Journal of Criminal Law and Criminology

Volume 22
Issue 3 *September*

Article 6

Fall 1931

Briefer Contributions: The Consistency of Testimonial Accuracy

Alfred Kuraner

BRIEFER CONTRIBUTIONS

THE CONSISTENCY OF TESTIMONIAL ACCURACY[1]

ALFRED KURANER[2]

For possibly a quarter of a century, psychologists have been pointing out that the testimony of witnesses, be their intentions honorable or otherwise, is often unreliable and inaccurate. Furthermore, these pyschologists insist on a more scientific procedure in handling cases where the errors of testimony due to human weaknesses may be present. In pursuance of this phase of legal psychology, hundreds of experiments have been performed to prove the unreliability of testimony, and the percentages of accuracy and completeness of the testimony of groups of witnesses have been worked out to a rather fine degree.

The attitude of the courts toward psychology is indicated in *Strand* v. *State*,[3] where it was held that evidence of the intelligence quotient of a ten year old prosecutrix was objectionable and unnecessary.

Wigmore probably expresses the general attitude of the legal profession when he says, "But where are these practical psychological tests, which will detect specifically the memory failure and the lie on the witness stand? . . . If there is ever devised a psychological test for the valuation of witnesses, the law will run to meet it. . . . Whenever the psychologist is really ready for the courts, the courts are ready for him."[4]

One of the faults in working with experiments in testimony in the past has been that, although group accuracy and completeness have been determined on numerous occasions, the individual himself has been neglected. That is, we do not know whether or not a particular individual's testimony is consistently accurate or inaccurate; whether an individual who is unusually accurate in reporting one incident is neces-

[1]The experiment herein reported was performed in the Psychology Laboratory of the University of Kansas. The writer wishes to thank Doctor Beulah M. Morrison for her advice and assistance in performing the experiment and in preparing this report.

[2]State University, Lawrence, Kansas.

[3]36 Wyo. 78, 252 Pac. 1030.

The words of the court follow:

"Before the prosecutrix was put on the stand, the state examined two witnesses to show her mental capacity. One of these witnesses gave some hearsay testimony as to a mental test, stating that the report of the test showed that the prosecutrix had an intelligence quotient of 104 (100 being the average) and that this indicated that she was above the average in mental ability. This evidence was probably objectionable. . . . It was also unnecessary. . . . When the jury had seen the prosecutrix on the stand, and heard her give all her testimony, they then had a so much better way of judging of her intelligence, that we are sure they could not have been influenced by the objectionable evidence about the mental test."

[4]J. H. Wigmore, Wigmore on Evidence (1923), Vol. 2, § 875.

sarily accurate in reporting the next. In this paper we speak only of the witness who honestly tries to report things as he saw them, and not of the witness who deliberately falsifies.

Of great significance is Hugo Munsterburg's experiment of more than twenty years ago,[5] wherein he showed that of the eighteen men (out of a class of about one hundred) who did not see the very obvious movements of his left hand, once he had centered their attention on his right, fourteen were the the same witnesses who reported, in all seriousness, that of the two colors held before them, the light grey was darker than the dark blue. The possibilities indicated by Munsterburg's experiment are as important as they are obvious. If some witnesses are consistently far below or far above average in their testimony, despite their honest intentions, and if their percentage of accuracy could be determined by some sort of test (see excerpt from Wigmore on Evidence, quoted above) before they enter the court room, then their testimony could be objectively evaluated according to their test scores.

Several authorities have expressed their faith in the possibility of such a procedure. For instance, Chafee says: "Although the law has refused to admit lay evidence that a witness' mentality is low, except when it approaches insanity, because such evidence is too uncertain, the report of a Binet-Simon or other intelligence test would be of distinct value to a trained judge in weighing testimony."[6]

As a step toward the ultimate possibility of measuring the value of testimony by means of test-scores of witnesses, the experiment herein reported was performed. It consisted of enacting three incidents, several weeks apart, before a group of sixty-five students of general psychology, most of them sophomores in the university, and having them report on the incidents later. The reports consisted of giving a free report of the incident, followed by answering a set of detailed questions concerning the incident.

In selecting the incidents, two important rules, suggested by Marston,[7] were followed; first, the incident must not be wholly foreign to the conscious content of the individual at the moment the incident occurs; and, second, the incident must have a logical meaning of its own and must not contain any tricks to increase the inaccuracy of the testimony. The incidents selected were similar to several used often in experiments in testimony.

The first incident, intended to cause a slight emotional shock to the witnesses, was enacted as follows: The instructor was lecturing as usual to her 11:30 elementary psychology class. She was standing behind a long table at the front of the room. At the north end of the table was a stack of large charts which had been used earlier in the year in the same class, and on the south end of the table was a front-heavy device specially constructed for the experiment, which presented the appearance of bona fide apparatus. It was made mostly of wood, wire and electric switches, and on

[5]Hugo Munsterburg, On the Witness Stand, 1908, pp. 28-31.

