Accid. Anal. & Prev., Vol. 9, pp. 167–176. Pergamon Press 1977. Printed in Great Britain

A COMPARATIVE EVALUATION OF TWO ROADSIDE BRAKE TESTING PROCEDURES†

RICHARD F. CORN‡
Analytic Services Inc., Falls Church, VA 22041, U.S.A.

J. RICHARD LANDIS
Department of Biostatics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

and

JAIRUS D. FLORA
Highway Safety Research Institute and Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

(*Received* 8 *October* 1976; *in revised form* 18 *January* 1977)

Abstract—A field survey was conducted to evaluate two procedures designed to measure the effectiveness of motor vehicle braking systems. Selected failure rates and agreement measures were computed using a recently developed unified approach to the analysis of multivariate categorical data. It was found that the procedures agree only weakly, and that the agreement varied with certain pass/fail criteria. On the basis of conditional arguments, a moving-stopping test was found to be more stringent than a wheel removal inspection.

INTRODUCTION

During the summer of 1975, the Highway Safety Research Institute (HSRI) in conjunction with the Michigan State Police and the Michigan Office of Highway Safety Planning conducted a field survey in order to evaluate two separate screening procedures which were each designed to measure the effectiveness of motor vehicle braking systems. The survey was part of a larger project designed to evaluate the Michigan checklane motor vehicle inspection program. The sample of vehicles was obtained from approximately thirty sites in two counties. The checklane teams visited the sites on different days and at different times of the day according to a randomized rotating schedule. At each site, a systematic sample with a random start was used to select vehicles from the traffic stream which were then subjected to the first screening procedure. These were further subsampled using a systematic sample in time with a random start to obtain the vehicles for the second screening procedure. For further details, the reader is referred to a complete description of the design which may be found in Flora, Corn and Coop [1976]. For the purposes of this paper, attention was limited to those 2,465 vehicles that were subjected to both screening procedures which will be denoted as:

(i) the Moving-Stopping Test (MST),
(ii) the Wheel Pull Inspection (WPI).

The MST was administered by a Michigan State Police trooper who accelerated the given vehicle to twenty miles per hour and then attempted to stop it in a specified lane which was twenty-five feet long and ten feet wide. Subsequently, the trooper classified the vehicle as unsafe if any of four conditions were observed: (1) the vehicle failed to stop, (2) it pulled to one side, (3) there was a metal-on-metal sound from the brakes, or (4) the brake pedal pressure required to stop the vehicle was not within safe bounds. On the other hand, the WPI was conducted by HSRI automotive technicians who removed the right front wheel in order to permit a visual inspection of the brake components. This inspection was performed separately and without knowledge of the results of the MST.

†This research was supported in part under contract MVI-75-001A with the Michigan Department of State Police with funds provided by the Office of Highway Safety Planning. The opinions, findings, and recommendations contained herein are those of the authors alone and do not necessarily represent those of the sponsoring agencies.

‡Formerly at Highway Safety Research Institute, University of Michigan, Ann Arbor, Michigan.

The basic objectives of this research were to investigate the extent to which the two procedures identically classified a vehicle as safe (pass) or unsafe (fail), and the efficiency of each screening procedure relative to the other. For this purpose, comparisons between the MST and the WPI were made using three different pass/fail criteria for the WPI which are labelled as WPI-1, WPI-2 and WPI-3. These three successively more stringent criteria for passing a vehicle are defined by:

$$\text{WPI-1} = \begin{cases} \text{fail, if shoe/pad fails, or cracked rotor/drum, or wheel cylinders fail,} \\ \text{pass, otherwise;} \end{cases}$$

$$\text{WPI-2} = \begin{cases} \text{fail, if WPI-1 fail or low master cylinder fluid,} \\ \text{pass, otherwise} \end{cases}$$

$$\text{WPI-3} = \begin{cases} \text{fail, if WPI-2 fail or worn rotor/drum,} \\ \text{pass, otherwise.} \end{cases}$$

Due to the nested nature of these criterion sets, a vehicle that failed WPI-1 must necessarily fail WPI-2, and a vehicle that failed WPI-2 must necessarily fail WPI-3. Converserly, a vehicle that passed WPI-3 must necessarily pass WPI-2, and a vehicle that passed WPI-2 must necessarily pass WPI-1. All possible classification outcomes arising from these pass/fail criteria, together with their respective observed frequencies, are shown in Table 1.

