Joseph L. Peterson,[1] *D.Crim. and Penelope N. Markham,*[2] *Ph.D.*

Crime Laboratory Proficiency Testing Results, 1978–1991, II: Resolving Questions of Common Origin

REFERENCE: Peterson, J. L. and Markham, P., **"Crime Laboratory Proficiency Testing Results, 1978–1991, II: Resolving Questions of Common Origin,"** *Journal of Forensic Sciences,* JFSCA, Vol. 40, No. 6, November 1995, pp. 1009–1029.

ABSTRACT: A preceding article has examined the origins of crime laboratory proficiency testing and the performance of laboratories in the identification and classification of common types of physical evidence. Part II reviews laboratory proficiency in determining if two or more evidence samples shared a common source. Parts I and II together review the results of 175 separate tests issued to crime laboratories over the period 1978 to 1991. Laboratories perform best in determining the origin of finger and palm prints, metals, firearms (bullets and cartridge cases), and footwear. Laboratories have moderate success in determining the source of bloodstains, questioned documents, toolmarks, and hair. A final category is of greater concern and includes those evidence categories where 10% or more of results disagree with manufacturers regarding the source of samples. This latter group includes paint, glass, fibers, and body fluid mixtures. The article concludes with a comparison of current findings with earlier LEAA study results, and a discussion of judicial and policy implications.

KEYWORDS: forensic science, criminalistics, proficiency testing, crime laboratories

This is Part II of a review of crime laboratory proficiency testing results covering the period 1978 to 1991. Part I has previously described the history of the proficiency testing program and results of laboratories in identifying and classifying substances.

This article reviews the performance of laboratories in determining if one or more unknown/questioned samples could have shared a common origin with a known sample. This question of common origin is commonly asked of crime laboratories in an effort to associate a suspect/defendant with a crime scene or victim. The results of laboratories were compared with target values supplied by the manufacturer. The criterion/unit of measurement for judging these results is the *comparison,* whereby one or more questioned (unknown origin) samples were compared with one or more standard (known) samples. Comparison results are placed in one of several categories:

Received for publication 11 Oct. 1994; accepted for publication 17 April 1995.

This research was supported, in part, by a grant from the Office of Social Science Research, University of Illinois at Chicago.

[1]Professor, Department of Criminal Justice, University of Illinois, Chicago, IL.

[2]Research Associate, Department of Medicinal Chemistry and Pharmacognosy, University of Illinois, Chicago, IL.

- Agree (+) Includes responses where laboratory responses reporting two or more items could have shared a common source agreed with the manufacturing laboratory.
- Agree (−) Includes replies where laboratory responses stating two or more items could *not* have shared a common origin agreed with the manufacturer.
- Disagree (+) Includes results where laboratories reported two or more items shared a common origin when, in fact, they originated from different sources.
- Disagree (−) Includes results where laboratories reported two or more items did not share a common source when, in fact, they did.
- Inconclusive (+) Includes responses where laboratories reported inconclusive results and where the items shared a common source.
- Inconclusive (−) Includes responses where laboratories reported inconclusive results and where the items came from different sources.
- Unjustified exclusions—This category was used in the toolmarks area and included responses where laboratories reported exclusions without knowing if the reverse side of a blade may have made the mark.

In the report issued in 1978 at the close of the LEAA study, an "unacceptable proficiency" rate was calculated to evaluate results as follows: number of unacceptable responses/number of laboratories responding with data for that particular examination. Here, the definition of "unacceptable response" included incorrect responses, as well as inconclusive responses which the PAC did not feel were supportable. We believe the categorization scheme employed in the present article is superior because the divisor in each of the tests is the total number of comparisons made for a given test (not the number of labs submitting data). The earlier "unacceptable proficiency" rate did not take into account the true scope of the exam (for example, the number of knowns and unknowns) and the number of comparative steps the examiners were required to perform. The new system also breaks out inconclusive responses as a separate category. An inconclusive response may indeed be the most appropriate response in a situation in which the sample, the test results, lab policy, and/or examiner capabilities do not permit a firm conclusion. In subsequent sections of this article when we compare performance of laboratories in this early testing with more recent results, we have attempted to translate the 1978 results to the present comparison-based scheme.

Common origin results are tabulated and reviewed for the following categories:

- Firearms
- Toolmarks

- Hair
- Footwear
- Physiological fluids
- Glass
- Paint
- Fibers
- Latent fingerprints
- Questioned documents
- Metals

The data for the eleven evidence categories in which common origin questions were posed are summarized in Table 1. It can be readily seen the success of laboratories in making comparisons varied considerably, ranging from latent prints where a high of 99.6% of responses associated latent prints of unknown origin with the proper fingerprint card, to a low of 64% of responses associating known handwriting with samples of unknown origin. Similarly, the percent of inconclusive responses varied from virtually no latent print cases to 32% of document comparisons. In addition to fingerprints, laboratories had a high rate of success in resolving questions of common origin in the categories of metals, firearms, and bloodstains. The highest percent of comparisons which disagreed were recorded in the categories of paints (household and automotive), glass, fibers and physiological fluid mixtures. Inconclusive responses were most common in the questioned documents, toolmarks, and human hair categories.

The following sections will review these results in greater detail.

Latent Prints

Between 1983 and 1991 there were a total of nine latent (finger and palm) print tests administered to crime laboratories. There were no latent print examinations under the previous LEAA study. The number of laboratories subscribing to the tests increased from 38 to 141 over this period, while laboratories that responded with results rose from 24 to 88, almost a fourfold increase in both categories. Overall the participation rate was 65%, a figure that did not change appreciably during the course of these exercises. Typically, participants were provided with several cards on which there were various latent prints of unknown origin and several sets of (known) inked finger and palm prints belonging to various "suspects." The manufacturer(s) of the latent prints attempted to simulate actual conditions by creating smudged, elongated, compressed and other irregular latent print specimens. Scenarios accompanying the prints described the circumstances in which the latent prints were recovered from property and/or crime scenes. In many instances, the manufacturers used latent prints which had been recovered in actual criminal cases. The manufacturers also attempted to use inked prints which represented the range of quality typically experienced by latent examiners. Customarily, the inked and latent impressions were reproduced photographically at the standard 1:1 size.

For each latent print card issued, the laboratories were asked to answer four questions:

- Is the latent print of value for identification?
- If of value, is (are) the print(s) identifiable; in other words, identified to one of the known inked prints?
- If identifiable, with which particular suspect card?
- Which finger(s) or palm identified?

Table 2 summarizes the results on the nine exercises. As with other tests we have used the "comparison" as the fundamental unit of analysis in reporting laboratory results. For "prints of value," we tabulated the fraction and percent of responses which reported prints to be of value (+) when the manufacturer said they were, and the percent of responses which reported they were not of value (−) when the manufacturer said they were not. There is considerable subjectivity involved in this "value" designation. Similarly, under "identifiable," we counted reports which agreed (+) (reporting prints identifiable when they were), and disagreed (−) (reporting prints not identifiable when they were not) ratios. Under "card" the percent of positive identifications of a suspect card to total identifications is tabulated. Under the column headed "assignment," we computed the percent of identifications of a particular finger or palm to the total number of identifications which agreed with the manufacturers' specifications.

Overall, laboratories enjoyed a high degree of success in making these comparisons. In terms of assessing the value of the latent prints, respondents agreed with the manufacturer in about 98% of cases where the prints were of value, and in about 92% of instances where the manufacturer stated they were *not* of value. Conversely, respondents reported latent prints to be of value, when they were not, in about 8% of determinations, and incorrectly reported they were not, when they were, in about 2% of responses on this question. Laboratories made positive identifications in 92% of situations where an identification was possible; they reported there was none in 98% of responses where the print was not identifiable. Conversely, laboratories called prints identifiable when they were not in 2% of responses, and labelled them unidentifiable, when they were, in 8% of cases. A distinct conservatism in calling identifications is evident. Because there were many more occasions where an identification was possible than not, and the referenced conservatism of examiners, the overall rate of identifications agreeing with the manufacturer averaged 93%.

Laboratories selected the proper card in 99.6% of responses where a card was identifiable and reported the proper finger or palm print on these cards in about 98% of their responses.

It is helpful to discuss each of these results in greater detail to review problem areas and hopefully clarify the meaning of these percentages. In terms of assessing whether the prints were of value or not, the labs made *no* positive errors (reporting a latent of value

TABLE 1—*Summary of comparison results.*

Evidence Type (*n* Comparisons)		Agree, %	Disagree, %	Inconclusive, %
Latent prints				
a) Card	(*n* = 4698)	99.6	0.5	...
b) Finger	(*n* = 4735)	98	2.0	...
Metals	(*n* = 147)	93	2	5
Footwear	(*n* = 1767)	87	0.7	12
Physiological fluids				
a) Blood	(*n* = 815)	89	6	5
b) Mixtures	(*n* = 3302)	83	11	6
Glass	(*n* = 765)	85	13	2
Firearms	(*n* = 2106)	88	1.4	10
Paints				
a) Automotive	(*n* = 664)	74	23	2
b) Household	(*n* = 566)	85	11	4
Fibers	(*n* = 925)	83	11	6
Hair	(*n* = 1609)	74	8	18
Toolmarks[a]	(*n* = 1961)	74	4	17
Questioned documents	(*n* = 938)	64	3	32

[a]An additional 5% of replies were classified as unjustified exclusions.

TABLE 2—*Latent prints.*

Report	Participation Rate	Unknown Latent Cases	Known Sets Prints	Prints of Value (Agree)		Identifiable (Agree)		Card (Agree)	Assignment Finger/Palm (Agree)
				+	−	+	−		
83-4	24/38 (63%)	21	41	379/418 (91%)	72/72 (100%)	279/328 (85%)	44/44 (100%)	277/279 (99%)	263/276 (95%)
84-5	28/45 (62%)	21	31	544/560 (97%)	26/28 (93%)	439/464 (95%)	79/81 (98%)	432/439 (98%)	526/546 (96%)
85-7	37/51 (73%)	21	30	718/740 (97%)	0/0 (0%)	603/650 (93%)	36/37 (97%)	602/604 (99%)	598/602 (99%)
86-7	43/65 (66%)	25	18	1006/1020 (99%)	43/43 (100%)	856/921 (93%)	83/85 (98%)	853/856 (99%)	847/857 (99%)
87-7	52/70 (74%)	13	3	654/676 (97%)	0/0 (0%)	511/557 (92%)	51/52 (98%)	555/558 (99%)	545/557 (98%)
88-7	62/97 (64%)	12	5	680/682 (99%)	58/62 (94%)	605/620 (98%)	122/124 (98%)	543/543 (100%)	543/543 (100%)
89-7	56/?	12	5	600/682 (99%)	32/56 (57%)	427/442 (97%)	166/166 (100%)	427/427 (100%)	421/427 (99%)
90-7	74/117 (63%)	12	5	1459/1477 (99%)	143/145 (99%)	992/1163 (85%)	290/296 (98%)	988/992 (99.6%)	915/927 (99%)
91-8	88/141 (62%)	12	3	1657/1723 (96%)	0/0 (0%)	(not requested on form)		1600/1657 (97%)	1524/1600 (95%)
Total[a]	408/624[b] (65%)	149	141	6040/6181 (98%)	374/406 (92%)	4712/5145 (92%)	871/885 (98%)	4677/4698 (99.6%)	4658/4735 (98%)
				6414/6587 (97%)		5583/6030 (93%)			

[a]Except for column totals for participation rate, unknown latent cases, and known sets prints, the remaining six columns *exclude* 91-8 results due to missing values for the "Identifiable" column.

