

1924

Studies in Testimony

William M. Marston

Follow this and additional works at: <https://scholarlycommons.law.northwestern.edu/jclc>

 Part of the [Criminal Law Commons](#), [Criminology Commons](#), and the [Criminology and Criminal Justice Commons](#)

Recommended Citation

William M. Marston, *Studies in Testimony*, 15 *J. Am. Inst. Crim. L. & Criminology* 5 (May 1924 to February 1925)

This Article is brought to you for free and open access by Northwestern University School of Law Scholarly Commons. It has been accepted for inclusion in *Journal of Criminal Law and Criminology* by an authorized editor of Northwestern University School of Law Scholarly Commons.

STUDIES IN TESTIMONY

WILLIAM M. MARSTON,¹ Washington, D. C.

1. PROBLEMS

It seems to be a regrettable fact that little systematic psychological experimentation is being carried on in the field of normal adult testimony. Much valuable material is being produced by psychiatrists, sociologists, and criminologists from time to time; but the subjects of such studies are, for the most part, either psychopathic or criminal variants from the mental or social norms. On the other hand, much constructive work is being done by statistical and educational psychologists toward the development of intelligence and fitness tests; but the direct application of such psycho-statistical procedures to everyday legal problems of testimony seems a long way off.

Such work as has been reported in the legal field proper lies almost wholly in the line of the "Aussage," or "fidelity of report" tests.² An incident is performed, or an object presented, before experimental subjects whose immediate, written report is scored as to accuracy and completeness by the experimenter. "The question has been asked," says Wigmore,³ "whether the alleged percentages of testimonial error, as found in the laboratory experiments, do really, in trials, produce misleading results in the verdicts. The way to answer this is to include a jury (or judge of fact) in the experiment, and observe whether the findings of fact follow the testimonial errors or whether they succeed in avoiding them and in reaching the actual facts." The problem, thus stated, is, from a psychological point of view, such a broad one that preliminary studies are indicated for the purpose of roughly determining, first, where lie the possible psycho-legal causes of testimonial error; and, secondly, as suggested by Wigmore, what possible psychological relations may exist between the findings of judge and jury and the original testimony upon which such findings were based. Such were the purposes, then, of the preliminary studies herein reported.

¹Member of the Boston bar, and of the American Psychological Association.

²For summary, see Wigmore: "Principles of Judicial Proof," p. 575 ff.

³Wigmore, III. Law Review, 1909, p. 426. Cited in summary, reference (2), above.

2. COMPARISON OF METHODS OF ELICITING TESTIMONY

The most obvious analytical modification of the older method, which consisted of testimony written without questioning, and of written responses to questions, would seem to be a closer approximation of the typical conditions obtaining in actual cases. Three different methods of eliciting testimony for use in the courtroom exist. First, under certain conditions, the witness may refresh his memory from notes made by himself. Such notes are written, according to usual rule, immediately following the occurrence. Psychologically, it is safe to assume, as a working premise, that perusal of such notes, on the stand, substitutes the notes for present memory. Therefore, an account of the incident witnessed, written by the observer himself without questioning, immediately after the occurrence in issue, almost exactly reproduces the actual condition of notes of this sort used on the witness stand. I have called this "Free Narration."

Secondly, the witness is subjected to direct examination. No leading questions may be asked, and the witness does not fear legal trickery, or attempts to entangle his testimony by "catch" questions, since he is examined by an attorney whose purpose is to co-operate in every way possible with his own witness. To be sure, in actual court procedure, the testimony is orally given, and is recorded verbatim by the court stenographer—a condition not practicable in experiments where a number of witnesses must be examined simultaneously. Moreover, no trial is ever held immediately after the occurrence of the incidents in issue; whereas, for the sake of experimental uniformity between witnesses in Aussage tests, it is advisable to take all testimony immediately upon conclusion of the incident, before the witnesses leave the room. In the present experiment, I was able to frame questions conforming, with sufficient accuracy, to a comprehensive direct examination of witnesses in actual practice; but it should be borne in mind that the answers were written by each witness, instead of being given orally; and that the writing was done immediately, instead of after the usual interval intervening between incident and trial. This series of immediate, written responses to non-leading questions I have called "Direct Examination."

Thirdly, witnesses actually on the stand are subjected to cross-examination. The witness knows that the examining counsel is antagonistic, and that every device will be used to entangle and obscure his testimony. Leading questions may be asked freely, and queries of a leading nature based upon deliberately erroneous suggestions are commonly introduced for the purpose of influencing the jury, and of

suggesting a doubt, or a contradictory answer, to the witness. These conditions may readily be reproduced in experimental procedure, subject to the qualifications of answers being written immediately, by the witnesses, instead of being given orally, after an interval, as in actual trials. I introduced several deliberately suggestive leading questions of the nature above mentioned into this portion of the test, and 40 questions were propounded, as against 10 in the Direct Examination. This longer series, as described, I have called "Cross-Examination."

METHOD

The incident to be observed and reported was performed before a group of 18 students, who were attending a lecture in Legal Psychology at American University, Washington, D. C. All 18 subjects were lawyers, either practising in the District of Columbia or employed by the Government. None of the subjects received any intimation that an experiment was to be performed, nor were they aware that any test was in progress until the incident had been concluded and the experimenter began to issue instructions for recording of testimony.

In constructing an incident to be reported, three modifications of the usual *Aussage* types of action were attempted, with a view to eliminating psychological sources of error obviously occurring in the older methods. First, many incidents performed before students, or other subjects, have been of a nature deliberately deprived of logical meaning, but designed to suggest a false or sham meaning. Thus an actor might be armed with a shining wrench, intended to suggest a revolver. When the glimmering steel did suggest a gun, and was so described by the witnesses, a serious error was recorded. But murderers or holdup men do not, as a rule, deliberately assume stage properties in order to lay a heavier weight of evidence against themselves; nor do most innocent persons voluntarily go through the motions of a holdup, in the presence of the real criminal, in order to divert suspicion from guilty to innocent persons. Although I have never seen this criticism in the literature, it seemed to me a most potent one from the standpoint of psychological method. I therefore took pains to prepare an incident with a logical meaning of its own, and devoid of a false or hidden one likely to trick the witnesses into error.

