

UNCLASSIFIED

AD 93165

Armed Services Technical Information Agency

Reproduced by

DOCUMENT SERVICE CENTER

KNOTT BUILDING, DAYTON, 2, OHIO

This document is the property of the United States Government. It is furnished for the execution of the contract and shall be returned when no longer required, or upon recall by ASTI to the following address: Armed Services Technical Information Agency Document Service Center, Knott Building, Dayton 2, Ohio.

NOTICE: WHEN GOVERNMENT OR OTHER DRAWINGS, SPECIFICATIONS OR OTHER DATA ARE USED FOR ANY PURPOSE OTHER THAN IN CONNECTION WITH A DEFINITELY RELATED GOVERNMENT PROCUREMENT OPERATION, THE U. S. GOVERNMENT THEREBY INCURS NO RESPONSIBILITY, NOR ANY OBLIGATION WHATSOEVER; AND THE FACT THAT THE GOVERNMENT MAY HAVE FORMULATED, FURNISHED, OR IN ANY WAY SUPPLIED THE SAID DRAWINGS, SPECIFICATIONS, OR OTHER DATA IS NOT TO BE REGARDED BY IMPLICATION OR OTHERWISE AS IN ANY MANNER LICENSING THE HOLDER OR ANY OTHER PERSON OR CORPORATION, OR CONVEYING ANY RIGHTS OR PERMISSION TO MANUFACTURE OR SELL ANY PATENTED INVENTION THAT MAY IN ANY WAY BE RELATED THEREWITH.

UNCLASSIFIED

**Best
Available
Copy**

93765

WADC TECHNICAL REPORT 56-143

**A SCREENING SEMI-QUANTITATIVE METHOD FOR THE
DETERMINATION OF
CARBON MONOXIDE IN BLOOD**

JACK MAYER, MAJOR, USAF (MSC)

AERO MEDICAL LABORATORY

FC

APRIL 1956

WRIGHT AIR DEVELOPMENT CENTER

WADC TECHNICAL REPORT 58-143

**A SCREENING SEMI-QUANTITATIVE METHOD FOR THE
DETERMINATION OF
CARBON MONOXIDE IN BLOOD**

JACK MAYER, MAJOR, USAF (MSC)

AERO MEDICAL LABORATORY

APRIL 1956

PROJECT No. 7159

**WRIGHT AIR DEVELOPMENT CENTER
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE
WRIGHT-PATTERSON AIR FORCE BASE, OHIO**

FOREWORD

The work reported herein was performed in support of Project 7159 entitled, "Health Hazards of Air Force Materials," - Task 71803, which is administered by the Aero Medical Laboratory, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio. The author, Major Jack Mayer, USAF (MSC), developed the test procedure and directed the application of the test. Lieutenant Colonel Fred E. Holdrege, USAF, and the authors of the appendix, Lieutenant Richard L. Deininger, USAF (MSC) and Mr. James Smithson, guided the statistical studies. Dr. George Kitzes contributed technical assistance. Mr. Roscoe Logsdon of the Technical Photographic Division, Wright Air Development Center, prepared the standard diffusion reaction chart in color. The clinical and pathological laboratory technical personnel of the Wright-Patterson Air Force Base Hospital cooperated with the practical application of the proposed test. The initial research was started in November 1954 and evaluation of the test was completed in January 1956.

ABSTRACT

A simple diffusion technique is presented for the determination of carbon monoxide in blood. Statistical evaluation of the data derived from quantitative estimation of the carboxyhemoglobin by matching the unknown reaction with a standard chart proves the definite reliability of this procedure.

PUBLICATION REVIEW

This report has been reviewed and is approved.

FOR THE COMMANDER:



JACK BOLLERUD
Colonel, USAF (MC)
Chief, Aero Medical Laboratory
Directorate of Research

TABLE OF CONTENTS

	Page
INTRODUCTION	1
MATERIALS AND METHODS	2
RESULTS	6
DISCUSSION	7
REFERENCES	8
APPENDIX - <u>Statistical Evaluation of the Palladium Precipitation Test</u> <u>of the Percent Carbon Monoxide Saturation in Blood.</u>	
Richard L. Deininger, 2/Lt, USAF (MSC), James E. Smithson	9