[b]This figure excludes test 89-7 since the report did not state the number of labs receiving tests.

when it wasn't) in five of the exercises. Of the 149 different latent print cards issued, only 10 were of no value (as judged by the manufacturer). Consequently, there were comparatively few opportunities to (improperly) report these prints to be of value, and few instances (32 of 406 reports) where they did. Here the ratio of correctly assigned "no values" would have been much higher were it not for exercise 89-7 in which 24 laboratory responses, constituting 75% of all disagreeing responses on this question in the nine exercises, were at odds with the manufacturer's judgment that the prints were of no value. The PAC deliberately chose a latent of questionable value and even though the manufacturer stated it was *not* of value, the PAC commended the laboratories for rendering these judgments, stating there was no "distinctly right or wrong answer."

There were 139 latent print cards that were of value and although in only 2% of responses did laboratories mistakenly judge these prints to be of no value, given the 4704 total responses to this question, this amounted to 123 values where there was disagreement. Here, exercise 83-4 stands out, with 9% of responses at odds with the manufacturer's assessment—a figure three times the average for all exercises combined. The PAC offered no other explanation for the high percent on this first exercise other than the lack of universal agreement of an objective standard for judging the value of a print. Judging by the lower rates of disagreement in subsequent tests the results in 83-4 may have been a function of the newness of this type of proficiency test and a tendency for examiners to be overly conservative on the initial exercise.

In terms of identifying the latent prints, only 2% of responses reported an identification when none was present; but, 7% failed to make an identification when one was possible. Throughout the nine tests, there were only 14 instances (out of 885 possible responses) where examiners reported an identification when none was present. Judging from the PAC's reports and comments, it appears a majority of these failures to identify occurred in situations when there were two latent prints on a card capable of identification, but the respondent identified only one of them. In other cases, the PAC called the prints "reasonably difficult" to identify, but that they "should have been identified."

In terms of matching the unknown latents with known inked cards, the laboratories had great success with only .4% of laboratories making responses that disagreed with the manufacturer. The reason for these improper card assignments is speculative, but in many of the reports the PAC suggests they may have been due to carelessness or typographical errors on the part of the examiner. About four times the number of laboratories that misidentified the correct set (card) of inked impressions, failed to assign the proper finger or palm print. It appears, too, that the best explanation for these errors was carelessness in recording the response. The highest rates of improper card and finger/palm assignment occurred on the first two tests (83-4 and 84-5) which, again, may be attributable to the laboratories' unfamiliarity with the testing process. The identification of one or more laboratories/examiners that were contributing a disproportionately high number of the incorrect responses early in the series (as in 84-5, where a single respondent accounted for almost half of all misidentifications) probably also led to remedial/corrective measures in those labs. Other examiners also were undoubtedly more careful in making such assignments in future exercises.

The PAC's comments in other instances provides additional insight as to the reasons for improper/inadequate answers. In test 87-7, two of the laboratories experiencing difficulty reportedly had trainees conduct the tests. In tests 85-7, 86-7, 87-7, 89-7 and 90-7, and 91-8, laboratories were asked to report years of experience of the analyst, percent time devoted to the examination of prints and if the examiner was certified by the International Association for Identification. In 85-7, the PAC concluded that examiners who were certified and devoted 100% of their time to prints performed better, and in 89-7 the PAC cited four laboratories whose examiners were primarily part time, had minimal experience, and lacked certification as those that failed to make "relatively uncomplicated comparison identifications." In 91-8, the PAC concluded experience was not a factor in examiner success in identifying prints. Such relationships were not as clear in the three other tests.

Metals Analysis

There were three metal analysis exercises during the testing period. In two tests laboratories were asked if two samples of metal scrapings could have shared a common origin. In 83-12 they did not and in 90-9 they did. In test 91-10 laboratories were issued samples of lead bullets and were asked if the comparison of those coming from a suspect's residence "matched" those originating from a firing range. The manufacturer stated that they matched those coming from one range but not the other.

Laboratories had good success in qualitatively reporting the presence of major elements present in the samples. In the initial test (83-12), however, the PAC expressed concern over the collective reporting of six elements not contained in the NBS certified samples, and the low percentage of laboratories reporting quantitative results. In 90-9 the PAC was concerned over two labs that detected only the major elements Cu and Zn, and seven laboratories that failed to report Fe and Al, although present in detectable concentrations. Though showing wide variations in methods and quantitative results, most laboratories in 91-10 performed relatively well.

Laboratories were also asked to report methods used to determine results and detail any qualitative and quantitative data developed in their analysis. In the common origin portion of the tests, laboratory results agreed with the manufacturer in 93% of comparisons reported (see Table 3). Two percent disagreed and five percent were inconclusive. Only a single response out of sixty-seven reported comparisons disagreed with the manufacturer's specifications in the first two exercises. In 91-10, two responses improperly excluded bullets from one of the firing ranges.

The rate of successful comparisons in these three tests (93%) is comparable to the rate reported in a single exercise in the earlier LEAA sponsored program.

Footwear Impressions

Laboratories were issued seven footwear impression proficiency tests between 1985 and 1991. No exams of this variety were prepared under the LEAA project. The base number of participating laboratories increased by about 90% during the period. The participation *rate* remained very steady throughout the tests, except for the second-to-last test in which the rate dropped off by about 10 percentage points. This might be attributed to the degree of difficulty of the test, which was substantially greater than the other six.

Scenarios typically involved situations where investigators discovered footwear impressions at the scene of a crime that were photographed and submitted to the laboratory along with the impressions of a suspect's shoes. Laboratories were asked if the evidence impressions had been made by the suspect's shoes. They were also asked what methods they used to conduct their comparisons and the time required to complete the examinations.

The results of the comparative analyses are presented in Table 4. Laboratories offered the results of a total of 1745 comparisons, typically comparing one or two suspect shoe impressions with two or three crime scene impressions. Responses are classified using the system described earlier in which laboratory results either agreed or disagreed with common origin specifications, or were inconclusive. Where a laboratory gave no response to a requested comparison, it was not included in the base number of comparisons made.

Overall, about 87% of the comparisons reported agreed with target values, only 0.7% did not, and 12% were inconclusive. Tests 88-11 and 90-11 lowered the comparison average since only 66

TABLE 3—*Metals analysis.*[a]

Report	Participation Rate	Total Number Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
83-12	18/40 (45%)	18	0	17 (94%)	17	1	1 (6%)	0	0	0 (0%)	0
90-9	49/99 (49%)	49	44	44 (90%)	0	0	0 (0%)	0	5	5 (10%)	0
91-10	40/103 (39%)	80	36	76 (95%)	40	0	2 (3%)	2	2	2 (3%)	0
Total	107/242 (44%)	147	80	137 (93%)	57	1	3 (2%)	2	7	7 (5%)	0

[a]The reader should consult page 1009 before examining this and subsequent tables. Using Report 91-10 as an example, 40 (39%) of the 103 labs receiving samples responded with data. These 40 labs reported on a total of 80 comparisons made between bullets taken from a suspect's residence and batches of bullets taken from two firing ranges. Seventy-six (76) comparisons were in agreement with the manufacturer (36 reporting bullets could have shared a common origin when they did and 40 reporting bullets could not have when they did not). There were two comparisons reported that disagreed with the manufacturer, both stating bullets could not have shared a common origin when, in fact, they did. There were two inconclusives reported in which the bullets were actually of the same origin. All columns are summed and average percentages computed for the three metals tests.

TABLE 4—*Footwear impressions.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
85-13	56/84 (67%)	336	103	326 (97%)	223	0	0 (0%)	0	9	10 (3%)	1
86-13	60/93 (65%)	180	58	177 (98%)	119	0	0 (0%)	0	2	3 (2%)	1
87-11	76/112 (68%)	302	0	282 (94%)	282	0	0 (0%)	0	0	20 (6%)	20
88/11 (resoling case)	56/106 (53%)	112	74	74 (66%)	0	0	5 (4%)	5	33	33 (29%)	0
89-11	79/126 (63%)	237	77	223 (94%)	146	2	3 (1%)	1	1	11 (5%)	10
90-11	91/157 (58%)	478	172	344 (72%)	172	0	0 (0%)	0	67	134 (28%)	67
91-12	100/160 (63%)	100	0	91 (91%)	91	4	4 (4%)	0	0	5 (5%)	5
Total	518/838 (62%)	1745	484	1517 (87%)	1033	6	12 (.7%)	6	112	216 (12%)	104

and 72%, respectively, of the comparisons reported agreed with cited values, and were also responsible for a high percentage of inconclusive responses (29 and 28%). Overall, about 12% of the comparisons were inconclusive.

The first three tests were "fairly easy and straightforward" according to the PAC (test 85-13) and "overall results were excellent." As with other test areas, however, a number of laboratories complained about the absence of actual "knowns" (shoes) with which they could make their own exemplars. Due to this shortcoming, many laboratories stated they could not offer firm conclusions. In test 88-11 the PAC attempted to overcome this limitation by providing the laboratories with a variety of suspect shoe impressions formed with heavy and light inking, as well as ones created with magna powder. In spite of these efforts, laboratories performed poorest on this test, quite likely due to how the evidence prints were created. After the evidence impressions were made, the shoes were *resoled* and then the suspect impressions were made. Many of the labs that made positive identifications specifically cited the heel impressions as the basis for their conclusions. Noting the differences in the foresole impressions, other labs correctly raised the possibility that the shoes had been resoled.

Test 90-11 provided laboratories with two sets of two evidence photographs of faint shoe prints (normal and UV light) from a crime scene and were asked to compare them with photographs of shoe prints from two suspects using fingerprint powder and tape lifts. The evidence photos taken with regular light were generally not suitable for comparison and yielded a very high rate (90%) of "no reports" and inconclusive responses. There were no incorrect responses, however. The final footwear exercise entailed the comparison of photographs of an "unknown" print left at a crime scene with two sets of known impressions. Ninety-one percent of 100 reported comparisons agreed they were not of common origin, five percent were inconclusive and four percent incorrectly reported they could have originated from the same source.

Blood and Body Fluid Analysis

Blood and body fluid analysis was the second largest area of testing during this period of proficiency testing with 28 samples issued to subscribing crime laboratories. The number of laboratory subscribers increased from 75 to 181 from 1978 to 1991, and the number of labs returning data rose from 32 to 104 (for the mixture exercise in 1991). Overall, the rate of participation averaged 45% during the course of testing which did not change appreciably over time. What appeared to affect this rate, however, were the types of scenarios offered. The highest rates of participation (49%) were found for scenarios asking labs to compare bloodstains (for example, 80-2, 82-9, 86-11, and 90-2). Lower rates of participation (44%) resulted when laboratories were asked to compare mixtures of blood and body fluids (81-2, 81-9, 82-3, and others) and laboratories were asked to test for possible associations.

The straight typing exercises were discussed in an earlier article. The results of comparative examinations follow and are grouped in one of three categories:

- a group of tests in which laboratories were given scenarios and asked to type bloodstains and to determine if they could have shared a common source (Table 5).
- two (nonscenario) tests introducing nonblood (semen, saliva) body fluids; and
- a final group of scenarios in which laboratories were issued singular and mixed blood and body fluid stains, were asked to identify and type them, and answer questions of common origin (Table 6).