Secondly, most of the *Aussage* tests previously reported have used an incident or observational material wholly foreign to the conscious content of the subjects at the moment the incident occurs. A sensational intrusion of blood (or paint) smeared actors, shouting and

gesticulating causes an abrupt shift in the whole motor attitude of the subjects and often evokes more or less violent emotion. This shifting of motor setting, and the accompaniment of the incident by unusual emotional experiences, inevitably tends to disturb normal memory. Often emotion may virtually hold connected perception in abeyance, with the result that the subject has only isolated sensations to remember, instead of a logically connected unit perception. It is a well-known psychological fact that memory is greatly assisted by the logical connection of the remembered sensations. Therefore, it would seem that emotional and motor factors tending to break up such connection must be eliminated if experimental results are to be regarded as significant of the amount and types of error occurring under most natural conditions. Occasionally we find in the courtroom eye witnesses of murders or other violent crimes; but the great mass of actual testimony relates to matters observed in connection with ordinary, routine actions, and the Aussage experiments should test the rule rather than the exception. The incident used, therefore, in these studies was of such a character that no one of the 18 witnesses suspected anything unusual had occurred until so informed by the experimenter.

Thirdly, the incident selected should contain the usual sort of legal meaning. No individual witness, as a rule, is called upon to testify as to the interpretation or significance of the incident he saw. He is put on the stand to narrate certain details, and these details are most often used as individual pieces in the mosaic counsel is attempting to lay in the minds of the jury. While the witness' mind is usually aided, in its remembering of details, by an obvious logical unity of incident, it is not the unity, or meaning, which has legal significance, but rather the details which are thus strung together. An occurrence should be planned, therefore, containing as great a proportion as possible of details legally significant, for some predetermined trial purpose. Most of the incidents used in previous experiments have contained a large proportion of incidental significance, and a very small proportion of legally relevant meaning. If such is the case, the chances are all in favor of a minimum of possible legal meaning in the result. I tried to construct an incident, therefore, nearly all details of which might have a predetermined and proper legal significance.

The incident used was that of an unknown youth rapping on the door of the lecture room, soon after the beginning of the lecture; entering on the lecturer's request, and delivering an envelope to the lecturer. The latter removed a yellow paper from the envelope, pretended to read a message, and exchanged remarks with the stranger,

who thereafter left the room. There were no deliberately false suggestions in the affair to lead witnesses astray, and the actions were of such a usual nature that no emotion beyond mild surprise or curiosity could be evoked from the witnesses.

There were a large number of details, however, of possible and connected legal significance. The predetermined plot was trial of the strange youth for the knifing of a person of his acquaintance. In strict adherence to realism, the youth chosen was a Texan, very quick of hand and temper. He possessed a long, green-handled pocket-knife, the blade of which might well be used for stabbing purposes. With one hand the young Texan could draw and open this knife, all in the same motion. Under his left arm he carried three books, one red, one green, and one blue (colors to furnish psychological primaries), and the predetermined plot included the finding of these books at the scene of the crime. Besides the envelope handed the lecturer, the young man carried a second envelope in the same hand, which might have contained a taunting letter, just received from the murdered acquaintance. While the lecturer read his supposed message, the Texan faced the audience, drew and opened his knife, and scraped at his gloved thumb with it, in supposed embarrassment. The points of usual legal significance, therefore, would be all those details serving to identify the Texan, his books, his envelope, and his knife—a not unusual type of legal meaning for any testimony offered. I may anticipate the results by stating that not a single witness noticed the knife at all!

RESULTS

Each observable detail in the above outlined incident was allotted a score of one point, summing into a total of 147. Similarly, the number of possible points scorable in answering direct and cross-examinations were separately compiled, giving a total of 120 points for direct, and 107 points for cross-examination. Most of these points, of course, were included in the original 147 which might have been obtained, as a perfect score, in the free narration; but for the purpose of comparison of the three methods of eliciting testimony, it was necessary to score each part of the test separately. Table I, following, gives the results:

TABLE I
SCORES OF WITNESSES REPORTING AN UNEXPECTED INCIDENT

	Free Narration	Direct Exam.	Cross Exam.
Average of all witnesses—			
Completeness	23.2	31.2	28.7
Accuracy	94.05	83.2	75.7
Caution	40.1	51.8

	Free Narration	Direct Exam.	Cross Exam.
Highest witness—			
Completeness	31.9	48.7	43.9
Accuracy	100.	94.5	97.2
Caution	77.7	88.8
Lowest witness—			
Completeness	6.4	7.5	19.6
Accuracy	83.7	64.2	61.9
Caution	15.3	10.

The figures in Table I, above, represent percentages. The per cent "Completeness" means the completeness of the correct points given by the witness. Thus the first figure given in Table I, which is 23.2 per cent completeness as the average for all witnesses in free narration, means that the 18 witnesses testified, correctly, to an average of a little over 34 points during this portion of their testimony. Since the total they might have given correctly was $147, \frac{34+}{147}$, or 23.2%, represents the completeness of their correct findings.

The figure for these witnesses' average "accuracy" in free narration, which is 94.05%, signifies the percentage which their correct points are of their total points given. That means that, whereas these subjects, on an average, gave 34+ points correctly, they actually gave a total average of 36+ points 2 of which were erroneous. The handling of errors in figuring Aussage results seems to me another fertile source of possible confusion. If a witness gives 34 points correctly and 2 points erroneously, how can it have any meaning to figure his total, 36, as the number representing his completeness? Surely the 2 erroneous points are not, in any sense, a portion or fraction of the possible number of 147 points to be given. Rather, it seems to me, he has gotten $\frac{34}{147}$ of his job done; and, in addition, he has volunteered 2 extra, wrong facts. These 2 mistakes will not tend to impair the completeness of what he has found correctly; but they will tend to mislead the jury into possible additional erroneous findings. The two errors should, therefore, be used as a basis for figuring the percentage of said witness' reliability, or the accuracy of all his testimony taken as a unit, and should be given a negative rather than a positive value in the estimation of the worth of each witness' testimony.