[6]Z. Chafee, Jr., The Progress of the Law, 1919-1921, Evidence, 35 Harvard Law Review, 302, 308.

[7]W. A. Marston, Studies in Testimony, JOURNAL OF CRIMINAL LAW AND CRIMINOLOGY, Vol. 15, p. 5, 1924-1925.

its front were five light bulbs inserted in the same number of sockets. At 11:38 a graduate student in the psychology department entered the room, walked to the north end of the table, picked up the charts, and, holding them horizontally, started to walk out of the room. As he reached the end of the table, he struck the apparatus with the charts, as if by accident, so that it fell to the floor with a crash, breaking one of the bulbs. The student picked up the apparatus and placed it in full view of everyone in the class, following which the instructor rebuked him mildly for his carelessness.

Nothing more was said in class concerning the incident until one week later, when the instructor brought up the subject with the following remarks:

"Some questions have arisen concerning the incident that occurred in this class recently, at which time a piece of apparatus was knocked from the table. We are going to ask that each of you assist us by writing an independent account of what happened. Take a clean sheet of paper, write your name at the top, and then write the account."

After sufficient time had been allowed for each student to write a free report of the incident, the papers were collected and the instructor continued as follows: "Now we ask you to answer certain questions about that same incident. Number the questions as they are asked, and write the answers only." Each student wrote the answers independently.

There were forty-one questions. In order to keep the number of variables as small as possible, the proportion of questions relating to color, dimensions, time, and the like, was kept the same for all three incidents. The form of questions also was kept the same; that is, the same proportion of leading and direct questions, etc., was kept constant.

A few students, not quite certain of their opinions, suspected before all the questions had been asked, that the whole procedure was an experiment, but, on the whole, the papers indicated that the students were earnestly trying to help solve the difficulties which seemed to have arisen.

The results of the experiment conformed in general to the results of other experiments in testimony. The free reports were comparatively accurate, but also very incomplete. The answers to questions, which were intended to correspond to the direct and possibly the cross examination of the court room were very inaccurate. For instance, of the fifty-five students who answered the question, only nine remembered that there were five lights on the face of the apparatus. The other answers varied from two to twelve, thirteen of the witnesses declaring that there were six, and almost an equal number stated that there were three or four light bulbs on the face of the apparatus. The estimates as to the number of lights broken also varied from two to twelve, whereas the device had been placed on the table in such a way that every one in the class could see that only one had really been broken. The graduate student entered the room at 11:38. In answer to the question as to what time he entered the room, the reports varied from ten minutes to eleven to ten minutes after twelve. Two students answered that he entered before half past eleven, probably forgetting for the moment that they were not even in the class room before that time. On the

whole, however, the percentage of accuracy on the part of the students was about as high as is usually found in experiments in testimony.

The second incident was enacted about two weeks later under approximately the same circumstances as the first. The incident follows: As the instructor was lecturing to her 11:30 class, a telegraph messenger in uniform entered with a telegram and handed it to the instructor. He went through a number of little acts, such as stealing a fountain pen from the desk, making a notation on a writing pad, etc., all of which had been previously rehearsed, so that answers to the questions concerning his actions could be controlled.

A week later the students were asked to write free narrative reports of the incident, and then were asked to answer about fifty detailed questions in the manner of examination in the court room. Since this incident was one which might happen at any time, none of the students indicated in any way that they suspected it was an experiment until the questioning. They then knew, of course, that it was simply an experiment, but the papers indicated that they were seriously trying to do their best to write accurate reports and to answer the questions correctly.

A number of the questions were unreasonable, but some of the questions asked of witnesses in court are equally unreasonable. For instance, a question as to the number on the messenger's cap could quite conceivably have been asked in court. On the whole, the set of questions for the second incident was more difficult than that for the first. Of the sixty-one witnesses who handed in papers, twenty-eight said correctly that the incident happened one week before the questioning, on the preceding Monday, March 19. The remaining answers varied from March 6 to March 21. Curiously enough, twenty-one students remembered for an entire week that the messenger carried a pencil behind his ear. In response to the question, "Did the messenger at any time place his right hand on the table?" fifteen answered, "Yes," but only one student in the entire group saw him calmly pick up the fountain pen from the table and place it in his pocket.

The third incident was not an incident at all. Since a high degree of correlation was expected in the consistency of accuracy of testimony of individuals for the first and second incidents, this incident was included as a preliminary step in determining whether a difference in the amount of time elapsing between the incident and the giving of testimony would tend toward a lower degree of correlation in the consistency of the witnesses in their testimony between the third and the other two incidents. It consisted simply of showing to the members of the class a large cinema poster for seventeen seconds after which they were asked to write as complete a description of the poster as they could. They were then asked a number of questions, in which the proportion of questions relating to time and color and the like was the same as in the first two incidents.

