33 were eliminated by the false alarm algorithms, yielding a residual false alarm rate of 0.12%. Similarly, of the 51 false alarms (0.29%) in the unedited data, 32 were eliminated, yielding a residual false alarm rate of 0.11%. Again, in no case did the false alarm algorithms reset an artifact flag for an actual artifact.

The false alarm algorithms thus yielded a substantial reduction in the number of spuriously flagged beats that would otherwise necessitate further evaluation. The vast majority of residual false alarms came from a few subjects with small or negative MAD-MED differences. Importantly, the false alarm algorithms constitute conservative tests, which did not spuriously exclude any of the 1,494

actual or simulated artifacts in the present data sets. Although rendered less efficient by the presence of a large number of artifacts, this change in efficiency is in a conservative direction. Fewer false alarms are excluded in the presence of artifacts.

Further Considerations and Limitations

A critical requirement of the present algorithms is that beat differences come from successive heart periods. With discrete-trial studies, a single criterion could still be applied to the entire set of trials. In this case, however, the population of difference scores used to calculate the criterion must be based only on within-trial beat differences. An issue that arises in this context is the optimal epoch or time window for calculating the beat-difference statistics that are used in deriving the artifact criterion. Clearly, the selected time window must be sufficiently long to afford an adequate sampling of beat differences. On the other hand, heart period variability may increase with increasing heart period level (Porges, McCabe, & Yongue, 1982). This was confirmed by the significant positive correlation between baseline heart period and beat-to-beat differences observed in the present study ($r=.59$, $p<.001$). Although this suggests that experimental conditions yielding different baseline heart periods should perhaps be tested separately, the regression function relating beat differences to heart period level has a shallow slope:

$$\text{Median ABS(Beat Difference)} = .058 \cdot \text{Median Heart Period} - 14.6$$

Thus, separate criterion calculations should be necessary only over extreme within-subject shifts in heart period. Between-subject differences are already accounted for, because the criterion difference is based on individual data.

Although percentile-based estimates of heart period variability are much less sensitive to artifacts than the standard deviation, they are not immune to bias. The presence of a large number of artifacts in heart period records inflates the artifact criterion, and renders the differentiation of false alarms more difficult. The present algorithms performed well at the 7.8% artifact rate of the simulated artifact files, although they probably should not be applied to artifact levels much higher than this without further validation. This imposes minimal restriction, because artifact rates approaching or exceeding 10%, except under extreme conditions, likely reflect poor recording techniques, a marginal ECG acquisition system, or both.

Other caveats also arise in the application of the present algorithms. The performance of the algorithms was clearly superior when MAD-MED differences were large and positive (low heart period variability). Consequently, although the beat difference criterion is highly conservative, caution should nonetheless be exercised in applications to subjects with high heart period variability, as indexed by negative MAD-MED differences. This would be of particular concern in conjunction with a high artifact rate.

Although the present algorithms can greatly reduce the tedium, and likely increase the accuracy of artifact identification, no automated system should ever completely replace careful attention to the raw ECG records. Noisy baselines or a low-amplitude ECG may indicate inadequate electrodes, electrode placements, or site preparation. An abnormal ECG or dysrhythmia might be the basis for exclusion of the subject, because cardiac arrhythmias could appear as artifacts to the present algorithms. In an isolated instance, such an outcome would not be problematic. Indeed, abnormal beats probably should not be left in the data. On the other hand, an excessive number of such abnormal beats would raise a question about the appropriateness of the subject.

Finally, the present algorithms do not strictly identify artifacts, they flag extreme beat differences that might be associated with artifacts. An extreme beat difference score is not informative as to which of the two contributing beats is deviant. That may require reference to the surrounding heart periods. Related to this point, a special consideration applies to a specific class of successive artifacts. Spurious beat detections occurring at similar locations within successive heart periods, although unlikely, could yield a consecutive series of similar artifactual heart periods, with the beat differences among these artifacts not exceeding criterion. This set of artifacts would be detectable by the present algorithms, because the first of such artifacts would yield a criterion beat-difference, as would the last. However, if the present program is allowed to run unattended, it becomes important to also check beats immediately surrounding the flagged heart periods. A more optimal approach would be an interactive implementation, in which artifacts are resolved as they arise in the heart period data. This would ensure the integrity of the preceding beat, against which the subsequent target beat would be compared. Consequently, each artifact within the sequence would be appropriately flagged. An interactive implementation also has an additional advantage. If artifacts are resolved in the sequence in which they arise, they would not preclude the evaluation of an immediately following criterion difference as a potential false alarm. This could appreciably increase the elimination of spuriously flagged beats, especially in data that are highly infested with artifacts.