LIST OF ILLUSTRATIONS

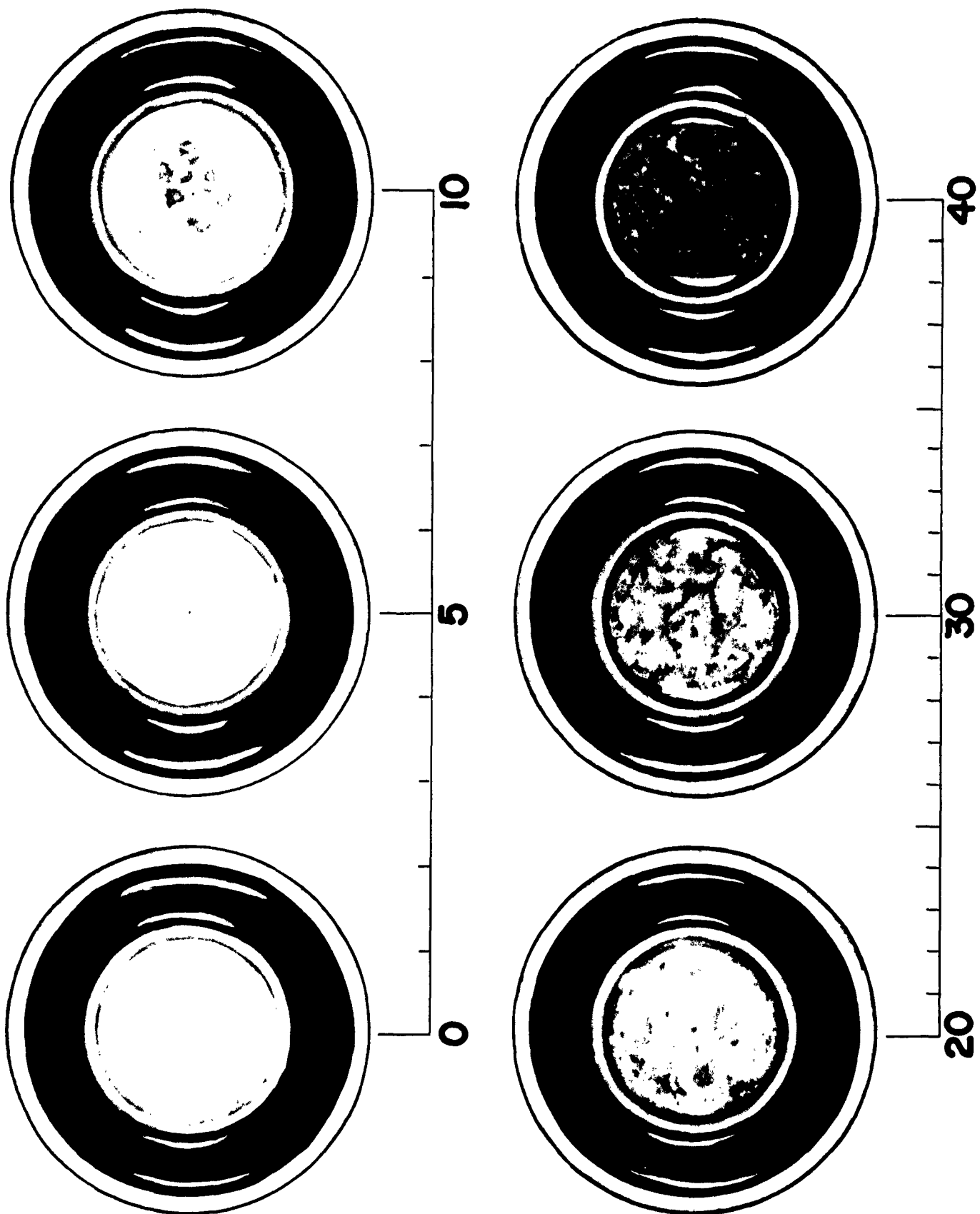
Figure	Page
Frontispiece Standard Color Chart - Percentages Carboxyhemoglobin Microdiffusion Technique	vi
1 Conway Type Diffusion Unit, Micro	1

LIST OF TABLES

Table	Page
I Comparative Ratings by Phases	5
II Statistical Characteristics of the Three Phases	6

APPENDIX TABLES

A A Summary of the Analysis of Variance Using the Ratings Obtained in Experiment 1	14
B A Summary of the Analysis of Variance Using the Ratings Obtained in Experiment 2	16
C A Summary of the Analysis of Variance Using the Ratings Obtained in Experiment 3	18



Frontispiece. Standard Color Chart — Percentages Carboxyhemoglobin
Microdiffusion Technique

INTRODUCTION

An Air Force requirement exists for a simple, available, and immediately adaptable technique for the screening of blood samples relative to concentrations of carbon monoxide. Questionable results are sometimes produced by untrained personnel using the Van Slyke manometric or volumetric techniques (1,2,3). Chinn reviewed other weaknesses of techniques in his recent article on a carbon monoxide determination (4).