Table 1. Response profiles and their frequencies

	WPI-1	vs WPI-2	vs WPI-3	
	+	+	+	-
	+	+	-	-
	+	-	-	-	Total
MST +	1749	24	52	62	1887
MST -	491	12	31	44	578
Total	2240	36	83	106	2465

(+) denotes pass; (-) denotes fail

The statistical issues concerning the differences in the pass/fail decision associated with each of these criteria can be summarized within the framework of the following questions:

(1) Are there any differences among the overall failure rates for the MST and each of the three WPI criterion sets?

(2) If the MST is regarded as the standard of braking ability, (i) What proportion of the safe vehicles are passed by the WPI? (ii) What proportion of the unsafe vehicles are failed by the WPI?

(3) If the WPI is regarded as the standard of braking ability, (i) What proportion of the safe vehicles are passed by the MST? (ii) What proportion of the unsafe vehicles are failed by the MST?

(4) To what extent do the MST and each of the WPI criteria agree on the specific pass/fail decision for individual vehicles?

(5) Is the agreement between the MST and each of the WPI criteria significantly different from chance agreement based on their overall crude distribution of passes and failures?

(6) Does the agreement on the classification of individual vehicles between the MST and the WPI increase as the WPI criteria for passing become more stringent?

In the following sections a general methodology for answering these questions is developed in terms of various agreement measures and corresponding hypothesis tests. These procedures are then illustrated with an analysis of the data in Table 1.

METHODOLOGY

Let the brake safety status of a given vehicle be measured separately by the MST and each of the WPI criteria using a two-point scale (pass (+), fail (−)). In general, this gives rise to $r = 2^4 = 16$ possible response profiles determined by the joint classification of the four measurement procedures. However, due to the hierarchical nature of the WPI criteria defined in the previous section, there are only 8 possible response profiles which can be labeled as shown in Table 2.

Table 2. General table of underlying probabilities

	WPI-1	vs WPI-2	vs WPI-3		
	+	+	+	−	
	+	+	−	−	
	+	−	−	−	Total
MST +	π_{+0}	π_{+1}	π_{+2}	π_{+3}	$\pi_{+\cdot}$
MST −	π_{-0}	π_{-1}	π_{-2}	π_{-3}	$\pi_{-\cdot}$
Total	$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\pi_{\cdot 3}$	1

(+) denotes pass; (-) denotes fail

Agreement between the MST and each of the WPI criteria in the sense of question (1) can be investigated by comparing the overall usage of the measurement scale (+, −) on the same vehicles by each of the four criteria. These comparisons can be made in terms of the overall failure rate for each criterion, which can be defined in terms of the joint probabilities in Table 2 by

$$\phi_{MST} = \pi_{-\cdot}; \tag{1}$$

$$\phi_{W1} = \pi_{\cdot 3}; \tag{2}$$

$$\phi_{W2} = \pi_{\cdot 3} + \pi_{\cdot 2}; \tag{3}$$

$$\phi_{W3} = \pi_{\cdot 3} + \pi_{\cdot 2} + \pi_{\cdot 1}. \tag{4}$$

If there are no differences among these crude distributions, the failure rates in (1)–(4) satisfy the hypothesis of first-order marginal homogeneity

$$H_0: \phi_{MST} = \phi_{W1} = \phi_{W2} = \phi_{W3}. \tag{5}$$

In contrast to comparisons among the failure rates, the MST and the WPI criteria can be evaluated in terms of agreement on the classification of individual vehicles. In this regard, a wide variety of agreement measures have been proposed for categorical (nominal or ordinal) data from contingency tables as reviewed in Landis and Koch [1975a, 1975b]. These measures are defined in terms of the cell probabilities in Table 2, in addition to the marginal probabilities used for the failure rates in (1)–(4). In particular, let the probability that both the MST and each of the respective WPI criteria pass a vehicle be denoted by

$$\pi_{++1} = \pi_{+0} + \pi_{+1} + \pi_{+2}; \tag{6}$$

$$\pi_{++2} = \pi_{+0} + \pi_{+1}; \tag{7}$$

$$\pi_{++3} = \pi_{+0}, \tag{8}$$

and let the probability that both fail a vehicle be denoted by

$$\pi_{--1} = \pi_{-3}; \tag{9}$$

$$\pi_{--2} = \pi_{-3} + \pi_{-2}; \tag{10}$$

$$\pi_{--3} = \pi_{-3} + \pi_{-2} + \pi_{-1} \tag{11}$$

for each of the WPI criteria, respectively.