TABLE 5—*Determining origin of bloodstains (80-2, 82-9, 86-11, 90-2).*

Report	Participation Rate	Total Number Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
80-2	32/75 (43%)	186	28	153 (82%)	125	17	18 (10%)	1	2	15 (8%)	13
82-9	33/71 (46%)	320	151	301 (94%)	150	6	11 (3%)	5	8	8 (3%)	0
86-11	60/123 (49%)	119	54	113 (95%)	59	0	1 (1%)	1	4	5 (4%)	1
90-2	95/177 (54%)	190	88	160 (84%)	72	18	20 (11%)	2	5	10 (5%)	5
Total	220/446 (49%)	815	321	727 (89%)	406	41	50 (6%)	9	19	38 (5%)	19

TABLE 6—*Determining origin of biological and bloodstain mixtures (81-2, 81-9, 82-3, 83-1, 83-9, 84-2, 84-11, 85-11, 89-13, 91-2).*

Report	Participation Rate	Total Number Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
81-2	22/66 (33%)	124	57	101 (81%)	44	18	23 (19%)	5	0	0 (0%)	0
81-9	16/66 (24%)	104	38	51 (49%)	13	7	17 (16%)	10	21	36 (35%)	15
82-3	17/45 (38%)	265	88	220 (83%)	132	15	45 (17%)	30	0	0 (0%)	0
83-1	32/72 (44%)	835	161	769 (93%)	608	36	61 (7%)	25	1	5 (.6%)	4
83-9	36/82 (44%)	477	106	432 (91%)	326	28	45 (9%)	17	0	0 (0%)	0
84-2	41/95 (43%)	53	31	50 (94%)	19	0	1 (2%)	1	0	2 (4%)	2
84-11	38/104 (37%)	408	97	379 (93%)	282	28	29 (7%)	1	0	0 (0%)	0
85-11	51/120 (43%)	153	43	113 (74%)	70	30	38 (25%)	8	0	2 (1%)	2
89-13	84/177 (47%)	679	192	456 (67%)	264	78	89 (13%)	11	34	134 (20%)	100
91-2	104/181 (57%)	204	93	183 (90%)	90	1	5 (2%)	4	5	16 (9%)	11
Total	441/1008 (44%)	3302	906	2754 (83%)	1848	241	353 (11%)	112	61	195 (6%)	134

Determining the Origin of Bloodstains

A group of blood and body fluid tests, 80-2, 82-9, 86-11, and 90-2, presented bloodstains to laboratories within the context of a scenario. Laboratories were typically asked to group the stains provided and to determine if one or more stains could have originated from the same source as known blood samples taken from suspects and/or victims. Test 88-14 is excluded from this tabulation and discussion because the test proved confusing in that it involved a fetal blood sample pooled from several donors that elicited many inconclusive responses. With respect to straight typing of these samples, laboratories were correct in about 99% of their results.

Overall, laboratories made a total of 815 comparisons in these four exercises and offered results in agreement with the manufacturer 89% of the time. They disagreed in 6% of comparisons and filed inconclusive results in 5% of cases. Whereas laboratories were showing distinct improvement over the years (from 1980 to 1986), the 90-2 results proved disappointing. About 80% of the responses in the disagree column improperly included parties when they should have been excluded, and three-fourths of these occurred in the first and final exercises. In the 1980 exercise, laboratories were asked to determine if two questioned stains from crime scenes could have originated from any of three different suspects. Mistaken conclusions resulted principally not from the failure of laboratories to correctly type the stains in the systems they employed, but from their failure to employ enough systems (GLO I, EsD, EAP and Hp) that would have distinguished among the samples.

In 90-2, most of the problems were due (again) to the failure to employ two systems (GLO and EAP) that would have discriminated the samples. Of the 18 laboratories that failed to exclude one of the suspects, almost three-quarters failed to perform either GLO or EAP. Two improper exclusions of another suspect were due to faulty typing. Overall, though, of the practically 2000 typing results reported in this exercise, less than 1% were inconsistent with the manufacturer's report.

Laboratories showed great improvement in these exercises compared with earlier test results published in the LEAA report. Test #8 in the LEAA study proved disastrous where about 70% of laboratories were unable to distinguish between two type O stains from different individuals because they had not begun employing isoenzyme and serum protein tests, which had been introduced into U.S. crime laboratories about five years prior to the issuance of this test. This situation changed dramatically over the next ten years to the point where laboratories made correct determinations in almost 90% of the bloodstain comparisons they attempted.

Nonblood Body Fluids

The next series of exercises involved nonblood body fluids, and we begin first with a review of two tests (80-9 and 87-13) that issued laboratories stains with no scenarios. In 80-9, laboratories were issued four "suspected physiological fluid" stains and laboratories were asked to examine them. All stains contained semen, but one consisted of a mixture of semen and saliva. All samples contained normal spermatozoa except for one which originated from a vasectomized donor and consequently was aspermic. Generally, all laboratories detected semen in all samples; only 13/24 laboratories searched for spermatozoa, however, and all that did correctly reported finding none in the sample from the vasectomized donor. Ten of 24 laboratories documented testing for amylase and 60% reported high levels in the sample containing saliva; other results were either negative or inconclusive. Three laboratories reported incorrect ABO grouping results, one laboratory misreported PGM results on a single stain, and another mistakenly reported a nonsecretor stain as a secretor.

Compared with the earlier physiological fluid test (#13) reported in 1978, the success of laboratories in identifying semen and saliva stains was comparable; that is, laboratories had little difficulty identifying semen and only modest success (in the 50 to 60% range) in identifying saliva.

In 87-13, laboratories were issued five stains and were asked to group them, plus to determine if the donor of one stain could be ruled out as the source of the other four. All stains could be ruled out except for one. One set of blood and saliva stains were from one person, and another set (blood, saliva, semen) were taken from another. Fifty-four laboratories responded with data and all correctly identified the stains in all samples, except for two that did not identify either saliva stain. About 97% (968/1000) of the grouping results agreed with the manufacturer with about half the 32 errors occurring in the GLO system. Also, about half the errors were reported by a single laboratory. The PAC had designed the exercise to test the proficiency of labs in performing PGM typing and laboratories did very well. Fifty of the 54 laboratories attempted PGM typing and only one reported (2) improper results. Only 6 of the 54 laboratories (21 results) attempted Lewis typing (secretor status), usually (18/21) on the semen or saliva stains, and the results were poor. Third-eight percent of the results were correct, 29% incorrect, and 43% inconclusive.

In terms of answering the common origin question, laboratories did not fare well, with only 77% of the 206 comparisons in agreement with the manufacturer: 13% of the results disagreed and another 10% inconclusive. The laboratories that failed to exclude two of the samples did so due to limited employment of systems other than ABO, and those that failed to exclude the saliva stain had not employed Lewis typing.

Body Fluid Mixtures Common Origin

The final group of ten tests involved scenarios with nonblood body fluids (NBBF), and mixtures of NBBF and blood, accompanied by questions as to the source of various stains. The exercises typically asked laboratories to identify several stains, to determine their species of origin, to group them, and determine if each could have originated from samples provided by victims, spouses or suspects. Typically, one or more stains in the exercises involved the mixture of blood, semen or saliva stains taken from more than a single individual (in test 81-9, breast milk was introduced as well).

On the ten tests where it was possible to compute an average, there were a total of 3302 possible source responses and approximately 83% of the conclusions agreed with manufacturers' values. Eleven percent of the comparative responses disagreed with the manufacturers' specifications and either mistakenly included a party as a possible donor of a stain, or improperly excluded a party. Inappropriate inclusions outnumbered improper exclusions by a margin of about 2 to 1. The remaining 6% of responses were inconclusive. These results are presented in Table 6. Other observations on the structure and results of these tests follow.

In 81-2, stains from a bedsheet included a mixture of the complainant's blood and the suspect's semen, and the suspect's blood and that of a third person; laboratories had to go beyond ABO grouping to correctly distinguish the stains. On this test, laboratories had the second highest percent of results that disagreed with the manufacturer largely because they did not employ an adequate number of grouping systems to distinguish samples with the same ABO type.

Test 81-9 included a mixture of semen and saliva from different donors as well as other blood, semen and breastmilk stains; the test was designed so that parties could be differentiated only on the basis of secretor status. Laboratories experienced great difficulty in determining secretor status and as a result this exercise produced the lowest percent (49%) of acceptable results, and the highest percent (35%) of inconclusive results.

Test 82-3 attempted to simulate a real world case and included a saliva stain, mixtures of blood from two donors and blood and saliva from the same source, and a bloodstain on an article already containing background ABO activity. Laboratories needed to be

able to detect and type saliva, employ a full range of typing procedures, and properly consider controls and possible background contamination. Failure of some labs to do the preceding resulted in a high percent of improper conclusions, mostly where they inappropriately excluded stains as being of a possible common source (particularly a sample containing a mixture of diluted saliva from one suspect and the blood of another suspect).

Test 83-1 again challenged laboratories with a realistic case through introduction of mixtures of stains, the presence of background contamination, and the need for selection of controls for determining secretor status. While 94% of the blood grouping results were correct, laboratories experienced the greatest problems in the "longest standing system"—ABO testing: nonetheless, laboratories provided a high (93%) rate of conclusions that agreed with the manufacturer.

Test 83-9 once again focused on the need to determine secretor status to distinguish stains and the importance of ABH background activity in interpreting physiological fluid stains. Similar to the previous test, laboratories performed correct grouping results more than 96% of the time, and more than 80% of errors were with ABO results. Laboratories also experienced problems in interpreting their results as related to determination of source. Nonetheless, the percent of proper conclusions of source again exceeded 90%.

In Test 84-2, laboratories were presented with a blood, a saliva and a (aspermic) semen stain, but no mixtures. Laboratories performed their blood and semen analyses very well, but had great difficulty with the saliva stain where *no* laboratories correctly detected the saliva. About 94% of the source attributions were on target.

Test 84-11 was a relatively "uncomplicated" exercise designed to test the reliability of electrophoresis grouping on dried stains and involved three bloodstains and a single saliva stain from a nonsecretor. All stains originated from type O persons, but were distinguishable using other systems (for example, Hp, EAP, PGM). The report indicated that in excess of 97% of the grouping results were correct and that three labs were responsible for more than half of the incorrect responses. The majority of errors occurred in the Hp and Rh systems.

Test 85-11 was a test designed to emphasize the importance of PGM subtyping; it presented laboratories with a bedsheet containing a semen stain, and blood and saliva from the complainant, her boyfriend and the suspected assailant. Laboratories did not perform well on this exercise, with 25% of source conclusions at odds with the manufacturer. The great majority of these inappropriate responses were attributable to the failure of laboratories to perform PGM subtyping, which would have excluded the boyfriend.

Test 89-13 had the objective of demonstrating the value of Lewis typing in distinguishing dried saliva stains. Laboratories were issued three "evidence" cigarette butts that had been dipped in saliva, and saliva and blood samples from each of three "suspects." Laboratories were asked if the cigarette butts could have been smoked by any of the suspects. In fact, two of the butts had been dipped in saliva from one suspect (a type A secretor) and the remaining butt had been dipped in saliva from another suspect, an A *non*secretor. Laboratories were also asked to provide grouping data derived from the cigarette, saliva and blood samples. Only 67% of the comparison results properly distinguished the saliva stains on the three butts; 13% were not accurate, and 20% were inconclusive (a rate only exceeded by the results of test 81-9). As with test 81-9, determination of secretor status was crucial for a successful answer to the problem posed, particularly since the saliva of one of the suspects exhibited a high level of A substance, even though originating from a nonsecretor.

Only eight laboratories performed Lewis typing on the stains, with 58% of the responses correct, and the remaining responses equally divided between incorrect and inconclusive categories. In addition to these secretor status problems, laboratories also did not perform basic ABO tests on the saliva samples very well. ABO typing of the saliva on the butts yielded correct responses 76% of the time, while ABO typing of the "known" saliva on paper substrates was correct in 79% of responses. ABO typing of blood samples on tissue paper yielded correct responses in more than 94% of responses—less than 2% of these responses disagreed with the manufacturer's values. The typing of the bloodstains in additional systems was generally performed well with 96 to 98% of PGM, AK, and ADA results correct, 90% of EAP and Hp results on target, but only between 83 to 88% of PGMsub, EsD, and GLO results reported correctly.