The method presented in this paper is, in part, the procedure of Gray and Sandiford employing the diffusion technique in a Conway microdiffusion unit (5,6). A small blood sample placed in the outer chamber of the unit is the source of the carbon monoxide released by a weak acid. The gas reacts with a palladium chloride solution in the inner chamber (fig. 1). Metallic palladium forms, and bears a quantitative relationship to the amount of carbon monoxide reacting with the palladium chloride solution. Gray and Sandiford determined the carbon monoxide in the blood sample by a spectrophotometric method applied to the residual palladium chloride solution. This method, apparently, has not received wide attention in the literature although the actual diffusion reaction is very simple. Accordingly, this paper will show this diffusion reaction as a very practical means for the estimation of carboxyhemoglobin, since the palladium mirror bears a visible relationship to the amount of carbon monoxide in the blood sample. The percentage carboxyhemoglobin can then be estimated by matching the visible density of the unknown palladium mirror with one of the densities on a standard chart (frontispiece). A statistical evaluation demonstrated the practical utility of this approach. Apparently, the original spectrophotometric method has not been accepted by clinical laboratories, but the currently described modification using the inexpensive, commercially available palladium chloride and unit, might attract clinical laboratory personnel.

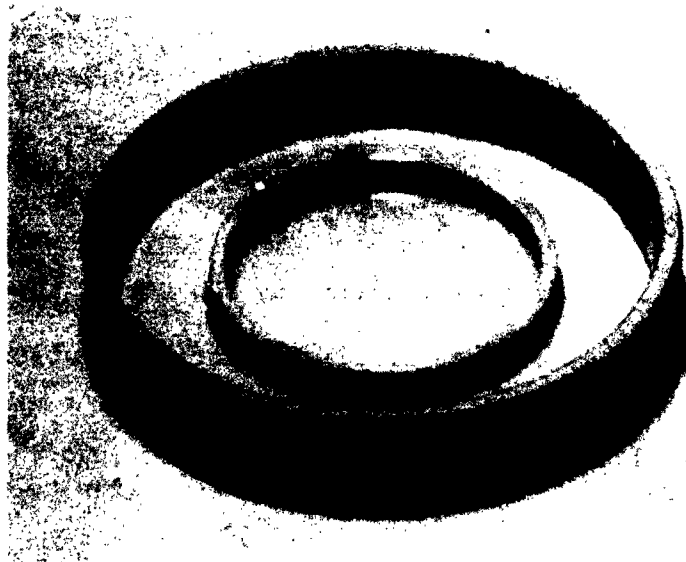


Figure 1. Conway Type Diffusion Unit, Micro

MATERIALS AND METHODS

Apparatus Required

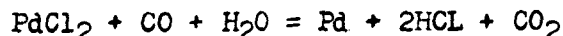
1. Diffusion unit, micro, Conway type, Coors USA porcelain, 68 mm diameter x 15 mm high, 3 oz in weight.*
2. Cover glass for unit.*
3. Pipette, 1.00 ml, blood.
4. Chart, standard showing palladium deposits relative to carboxyhemoglobin percentages.

Reagents Required

1. Palladium chloride, 1 gram dry, purified.**
Preparation of reagent: 0.5 gram of the palladium chloride is dissolved by heating 25 ml of 0.1 N hydrochloric acid. After cooling, the solution is made up to 500 ml with distilled water (6). There is no evidence of decomposition at room temperature of this palladium chloride acid solution. Accordingly, a 1 gram bottle could permit over 300 determinations.
2. Phosphoric acid, 4 normal.
3. Distilled water.
4. Vaseline or stopcock grease.

Methods of Analysis

1. The diffusion reaction:



- a. Place a thin ring of vaseline on the rim of the unit.
- b. Transfer 3 ml of the palladium chloride solution to the inner chamber.
- c. Measure 1.00 ml of anticoagulated blood into the outer chamber. Add about 2 ml of distilled water to the blood and mix gently by rocking the unit.
- d. Carefully place the glass lid on the unit leaving a space open between the plate and the unit.

* This unit with glass cover can be purchased from Arthur H. Thomas Co., Philadelphia, Pennsylvania, Catalog 1950 ed. lists the unit as \$1.97 ea.

** This reagent can be purchased from Fischer Scientific Co., New York, N. Y., at \$1.35 per gram.

- e. Quickly introduce about 0.3 ml of the 4 N phosphoric acid into the hemolyzed blood in the outer chamber and cover the entire unit.
- f. Gently rock the entire covered assembly so that the acid mixes with the blood. A darkening of the blood occurs when this condition is established.
- g. Permit the unit to stand for a period not less than one hour, rocking carefully at selected intervals.
- h. CAUTION: Do not allow any mixture of contents of the outer and inner chambers. If this occurs — start over.