The percentage of "caution," as used in Table I and succeeding tables, signifies that percentage of the total times the witness really didn't know a point, in answer to a question, when that witness declined to answer, or said, "I don't know." The figure is arrived at by totaling the number of errors, "don't know's," and no-answer re-

sponses, which total clearly represents the actual total of points unknown to the witness, and then figuring what fraction or percentage of this total is represented by the number of "don't know's" and no-answers, which clearly constitutes the actual total of points the witness knew he didn't know. To take an hypothetical case, suppose a given witness made 6 erroneous answers and replied "I don't know" to 4 queries. Then 4 plus 6 shows that the witness actually did not know the answers to 10 inquiries. But he took a chance on 6 answers and got them wrong, while on 4 he exercised caution and declined to respond. His index of caution, then, would be $\frac{4}{10}$, or 40%. Of course, there is always the possibility that the witness also took chances in answering some of the questions to which he responded correctly; but on this we can obtain no objective check. The value of the index, or percentage of caution, is largely interpretive, as may be seen by a glance at the results in Table I.

SUMMARY

Careful inspection of these results reveals data which may be summarized as follows:

A—Free narration is uniformly less complete and more accurate than either direct or cross examination.

B—Direct examination is both more complete and more accurate than cross examination. The "low" witness, put into the table to show the range of variation in results, does not conform to the average or to the "high" witness in this particular, and the explanation would seem to lie in an individual eccentricity which leads this man not to respond to any inquiry unless pressed or irritated. The "high" witness, on the other hand, proved even more accurate on cross examination than on direct, this increase in accuracy evidently being at the expense of completeness.

C—Cross examination shows greater caution than direct examination, though without any corresponding gain in either completeness or accuracy on the average. From study of individual records, however, I would suggest that the added caution in cross examination has a vital influence in holding the accuracy percentages up to a figure at all comparable with that obtaining in direct examination. Added caution does an important, though hidden, work, in short, in redeeming cross examination from a degree of inaccuracy which would render it almost wholly worthless.

D—The percentage of caution, in individual results, shows a close correlation with both completeness and accuracy. It seems a rough but certain indicator of the probable value of any witness' testimony.

3. COMPARISON OF UNEXPECTED AND EXPECTED INCIDENTS

The lawyer-witnesses in my first experiment were very much dissatisfied with their scores. It seemed to them incredible that not one subject had observed the large, green-handled knife which had been the crux or point of maximum legal meaning in the incident presented. They advanced the theory that the very usualness of the actions involved had cut their perceptions in half; and I countered with the judgment that if they were to see the same incident over again, fully cognizant that it was to occur, their scores would not be improved more than fifteen or twenty per cent. The point was worth investigating, however, since if knowledge that something to be especially observed is about to occur should noticeably increase completeness and accuracy of report, it would mean that the state of witnesses' attention at the time their observations were made is the controlling factor in the resulting value of their testimony. This was the problem, therefore, which I next sought to investigate.

METHOD

The same incident, with different actor, costume, knife, envelopes, and conversation, was performed before another class in legal psychology at American University, all members of which had been fully informed as to the details and results of the first experiment. This second class, moreover, which was composed of 12 lawyers, contained two of the same witnesses who had reported on the first incident. No special announcement was made when this second incident was performed that a test was in progress; but, according to later report, every member of the class knew from the moment knocks were heard upon the door of the lecture-room that an Aussage experiment was contemplated. In this way a general study of the importance of witnesses' initial state of attention could be made without interposing that artificial setting which must always arise when subjects are told: "Here is a picture, look at it." As nearly as the actual conditions usually obtaining in observations of actions especially attracting witnesses' attention, and on which they knew while making the observation that they would be called upon to report later, could be reproduced experimentally, this second experiment reproduced them. The two

men who had seen the same incident, with differing details, previously performed enjoyed, of course, a considerable advantage over witnesses in actual cases under otherwise corresponding conditions. A separate study of the results obtained by these two witnesses, therefore, was made, as well as a study of the average scores of the whole group, including the two men in question.

RESULTS

Since the plot of the second incident remained identical with that of the first, including all the actions of knocking, entering the room, giving an envelope to the lecturer, removing a yellow paper, reading a supposed message, stranger's putting second envelope in pocket, taking out knife and scraping at gloved thumb, exchange of almost identical remarks, and stranger's leaving the room, it might certainly be expected that the scores of the two witnesses who had seen the first incident, at least, would show a remarkable improvement. No tricks were introduced, even in the details changed, the new actor being an obviously different individual, with obviously different clothes; his second envelope being very large, and his knife a very large, white-handled affair. The colors of the books were kept the same as in the first experiment. Despite all these favorable conditions for an improved score, Table II, following, shows that the two witnesses reporting on both incidents, one of whom was next to the high man in the second group, failed to exhibit any significant improvement whatever, except in completeness of free narration.

A glance at Table II shows that the only notable improvement in score was obtained, by both subjects, in the completeness of their free narration, their average gain being 22%. This is slightly more of an increase in score than I had predicted, basing my guess upon the general psychological factors involved. When we continue our inspection, however, we soon discover that only one other gain was made, and that was an average improvement of only 7.2% in cross examination completeness. On the other hand, the most consistent difference between the scores of the first and second experiments seems to be a diminution in accuracy in all types of report, varying from .2% in Mr. H's cross examination to 15.9% in the same subject's free narration. This loss in accuracy is uniform except for a gain of 1.1% during the direct examination of Mr. H. The explanation clearly lies in the tremendous diminution of caution which both subjects showed, this loss running as high as 44.5% during the cross examination of Mr. W. In short, the gain which might have been expected to result

from repetition of identical perceptions and report of identical observations in the second experiment was largely counteracted, and, during a considerable part of the testimony was completely reversed by the overconfidence or self-assurance of subjects witnessing the same incident for the second time. This result once more emphasizes the apparent close relationship between the caution of the witness and the general reliability of his testimony; and I venture the suggestion that caution plays its part during the observation of an incident as well as during the reporting thereof.

Table III, following, gives the comparative scores of the 18 witnesses to the first incident under unexpected conditions, and the scores of the 12 witnesses to the second incident, who did expect it, from the initial knocks on the lecture-room door, and who knew it was a test incident throughout. The two witnesses who had also seen the first incident are averaged into the second, or "Expected" group, in order to emphasize any possible differences between the scores resulting from the difference in the state of witnesses' expectations.

SUMMARY

Analysis of Tables II and III, above, shows:

A—Messrs. W. and H., repeating witnesses, manifest great improvement in the completeness of their free narration when the incident is expected; they show slight, but consistent, loss of accuracy under these same conditions; and they show great loss of caution throughout all reports of the expected incident.