Unlike earlier proficiency exercises, laboratories had a substantially lower error rate with their EAP and Hp typing, with fewer than 1% of their results in error. However, problems arose in PGMsub and ESD typing (11% and 4% error rates respectively) where in earlier tests errors in typing were closer to the 1% level.

The determination of secretor status was necessary to identify the donor of saliva on the cigarette butts. Of the 78 improper comparisons, 59 (76%) could have been avoided had the proper secretor status been correctly determined.

Laboratories performed much better on test 91-2 than they had in previous physiological fluid exercises. In this test, laboratories were asked to determine if the source of a physiological fluid sample originated from the suspect or complainant. Only 2% of responses disagreed with the manufacturer, 90% agreed and 9% were inconclusive. Most laboratories successfully identified amylase (or saliva) on the questioned swab and performed the typing properly. Less than 2% of the typing results were incorrect.

Glass Analysis

There were a total of eleven glass tests issued during the project period. Over this time, the number of laboratories participating in the tests grew from 63 to 148. The participation rate averaged 51% over the eleven tests, ranging from a low of 26% on test #19 to a high of 58% on tests 83-10 and 91-16. Generally, the rate of participation increased over time. The tests covered a variety of glass types, including bottle, optical and float glass. Samples were compared using various tests, among them: UV light, refractive index, density, and elemental analysis.

Most of the exercises required laboratories to compare two or more glass chips and to determine if they could have shared a common origin. Laboratories were asked to indicate the procedures employed in their examinations, their sequence, the information developed and, in tests 88-13, 89-14, 90-14, and 91-16, the time required to complete the examination. The PAC introduced a "search parameter" in test 84-13 in an effort to make the test more realistic by asking laboratories to search for foreign materials on a glove and to compare any glass particles found with a control glass sample.

As with earlier exercises, we tabulated the number of comparisons made by laboratories and the number of inclusions and exclusions that agreed/disagreed with manufacturer specifications (Table 7). Only nine of the eleven test reports listed data in such a way as to permit this type of analysis. Overall, laboratories reported proper comparisons about 85% of the time. Laboratories filed

TABLE 7—*Glass.*

Report	Participation Rate	Total Number Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
10	24/63 (43%)	54	15	42 (78%)	27	0	10 (19%)	10	2	2 (3%)	0
19	20/78 (26%)	58	18	41 (71%)	23	15	17 (29%)	2	0	0 (0%)	0
82-10	31/59 (53%)	29	0	29 (100%)	29	0	0 (0%)	0	0	0 (0%)	0
83-10	34/59 (58%)	34	29[a]	29 (85%)	0	0	5 (15%)	5	0	0 (0%)	0
84-13	40/84 (48%)	40	37	37 (93%)	0	0	3 (8%)	3	0	0 (0%)	0
88-13	46/92 (50%)	92	43	89 (97%)	46	0	3 (3%)	3	0	0 (0%)	0
89-14	75/117 (64%)	150	74	149 (99%)	75	0	1 (1%)	1	0	0 (0%)	0
90-14	68/133 (51%)	136	52	59 (43%)	7	53[b]	63 (46%)	10	6	14 (10%)	8
91-16	86/148 (58%)	172	86	172 (100%)	86	0	0 (0%)	0	0	0 (0%)	0
Total	427/833 (51%)	765	354	647 (85%)	293	68	102 (13%)	34	8	16 (2%)	8

[a]Includes 4 inconclusive responses considered acceptable.

[b]Although these 53 conclusions disagreed with the manufacturer's statement of different sources, the PAC nonetheless called them "valid." Excluding this exercise, laboratories averaged 93% agreement on the other eight exercises. See text for discussion.

conclusions that disagreed with the manufacturer in 13% of their comparisons and inconclusives in about 2%. Except for test 90-14, the majority of results that disagreed with the manufacturers' specifications involved laboratories excluding samples as possibly having common origin that should have been included. Test 90-14 yielded very different results that will be explained as follows.

Test 19 proved to be a major challenge to many laboratories, producing a high percent of disagreement (29%). The challenge was to determine if three glass samples could have shared a common origin. All the samples of window glass had been manufactured by the same company using the same process, with two made at the same plant, at the same time, but the third manufactured at a different plant five years later. Ninety percent of the labs properly concluded the two samples could have shared a common origin, but about 40% improperly stated the third could also have shared a common origin. While all three samples proved indistinguishable using UV fluorescence and refractive index, several laboratories did find differences in their density and elemental content. The report of the referee laboratory determined the presence of barium and lithium in the first two samples but not the third.

Laboratories had good agreement on their refractive index measures in test 82-10, with 100% concluding the samples could not have shared a common origin. Some laboratories commented the test was "too easy." The PAC, however, was critical of laboratories' UV light results, observing that "defects in either procedure or observations persist." In test 83-10 the objective was to see if laboratories would find significant differences in glass samples taken from opposite ends of the same sheet of window glass. The scenario called for glass in a broken store front window to be compared with glass fragments found in the vehicle of a suspect. While the great majority of respondents concluded the samples could have shared a common origin, 15% reported they could not. Four of the five laboratories reaching this conclusion apparently based it on differences in thickness of the two samples, which would not be totally unexpected in glass broken from different locations of a large piece of glass.

In test 84-13, two pieces of window glass and two pieces of optical glass were placed on a glove; laboratories were given the glove and a comparison piece of glass from the same window. Laboratories were asked to locate foreign materials on the glove and determine if any glass found could have shared a common origin with the sample taken from the window. Ninety-three percent of the respondents agreed with the manufacturer and concluded there were one or more pieces of glass on the glove that could have shared a common origin. For simplicity purposes, Table 7 tabulates only one comparison per laboratory although, in actuality, the number of comparisons varied as a function of the number of particles found during the searching stage. The three laboratories that stated the samples could not have originated from the broken window only found one or two glass fragments on the glove and, therefore, possibly missed the fragments of similar origin and based their conclusions on the samples that were from a different source.

Although a high percentage of laboratories (97%) agreed with the manufacturer's specifications in test 88-13, the PAC nonetheless found practices needing attention. The PAC noted that the three laboratories that disagreed (see Table 7) based their conclusions on

differences in chemical composition/concentration of the samples. This and the overall wide variation of elements found in samples suggests laboratories need to devote greater attention to such procedures. Laboratories, as a group, also failed to show consistency in conducting and reporting results of UV fluorescence tests. The PAC noted with approval the fewer number of laboratories reporting refractive index measures to only three decimal places (compared with earlier tests) and introduced a graphical (YOUDEN) analysis to display how well a specific laboratory's results agreed with the total body of results reported. Three laboratories were found to be making systematic errors in their refractive index tests and would experience difficulty in using published refractive index data on glass populations.

A high percentage (99%) of results reported in 89-14 were in agreement with the manufacturer's specifications that two of three glass samples could have shared the same origin. Although only a single laboratory improperly excluded one of the samples, an interlaboratory comparison of refractive index measurements identified nine laboratories that produced systematic errors (either high or low). While not leading to improper conclusions in this exercise, such errors have the potential of not detecting differences in samples and difficulties in using published data on refractive indices.

Test 90-14 issued three mirrored glass samples and laboratories were asked if either of two samples could have the same origin as another. In fact, two samples were from the same mirror and the third from another. All, however, were produced by the same manufacturer, using the same process, two-to-three months apart, and possessed very similar refractive indices (differing only in the fifth decimal place), and density and elemental properties that were indistinguishable. This case is an excellent example of the dilemma faced by forensic laboratories in differentiating some types of mass produced products which, although technically originating from different sources, nevertheless possess very similar properties and cannot always be discriminated using conventional testing procedures.

In 91-16, laboratories recovered with 100% agreeing in their examination of three window glass samples, two of which shared a common origin and a third that did not. This was an easier exercise (a physical match was possible with two fragments) and although refractive index measures were generally performed well, the PAC recommended that nine laboratories check their analytical procedures.

When the results of these tests are compared with those of glass examinations (4 and 9) in the earlier LEAA study, we see rates of agreement are comparable (85%). The percent of inconclusive responses dropped by a percentage point. Qualitatively, the tests issued in the CTS program were substantially more challenging than those issued in the earlier LEAA study.

Firearms

Between the years 1978 and 1991, a total of fourteen firearms tests were issued to crime laboratories participating in the proficiency testing program. Over the span of this period the number of laboratories subscribing to this battery of tests increased more than four-fold: from 42 to 173. Two tests (83-3 and 87-3) are not included in the following discussion because they were limited to identification of ammunition components and did not involve determinations of common origin.

Most (12) of the tests issued involved scenarios in which laboratories were asked to compare (known) test fired bullets and/or cartridge cases with (evidence) projectiles found at the scenes of crimes. In three of the tests (80-4, 81-5, and 82-5) examiners were given both a set of bullets and cartridge cases, and determinations of common origin were asked for each set. One (91-4) asked laboratories if any of five sets of three cartridge cases were fired from the same weapon. Two scenarios (81-5 and 90-3) also asked laboratories to make a target/muzzle distance-of-fire determination. Three additional samples asked laboratories to examine ammunition components in order to tell investigators as much as possible about the type of ammunition, manufacturer, caliber, and so forth.

For the tests requiring laboratories to undertake a comparative analysis of test fired and evidence bullets and cartridge cases, the results appear in Table 8. Fifteen sets of comparisons were tabulated, including the three exercises which included both bullets (B) and cartridge cases (C). Examiners generally did very well in making the comparisons. For all fifteen tests combined, examiners made a total of 2106 comparisons and provided responses which agreed with the manufacturer responses 88% of the time, disagreed in only 1.4% of responses, and reported inconclusive results in 10% of cases. In several tests, the inconclusive responses may be appropriate because, as in test 85-3, some laboratories will not "positively exclude" a projectile in the absence of the firearm (as was the case in these proficiency tests). The very high percent (69%) of inconclusives in test 84-3 was in part the result of the difficulty of the exam, but also the policy followed in certain labs not to positively exclude projectiles without having the firearm.

In ten exercises, more than 90% of comparisons made by examiners agreed with the manufacturer. In one exercise (84-3), where none of the test fired projectiles matched the evidence projectile, only 29% of the comparisons properly excluded all four test fires. Almost 70% of the comparisons in this exercise resulted in "inconclusives." Though none of the other exercises compared with 84-3, there were a substantial fraction of inconclusive replies (10%) throughout all the tests. More than three-quarters of these inconclusive responses resulted where laboratories failed to report a nonmatch where none existed (as with 84-3). The percentage of inconclusives seemed to be a function of the difficulty of the test, the clarity of instructions in the test scenario, and the discomfort expressed by some labs in not being supplied the actual weapons with which they could conduct their own test firings.

In those exercises where the time invested on an examination was requested, examiners arriving at an incorrect result generally spent *less* time on the exam than persons who arrived at a correct answer (see 84-3). Except for exercises 89-3 and 90-3, it appears the examiners generally did better as time went on—more than half of the improper responses reported up through test 88-3 occurred in the first three exercises. Laboratories experienced considerable difficulty with the shotgun shell exercise (89-3), producing eight replies which disagreed with the manufacturer, and 90-3, yielding seven errant replies, and together representing about half of the incorrect comparisons in all fifteen tests. Even if 89-3 and 90-3 were excluded, we cannot necessarily infer that examiners were getting more proficient, however, since we can't be certain how the difficulty of the latter tests compared with the difficulty of the earlier ones. Still, the 88-3 test, where examiners made no errors, was considered one of the most challenging.