Using these quantities in (6)–(11), one approach to the evaluation of agreement between the MST and the WPI criteria is to arbitrarily consider one of the measurement procedures as a known standard, and to evaluate the other criteria with respect to that standard. For example, the probabilities that the respective WPI criteria correctly classify a randomly selected vehicle as pass (+), given that the MST has passed the vehicle, are

$$\xi_{W1} = \pi_{++1}/(1 - \phi_{MST}); \tag{12}$$
$$\xi_{W2} = \pi_{++2}/(1 - \phi_{MST}); \tag{13}$$
$$\xi_{W3} = \pi_{++3}/(1 - \phi_{MST}), \tag{14}$$

and the probabilities that the respective WPI criteria correctly classify a randomly selected vehicle as (−), given that the MST has failed the vehicle, are

$$\eta_{W1} = \pi_{--1}/\phi_{MST}; \tag{15}$$
$$\eta_{W2} = \pi_{--2}/\phi_{MST}; \tag{16}$$
$$\eta_{W3} = \pi_{--3}/\phi_{MST}. \tag{17}$$

These quantities in (12)–(14) are known as the *sensitivity* of the WPI criteria as addressed in question (2i) and the quantities in (15)–(17) are known as the *specificity* of the WPI criteria relative to the MST considered in question (2ii) (see Fleiss [1973, Ch. 1], Landis and Koch[1975b]). Conversely, if we regard the WPI criteria as the standard, the *sensitivity* of the MST as addressed in question (3i) is given by

$$\xi_{M1} = \pi_{++1}/(1 - \phi_{W1}); \tag{18}$$
$$\xi_{M2} = \pi_{++2}/(1 - \phi_{W2}); \tag{19}$$
$$\xi_{M3} = \pi_{++3}/(1 - \phi_{W3}), \tag{20}$$

and the *specificity* of the MST considered in question (3ii) is given by

$$\eta_{M1} = \pi_{--1}/\phi_{W1}; \tag{21}$$
$$\eta_{M2} = \pi_{--2}/\phi_{W2}; \tag{22}$$
$$\eta_{M3} = \pi_{--3}/\phi_{W3}, \tag{23}$$

relative to each of the respective WPI criterion.

An alternative to measures of *sensitivity* and *specificity* which combines agreement on both pass and fail between the MST and the WPI criteria, without assuming that one of the procedures is a standard, are quantities of the form

$$\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e}, \tag{24}$$

where π_0 is an observational probability of agreement and π_e is a hypothetical expected probability of agreement under an appropriate set of baseline constraints, such as total independence of the two measurement procedures. Ranging from $(-\pi_e/1-\pi_e)$ to $(+1)$, κ indicates the extent to which the observed probability of agreement exceeds the expected probability of agreement, thus giving rise to the term "chance corrected" or standardized agreement measure. As a result, the test of whether the observed probability of agreement is significantly different from the expected probability of agreement can be formulated in terms of the test of the hypothesis that $\kappa = 0$. In this context, let

$$\pi_{oj} = \pi_{++j} + \pi_{--j} \tag{25}$$

be the observational probability of agreement between the MST and the jth WPI criterion obtained from the corresponding probabilities in (6)–(8) and (9)–(11). In addition, let

$$\pi_{ej} = (1 - \phi_{MST})\ (1 - \phi_{Wj}) + \phi_{MST}\phi_{Wj} \tag{26}$$

be the corresponding expected proportion of agreement under the baseline constraints of independence. Combining these probabilities in (25) and (26), the standardized coefficient of agreement between the MST and the jth WPI criterion from (24) due to Cohen[1960], can be denoted by

$$\kappa_j = \frac{\pi_{0j} - \pi_{ej}}{1 - \pi_{ej}} \qquad \text{for } j = 1,2,3. \tag{27}$$