During the hour, the surface of the yellow palladium chloride solution darkens, and a silvery deposit forms. It will vary in density depending on the volume of carbon monoxide present. As the metallic reaction increases, the yellow color of the background palladium chloride decreases. Of course, absence of carbon monoxide results in the yellow solution remaining unchanged.

1. After one hour, remove the glass cover and compare the grey to silvery deposit of metallic palladium on the surface of the residual yellow palladium chloride solution with the standard chart in color (frontispiece). Read the percent carboxyhemoglobin from the scale consisting of 0, 5, 10, 20, 30 and 40% carboxyhemoglobin.
The standard chart in color is an artist's conception of a series of diffusion units containing deposits of metallic palladium related to percentages of carboxyhemoglobin.

2. A statistical evaluation of the diffusion reaction for the estimation of the percentage carboxyhemoglobin:

While the validity of the chemical reaction has been established by others (5,6,7,8), the question of reliability of the proposed test had to be answered since human judgments are subject to error. It was also necessary to establish the fact that technicians, relatively untrained, could reliably perform the reaction and make the judgment of percentage carboxyhemoglobin as well.

The statistical evaluation consisted of three phases, namely:

Phase I - All diffusion reactions were done by the author. Forty military and civilian employees of the Aero Medical Laboratory, possessing widely varied backgrounds, estimated percentages carboxyhemoglobin from 1 through 40% using the standard chart. Each judge rated four samples, and each sample was rated by four judges. Although 160 judgments were to be obtained in this phase, 153 were actually made. The seven ratings that were lost, however, in no way influenced the experiment.

In this series of blood samples, all percentages were established from an analyzed initial volume of blood containing 80% carboxyhemoglobin. This was analyzed

by the Van Slyke volumetric technique (3). For this experimental study samples of blood consisting of the appropriate relative fraction of a milliliter were used for each percentage. For example, a 40% sample consisted of 0.5 ml of the 80% stock blood, and this was introduced into the dish.

Phase II - Again, all diffusion reactions were done by the author, but in this case, the tests were conducted at the clinical and pathological laboratories of the Wright-Patterson Air Force Base Hospital. The judges in this phase consisted of a group of 10 clinical and pathological laboratory personnel. In addition, the blood samples were not prepared blood with carbon monoxide. Instead, the samples consisted of clinically obtained blood ranging from 0 to 35%. The 10 judges rated each of the following samples of blood:

1. A normal nonsmoker's blood (0%);
2. A normal heavy smoker's blood (8%);
3. An autopsy blood sample from a fatal case of carbon monoxide poisoning (sample diluted 50%, final concentration 35%);
4. A sample of the autopsy blood diluted to 18% (representing a level possibly encountered in motor vehicle garages).

The second phase should give information concerning the ability of technicians to estimate percentages carboxyhemoglobin in actual clinical samples.

Van Slyke analyses of each sample were performed to establish the concentration. One ml of blood was used in each test.

Phase III - The third and final experiment investigated the ability of clinical laboratory personnel to perform the reactions and then to estimate the percentages carboxyhemoglobin in four samples of blood. In this statistical evaluation, two 4 x 4 Latin Square designs were employed with eight persons serving as the subjects (9,10,11).

The four saturations planned included 1.6, 3.0, 10.5 and 24.8%. Each sample was placed in a test tube and identified by a code number; one digit indicating the subject and another digit indicating whether the sample was to be used first, second, third, or fourth. A total of 32 analyses were done. Each saturation was reacted and judged once by each judge but at different periods.

The phases will provide the following information:

1. The standard deviation of the variable errors in rating;
2. The reliability of the ratings;
3. The number of usable categories in the ratings.

These are statistical indices describing the adequacy of the test (10,11). Because these indices were used, the data in each phase had to be gathered systematically according to a particular experimental design. The three experimental designs referred to in this report are frequently used in agricultural, biological, and psychological research (9).