B—The entire group of witnesses under expected conditions, as contrasted with the whole group under unexpected conditions, shows slight gain in completeness during all types of report (ranging from 4.5% to 8.8%); they show slight loss in accuracy in all types of testimony (averaging 7.9%); and they show considerable loss of caution in all three types of report (ranging from 12.6% to 52.9%). It is to be noted that most of the gain in completeness in the total group scores, just as in the repeating witnesses' scores, was achieved during free narration, only average gains of 2.5% and 3.6% being registered in direct and cross examinations, respectively.

4. COMPARISON OF FINDINGS OF FACT BY JUDGE, MALE JURY, AND FEMALE JURY

The second part of the original problem, it will be remembered, consisted of opening up the psychological analysis of relations existing

between the findings of judge or jury, and the testimony upon which such findings were based. A preliminary complication immediately made itself manifest. Which finder of fact can produce most accurate results for purposes of comparison with witnesses' testimony, judge or jury? And, since we have both male and female jurors in most jurisdictions since the adoption of the Nineteenth Amendment, which sex shall we assume to be most reliable in the exercise of this function? It would seem, in light of these complications, that a preliminary comparison of judge, male jury, and female jury, as finders of fact, was indicated before any attempt could be made to compare the respective values of findings of fact and testimony.

METHOD

The testimony of each witness in both experiments was, it will be remembered, written out by the witness himself. It was only necessary, therefore, to have this testimony typed, with individual copies for each judge and juror. Diagrams of the lecture-room were also prepared for the judge in the first experiment, and for two of the juries in the second experiment, who were unable to make a personal inspection of the scene where the incident occurred.

Dean J. H. Wigmore, of Northwestern University Law School, very obligingly consented to serve in capacity of court for the first experiment, and I wish to take this opportunity of acknowledging my obligation to Dr. Wigmore for his courtesy. It seemed eminently advisable to select as judge a lawyer especially experienced in dealing with evidence, and it will be generally agreed that no American member of the bar more exactly satisfying this experimental requirement could have been chosen. Both the male and female juries who rendered findings upon the first incident were students at George Washington University, Washington, D. C.; and I am greatly indebted to Dean W. C. Ruediger for his careful selection and supervision of these juries. Views of the lecture-room where the incident occurred were arranged for both juries, and each individual juror was furnished with a copy of all the evidence several days in advance of the meeting at which they made their final findings. Notes on the methods and general procedure in the jury-room were kept by the foreman and forewoman of male and female juries, respectively. I wish to thank all the students who thus participated in this work.

In the second experiment I decided to select as judges persons of each sex experienced in some practical method of finding facts, but not possessed of legal training. In this way it seemed feasible to open up

the possibilities, if any, in problems of the most efficient type of training for individual finders of fact. Miss Emily Davis, newspaper woman and correspondent, consented to review the testimony of the second experiment and to report her findings exactly as she would do in the actual pursuit of her profession; and Dr. Charles C. Tansill, Assistant Chief of the Legislative Reference Service, Library of Congress, whose field of specialization is American history, agreed to make similar findings in accordance with the methods customarily employed by historical research workers. Expression of warm appreciation is due both Miss Davis and Dr. Tansill for their painstaking efforts. Director of Rehabilitation Edward F. New, of the University of Maryland, selected and instructed male and female juries from the college students under his charge; and I wish also to thank Mr. New, who is himself an attorney, for the care and trouble taken both by himself and by his students. Exactly the same procedure was followed by judges and juries in studying testimony and making final findings of fact as in the first experiment.

RESULTS

In the first experiment, both judge and juries, in rendering their findings, followed the three forms in which the testimony was elicited; that is, they made one finding in the form of free narration, another in form of answers to direct examination, and a third in form of answers to cross examination questions. It is impossible to state, with complete accuracy, to what extent the direct and cross examination testimony influenced the free narration findings; but it may be judged, from careful inspection of all results, that the finders of fact adhered rather strictly to the actual content of testimony in each part to which their report corresponded. In Table IV, following, the completeness, accuracy, and caution of the judge, of the male jury, and of the female jury are compared separately for free narration, direct examination, and cross examination.

TABLE IV
SCORES OF JUDGE, FEMALE JURY, AND MALE JURY

	Completeness	Accuracy	Caution
Free narration—			
Judge	31.2	88.4	...
Female jury	16.3	92.3	...
Male jury	20.4	75.	...
Direct examination—			
Judge	40.8	72.	17.3
Female jury	33.3	71.4	23.8
Male jury	21.6	59.	14.2

	Completeness	Accuracy	Caution
Cross examination—			
Judge	38.3	82.	30.7
Female jury	30.8	71.7	23.5
Male jury	27.1	68.1	36.3
Total averages—			
Judge	36.7	80.8	24.
Female jury	26.8	78.4	23.6
Male jury	23.4	67.3	25.2

The interesting, and to many jurists unexpected, result of superiority of female jury findings over those of the male jury is manifest in Table IV. The judge, as might, perhaps, be predicted, scored approximately 10% higher, on the average, than the highest jury in the completeness of his findings, while the judge's average accuracy was also slightly higher than that of the juries. For the first time, however, in these experiments, we find that the index of caution seems to have little or no correlation with successful results. The differences between the percentages of caution manifest in the reports of these three finders of fact was negligible, and the very slight difference which existed, on the average, was shown to be in favor of the least successful jury. The explanation of the judge's superiority over both juries, therefore, and of the superiority of the women's jury over that composed of men, is not to be sought in native conservatism, or in added suspiciousness of attitude on the part of those finders of fact who turned in the best results. It is frequently assumed by trial judges and lawyers that the "hardboiled" attitude so quickly acquired in the courtroom is conducive to increased success in evaluating witnesses' testimony; but such a conclusion is not indicated by the results of this preliminary experiment.

Since the percentage of caution seemed to have little bearing upon results, and since, in actual cases, judge, jury, or master in chancery would make findings of fact in the form of unified statements based almost altogether upon direct and cross examination of witnesses, I adopted this form of finding for the judges and juries in my second experiment. Instead of scoring the reports of finders of fact under the separate headings of free narration, direct examination, and cross examination, the maximum number of different points which it was possible to score in the entire report was used as the basis from which percentages were reckoned, and a single score, therefore, resulted for each entire set of findings. In such a method of scoring, of course, the index of caution could not be figured, since this index depended upon specific answering of questions. In Table 5, below, percentages of completeness and accuracy are presented, in order of merit, as obtained by the findings of male and female judges, not lawyers, and by male and female juries, as in the preceding experiment.