In practice, these agreement measures can be obtained directly from the three marginal tables of Table 2 determined by the cross-classification of the MST with each of the WPI criteria as shown in Table 3. Here, we note that the sensitivity, specificity, and kappa measures in (12)–(27) can all be expressed in terms of the main diagonal and marginal probabilities in Table 3. As a result, point estimates for these quantities can all be computed directly from their corresponding sample proportions shown in Table 4. In particular, estimates of the failure rates in (1)–(4) are

$$\hat{\phi}_{\text{MST}} = p_{-\cdot} \tag{28}$$

$$\hat{\phi}_{\text{W}j} = p_{\cdot j} \qquad \text{for } j = 1,2,3; \tag{29}$$

the estimates of sensitivity and specificity for the WPI criteria in (12)–(17) are

$$\hat{\xi}_{\text{W}j} = p_{++j}/p_{+\cdot} \tag{30}$$

$$\hat{\eta}_{\text{W}j} = p_{--j}/p_{-\cdot} \qquad \text{for } j = 1,2,3; \tag{31}$$

and the corresponding estimates for the MST in (18)–(23) are

$$\hat{\xi}_{\text{M}j} = p_{++j}/(1 - p_{\cdot j}) \tag{32}$$

$$\hat{\eta}_{\text{M}j} = p_{-\,-j}/p_{\cdot j} \qquad \text{for } j = 1,2,3. \tag{33}$$

Furthermore, if we let

$$p_{0j} = p_{++j} + p_{-\,-j} \tag{34}$$

estimate π_{0j} in (25), and

$$p_{ej} = p_{+\cdot}(1 - p_{\cdot j}) + p_{-\cdot}\, p_{\cdot j} \tag{35}$$

estimate π_{ej} in (26), then a consistent estimate of κ_j in (27) is

$$\hat{\kappa}_j = \frac{p_{0j} - p_{ej}}{1 - p_{ej}} \qquad \text{for } j = 1,2,3. \tag{36}$$

Moreover, the estimated asymptotic variance for $\hat{\kappa}_j$ can then be obtained by linearized Taylor series approximations as reported in Fleiss, Cohen and Everitt[1969].

Table 3. Second-order margin of table 2 comparing MST with WPI-j

		WPI-j		
		+	-	Total
MST	+	π_{++j}	π_{+-j}	$\pi_{+\cdot}$
	-	π_{-+j}	π_{--j}	$\pi_{-\cdot}$
Total		$1-\phi_{\text{W}j}$	$\phi_{\text{W}j}$	1

(+) denotes pass; (-) denotes fail

Table 4. Observed proportions comparing MST with WPI-j

		WPI-j		
		+	-	Total
MST	+	p_{++j}	p_{+-j}	$p_{+\cdot}$
	-	p_{-+j}	p_{--j}	$p_{-\cdot}$
Total		$1-p_{\cdot j}$	$p_{\cdot j}$	1

(+) denotes pass; (-) denotes fail

Although these estimates can all be obtained directly from subtables as shown in Table 4, their estimated covariance structure is considerably more complex, since the proportions comprising these three subtables are actually obtained as different sums of the same component observed proportions in precisely the same manner as Table 3 is constructed from Table 2. For this reason, we propose that these estimates of failure rates, sensitivity, specificity, and kappa be generated with the framework of a unified approach to the analysis of multivariate categorical data originally outlined in Grizzle, Starmer, and Koch[1969]. This methodology has recently been expanded to include the measurement of agreement for categorical data in Landis and Koch[1977]. In particular, the separate kappa statistics in (36), measuring the standardized agreement between the MST and the jth WPI criteria, can be computed simultaneously, thus permitting testing for differences among these three correlated kappa statistics. The computations can all be performed by a recently developed computer program ((GENCAT) which is documented in Landis *et al.*[1976]. This program may also be used to obtain simultaneously the point estimates of sensitivity and specificity in (30)–(33) and their covariance structure, and to test the hypothesis of marginal homogeneity in (5).