For the four tests asking for qualitative information, the laboratories performed very well. All laboratories (17) reporting the target to muzzle distance in 81-5 produced results in conformity with the manufacturer's specifications. Approximately one-quarter of these, however, reported distances that were at the outer limits of acceptability. In addition, about one-quarter of the laboratories did not attempt or report results on this test. In 90-3, while about

TABLE 8—*Firearms.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
6	24/42 (57%)	168	45	161 (96%)	116	0	1 (.5%)	1	2	6 (3%)	4
16	22/57 (38%)	66	21	63 (95%)	42	2	3 (5%)	1	0	0 (0%)	0
80-4	35/72 (49%)	B) 105	30	82 (78%)	52	0	3 (3%)	3	2	20 (19%)	18
		C) 35	23	23 (66%)	0	0	2 (6%)	2	10	10 (29%)	0
81-5	21/79 (26%)	B) 42	42	42 (100%)	0	0	0 (0%)	0	0	0 (0%)	0
		C) 84	80	80 (95%)	0	0	0 (0%)	0	1	4 (5%)	3
82-5	24/52 (46%)	B) 24	0	17 (71%)[a]	17	2	2 (8%)	0	0	5 (21%)	5
		C) 24	22	22 (92%)	0	0	0 (0%)	0	2	2 (8%)	0
84-3	36/69 (52%)	134	0	40 (29%)	40	2	2 (1%)	0	0	92 (69%)	92[b]
85-3	50/81 (64%)	150	95	137 (91%)	42	0	1 (.7%)	1	4	12 (8%)	8[b]
86-3	56/93 (66%)	168	51	159 (95%)	108	0	0 (0%)	0	5	9 (5%)	4[b]
88-3	66/123 (53%)	187	56	154 (82%)	98	0	0 (0%)	0	9	33 (18%)	24
89-3	79/140 (56%)	223	70	204 (91%)	134	3	8 (3%)	5	3	11 (5%)	8
90-3	86/153 (56%)	172	81	159 (92%)	78	3	7 (4%)	4	1	6 (3%)	5
91-14	99/173 (57%)	524	289	517 (99%)	228	0	0 (0%)	0	4	7 (1%)	3
Total	598/1134 (53%)	2106	905	1859 (88%)	954	12	29 (1.4%)	17	43	218 (10%)	175

[a]Includes eight inconclusives considered correct.
[b]Result of lab policy, may be considered correct.

90% of the mean target value estimates were within a "reasonable" range of 6″–12″ (9″ was the actual muzzle-to-target distance), in about half these replies one of the estimates (near or far) was outside the reasonable range. The laboratories performed very well on the two tests (83-3 and 87-3) where they were asked to supply information about the ammunition. About 95% of the laboratories supplied the desired information in 83-3, and 92% supplied complete descriptions in test 87-3.

The performance of laboratories in the firearms tests was comparable to that under the earlier LEAA study, although the rate of successful identifications actually was slightly lower—88% vs 91%. Laboratories cut the rate of errant identifications by half (3% to 1.4%) but the rate of inconclusive responses doubled, from 5% to 10%. This substantial increase is primarily attributable to a single test (84-3), which accounted for almost half of all inconclusive responses recorded throughout nine tests. Were it not for this exam the rate would have been about 6%.

Paint Analysis

Between 1978 and 1991 a total of 18 paint tests were issued to crime laboratories. The results of the first three tests (#6, #8 and #15) were presented in such a way as to preclude a summary analysis. The number of laboratories subscribing to the tests increased two and one-half fold (from 67 to 164) over this period and the number of laboratories responding with data also more than tripled (from 31 to 96). As with other testing categories, the participation rate increased over time, with 38%

of laboratories responding with data for the first seven tests and 54% submitting results in the final eight tests, for an overall average of 49%.

For the fifteen tests included in this summary, ten (10) consisted of automotive paint coverings and five (5) were interior household paint samples. Laboratories were typically issued two or three paint samples and asked if two samples could have shared a common origin or, in the case of exercises with three samples, if either or both of two samples could have shared common origin with a third. Laboratories were also asked to indicate the methods and sequence of usage and, in latter tests, the specific information developed from the methods leading to answering the question of common origin. In some exercises laboratories were given samples having identical primers but topcoats from different suppliers. In one test of automotive paints (86-4) laboratories were initially asked to supply as much information as possible about the suspect vehicle that had presumably yielded the questioned paint chips.

As with earlier samples, the number of comparisons have been tabulated as was the number of agree, disagree and inconclusive determinations. The results of the automotive paint comparisons are presented in Table 9. The majority (about $^3/_4$) of the tests involved samples which, although similar in appearance and other respects, were actually of different origin. When conclusions differed from the manufacturers' specifications, therefore, chances were they would be errors of mistaken inclusion, rather than exclusion. Overall, laboratories were in agreement in 79% of the comparisons reported and disagreed 18% of the time. Inconclusive results accounted for only 3% of the responses. Were it not for test 90-1, the average percent of responses in agreement with the manufacturer would have been about the same (or 85%) for automotive and household paints; however, the 90-1 results pulled down the overall automotive paint results by about 10 percentage points. There was also greater variability among automotive paint results.

For example, in four automotive paint tests, in excess of 92% of comparisons agreed with the supplier. However, in another five tests (21, 80-7, 84-6, 86-4, and 90-1) fewer than 75% of the comparisons were in agreement. The majority of the errors reported in these tests were in scenarios where laboratories reported inclusive results, stating samples could have had a common origin when, in fact, they did not.

Test 90-1 illustrates well the dilemma facing forensic laboratories in trying to distinguish among materials covered with paint from different batches off the same production line. Quality control procedures in place in the present day paint industry makes detection of small differences extremely challenging. While the paint samples in this exercise were from different batches (and therefore different origin) they were of virtually identical formulation. More than three-quarters (78%) of the responses mistakenly concluded the two samples could have come from the same automobile. The PAC concluded that while the "correct" answer was "no," many of the laboratories reporting yes or inconclusive responses would be considered "acceptable."

Among the problems noted by the Project Advisory Committee leading to erroneous conclusions were 1) inadequate test selection, 2) faulty solubility determinations, and 3) poor pyrolysis gas chro-

TABLE 9—*Automotive paint.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
21	20/85 (24%)	20	14	14 (70%)	0	0	4 (20%)	4	2	2 (10%)	0
80-7	25/82 (30%)	50	25	37 (74%)	12	11	12 (24%)	1	0	1 (2%)	1
80-10	32/82 (39%)	64	0	62 (97%)	62	2	2 (3%)	0	0	0 (0%)	0
81-3	29/67 (43%)	58	0	55 (95%)	55	3	3 (5%)	0	0	0 (0%)	0
81-11	26/67 (39%)	26	0	24 (92%)	24	2	2 (8%)	0	0	0 (0%)	0
82-2	30/59 (51%)	60	0	59 (98%)	59	1	1 (2%)	0	0	0 (0%)	0
84-6	49/64 (52%)	98	0	72 (73%)	72	24	24 (24%)	0	0	2 (2%)	2
86-4	54/120 (45%)	46	31	31 (67%)	0	0	14 (30%)	14	1	1 (2%)	0
88-4	73/143 (51%)	146	67	124 (85%)	57	14	19 (13%)	5	1	3 (2%)	2
90-1	96/172 (56%)	96	0	16 (17%)	16	75	75 (78%)	0	0	5 (5%)	5
Total	434/941 (46%)	664	137	494 (74%)	357	132	156 (23%)	24	4	14 (2%)	10

TABLE 10—*Household paint.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
83-5	33/70 (47%)	66	0	54 (82%)	54	11	11 (17%)	0	0	1 (1%)	1
85-5	54/102 (53%)	108	0	90 (83%)	90	11	11 (10%)	0	0	7 (6%)	7
87-4	69/121 (57%)	136	0	113 (83%)	113	16	16 (12%)	0	0	7 (5%)	7
89-1	65/138 (47%)	64	51	51 (80%)	0	0	10 (16%)	10	3	3 (5%)	0
91-1	96/164 (59%)	192	81	173 (90%)	92	2	13 (7%)	11	4	6 (3%)	2
Total	317/595 (53%)	566	132	481 (85%)	349	40	61 (11%)	21	7	24 (4%)	17
Grand Total (Auto plus household)	751/1536 (49%)	1230	269	975 (79%)	706	172	217 (18%)	45	11	38 (3%)	27

matography and/or IR testing and interpretation. Test 84-6 is a good example where in this test of three auto paints, 24% of the comparisons resulted in improper inclusions. In 83-5, (Table 10) laboratories were given three samples of similar white latex interior paint, but all of a different origin. Two of the samples, however, were very similar and about one-third of the labs mistakenly reported a possible common source for these two. Test 88-4 challenged laboratories to analyze both the clear top coat and base coat in an automotive paint exercise. Inadequate test selection and failure to perform pyrolysis GC were the primary reasons for the majority of erroneous inclusions. In 86-4, two samples were taken from different locations on the same repainted vehicle in a wrecker's yard. Thirty percent of comparisons were not in conformance with manufacturer's specifications; that is, they concluded the samples could not have originated from the same vehicle.

Laboratories were warned that microscopic and solubility tests were primarily screening tests, usually required confirming analyses, and were not adequate for differentiating among samples with small quantitative differences. The solubility tests keyed on the selection of proper solvents and results sometimes were misinterpreted and other times led to proper conclusions but for the wrong reasons. It also became clear that pyrolysis GC was an indispensable instrument in making many of the requested analyses, but that proficiency with this tool varied widely, A number of laboratories erroneously reported samples to be of common origin either because they failed to employ pyrolysis GC or, if they did, did not apply the technique properly or did not interpret the data correctly. It became clear to the PAC that labs employed different decision criteria "in the significance attributed to small differences between paint samples" (87-4).

Test 89-1 provided an interesting and challenging housepaint exercise in which two paint chips of common origin were to be compared, but where one of the chips had been exposed to the out-of-doors for $3^1/_2$ years, and the other had been masked during this exposure. Eighty percent of the laboratories were able to account for the differences in the surface characteristics of the samples and concluded they were of common origin. In spite of test results indicating likely common origin, 20% concluded they were either not of common origin or reported inconclusive findings.

Overall, paint results in this period of testing were no better than the performance of laboratories on the paint tests in the LEAA study. Combining house and automotive results, laboratories achieved success in about 79% of their comparisons in the new tests, compared with 81% under the old.

Fibers

Subscribers were issued a total of fourteen fiber tests during the period covered by this review (culminating in 91-7). Five of the tests (7, 83-6, 86-8, 88-10, and 89-6) asked laboratories to identify fibers, four requested that laboratories perform comparative examinations (11, 17, 90-6, and 91-7), and five (22, 80-6, 84-4, 85-4, and 87-10) asked that laboratories do both. In the first comparative analysis (11), no individual laboratory data were reported and the results are not included in Table 11. The number of laboratories subscribing to the tests rose from 67 in the early years (1978) to 199 in 1991, which represents almost a three-fold increase. The percentage of subscribing laboratories that actually responded with data increased over time, averaging 46% for the years covered.

Beginning with test 22, many comparative exercises not only posed a common origin scenario—could fiber w have a common origin with fibers x, y or z—but laboratories were also asked to identify the fibers. As with earlier comparative exercises, we first calculated the total number of comparisons reported by laboratories and then computed the percent of comparisons which agreed, disagreed, and were inconclusive. Where laboratories failed to report the results of a particular comparison, such reports were not included in the base number of comparisons made.