TABLE V

SCORES OF FEMALE JUDGE AND JURY, AND MALE JUDGE AND JURY

	Female Judge	Male Judge	Female Jury	Male Jury
Completeness	61.7	45.9	48.9	44.8
Accuracy	88.9	85.7	77.4	92.6

The first notable result shown by comparison of Tables IV and V is the marked increase in the absolute values of the scores of finders of fact in the second experiment. The absolute average for the completeness of findings in the first experiment was only 28.8%, while in the second it rose to 50.3%, an increase of 21.5%. Similarly, the first total average accuracy of report was 75.5%, as against an average of 86.1% in the second experiment, or a gain for the latter of 10.6%. There was, as we noted above, a slight gain in completeness of testimony by the witnesses in the second experiment; but since there was also manifest a slight loss in accuracy, we cannot look to this source for an explanation of the marked increase in absolute scores of finders of fact for the second incident. Rather, it seems, the main cause for this difference is to be sought in the different methods of computing the scores in question. Although the second method, which resulted in a single general score for each set of findings as a unit, would seem, at first thought, nearer to actual conditions obtaining in court than would the first method; as a matter of fact, its sharp increase in absolute score values is largely artificial. For instance, had the second method been applied in figuring witnesses' scores in the second experiment, a witness might have testified in direct examination that one book carried by the stranger was red, another green, and the third blue, receiving three points in his score for these correct answers. But in cross examination he might have been asked if one of those books wasn't, as a matter of fact, yellow. If he answered, "No," he must, on the second method of scoring, receive an additional point, since he at no time had previous opportunity to reject this erroneous suggestion. But he has not added any real value whatever to his testimony. He has merely repulsed an attack upon his credibility as a witness. If he hesitated in repulsing this attack, a jury in actual practice might well lessen their estimates of the value of his testimony; but if he replied firmly and consistently, they would not have any really new matter to add to what he had already given. Similarly, the finders of fact in the second experiment were enabled to gain many points in their absolute scores by making correct findings upon matters placed in issue by leading cross examination inquiries which, by the second method of scoring, must be reckoned in as new points. If the first method were employed, on the other hand, each type of testimony

would be scored separately, testimonial reduplications like that cited above would be merely averaged together in the final result, and each successful finding would, therefore, receive its logically proper weight. We may conclude, then, that the second method of scoring, though more legally realistic, did not prove logically or psychologically sound.

In light of this analysis, then, the absolute increase in fact-finding scores of the second experiment should not be given appreciable weight, although the comparative relations between scores of different finders of fact retain the same value as in the previous scoring method. In considering this comparison of scores, as shown in Table V, it becomes evident that women again succeeded in carrying off the laurels. Not only did the second female jury excel the male jury, by a slight margin, in the completeness of their correct findings, but the woman judge, who was, it will be remembered, a newspaper correspondent, bettered the completeness scores of both juries and that of the male historian judge by very considerable margins. The male judge, on the other hand, obtained a better score than the male jury in completeness of findings, but not quite as high a score in this particular as that of the female jury. Accuracy scores do not appear to follow completeness scores, the male jury being highest in accuracy, though lowest in completeness. It may be said, in explanation of this rather odd reversal, that members of this particular male jury were well acquainted with one of the witnesses, who happened to be most accurate in his testimony of all those who participated in the second experiment. It would seem, from careful review of the findings and testimony, that the male jury in question followed closely the testimony of the witness they knew personally, as an accuracy criterion for all the rest, whom they did not know. This is not a condition likely to occur in court procedure, and might, in fact, if proved, furnish grounds for the granting of a new trial. Unfortunately, it was impossible for me, after discovering this peculiarity of result, to repeat the experiment under comparable conditions with a new male jury. The artificially high accuracy score of the first male jury may be disregarded, however, without impairing the general comparative values of the results. The low accuracy score of the female jury should receive, as I suggested above, a negative value in comparative rankings of finders of fact.

The woman judge, then, takes first place by a significant margin. The man judge follows, excelling by a small and rather dubious percentage the female jury. The woman jury excels the male jury by a slight margin also if relative accuracy results under uniform experimental conditions may be regarded as following the indications of the

first experiment, and the results of the supplementary experiment, reported below.

Table VI, following, shows a comparison of the amounts by which each of the three judges excelled the most successful jury reporting in the same experiment. Because of the difference of method in scoring the findings of fact in the second experiment, no comparison of absolute scores could be made as between the eminent lawyer judge, who made findings for the first incident, and the reporter and historian judges, who made findings of fact for the second incident. It seemed equable, however, to make such comparison upon the basis of comparative excellence in the separate experiments, since difference in scoring methods did not, as above pointed out, affect comparative relations between scores in the least degree. The purpose of this comparison is to open up the problem of whether or not legal training and experience in dealing with evidence is more or less effective, psychologically, than other types of fact-finding experience. No suggestion of a final solution of this problem, or of the problem of possible sex differences in fact-finding ability of judges is intended to result from this preliminary analysis, but merely a tentative presentation of promising avenues for psycho-legal investigation.

TABLE VI

COMPARATIVE SUPERIORITIES OF LAWYER JUDGE, REPORTER JUDGE, AND HISTORIAN JUDGE OVER THEIR RESPECTIVE, HIGHEST JURIES

	Reporter Judge	Lawyer Judge	Historian Judge
Completeness	+12.8	+ 9.9	-3.
Accuracy	+ 3.7	+ 2.4	+8.3
Net, or key score.....	+16.5	+12.3	+5.3

Table VI, above, clearly indicates that considerable differences in fact-finding proficiency may exist as between different professions, and as between different methods used. A general guess might be hazarded, as an hypothesis for further investigation, that those callings which depend for their factual bases upon direct eliciting of data from human sources might be found to secure greatest success in completeness and accuracy of result, while it would not be psychologically surprising if legal training and experience were found to rank comparatively low in the scale of efficient disciplines for successful finding of facts.

The net or key score, as given in Table VI, is merely the algebraic sum of the completeness and accuracy differences between each judge and the highest jury in that experiment; and this key score may be used as a rough indicator of the net comparative proficiency of each

judge as a finder of fact. In a more extended series of experiments which might seek actually to establish statistical ratings for fact-finding ability in different professions, a more elaborate and exact method of arriving at the final key score should be devised.