ANALYSIS AND RESULTS

This section is concerned with the analysis of the brake testing data presented in Table 1 within the scope of the methodology developed in the previous section. Because a considerable number of tests are required in order to investigate the full range of hypotheses suggested in questions (1)–(6), the issue of multiple comparisons involving correlated test statistics must be considered. If the hypothesis testing is limited to a few prespecified contrasts, the Bonferroni method of multiple comparisons may be the preferred procedure. However, in general, if many contrasts are of interest, or if certain contrasts are chosen after inspection of the data in some post hoc manner, the Scheffé type procedure discussed in Goodman[1964] and Grizzle, Starmer and Koch[1969] may be more appropriate. As a result, we will utilize the Scheffé approach to propose appropriate critical values for significance testing. This method consists of comparing individual χ^2 statistics with (usually) one degree of freedom to the critical value obtained from the χ^2 distribution with degrees of freedom equal to the total for the composite hypothesis.

The comparisons required to answer the questions associated with the data in Table 1 can be described more clearly within the context of the three 2×2 subtables of observed frequencies corresponding to Table 4 as shown in Table 5. For example, the functions required to test the hypothesis of first order marginal homogeneity in (5) discussed in question (1) can be obtained directly from the corresponding margin of the subtables in Table 5. In particular, the MST

Table 5. Observed frequencies comparing MST with each of the WPI-j

		WPI-1 Pass	WPI-1 Fail	Total
MST	Pass	1825	62	1887
	Fail	534	44	578
	Total	2359	106	2465

		WPI-2 Pass	WPI-2 Fail	Total
MST	Pass	1773	114	1887
	Fail	503	75	578
	Total	2276	189	2465

		WPI-3 Pass	WPI-3 Fail	Total
MST	Pass	1749	138	1887
	Fail	491	87	578
	Total	2240	225	2465

failure rate in (28) is

$$\hat{\phi}_{MST} = {}^{578}/_{2465} = 0.23, \tag{37}$$

and the WPI-j failure rates in (29) are

$$\hat{\phi}_{W1} = {}^{106}/_{2465} = 0.04; \tag{38}$$
$$\hat{\phi}_{W2} = {}^{189}/_{2465} = 0.08; \tag{39}$$
$$\hat{\phi}_{W3} = {}^{225}/_{2465} = 0.09. \tag{40}$$

The test statistic for H_0 in (5) is $\chi^2 = 529.81$ with d.f. = 3, which implies that there are significant ($\alpha = 0.01$) differences among the four failure rates. In particular, the estimated pairwise failure rate differences between the MST and each of the WPI criterion, together with the resulting test statistics, are displayed in Table 6. These results suggest that the overall failure rate for the MST is significantly different ($\alpha = 0.01$) from that of any of the WPI criteria.

The issues addressed by questions (2) and (3) can be discussed in terms of sensitivity and specificity measures of agreement as outlined in the previous section. These estimates can be obtained from Table 5 by arbitrarily regarding one of the test criteria as a known standard using the results in (30)–(33). These estimates of sensitivity and specificity, together with their estimated standard errors, are summarized in Table 7.

Table 6. Hypothesis test involving pairwise failure rate differences between MST and each of the WPI-j

Test Criterion	Estimated Failure Rate	Hypothesis	d.f.	Estimate of Differences	Estimated Standard Error	χ^2 Test Statistic
MST	0.23					
WPI-1	0.04	$\phi_{MST}=\phi_{W1}$	1	0.19	0.009	440.71**
WPI-2	0.08	$\phi_{MST}=\phi_{W2}$	1	0.16	0.010	272.34**
WPI-3	0.09	$\phi_{MST}=\phi_{W3}$	1	0.14	0.010	215.40**

**denotes statistical significance at an overall $\alpha = 0.01$ level determined by the critical value $\chi^2_c = 11.35$ with 3 d.f.

Table 7. Measures of sensitivity and specificity and their estimated standard errors.