For all tests combined, examiners made a total of 925 comparisons and reported results that agreed with the manufacturer 83% of the time. Laboratories were at odds with 11% of comparisons and reported inconclusive results in 6%. These tests are also noteworthy in that about three-quarters of the improper comparisons

TABLE 11—*Fibers.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
17	27/71 (38%)	54	22	38 (70%)	16	9	12 (22%)	3	2	4 (8%)	2
22	22/71 (31%)	44	0	27 (61%)	27	16	16 (36%)	0	0	1 (3%)	1
80-6	30/82 (37%)	90	26	72 (80%)	46	13	14 (16%)	1	3	4 (4%)	1
84-4	52/93 (56%)	104	0	102 (98%)	102	2	2 (2%)	0	0	0 (0%)	0
85-4	48/105 (46%)	191	35	143 (75%)	108	21	27 (14%)	6	9	21 (11%)	12
87-10	54/130 (41%)	162	88	142 (88%)	54	0	11 (7%)	11	9	9 (5%)	0
90-6	94/175 (54%)	94	0	84 (89%)	84	9	9 (10%)	0	0	1 (1%)	1
91-7	93/199 (47%)	186	161	161 (87%)	0	0	8 (4%)	8	17	17 (9%)	0
Total	539/1184 (46%)	925	332	769 (83%)	437	70	99 (11%)	29	40	57 (6%)	17

reported fibers could have been of common origin when they were not. Laboratories had much more difficulty with the six exercises issued in this testing compared with the single fiber examination (#12) reported in the 1978 study. Clearly, the exams in the 1978–91 period were more difficult than the relatively easy exam issued in 1976, which resulted in a 99% correct comparison rate. The percent of acceptable responses ranged from a high of 98% in test 84-4 to a low of 61% in test 22. In 84-4, the PAC observed that the fibers in question were rather easily differentiated by their microscopical appearance. Exercise 22 posed a much greater challenge as indicated both by the low response rate (31%) and high percent of improper comparisons (36%). In this exercise, the questioned fiber was actually used in the manufacture of one of the two unknown fibers, but its length and crimp were changed in the process. In test 17 which had the next highest percent of comparisons which disagreed (22%), the fibers were of similar composition but had different cross sectional shapes, which many laboratories failed to detect.

Prompted by wide variations in the types of tests performed and the analytical data reported, the program focused on the analytical procedures employed by labs in the course of making comparative determinations. In test 86-8, wide variations were noted in the melting points and refractive index measurements, indicating some laboratories had not properly calibrated their instruments. Even measurements of diameters of fibers were not uniformly consistent. The PAC also noted variation in the terminology used by laboratories in describing the same phenomena. Other times laboratories seemed to be unclear as to the condition they were observing. In exercise 87-10 about 10% of the fibers were misidentified as a result of difficulties in interpreting flame, solubility and RI tests. Several laboratories incorrectly excluded fibers as having common origin due to difficulties in interpreting pyrolysis gas and thin layer chromatography results and macro and microscopic appearances. The PAC did note that laboratories' refractive index and melting point results were closer to target values than in the previous test.

Problems in performing melting point analyses were noted in test 90-6; most of the laboratories that failed to distinguish the fibers could have done so had they performed this particular test correctly. Even some of the respondents who properly reported the fibers were not of common origin gave melting point data outside the range of acceptability. In 91-7, only 4% of responses mistakenly reported the fibers were not of common origin and here it appeared these laboratories placed too much emphasis on slight differences in microscopical observations. The PAC stressed the need to employ multiple techniques (microscopical, chemical, instrumental) in making comparisons of fiber evidence.

Hair

As noted in an earlier section on identification of species of hair, there were a total of eight hair proficiency tests issued to crime laboratories. There were five exercises in which laboratories were asked to answer questions regarding the origin of hair samples. There were no comparable exercises in the LEAA study. Throughout the testing the PAC warned readers that they needed to employ particular caution in interpreting the hair results given the virtual impossibility of achieving complete sample homogeneity. The PAC acknowledged the "insufficient samples" issued to laboratories and the variability of individual hairs taken from the same source. In many cases the PAC thought that inconclusive results may have been the proper answer even though such a response may have been at odds with the manufacturers' information. We have tabulated the following (see Table 12) inclusions, exclusions and inconclusives based upon the data contained in the individual test reports. The reader should read these results

TABLE 12—*Hair.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
81-6	21/79 (27%)	105	17	82 (78%)	65	12	14 (13%)	2	2	9 (9%)	7
85-6	52/109 (48%)	451	80	319 (71%)	239	14	41 (9%)	27	45	91 (20%)	46
86-6	61/122 (50%)	364	114	250 (69%)	136	6	21 (6%)	15	54	93 (26%)	39
88-6	67/147 (46%)	391	103	313 (80%)	210	13	27 (7%)	14	15	51 (13%)	36
89-9	53/141 (38%)	298	84	231 (78%)	147	19	28 (9%)	9	12	39 (13%)	27
Total	254/598 (42%)	1609	398	1195 (74%)	797	64	131 (8%)	67	128	283 (18%)	155

conservatively, recognizing the inherent limitations of this particular exercise.

The five separate hair proficiency tests resulted in a total of 1609 reported comparisons. The tests of common origin usually portrayed a scenario in which hair(s) of unknown origin were found in connection with a crime and laboratories were asked to compare these hairs with standards taken from the victim and one or more suspects. Laboratories reported inclusions and exclusions which agreed with the manufacturer in approximately 74% of their comparisons. About 18% of the responses were inconclusive, and 8% in disagreement with the manufacturers' information. The high percentage of inconclusive results undoubtedly reflects the types and limitations of samples issued and the lack of consistency of individual hairs taken from the same source. Of the inconclusives, about an equal number represented an inability to find a match when one was present, and a failure to report a nonmatch when a match was not present.

In test 81-6, laboratories were presented with a scenario in which they were asked to compare hairs found in the grasp of a homicide victim with known head hair from the victim and four suspects. In fact, the hair originated from only one of the suspects, which was confirmed by 78% of the responses. The majority of the 13% improper responses resulted from laboratories stating the hair could have originated from the victim or, to a lesser extent, one of the other suspects. Thirteen of the twenty-one participating laboratories provided completely correct results.

In 85-6, laboratories were asked to compare five hairs recovered from the crime scene with known hairs from the victim and two suspects. Two of the crime scene hairs proved extraneous, while one was a pubic hair taken from the victim and the two others were pubic hairs from one of the suspects. What distinguished the results in this exercise was the lower than average percent of "agree" results (71%) and the high percent of inconclusive responses (20%). The majority of results that differed with the manufacturers' specifications (9%) failed to find a match between the two unknown hairs and knowns from the proper suspect, and another unknown hair that matched with the victim. The high percentage of inconclusives can be explained, at least in part, by the variation within standards and the possibility of "overlapping characteristics" between the victim and suspect samples.

Test 86-6 was similar to the above exercise in that more than a quarter (26%) of the responses fell into the inconclusive category. Laboratories were asked to compare three hairs found in the hand of a homicide victim with standards taken from the victim and a single suspect. All unknown hairs came from the suspect. As in test 85-6, the PAC acknowledged the "inherent problems" in conducting a hair proficiency test given the inconsistency of samples. Two-thirds of the "disagree" responses were improper exclusions, and one-third improper inclusions; according to the PAC, these problems might have been averted had the laboratories been provided larger and more representative hair samples.

Test (88-6) issued laboratories three unknown hairs and known head hair samples from the victim and a suspect. Two of the unknown hairs matched the suspect and the third originated from neither the suspect nor victim. This proved to be one of the most successful exercises for laboratories with about 80% of comparisons being on target, 7% in error and 13% inconclusive. As in previous exercises, the PAC advised that inconclusive responses may in fact be the proper response given the problems of variability among samples. The PAC was most concerned with the thirteen improper inclusions and, of these, the five which mistakenly "matched" the questioned hair with the suspect.

In the final test (89-9) in this series, the scenario described a situation where three hairs were removed from the clothing of a homicide victim; these were submitted to laboratories along with known head hairs from the victim and the suspect. One hair originated from the suspect, one from the victim, and one from neither the suspect nor the victim. Laboratory responses agreed with the above sources in 78% of their comparisons. Consistent with other tests, laboratories expressed inconclusive results in 13% of their responses and were at odds with the manufacturer's specifications in 9% of replies. The Proficiency Advisory Committee again cautioned that given the variability of hairs taken from the same source, in many cases an inconclusive response may be the most appropriate reply.

Toolmarks

There were a total of twelve (12) toolmark tests issued to participating laboratories between the years 1980 and 1991. As with the

TABLE 13—*Toolmarks.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive			Unjust. Exclusion
			+	Total	−	+	Total	−	+	Total	−	
81-12	17/79 (22%)	34	17(14)	31 (88%)	0	0	0 (0%)	0	(14)	0 (0%)	0	3 (12%)
82-8	29/32 (90%)	116	28(34)	62 (53%)	0	0	1 (1%)	1	(34)	0 (0%)	0	53 (46%)
83-8[a]	25/53 (47%)	100	10	53 (53%)	43	0	4 (4%)	4	11	43 (43%)	32	
84-12[a]	42/73 (57%)	132	72	93 (70%)	21	0	4 (2%)	4	12	35 (27%)	23	
85-12	49/82 (60%)	98	17	50 (51%)	33	0	13 (13%)	13	19	35 (36%)	16	
86-12[a]	43/100 (43%)	129	18	68 (53%)	50	4	13 (10%)	9	16	48 (37%)	32	
87-12	65/106 (61%)	260	63	214 (82%)	151	0	0 (0%)	0	2	2 (1%)	0	44 (17%)
88-12	56/114 (49%)	155	45	130 (84%)	85	8	13 (8%)	5	5	12 (8%)	7	
89-10[a]	56/124 (45%)	280	94	140 (50%)	46	16	20 (7%)	4	14	120 (43%)	106	
90-10	91/152 (59%)	364	89	337 (93%)	248	1	1 (.2%)	0	2	26 (7%)	24	
91-11[a]	98/163 (60%)	293	193	271 (92%)	78	1	2 (.7%)	1	2	20 (7%)	18	
Total	601/1150 (52%)	1961	646(48)	1449 (74%)	755	30	71 (4%)	41	83	341 (17%)	258	100 (5%)

[a]Tool provided.

firearms exercises, the number of laboratories participating in these tests increased over the years, beginning with 72 laboratories and ending with 163—an increase of 126% (see Table 13). The rate of participation, the percentage of laboratories receiving samples that returned data, also increased; the mean for the twelve years was 52%, with a 48% response rate for the first five tests and 54% for the final six tests.

The tests offered the laboratories a wide range of toolmarks—five made by single bladed tools (screwdrivers), two involving bolt/wire cutters, a stapler, fingernail clipper, crimping tool, and die stamp. The test and evidence marks were provided in seven tests, while in five the tools (screwdriver, pairs of pliers, fingernail clippers, and a die stamp) were provided along with the marks. In the scenarios without tools provided, examiners were asked if the test (known) toolmarks were made by the same tool as that made any of the (2, 3, or 4) evidence marks. In three cases where a single tool was supplied, laboratories were asked if it made one or more of the marks provided. In another exercise, laboratories were asked which (if any) of three fingernail clippers was used to cut the questioned wire.

For the initial test (80-11) no data were provided in the summary report, other than the observations that no "incorrect origins" were reported, and is not tabulated in the above table. In three tests (81-12, 82-8, and 87-12) we also created a fourth category of "unjustified exclusions." In 81-12, laboratories were provided with three toolmarks (two unknown, one known) impressions all made with the same boltcutter; however, different sections of the cutting blade were used to simulate different cutting instruments. They were asked if the same tool that was used to cut the known was used to cut the evidence wires. All seventeen laboratories correctly responded "yes" for one of the wires and 14 labs gave inconclusives on the second wire, which are tabulated in the agree column since a different portion of the cutting blade was used. Three (3) unqualified "no" responses were considered "unjustified exclusions."

In 82-8, laboratories were provided with one toolmark made with a suspect's screwdriver and were asked if the same tool was used to make any of four other questioned marks. All laboratories but one properly identified the second mark as being made by the same tool. Fifteen labs reported 34 inconclusive comparisons which were tabulated in the agree column because laboratories did not know if the reverse side of the blade may have been used to make those marks. Thirteen labs reported 53 additional comparisons what we (and the PAC) concluded were improper or unjustified exclusions.