In allowing a week to elapse between the incident and the testimony in the first two incidents, the effect was the same as in actual litigation or prosecution; that is, seldom any less, and in most cases more than a week's time elapses between an incident and subsequent testimony concerning it. Each incident was rehearsed before being

given in class, and the answers to questions were noted then, and subsequently checked when the incident was presented in class.

The system of grading papers was somewhat difficult to devise, and required careful consideration. It is to be noted that an answer of "I don't know" is preferable to an incorrect answer[8] A witness who is cautions enough to avoid guessing is certainly more helpful in arriving at the truth of the matter than one who answers indiscriminately both questions he knows and those which he does not know. It was therefore decided to give the witness credit for the percentage of questions answered correctly out of those which he answered, rather than the percentage of questions answered correctly out of the entire number asked. An objective criterion for the grading of free reports was even more difficult to devise, but when it was found that virtually every point brought out in any of the papers was covered by the questions pertaining to that same incident (the questions in each case were asked after the free reports had been handed in, in order not to refresh unduly the witness' recollection) this solution presented itself: (1) Count the separate facts in each free report which would serve as answers to the questions concerning the incident in question. (2) Divide this figure by the total number of questions on that incident to obtain a percentage of completeness. (3) Then find the proportion of correct answers to all the answers (on facts covered by the questions) brought out in that free report to determine the percentage of accuracy. (4) Average the accuracy and completeness percentages for the free report score. These methods of grading the answers to questions and the free reports may or may not be sound; but since they were used consistently throughout the experiment, the correlations of the accuracy of the various individuals' testimony should not be greatly affected by the grading system.

The Pearson product - moment[9] method of correlation was used in treating the data statistically. The object of the experiment was to show the consistency of the accuracy or inaccuracy of the testimony of individuals. Thus, if the witness who received the highest grade in answering the questions of the first incident, had received the highest grade in answering the questions of the second incident, the next highest in the former had received the next highest in the latter, and so on down the line, there

[8]Robert M. Hutchins and Donald Slesinger, Some Observations in the Law of Evidence—Memory, 41 Harvard Law Review, 860.

"In other words, it is safe enough to trust absolute subjective certainty as an indication of objective accuracy, but anything less than absolute is no better (and it may be worse) than absolute uncertainty."

[9]The Pearson product-formula is $r = \frac{\Sigma xy}{N\sigma x\sigma y}$ (See H. E. Garrett, Statistics in Psychology and Education, 1926, pp. 168 ff.) A perfect positive correlation (+ 1.00) between two variables, as intelligence and testimonial accuracy, indicates that as one increases the other does also in exact proportion. A perfect negative correlation (— 1.00) indicates that as one variable increases, the other decreases in exact proportion. A zero correlation (0.00) indicates no relationship between the variables. Actual correlations obtained from scientific data range between + 1.00 and — 1.00.

would be a perfect positive correlation. Quite the opposite was found to be true. The table of correlations follows:

Variables or Factors	Coefficient of Correlation	Probable Error
Free report, incident No. 1 with Answers to questions, No. 1	.08	± .095
Free report, incident No. 1 with Free report, incident No. 2	.35	± .084
Free report, incident No. 1 with Free report, incident No. 3	— .04	± .095
Free report, incident No. 2 with Free report, incident No. 3	.04	± .095
Free report, incident No. 2 with Answers to questions, incident No. 2	.03	± .093
Free report, incident No. 3 with Answers to questions, incident No. 3	.37	± .083
Answers to questions, incident No. 1 with Answers to questions, incident No. 2	.10	± .095
Answers to questions, incident No. 1 with Answers to questions, incident No. 3	.08	± .096
Answers to questions, incident No. 2 with Answers to questions, incident No. 3	.10	± .095
Average of all scores for all incidents with scores in Intelligence Test.	— .03	± .093

The first requisite for the reliability of a correlation is that it exceed the product of at least four times its probable error.[10] The explanation of the reason for this rule is not within the scope of this

[10]The formula used here for computing the probable error of the Pearson coefficient is $PEr = .6745x \frac{(1-r^2)}{\sqrt{n}}$. (See Garrett, op. cit. p. 170.) A simple illustration may indicate the significance of the probable error. If we were to attempt to find the average salary of a million wage earners, it would be somewhat impractical to handle a million numbers. We therefore take 500 cases at random, and find that the average salary is $30 per week with a probable error of ± $5. This means that if we were

paper, but it is to be noted that all the correlation figures were low and all the probable error figures were high. In fact, of the ten correlations above, only two exceed the product of four times the probable error, and these two do so only by a narrow margin. In several of the remaining instances, the probable error even exceeds the correlation. Furthermore, a correlation figure of less than .65 is too small to be indicative of any substantial correlation, so that taken as a whole, there is almost perfect lack of correlation which indicates little consistency in the individual's testimony regarding these three incidents.