Estimates of Agreement	Performance of WPI Criteria Relative to MST		
	WPI-1	WPI-2	WPI-3
Sensitivity ($\hat{\xi}_{Wj}$)	0.967	0.940	0.927
	(0.004)	(0.005)	(0.006)
Specificity ($\hat{\eta}_{Wj}$)	0.076	0.130	0.151
	(0.011)	(0.014)	(0.015)
	Performance of MST Criterion Relative to WPI-j		
	WPI-1	WPI-2	WPI-3
Sensitivity ($\hat{\xi}_{Mj}$)	0.774	0.779	0.781
	(0.009)	(0.009)	(0.009)
Specificity ($\hat{\eta}_{Mj}$)	0.415	0.397	0.387
	(0.048)	(0.036)	(0.033)

Standard errors indicated in parentheses

Furthermore, a preliminary investigation of the observed agreement on the classification of individual vehicles between the MST and each of the WPI criteria as mentioned in question (4) can be performed by computing p_{oj} in (34) for each of the 2×2 subtables in Table 5. The resulting crude agreement statistics between the MST and the WPI-j are (0.76, 0.75, 0.74) for $j = 1,2,3$ respectively. These estimates indicate that the MST and each of the WPI criteria identically classified approximately 3/4 of the vehicles, although not necessarily the same vehicles. In addition, the expected proportion of agreement under the baseline constraints of independence can be obtained by computing p_{ej} in (35) for each of the 2×2 subtables in Table 5. This yields (0.74, 0.72, 0.72), respectively. Consequently, these observed and expected proportions of agreement can be used to create the standardized agreement statistics in (36) between the MST and each of the WPI criteria. These estimates of agreement, together with the estimated standard errors of kappa and the test statistics associated with question (5), are displayed in Table 8.

Table 8. Agreement statistics and corresponding hypothesis tests

Observed Crude Agreement (P_{oj})	Expected Crude Agreement (P_{ej})	Estimated Kappa Statistic ($\hat{\kappa}_j$)	Estimated Standard Error	Hypothesis	d.f.	χ^2 Test Statistic
0.76	0.74	0.060	0.0163	$\kappa_1 = 0$	1	13.80**
0.75	0.72	0.090	0.0193	$\kappa_2 = 0$	1	21.90**
0.74	0.72	0.098	0.0201	$\kappa_3 = 0$	1	23.89**

**denotes statistical significance at an overall $\alpha = 0.01$ level determined by the critical value $\chi^2_c = 11.35$ with 3 d.f.

These results suggest that even though the kappa statistics are significantly non-zero ($\alpha = 0.01$), the strength of the agreement between the MST and WPI criteria is only slight ($\hat{\kappa}_j \leq 0.10$ for $j = 1,2,3$) when adjusted for the expected agreement on the basis of the overall failure rates.

Finally, as mentioned in question (6), it is of interest to investigate the extent to which the agreement between the MST and WPI changes as the criteria for passing the WPI become more stringent. This issue can be investigated by testing for the equality of the three kappa statistics displayed in Table 8 by means of the pairwise differences which are shown together with their estimated standard errors and test statistics in Table 9.

These results suggest that the agreement between the MST and the WPI is different for WPI-1 than for WPI-3, but that WPI-1 and WPI-2 exhibit essentially the same agreement with the MST, as do WPI-2 and WPI-3. However, as mentioned previously, all these statistics reflect only slight chance-corrected agreement between these measurement procedures.

Table 9. Hypothesis test for differences among kappa measures of agreement.

Hypothesis	d.f.	Estimate of Statistic	Estimated Standard Error	χ^2 Test Statistic
$\kappa_1 - \kappa_2 = 0$	1	-0.030	0.0131	5.24
$\kappa_1 - \kappa_3 = 0$	1	-0.038	0.0150	6.32*
$\kappa_2 - \kappa_3 = 0$	1	-0.008	0.0082	0.89

* denotes statistical significance at an overall $\alpha = 0.05$ level determined by the critical value $\chi^2_c = 5.99$ with 2 d.f.

DISCUSSION

The significant differences between the overall MST failure rate and those of the WPI criteria provide a preliminary indication that these procedures may not evaluate the same vehicle characteristics. Despite strong crude agreement between the two procedures (MST and WPI-2) only weak—although significantly non-zero—chance-corrected agreement was observed. The chance-corrected agreement measures differed significantly only between the least and the most stringent of the WPI criteria.

The MST is a measure of the vehicle's performance at the time the test is performed. The results of the MST reflect a variety of factors including all four brakes and their interaction, all four tires and their inflation balance, and the road surface. On the other hand, the WPI measures the vehicle's condition. It only reflects the components of the inspected brake. The poor mechanical condition of a brake may lead to performance deficiencies in the near future, even though these may not manifest themselves at the time the vehicle is inspected. The lack of a strong chance-corrected agreement between the WPI and the MST may be an indication of their inherently different characteristics.