A similar situation arose in 87-12 where laboratories were asked if a mark obtained from one crime scene was made by the same tool as any of the casts recovered from four other scenes. All

but two of the responding laboratories identified the other single mark made by the same tool as that labeled suspect. The proper response for the other marks would be inconclusive since laboratories did not know if some other area of the same tool may have been used in making those marks. Those that answered "no," a total of 44 responses, were classified as unjustified exclusions.

Overall, laboratories performed not as well on the toolmark tests as they did on the firearms tests. A total of 1961 comparisons were reported and 74% of them agreed with the manufacturers' specifications (compared with 88% of the firearms comparisons). The percent of comparisons which disagreed with the manufacturer was also substantially higher—4% compared to 1.4%, as was the percent of inconclusive comparisons, 17% versus 10%. Five percent of the comparisons were placed in the unjustified exclusion category which, if they had been placed in the agree category (which some might argue they should be), would bring toolmarks closer into line with firearms results—at least in the percent of correct comparisons. There were no clear trends over time, with laboratories starting out well, variations occurring on the middle group of five tests, then ending with five tests, the last two of which had very strong results. Nor did many of the laboratories learn from the experience in the 1981 and 1982 tests with the unjustified exclusions, only to fall victim to the same problem in a 1987 exercise.

Exercises that gave laboratories the greatest difficulty included 85-12, which called for the comparison of two sets of staples of known origin with staples recovered from a crime scene. The test proved to be a great challenge given the extremely small markings on the staples. The high rate of improper exclusions (13%) in this exercise may also have been reduced had the staplers themselves been provided, along with a sample of unejected staples. Test 86-12, which required the comparison of a wire and three fingernail clippers, also resulted in a high percentage of results which disagreed with the manufacturer or were inconclusive. The PAC noted that some laboratories inappropriately relied on the presence or absence of copper on the cutting edges of the clippers as a basis for their conclusions.

There was also a higher than average percent of improper conclusions in 88-12, an exercise involving crimp terminals, in which laboratories evidently confused class and individual characteristics. Test 89-10 was noteworthy in that a set of pliers was provided along with five cut wires and more than 40% of responses fell into the inconclusive category—with most being situations where the pliers should have been excluded, but laboratories provided inconclusive results.

Participants finished with two strong performances (in 90-10 and 91-11). One involving markings from a screwdriver and the other where laboratories were asked if any of three metal identification tags had been stamped with a given die stamp. Comments from participants in both exercises indicated many thought the tests too easy.

Several of the laboratories expressed displeasure with not being provided with the tool with which they would make their own test marks. Many reported they were obliged to report inconclusive results where the tool was not supplied. It is interesting, though, that in the five exercises where tools were provided, rates of erroneous comparisons were (on average) about the same as for other tests, and that the percent of inconclusive results were substantially greater (28%), particularly on exercise 89-10 where the inconclusive response rate was 43%.

Questioned Documents

Between the years 1984 and 1991 a total of eight questioned documents exercises were issued to participating laboratories. The number of laboratories subscribing to the documents testing increased from 41 in 1984 to 117 in 1991, with an average of about 68% of laboratories which received samples responding with results.

The first three tests issued had major problems as the Project Advisory Committee attempted to develop a suitable test design. Tests were restricted by the inability to supply original materials and exemplars to all participants and many laboratories found the photocopies unsuitable for examination. The final five tests, however, proved more realistic/acceptable to laboratories and the results more amenable to analysis.

The first of the initial group of three tests (84-7) was comprehensive. It asked laboratories to intercompare three questioned photocopied letters, and then to compare each of the questioned documents with specimen writings of six individuals. Two of the questioned writings were written by the same author, but whose writing was not included in the specimens. The third questioned writing was a simulation of the other two, and this author's handwriting had been included among the specimens. None of the 23 laboratories responding with data was able to make this association. Most (57%) laboratories, however, were able to identify which of the three questioned documents were of common authorship. About a quarter, however, mistakenly concluded, either with certainty or probability, that all three were written by the same person.

The second test (85-8) involved the intercomparison of twelve signatures in the absence of known writings. Tracings and simulations were included in the specimens. Two of the samples were written by the same person, and one was a simulation and another was a tracing of either of these two. About 41% (13) of the 32 laboratories responding with results (24) gave completely correct responses. An additional nine laboratories properly associated the two signatures written by the same person, but mistakenly included either or both the simulated and traced signatures. Ten improper associations were reported, eight of which involved simulated signatures. One third (10) of the respondents reached inconclusive results.

In the third test (86-5), participants were asked to examine one note and three specimen writings, one of which was written by a document examiner deliberately simulating the writing in the note. Three (10%) of the thirty-one laboratories identified the true author and three additional gave proper "probable" answers. Fourteen (45%) of 31 laboratories, however, misidentified the simulation, with an additional four indicating a probable misidentification. The simulation was noted by eleven laboratories. The PAC, in its comments, acknowledged the difficulty that inclusion of the simulation added to the examination but noted that prior to issuance of the test they submitted the exercise to three examiners who arrived at the correct answer. The PAC also noted that most laboratories recognized the "signs of simulation or disguise" but overlooked the high percentage of labs which made misidentifications.

The experience from these three exercises led to subsequent tests being of a more clearcut design and amenable to tabulation (Table 14). In the words of the PAC, test 87-5 was "designed to be relatively easy and straightforward." Participants were asked to determine if an extortion note of unknown authorship was written by any of four suspects, each of whom supplied both requested and nonrequested writings. All the writings were natural and free of disguise. About one-half (17/33) of the laboratories

TABLE 14—*Questioned documents.*

Report	Participation Rate	Number of Comparisons	Agree			Disagree			Inconclusive		
			+	Total	−	+	Total	−	+	Total	−
87-5	33/53 (62%)	128	19	91 (71%)	72	0	1 (1%)	1	12	36 (28%)	24
88-5	49/73 (67%)	460	198	290 (63%)	92	3	8 (2%)	5	41	162 (35%)	121
89-5	53/71 (74%)	207	0	110 (53%)	110	17	17 (8%)	0	0	80 (29%)	80
90-5	60/92 (65%)	60	43	43 (72%)	0	0	2 (3%)	2	15	15 (25%)	0
91-6	83/117 (71%)	83	0	71 (86%)	71	1	1 (1%)	0	0	11 (13%)	11
Total	278/407 (68%)	938	260	605 (64%)	345	21	29 (3%)	8	68	304 (32%)	236

provided entirely consistent responses and only one made an improper exclusion; there were no mistaken inclusions. About 71% of the comparisons agreed with the manufacturer, and only 1% disagreed. However, given the ease of the test, an unacceptably high percentage (almost 25%) of laboratories reported inconclusive results for three or all four of the suspects.

The next test (88-5) involved the comparison of six questioned signatures with five known handwriting samples. Authors of five of the six questioned documents were represented among the five known writings. The exercise tested the ability of laboratories both to exclude and include (identify) writers of questioned material, as well as to conclude the known standards were insufficient to make a positive identification. On the first point, only about half the participants properly concluded that none of the writers of the five known specimens could have produced the questioned writing. Forty-one percent expressed their results as "no conclusion" and about 10% improperly suggested it may have been written by one of two known writers.

On the issue of identity, in excess of 90% of the respondents correctly associated three of the questioned writings with the proper knowns. For a fourth questioned writing, only 67% associated it with the known writing. Remaining laboratories either reported inconclusive results or improperly excluded the actual knowns. In all, of a total of 245 possible associations in this exercise (49 × 5), 199 (80%) agreed with the manufacturer. Six inclusions (two definite, four probable) and five exclusions (two definite and three probable) disagreed with the manufacturer.

There were ten comparisons singled out in Report 88-5, Table 1, which merit further examination. With a total of 49 laboratories responding with data, this yielded 460 possible comparisons (after subtracting 30 non-responses). Using a fairly liberal scheme of interpretation, that is, counting both certain and probable associations as "agree" inclusions, and including both certain and probable no associations as proper exclusions, laboratories were found to have agreed with the manufacturer in 63% of their responses, and disagreed in only 2% of their responses. A high percentage (35%) of responses were inconclusive, most (75%) of which occurred where exclusions should have been reported.

Test 89-5 involved comparison of a handwritten note found at the scene of a tire slashing (presumably by a high school student) and five known samples of handwriting selected from among students whose writing the teacher thought similar to the questioned note. Examiners were asked if the questioned note was written by any of the five writers. In fact, none of the five exemplars was written by the female who wrote the threatening note. Fifty-three laboratories responded with data. The vast majority of the respondents either successfully ruled out three of the writers or arrived at no conclusion. However, for two of the exemplars (K1 and K2), sixteen laboratories thought it at least probable that one of two writers of specimens had written the questioned note. Overall, then, while just 8% of results may be considered improper, only 53% were on target since such a high percent of results (39%) were inconclusive.

Test 90-5 challenged laboratories to determine if copies of anonymous letters had been made by the same machine copier, and if these copies corresponded to samples made with any of four different machines. Both exhibits (Q_1 and Q_2) were copied from the same photocopying machine. Only 2 (3%) of the responses disagreed with the manufacturers' specifications, 25% were inconclusive and 72% were in agreement. Virtually all the inconclusives were to the question if Q_1 and Q_2 were made with the same copier.

Exercise 91-6 asked respondents to examine bank deposit tickets and to determine if a teller's stamp on the tickets had been made with a particular rubber stamp, for which they were issued standards. The standard and questioned items were made with different stamps. Only one response disagreed and improperly reported both impressions were made with the same stamp. Laboratories were also asked if there had been alterations to the deposit slip and all laboratories' responses agreed there had been changes. Two of eighty-three responses, however, did not report accurately what those changes were.

Conclusion

How Do the Recent Results Compare with Those of the Earlier LEAA Study?

In this effort to compare the earlier LEAA data with the more recent results, the reader must keep in mind differences in the way these data have been tabulated and reported. The LEAA report expressed results in the percent of *laboratories* returning data on

a given test judged to exhibit "unacceptable proficiency." This scheme is different from the present review (particularly in the common origin exercises) in two fundamental ways. In the present paper the unit of analysis is each *comparison* between known and unknown samples, while the LEAA study consolidated all such comparisons on a given exercise into a single *laboratory* response. Here, an error in answering one part of the exercise might render the entire response "unacceptable." Secondly, in the present report "inconclusives" are broken out separately while in the LEAA study these often were placed in the unacceptable proficiency category (particularly if the reviewers felt such a response was not properly supported or justified). As noted earlier, an inconclusive may be a justifiable response depending upon the condition of the sample and the test results obtained.

Therefore, in the discussion that follows, we first attempted to translate the LEAA study's unacceptable, laboratory-based results into a form roughly comparable to the present study's comparison-based scheme. While this translation is admittedly imperfect, it does allow a reasonably good basis for comparing these results. Other limitations include:

• The 1978 study, based on tests conducted over a three year period between 1974 and 1977, focused on the development and mechanics of the proficiency testing procedure and as a result experienced problems of test administration (for example, sample manufacture, question formulation), which undoubtedly influenced test results. The present group of tests, conducted over a 14-year period between 1978 and 1991, took advantage of the lessons learned in the initial cycle of testing, covered many more tests and, notwithstanding its deficiencies, probably serves as a more reliable indicator of current laboratory performance.

• There has been no control over the difficulty of tests during these years so it is hard to say if improvements (or declines) in performance reflects enhanced proficiency or easier tests. A part of this issue is the relative levels of difficulty of examinations between, for example, the paint area and fingerprints. Paint comparisons are much more difficult in today's climate given the added quality control efforts of manufacturers and the resulting homogeneity of samples produced. There was no acknowledged attempt by manufacturers to make one evidence category of testing any more challenging than another, or to adjust the level of difficulty of tests as time went on.