The last correlation figure in the table refers to the correlation between the average of all the grades of each individual, and his grade in the psychological examination which he took upon entering the university. It has been found that the psychological entrance examinations (intelligence tests) to the University of Kansas, agree quite well with the subsequent scholastic records of the students; nevertheless, in this experiment as shown above, there was no correlation between the scores in the psychological examinations and the averages of the grades for each individual for all three incidents.

It would seem, then, that if the proper method has been used in carrying out the experiment, that there is no reason to believe that those who testify accurately on one incident, will necessarily testify accurately on another. Those who scored high in one part of the test failed in the next. The noticeably higher correlation figure in two instances in the above correlation table is unexplained, but it is to be noted that they are still too low to be of any value, particularly since they exceed the product of four times their probable error by such a narrow margin.

Giving testimony is commonly said to consist of observation, recollection, and communication or narration.[11] Although it is probable that neither observation nor narration is entirely at fault for the lack of correlation in the present experiment, neither is it accurate to

to take another 500 cases at random, the chances are exactly even that the average would fall between $25 and $35, but not necessarily at $30. On the other hand, if our probable error had been ± $25 instead of ± $5, the chances would be even that the average would fall between $5 and $55 for the next 500 cases. Obviously, although the average is the same in both cases, the first one, having a smaller probable error, is more reliable than the second.

Applying these remarks to the case at hand, we find, for example, a correlation between the free report of the first incident with the answers to questions of that incident of + .08 with a probable error of ± .095. This means that if we were to perform the experiment in exactly the same way with another group of the same number and about the same ability, the chances are exactly even that the correlation would fall between — .015 and + .175 which is obviously so wide a range that we are forced to conclude that our correlation is not reliable for that reason, in addition to its being so small as not to indicate a substantial relationship.

It follows, then, that the smaller the probable error, the more reliable is the figure in question. Statisticians have reached the rule that a correlation must exceed the product of four times its probable error to be reliable. The reason for that rule, and the means for reaching the probable error formula cannot be given in the space available here. See F. C. Mills, Statistical Methods, 1924, p. 160.

[11] J. H. Wigmore, Wigmore on Evidence (1923), Vol. I, § 478.

assume that they were not of some influence in the final result. It is highly probable, however, that the strongest variable leading to the lack of correlation is recollection, but there is no way of proving it from this experiment. The difficulty with experiments in testimony is fundamental and inherent. We say testimony consists of observation, recollection, and narration, but these constituent elements are not mutually exclusive, and therefore not separable. Each of these elements in turn contains an infinite number of elements, varying with the incident. The fact is that each situation is a whole involving so many variables that any analysis of it borders on the impossible. It is hopelessly impracticable to isolate the objective elements which affect an individual's testimony; it is equally impossible to isolate the subjective elements which ultimately result in testimony. Disregarding these practical difficulties, we blindly proceed, in this experiment and similar ones, to attempt to place objective values on intricate subjective concepts; the "assignment of numerical values to non-quantitative material" is almost certain to be misleading.[12]

Other criticisms to the experiment suggest themselves. For instance, there was too little difference between the subjects. To be sure, they range from the highest ten per cent to the lowest ten per cent in their scores on their psychological examinations taken upon entering the university, but even so, there are not the tremendous differences among them that we find among witnesses actually before the court. There are usually no morons in the university, but morons are occasionally asked to testify in court. There are differences in intelligence among university students, but these differences are small when compared to those among the thousands of individuals who come before the courts as witnesses. If an experiment similar to the one herein described were performed on a group wherein the differences in intelligence were greater than they are in a university class, there might be found a greater consistency of testimonial accuracy.

The fact that the subjects were too nearly of the same intelligence not only explains the lack of consistency in accuracy, but also the lack of correlation between the average of testimony scores of each subject with his intelligence test score. Thus if the subjects had represented all grades of intellectual ability, from the imbecile to the genius, it is quite probable that the testimony grades would have correlated at least slightly with intelligence test scores. For this reason, the following paragraph of Hutchins and Slesinger may be sound if it is understood to apply generally, and not to a group of people of almost the same intelligence:

"In evaluating the memory of a particular person in a particular situation, psychology has developed a number of objective tests which the courts are reluctant to admit. The intelligence tests which have been most widely used and are therefore the best standardized may be admitted in evidence without hesitation. These tests, since they have a high correlation with recall,

[12]Donald Slesinger and E. Marion Pilpel, Psychological Bulletin, December, 1929, Vol. 26, p. 679.