The purpose of questions (2) and (3) was to determine the relative performance characteristics of the WPI and the MST. This information can be useful in policy decisions of which type of brake test should be used in an inspection program. However, other information is also necessary for these decisions. For example, the relative cost of the two procedures is important. This would include the time and equipment needed to perform the test as well as the personnel costs of conducting each type of test. The desired balance between true pass (sensitivity) and true fail (specificity) rates must be determined. It is desirable to pass a large proportion of safe vehicles, and it is also desirable to fail a large proportion of unsafe vehicles. Requiring unnecessary expensive repairs would seriously adversely affect the public acceptance of an inspection program. The MST can be easily implemented by State Police troopers using readily available equipment, while the WPI requires one or more mechanics as well as special equipment such as air compressors, air wrenches, hydraulic floor jacks, and assorted tools and parts.

The primary criterion of comparison of the two procedures is the specificity of the procedure. In this situation this refers to the ability of the procedure to detect the vehicles with defective brakes (as judged by the other procedure). This assumes implicitly that failing to detect a vehicle with defective brakes is qualitatively more serious than inadvertently failing a vehicle with good brakes. In making policy decisions, this assumption should be critically evaluated. It is generally necessary to balance the two types of errors in arriving at the preferred decision.

Using the specificity as the criterion for comparison, it can be seen from Table 7 that the MST has a better rate than any of the WPI criteria. For example, the MST detects 39.7% of the vehicles judged as defective by WPI-2, while the WPI-2 would detect only 13.0% of the vehicles judged defective by the MST. This same advantage accrues to the MST for each of the three WPI criteria.

On the other hand, the MST is somewhat less sensitive than are the WPI. In this context, sensitivity means the ability of a procedure to correctly identify the "safe" vehicles. The three WPI criteria have sensitivities of above 92% for correctly passing the vehicles which pass the MST. However, the MST has a sensitivity of only about 78% for passing vehicles approved by the WPI criteria. An improvement of from 13 to 40% in the ability to detect vehicles with defective brakes with a loss of sensitivity from 94 to 78% may be viewed as showing an advantage for the MST. However, the actual determination of which procedure to select would depend on the relative importance attached to the sensitivity and specificity and the ease of implementation. Since the MST is more stringent and, less costly, it may be preferred by many.

Acknowledgements—The authors wish to thank the Michigan State Police and the Highway Safety Research Institute for their cooperation and support. The authors are particularly grateful to Mr. Ronald L. Copp for his participation in the data collection and reduction, Mr. Eugene R. Heyman for his assistance in the data analysis, and to Ms. Elizabeth L. Ng for her typing of the final revisions. The authors are also indebted to Prof. Gary G. Koch for many helpful comments on a previous version of this manuscript.

REFERENCES

Cohen J., A coefficient of agreement for nominal scales. *Ed. Phsychol. Meas.* **20**, 37–46, 1960.

Fleiss J. L., Cohen J. and Everitt B. S., Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **72**, 323–337, 1969.

Fleiss, J. L. *Statistical Methods for Rates and Proportions*. Wiley, New York, 1973.

Flora J. D., Corn R. F. and Copp R. L., The Michigan trial substitute vehicle inspection program: first year final report. HSRI Report No. UM-HSRI-76-9, 1976.

Goodman L. A., Simultaneous confidence intervals for contrasts among multinomial populations. *Ann. Math. Statist.* **35**, 716–725, 1964.

Grizzle J. E., Starmer C. F. and Koch G. G., Analysis of categorical data by linear models. *Biometrics* **25**, 489–504, 1969.

Landis J. R. and Koch G. G., A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statistica Neerlandica* **29**, 101–123, 1975a.

Landis J. R. and Koch G. G., A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). *Statistica Neerlandica* **29**, 151–161, 1975b.

Landis J. R. and Koch G. G., The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174, 1977.

Landis J. R., Stanish W. M., Freeman J. L. and Koch G. G., A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). *Comput. Programs Biomed.* **6**, 196–231, 1976.