Overview

Based on the distribution characteristics of successive heart period differences, the present artifact detection algorithms were able to effectively identify artifactual heart periods embedded in heart period records. Each of the 1494 simulated and actual artifacts in the present data sets were appropriately flagged by the artifact criterion. The general approach is somewhat similar to that previously described by Cheung (1981). However, because the artifact criterion is tailored to the individual subject, the present approach does not require assumptions concerning normal beat-to-beat variance, and is applicable over a wider range of heart period variability. Also critical to the performance of the present algorithms is the use of percentile-based distribution indexes, which are less sensitive to corruption by the presence of artifacts than are least-squares estimates.

Although the criterion was highly effective in identifying artifacts, it yielded a low rate of false alarms (0.3% to 1.0%). Additional algorithms, however, were able to further reduce the percentage of false alarms to between 0.1% and 0.3%. Thus, beyond actual artifacts, only a few beats per thousand would require additional evaluation, and the majority of these would come from a small number of subjects having high basal heart rate variability.

In summary, the outlined algorithms are highly efficient in identifying artifacts in heart period records, while at the same time yielding only a low rate of false alarms. The consistent performance of the algorithms has been documented over a range of subjects, experimental conditions, and criterion placements. The MAD and MED values, which are based on individual heart period level and heart period variability, provide meaningful indexes that relate these variables to expected performance of the algorithms.

REFERENCES

Berntson, G.G., & Boysen, S.T. (1989). Specificity of the cardiac response to conspecific vocalizations in the chimpanzee. *Behavioral Neuroscience, 103*, 235–245.

Berntson, G.G., Boysen, S.T., Bauer, H.D., & Torello, M.T. (1990). Conspecific screams and laughter: Cardiac and behavioral reactions of infant chimpanzees. *Developmental Psychobiology, 22*, 771–787.

Cacioppo, J.T., & Dorfman, D.D. (1987). Waveform moment analysis in psychophysiological research. *Psychological Bulletin, 102*, 421–438.

Caldeira, J.D. (1980). Parametric assumptions of some "nonparametric" measures of sensory efficiency. *Human Factors, 22*, 119–120.

Cheung, M.N. (1981). Detection and recovery from errors in cardiac interbeat intervals. *Psychophysiology, 18*, 341–346.

Craig, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. *Human Factors, 21*, 69–78.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Heslegrave, R.J., Ogilvie, J.C., & Furedy, J.J. (1979). Measuring baseline-treatment differences in heart rate variability: Variance versus successive difference mean square and beats per minute versus interbeat intervals. *Psychophysiology, 16*, 151–57.

Jennings, J.R., Stringfellow, J.C., & Graham, M. (1974). A comparison of the statistical distributions of beat-by-beat heart rate and heart period. *Psychophysiology, 11*, 207–210.

Rompelman, O. (1986). Investigating heart rate variability: Problems and pitfalls. In P. Grossman, K.H.L. Janssen, & D. Vaitl (Eds.), *Cardiorespiratory and cardiosomatic psychophysiology* (pp. 226–235). New York: Plenum Press.

Linden, W., & Estrin, R. (1988). Computerized cardiovascular monitoring: Method and data. *Psychophysiology, 25*, 227–234.

Norman, D.A. (1964). A comparison of data obtained under different false alarm rates. *Psychological Review, 71*, 243–246.

Porges, S.W., McCabe, P.M., & Yongue, B.G. (1982). Respiratory-heart rate interactions: Psychophysiological implications for pathophysiology and behavior. In J. Cacioppo & R. Petty (Eds.), *Perspectives in cardiovascular psychophysiology* (pp. 223–264). New York: Guilford Press.

(Manuscript received January 28, 1989; accepted for publication October 16, 1989)