SUMMARY

Review of Tables IV, V, and VI shows the following uniform results:

A—A single trained individual, sitting as judge, is more successful as finder of fact than is a male jury or a female jury, the judge excelling the jury more significantly, on the average, in the completeness of correct findings than in the accuracy of total report.

B—Female juries excel male juries, on the average, both in completeness of findings and accuracy of report.

C—Professional training and experience, and possibly also sex, appear to have some psychological bearing upon success in individual finding of facts.

5. COMPARISON OF FINDINGS OF FACT BY JURIES, BASED UPON WRITTEN AND ORAL TESTIMONY

It is frequently asserted by presiding justices at actual trials that much fact-finding value is to be derived from the opportunity given a jury, in the court-room, of observing each witness, face to face, as he gives his testimony on the witness stand. This value, if any, must have been totally lacking in the findings of fact made by judges and juries in the experiments reported above, since all testimony was presented to them in typewritten form, and the witnesses were not even known by sight to the finders of fact, except in the accidental instance above mentioned. Such, indeed, is the usual condition obtaining in Aussage tests, since it is difficult and often impracticable to arrange for a dozen or more witnesses to give their testimony, orally, before an experimental court. However, this problem should not be overlooked in outlining the field for further psycho-legal investigation, since if a material loss of testimonial value were indicated as a result of failure on the part of the finders of fact to observe the witnesses while they testified, routine Aussage procedure must be radically revised to permit such observation during testimony orally given. This was the final problem, therefore, which I undertook to investigate before attempting a comparison of the values of findings of judges and juries with the value of the original testimony upon which such findings were based.

METHODS

Two juries, male and female, were assembled simultaneously, with the aid of Mrs. Ann Severance Hartman, Mrs. E. H. Marston, and Captain F. M. Van Natter, whose assistance I gratefully acknowledge. The jurors were all of mature age, advanced education, and unimpaired perceptual abilities. The 12 witnesses to the second experimental incident were brought before the juries, one at a time, in the same lecture-room where the incident had originally occurred. Each witness was asked, first, to tell what happened in his own way, without questioning. He was thereafter examined in both direct and leading form, just as he had been questioned when he wrote out his report, immediately after the occurrence in issue. All witnesses remained in a separate witness-room, across the hall from the lecture-room, until called to the stand, in accordance with strictest trial procedure. Jurors were allowed to take such notes as they wished during testimony, and the testimony was taken down verbatim by stenographers who were present. The transcript of testimony was not furnished the juries during their deliberations, since this would have necessitated impracticable delay, and since juries in actual trials usually make little, if any, use of the testimonial record.

Necessarily, as in actual cases, court did not sit, as above described, until three months after the incident in issue occurred. It is to be noted, however, that every witness had written out his complete testimony immediately after the incident. This might well tend, psychologically, to fix the memory in his mind, since kinæsthetic and visual imagery of the testimony, thus set down would be added to perceptual imagery of the original incident. Moreover, the real incident was reviewed at the lecture following its occurrence, and witnesses spent much time discussing together their individual errors. In view of these memory reinforcements and corrections, and in light of the fact that actual trials often take place several years after the occurrences which are testified to thereat, a higher percentage of oral testimonial value might well be expected than that obtaining in actual cases. E. H. Marston⁴ sat, as an observer, with the female jury after it had retired to the jury-room to make its findings; and I sat, similarly, with the male jury while its report was decided upon. Results of these observations proved valuable in suggesting a possible explanation of sex differences in fact-finding success.

⁴Member of the Boston Bar.

RESULTS

One of the most interesting details of the results of this experiment was the extraordinary testimony given by one of the witnesses, based, as was afterward revealed, upon deliberate suggestions made to him in the witness-room by the other witnesses awaiting their turn to testify. The men were evidently comparing their memories of the incident, to start with, and began to exaggerate and make fun of each other's alleged mistakes. Finding Mr. A. hypersuggestible, they combined to put into his mind the most absurd and improbable details. When called before the juries, in fact, Mr. A. actually testified that the actor in the incident in issue wore one high, black shoe, and one low, tan shoe; that he had some sort of gold medal in the buttonhole of a bright scarlet coat, and that he wore a flaming red tie, knotted in spread-eagle fashion, outside coat and vest! More significant than the mere fact of this witness' unusual suggestibility to mischievously motivated "coaching," however, was the resulting effect upon certain jurors. One woman juror was so impressed by the extreme self-confidence shown by Mr. A. on the stand that she came to the conclusion that he was the only reliable witness out of the 12 men called. As a result of her insistence upon this point, at least two erroneous findings were introduced into the female jury's report. A member of the male jury argued, also, that such extraordinary statements would never have been made by any witness unless he had actually seen details which the other witnesses were not in position to observe. After extended discussion, several of Mr. A.'s statements were adopted as correct by the male jury, thus materially reducing both their completeness and accuracy ratings. Such disproportionate and erroneous importance given by jurors to prejudiced witnesses' self-assertiveness on the stand may well play an important role in many a jury's final findings.

Table VII, following, compares the scores of male and female jury findings of fact based upon oral testimony, as above described, with similar scores of findings of fact reported by the University of Maryland juries, and based solely upon the written testimony of the same witnesses to the same incident. In order to make these jury scores fairly comparable, both were figured by the second method described above.

TABLE VII
COMPARATIVE SCORES OF JURY FINDINGS BASED UPON ORAL AND
WRITTEN TESTIMONY

Completeness—	Oral	Written	Superiority when based on written testimony
Female jury	38.6	48.9	10.3
Male jury	33.6	44.8	11.2
Average	36.1	46.8	10.7

Accuracy—	Oral	Written	Superiority when based on written testimony
Female jury	80.6	77.4	—3.2
Male jury	68.8	92.6	23.8
Average	74.7	85.	10.3

It is to be noted, in Table VII, that the accuracy percentage of the male jury, which was discussed at length above, is actually based upon the written testimony of a single witness, instead of upon the written testimony of all 12 witnesses. While this accidental variation must invalidate direct comparison of male jury accuracy with the accuracy scores of corresponding finders of fact, as noted in the foregoing discussion, it has no serious bearing upon comparison of average accuracy of written testimony method with average accuracy of oral testimony method. Since all the testimony furnished both Maryland University juries was written, all their findings must have been based solely upon material received in written form. It may well be, however, that accidental outside knowledge of a single witness' reliability might also have raised the total average accuracy percentage of findings based upon oral testimony. I suggest, nevertheless, that such previous confidence in a witness would probably be found to have less effect, in the oral procedure, since fresh perceptions of each witness as he testifies might be expected, psychologically, to supplant or materially modify previous estimates of the particular witness' accuracy.