As a member of the earlier LEAA project, Peterson believes the effort to get laboratories enrolled in the project and to keep them involved probably led to more straight forward tests in the earlier project than the continuing, fee-based one. A review of examinations administered in this latter period also leads us to conclude the manufacturers have generally made the tests more challenging, particularly in biological fluid mixtures, toolmarks, questioned documents, as well as other areas.

• Continuing changes in test design and sample makeup also make the comparison of results difficult, because there are very few instances in which the manufacturers attempted to replicate tests that had been issued in the LEAA program.

Notwithstanding these limitations, we offer the following general observations:

Bloodstains—The typing of bloodstains has improved substantially as forensic laboratories incorporated electrophoresis procedures (routinely) in the identification of isoenzyme and serum protein systems. There are areas, such as the determination of secretor status of stains, however, where laboratories still experience difficulty.

The described advancements have greatly improved the ability of forensic laboratories to answer common origin questions surrounding bloodstains. The frequently cited bloodstain comparison exercise in the LEAA study (#8), where about 70% of laboratories submitted unacceptable responses for failing to distinguish two type O stains taken from different individuals, would be handled competently by most all present-day crime laboratories.

Nonblood Body Fluids—The identification of semen and saliva stains is performed at about the same level as in the 1978 report. The typing of these stains is generally done well, but this is an area largely untested in the first LEAA study. The determination of secretor status of these nonblood body fluids is not being performed well and merits attention. DNA typing of body fluids and tissue is, of course, revolutionizing this area and in the not-too-distant future may replace most or all tests for serum proteins and isoenzymes.

When presented with mixtures of blood and other body fluids, laboratories are understandably far less successful in answering questions of possible common origin. Such mixtures were not included in the earlier LEAA testing. As noted, deficiencies in secretor status testing is one of the primary problem areas that can lead to improper determinations of common origin.

Drugs—Generally, the performance of laboratories in identifying drugs of abuse is quite good and exceeds that recorded in the initial LEAA testing. The quantitative testing of drugs is problematic but this is an aspect of drug testing not routinely performed by many forensic laboratories because of variations in legal requirements from jurisdiction to jurisdiction.

Latent Prints—Although not included in the LEAA project, examiner performance in determining the origin of latent finger and palm prints is done with the highest level of accuracy of any physical evidence area. The only difficulties appear to be where examiners occasionally are careless and mistakenly link the questioned latent print to the wrong finger position on the fingerprint card.

Glass—Comparison of fragments of glass in establishing possible common origin has not changed appreciable from the earlier study and remains problematic.

Paint—Paint comparisons also have not improved since the earlier LEAA study and represent one of the most troublesome evidence categories. As noted earlier, the great improvements in paint production technology and the very small differences present in automobiles or other materials painted with a manufacturer's product can make discrimination among paint samples extremely difficult.

Firearms—Although the rate of proper comparisons did not increase, the percent of improper comparisons was reduced by half. There was, however, a substantial increase in the percent of inconclusive results.

Fibers—Although the initial fiber examination in the LEAA study was easy and straightforward, the performance of laboratories

in the recent testing dropped substantially, resulting in one of the highest percent of comparisons that disagreed with the manufacturers' values of any evidence category.

Performance: 1978–1991

We have placed the results of testing into three general performance categories:

1. The first category includes evidence types where laboratories are performing very well as reflected by high rates of proper responses, for both straight identifications as well as comparative exercises. These areas also have according low rates of misidentifications, inconclusives, and responses that were otherwise at odds with the correct values.

   The four types of evidence where laboratories are performing the best—as expressed in terms of high rates of proper common origin determinations and low rates of improper comparisons and inconclusives are: finger and palm prints, metals, firearms, and footwear. While rates of successful comparisons of footwear are not as high as the other three categories, it had the second lowest rate of improper comparisons of all evidence areas tested. Also included in this category are those evidence types (bloodstains and drugs) in which laboratories correctly identify/type substances in 94% or more of their attempts, plus exhibit low rates of misidentification.
2. A second category is where (in the comparative testing area) due to the nature of the evidence, limitations in the data and/or examination techniques, laboratories have higher rates of inconclusive responses and lower percentages of correct comparisons. Also placed in this category would be those types of evidence where rates of correct identification are in the 80% to 90% range.

   Questioned documents, toolmarks, hair, and bloodstains fall in this category. Three of these evidence types—all except for bloodstains—had the lowest rates of proper comparative responses of all categories tested. For evidence areas where the goal is identification, and success was attained in about 80% to 90% of attempts, are the categories of flammables, fibers, and explosives (at the low end). Flammables are placed in this category for, although having a slightly higher rate of identification ("Is it a flammable?"), laboratories only identified the correct class of flammable about 65% of the time, and had a false positive identification rate exceeding 10%.
3. The third category would be evidence categories of serious concern where laboratories are regularly reporting higher rates (in excess of 10% of their results) of improper comparative examinations. Also included in this category would be evidence where laboratories have difficulties successfully identifying the material.

   Fibers, paints, (automotive and household), glass and body fluid mixtures all have improper comparison rates exceeding 10%. In terms of identifications, animal and human hair body area identification was clearly the most troublesome area.

Judicial Implications

One of the important questions that these proficiency testing results demands asking is: "Assuming these results reflect the type of casework typically performed by present-day laboratories, what might the implications be for cases being adjudicated in the criminal justice system?" Given the many limitations we have identified previously, it is not possible to say for sure. However, to put these results in some perspective, it would be helpful, at minimum, to estimate the frequency that laboratory results in these different categories are actually used by courts of law. This question we can answer with a substantial level of confidence.

Studies completed about eight years ago by one of the authors [*1,2*] set out to determine the frequency of use of scientific evidence and its impact on case investigations and prosecutions. These national studies were conducted in several jurisdictions and attempted to trace the movement of cases through the investigation, prosecution, and sentencing stages and to document the presence and assess the impact of different types of evidentiary and nonevidentiary information. Major effort was devoted to determining if scientific testing of physical evidence took place and the results of this testing. The present discussion focuses principally on the baseline question, "Was a report of scientific evidence reflected in the case files?"

In general, scientific laboratory reports were found in from about one quarter to one third of the 4500 felony cases sampled where charges had been filed by the local prosecutor. Rates varied greatly as a function of crime type; that is, laboratory reports were virtually always present in drug and murder prosecutions, but seldom (10% or less) in thefts. These rates were a function of many factors, including the seriousness of the crime, the amount of physical clues generated in the course of committing the crime, the resources of the local agencies, and the necessity of having a laboratory report present to prosecute the case (as with drug possession). On average, around 50% of rapes yielded laboratory reports (usually centering on the identification of semen), but with wide variation among jurisdictions. Laboratory reports appeared next most often (usually of fingerprints) in burglaries—in about a quarter of these prosecutions. Only about 10 to 20% of robberies and attempted murder prosecutions had forensic laboratory reports.

Given these usage patterns, it is not surprising that drugs were the single most common type of forensic evidence, appearing in about 12% of the 4500 case files reviewed. Taking the sample another way and looking just at the cases having some form of scientific evidence, drugs represented almost half of the different types of lab reports present. Fingerprints were present next most often, appearing in about 7% of the felony case filings, followed by semen (3%), firearms (2%) and blood and bloodstains (2%). We see, therefore, that scientific evidence utilization in felony prosecutions remains at a minimal level when the total body of cases is considered.

Other types of scientific evidence were utilized at even a lower level. Hair comparison reports were present overall in about 1% of all cases, but in a quarter of rapes and about 10% of murders. Other evidence categories barely registered; impressions and imprints, toolmarks, paint, glass, and fibers *collectively* were present in less than 1% of the cases reviewed.

In sum, there may be some comfort in the fact that the two evidence categories that are present most often in actual prosecutions, making up more than 70% of the laboratory reports in the case files, are drugs and fingerprints. The reader will recall these two evidence categories have the highest rates of successful identification/comparison. Looking at the middle group (in terms of frequency of usage) of evidence categories—semen, blood and bloodstains, and firearms—and which are most closely associated with crimes of violence, laboratory performance is mixed, with

straight bloodstain identification and comparison falling into the superior category. Body fluid mixtures, however, do represent an area where laboratories experienced difficulties in sorting out their origin. Human hair examination also had a relatively low rate of successful comparison. But it is in the final category of evidence types infrequently found in felony case files where laboratories performed the poorest. This would include such categories as paint, glass, and fibers. Laboratories did perform moderately well in toolmark comparisons and very well in comparing footwear impressions.

This apparent relationship between frequency of appearance of laboratory reports and laboratory performance may be a reflection of several factors, including two primary ones: 1) lab performance may be wanting because there is little (prosecutorial) demand for these evidence types and laboratories accordingly place a lower priority on developing suitable methods and training of examiners; as well as the converse, 2) prosecutors don't demand the results of examinations of such evidence because they lack confidence in laboratory results, both in terms of the specificity and reliability of results.

Proficiency Testing Policy Questions

It has become clear to most forensic examiners that proficiency testing should become a routine requirement for all laboratories serving the criminal justice system. Lucas et al. [*3*] have concluded that the burden principally rests with forensic scientists to prove their competence and that it cannot be assumed. There are several important issues which need to be addressed, however, in accomplishing such an objective.

• Optimally, what form should proficiency testing take—a national, regional or local program? What should the level of difficulty of the tests be—should they be fairly basic, straight forward exercises that all examiners should be able to answer, or should they also challenge laboratories with more complex problems that occasionally arise in casework? Should laboratories know they are being tested or should efforts be made to introduce blind tests where examiners are not so informed? The clinical and urine screening fields appear to have reached consensus that blind testing is desirable whenever possible.

• Should efforts be made to identify the true source of poor laboratory results by controlling better for qualifications of the examiner, methods followed and types of instruments and reagents employed? In so doing, the focus of the testing would be primarily on improving laboratory performance, rather than simply identifying laboratories that fail to measure up.

• Should proficiency testing be made *mandatory* and should the results be available for public review? As noted earlier in this article, Jonakait [*4*] has published a detailed review of the forensic laboratory area and has advocated that such a program be instituted. Should the forensic field attempt to operate such programs itself or should a disinterested public/private organization oversee the program?

• Need there be sanctions for laboratories that do not meet a satisfactory level of proficiency? And who is to determine what this satisfactory level of performance is? Should the profession oversee such sanctions or should they be placed in the hands of an external judicial or regulatory agency? These are all bona fide questions which merit serious consideration.

These are very challenging issues but ones which call for answers. How the profession chooses to address them will profoundly affect the course of forensic science in the coming century.

References

[*1*] Peterson, J. L., Mihajlovic, S., and Gilliland, M., *Forensic Evidence and the Police,* U.S. Government Printing Office, Washington, DC, 1984.

[*2*] Peterson, J. L., Ryan, J. P., Houlden, P. J., and Mihajlovic, S., "The Uses and Effects of Forensic Science in the Adjudication of Felony Cases," *Journal of Forensic Sciences,* Vol. 32, No. 6, Nov. 1987, pp. 1730–1753.

[*3*] Lucas, D. M., Leete, C. G., and Field, K. S., "An American Proficiency Testing Program," *Forensic Science International,* Vol. 21, 1985, pp. 71–79.

[*4*] Jonakait, R., "Forensic Science: The Need for Regulation," *Harvard Journal of Law and Technology,* Vol. 4, Spring 1991, p. 178.

Address requests for reprints or additional information to
Joseph L. Peterson, D.Crim.
Department of Criminal Justice
University of Illinois—Chicago
1007 W. Harrison St.
Chicago, IL 60607