SUMMARY

The results and observations in this experiment may be summarized as follows: .

A—Jury findings based upon testimony written out by witnesses immediately following the occurrence are superior in both completeness and accuracy to jury findings based upon oral testimony of the same witnesses heard by a jury some time after the event in issue. What proportion of this superiority, if any, is due to the mere lapse of time between incident and oral testimony is not indicated by this experimental procedure, and should be made the subject of subsequent investigation.

B—The self-confidence, or self-assertiveness, of a witness on the stand, even though that witness be the most obviously unreliable and improbable of all those heard, may have a greater influence upon jurors of both sexes than does the logical or psychological probability of other testimony.

C—Careful comparison of the observations of E. H. Marston and myself upon the jury-room deliberations of female and male juries, respectively, gave the following results:

(1) The female jury exercised much greater care in considering detailed testimony than did the male jury. The women jurors were more painstaking, and manifested a thoroughness and a willingness to put themselves to more trouble and to do more work in considering all the testimony submitted, no matter how trivial or obviously erroneous it seemed, than the men jurors.

(2) Both juries pursued, without previous discussion or consultation, the same method of compiling their final findings; that is, separate consideration, discussion, and vote upon each moot point, with subsequent revision of previously agreed on findings, where this became necessary.

(3) Individual jurors of both sexes appeared equally prejudiced and illogical in discussion of certain points; and both juries, taken as units, seemed equally unable to "size up" or psychologically analyze the behavior and responses of witnesses on the stand which had been especially noted.

6. COMPARISON OF FINDINGS OF FACT WITH TESTIMONY ON WHICH FINDINGS WERE BASED

Having examined the preliminary psycho-legal problems of the relationships existing between the findings of judge and jury, and between the findings of juries of different sexes, we may finally compare the scores of all these findings of fact with the average score of the testimony upon which each finding was based. Table VIII, below, contains such a comparison. Necessarily, each experiment, or experimental incident, had to be treated separately, since all finders of fact in the second experiment were scored by the second or "lump-score" method; and all the witnesses in that experiment had, therefore, to be re-scored by the same method for the purpose of this final comparison with the scores of the finders of fact who passed upon their testimony.

TABLE VIII
COMPARATIVE SCORES OF JUDGES, JURIES, AND WITNESSES

	—Completeness—		—Accuracy—	
	Score	Superiority over witnesses	Score	Superiority over witnesses
Unexpected incident (first scoring method)—				
Witnesses	27.6		84.3	
Judge	36.7	9.1	80.8	-3.5
Female jury	26.8	-.8	78.4	-5.9
Male jury	23.	-4.6	67.3	-17.
Average of judge and juries.....	28.8	1.2	75.5	-8.8

	—Completeness—		—Accuracy—	
	Score	Superiority over witnesses	Score	Superiority over witnesses
Expected incident (second scoring method)—				
Witnesses	40.7		73.4	
Female judge	61.7	21.	88.9	15.5
Male judge	45.9	5.2	85.7	12.3
Female jury (written testimony).....	48.9	8.2	77.4	4.
Male jury (written testimony).....	44.8	4.1
Female jury (oral testimony).....	38.6	-2.1	80.6	7.2
Male jury (oral testimony).....	33.6	-7.1	68.8	-4.6
Average of judges and juries.....	45.5	4.8	80.2	6.8

The preliminary point of interest to be noted from the results in Table VIII, above, is the artificial rise in the average completeness score of the witnesses in the second experiment when the "lump-score" method of scoring is adopted. In Table VIII the average completeness of witnesses' testimony is 40.7%, as against the 33.7% average in Table III, figured by the original method of scoring. The average accuracy of these witnesses, on the other hand, appears to be 3% less when figured by the second method, a result which calls attention to the suggestibility element in any accuracy score over-emphasizing errors in cross examination.

The average accuracy superiority of finders of fact over the average witness was very much greater in the second, or expected incident, than in the first, or unexpected incident test. This is difficult to explain in view of the fact that the second group witnesses were more than 10% less accurate, on the average, than were the witnesses of the first group. That is, the witnesses who expected the occurrence suffered from overconfidence in reporting it, and they included 10% more erroneous surplussage to what they really knew, in their report, than did the witnesses who were not aware that they were watching an experimental incident. Yet in spite of this greater number of erroneous testimonial statements, the second group finders of fact were, on the average, 80.2% accurate in their findings instead of 75.5%, the average accuracy of the finders of fact in the first experiment. When this absolute increase in fact-finders' accuracy is taken in numerical conjunction with the decrease in average witnesses' accuracy in the second experiment, we find that the comparative average difference between accuracy scores of findings and testimony has passed from a minus quantity to a plus quantity, showing a total improvement in the second experiment average over the first of 15.6%. One might guess, in explanation of this rather odd result, that the finders of fact in the

second experiment were able to detect very gross inaccuracies in the testimony before them, while those in the first experiment were not able to detect ordinary, trifling errors; and that, as a consequence of detection of some gross errors, the second group of fact-finders were led to examine all the testimony more carefully. The data at hand, however, is insufficient to point, significantly, toward any definite solution, though it is wholly adequate to indicate the advisability of further investigation of the relation between accuracy of findings of fact and accuracy of the testimony on which such findings were based.

Table IX, following, presents a unified comparison of the findings of fact and the testimony on which the findings were based. No absolute value is to be attached to these averages, since they represent a mean made up of diverse elements in two dissimilar groups; but the comparative value of the averages is not in the least impaired by such dissimilarity, since each unit in each average is individually comparable with a corresponding unit in the contrasted average. The total average, or "indicator," score for judges, juries, and witnesses, made up as it is of a simple sum of accuracy and completeness averages, represents no fixed value, or attempt to create a compound measuring stick giving equal place to completeness and accuracy. It rather constitutes a simple means for showing, at a glance, the final, rough relationship between the success of judges and juries and the success of the average witness in reconstructing a given occurrence of an ordinary sort.

TABLE IX

COMPARATIVE AVERAGE SCORES OF JUDGES, JURIES, AND WITNESSES

	Average of 3 judges	Average of 6 judges	Average of 30 witnesses
Completeness	48.1	35.9	32.5
Accuracy	85.1	77.5	80.1
Indicator scores	133.2	113.4	112.6
Total superiority over average witness....	20.6	.8

SUMMARY

Consideration of the comparisons of results contained in Tables VIII and IX may be summarized as follows:

A—All finders of fact, averaged together, slightly excel the average witness in completeness of report, and, less clearly, in accuracy of report.

B—Accuracy of findings of fact does not seem to depend upon the accuracy of that testimony on which the findings were based.

C—Completeness of findings of fact does seem to parallel, roughly, the completeness of the basic testimony.

D—Findings of fact, by juries, based upon oral testimony, are consistently less complete than is the testimony of the average witness. Such jury findings do not differ significantly in accuracy from the average witness' testimony.

E—The total averages of six juries of both sexes show these juries to be slightly more complete, and slightly less accurate, in their results, on the average, than were the 30 witnesses on whose testimony the juries' findings were based. Indicator scores of juries and witnesses, made up of the sums of completeness and accuracy averages, differ by less than 1%.

F—The total average scores of three judges (two of whom had no legal training) significantly excel the average scores of the 30 witnesses, both in completeness and in accuracy. The judges' indicator score excels the witnesses' indicator score by more than 20%.

7. CONCLUSION

Experimentation upon the completeness and accuracy of testimony, and of the findings of fact which might be based upon that testimony, probably had its origin in an unconscious, scientific wish to prove that our present juristic system is incapable of achieving a decent degree of justice. Controversial discussion of Aussage results has always centered, therefore, upon the issue of whether or not witnesses are capable of perceiving and reporting accurately any considerable proportion of what takes place in their presence. But, if we admit that psychological experiments have proved that human testimony has a necessarily low value, what of it? That is, what can we do about it? Science of the future may, conceivably, substitute itself as a finder of facts to a large degree for judge and jury; but, after all, science can never artificially reproduce a perfect witness always at hand to observe and report everything that goes on in the world at all times. For such observations and reports, therefore, we must continue to depend upon human testimony. In light of this undeniable conclusion, it seems to me that the most profitable subjects of psycho-legal discussion and experiment are to be found in the various possibilities of practical improvement in the elicitation and use of normal, average testimony, rather than in over-emphasizing its futility. Moreover, our present trial system cannot be supplanted overnight, no matter how radically wrong its results may be proved by science. Psycho-legal investigation, therefore, should seek, primarily, to discover sources of error in the handling of human testimony in our courts which can be corrected, and gradually eliminated, without recourse to legal pro-

cedure too greatly at variance with the evidential system now in vogue. The present series of experiments sought not conclusively to determine such sources of error, but to indicate avenues of investigation which, when followed through, might scientifically establish fields for psycho-legal evolution.

In conclusion, I would point out various concrete examples of results herein reported which, if verified, might be utilized to bring about practical improvements in courtroom procedure without any prerequisite change in basic law. Judges, for instance, have wide discretion in limitation of cross examination, and granting of broad latitude in direct examination. If psycho-legal research should finally establish the fact that testimonial results in response to direct questioning have preponderantly greater completeness and accuracy, the bench would already possess sufficient power to put this conclusion into practice to a considerable extent.

Again, it is possible, under our present procedure, to take immediate, written statements from witnesses who have been present during accidents and other occurrences of ultimate trial importance. Such statements may be used in various ways during actual trial and are so used very frequently. If it were established as a definite, scientific fact that such immediate, written testimony has several times the value of oral testimony given later, at the trial, the court and counsel, knowing this fact, might add much emphasis and weight to the immediately written statements, in many ways, without departing materially from our present trial system. Indeed, it would not necessitate a very radical departure in legal procedure to provide official machinery for the eliciting and recording of crucial testimony, in important cases, as soon after the occurrence as the witnesses could be summoned; and the establishment of subsequent admissibility of such testimony, as a check upon the witnesses' later testimony, at the trial, would merely require an extension of the rule under which previous sworn testimony may be used at the present time to discredit a witness in cross examination. The official, immediate eliciting of testimony might, moreover, be conducted according to that form which psycho-legal research had proved productive of most complete and accurate results.

In the matter of women jurors, in jurisdictions where sex is no longer a bar to jury service, both counsel and court possess the power to influence the proportion of female representation on any panel. Excusing of jurors, disqualification for cause, and the arbitrary right of challenge may all be used, legitimately, to effect the result of securing women on every jury, if this result turns out to be desirable in the

interests of justice. Only continued psycho-legal investigation can establish the desirability of such procedure; but the preliminary experiments reported above certainly indicate a possible value of women as jurors which is wholly contrary to the older practice of deliberate exclusion of women from the jury, whenever possible.

If properly guarded scientific researches should establish the fact, which already seems obvious to many psychologists, that an ordinary jury has lamentable absence of ability and skill in analyzing, psychologically, the reliability of the witnesses appearing before it, and of analyzing, both logically and psychologically the testimony of those witnesses, upon which the case must be decided, much may be done in the way of producing expert analysis of testimony for the jury's assistance. Counsel may procure properly qualified experts in various aspects of testimonial analysis; the court has a wide discretion in admitting such expert testimony; and the jury may be required, by the court's charge, to give due consideration to all the testimony before it, with the possibility of a new trial should expert analysis of crucial evidence be patently disregarded. Nor is the plan of a testimonial expert officially retained in capacity of "friend of the court" so foreign to a modern practice which thus utilizes the sociological advice of probation officers and the psychiatric services of medical examiners.

Finally, if it can be conclusively demonstrated that a single, trained individual greatly excels any jury in fact-finding ability, parties to any cause, civil or criminal, who honestly desire efficient findings of fact may, in most instances, waive the right to jury trial and thus secure that result most in accordance with justice. Also, in the appointment of masters, commissioners, and other non-judicial fact-finding agents, the court or other appointing official may be directly guided by the results of adequate psycho-legal investigations indicating what types of training and experience are most conducive to completeness and accuracy of report upon the testimony submitted.

All these possibilities for practical improvement in the handling of human testimony as a result of psycho-legal experimentation are cited, not because of their supreme intrinsic importance, but merely to illustrate the possibilities of immediate co-operation between the psychological laboratory and the courtroom. Juristic theory and practice have profited enormously from scientific treatment of abnormalities; why should not a similar profit be derived from psycho-legal solution of normal problems?