TABLE I
COMPARATIVE RATINGS BY PHASES

PERCENT SATURATION	PHASE I*	PHASE II**	PHASE III***
	Average rating	Average rating	Average rating
None Found		1.4	1.38
1	1.50		
2	2.50		
3	2.75		4.00
4	2.75		
5	8.25		
6	4.75		
7	7.50		
8	6.25	6.6	
9	9.75		
10	13.25		
10.5	-----		12.8
11	9.75		
12	13.50		
13	15.75		
14	13.00		
15	12.99		
16	18.65		
17	20.13		
18	17.00	16.8	
19	17.50		
20	-----		
21	22.00		
22	18.00		
23	22.75		
24	22.00		
24.8	-----	----	25.1
25	25.75		
26	23.50		
27	21.75		
28	29.75		
29	27.00		
30	26.00		
31	25.50		
32	28.75		
33	30.00		
34	30.50		
35	29.50	33.0	
36	33.00		
37	35.00		
38	35.50		
39	37.25		
40	36.75		

* Average of 3 or 4 ratings
 ** Average of 10 ratings
 *** Average of 8 ratings

RESULTS

Agreement of average ratings with true percent carbon monoxide blood saturations is reflected in table I (p. 5). Generally, overrating was observed in the lower concentrations whereas underrating is found for the higher concentrations.

Table II considers data more descriptive of the adequacy of the test and ratings. Attention is directed to the standard deviations listed in the first row of table II because these are measures of the variable errors in the ratings. Such evaluations are frequently discussed in the precision techniques in analytical chemistry. Essentially, the greater the standard deviation, the greater the errors in the ratings. Thus, the errors in Phase I are slightly greater than those in Phase II, and the variable errors in Phase II are about twice those in Phase III.

TABLE II
STATISTICAL CHARACTERISTICS OF THE THREE PHASES

	PHASE I	PHASE II	PHASE III
Standard Deviation in Percent Saturation*	3.18	3.16	1.80
Reliability Coefficient	0.92	0.95	0.97
Number of Usable Categories	3.46	4.51	6.03

* The standard deviation is interpreted as follows: 68% of the ratings will fall in the interval defined by the true percent saturation plus and minus the standard deviation. Also, 95% of the ratings will fall in the interval defined by the true percent saturation plus or minus 1.96 times the standard deviation.

The reliability coefficient measures the degree to which the ratings are influenced by the true carbon monoxide saturation rather than by errors. This coefficient ranges from zero to unity, 1.00. In practice, values from 0.9 to 1.00 are considered very good. Results of the three Phases, therefore, point with emphasis to the proposed diffusion test for carbon monoxide in blood as highly acceptable.

Usable categories convey the extent of information gained from the ratings. That is, it becomes a measure of the distinctions a judge can make when estimating concentrations of carboxyhemoglobin. Ability to consistently distinguish between the nonsmoker, smoker and sublethal concentrations, indicates usable categories.

Usable categories also increase in number from Phase I through Phase III. This means that finer distinctions are made in Phases II and III than in Phase I. While a mathematical relationship ties the three statistical indices, each one presents a somewhat different picture of the results.

DISCUSSION

The diffusion reaction for the quantitative determination of carbon monoxide in blood by visual comparison has been justified through statistical experimentation and analysis. This study revealed all judgments were little influenced by the type of judge or the type of blood sample. Equal accuracy was noted when the judges were at least average in intelligence and acquainted somewhat with laboratory procedures. The ratings are considered very reliable since the coefficients of reliability are near the 1.00 maximum.

Establishment of usable categories clearly revealed the ability of judges generally to distinguish at least between the bloods of:

1. Nonsmoker;
2. Heavy smoker;
3. Sublethal carbon monoxide poisoned individual;
4. Fatally poisoned individual.

This characteristic is of definite value for Air Force clinical cases.

Experimental data revealed little training was needed in the performance of the proposed technique. This feature and the sole accurate measurement of 1.00 ml (blood) proved a definite attraction to the Wright-Patterson Air Force Base Hospital clinical laboratory personnel. Economy of investment in the commercially available palladium chloride and diffusion unit are advantages to consider.

Attention to the percentages on the standard chart will show a maximum reading of 40. This, of course, does not infer that the highest level to be measured with this chart is 40%. In the event the reaction is equal or perhaps heavier in palladium density than the 40% reaction, rerunning the blood sample with 0.5 ml or any accurate fraction of a ml will result in the final percent when the reading of the fraction ml is properly multiplied. For example, an autopsy sample of blood was used in one of the experiments. However, the final concentration of carboxyhemoglobin was 70%. Rerunning the sample with 0.5 ml gave a reaction that was declared 35%. This figure multiplied by 2 resulted in the 70% concentration.

Identification of carboxyhemoglobin qualitatively is important regardless of the quantitative method used subsequently. The diffusion reaction is specific from a practical point of view. While hydrogen sulfide will reduce palladium chloride to the metal, this is deemed very improbable when investigating the carbon monoxide level in blood. The proposed method, therefore, could be used wherever screening of blood samples is desired. Formation of the palladium metal is

in effect, qualitative evidence of the carbon monoxide. For example, when the reaction is interpreted as that of a nonsmoker or heavy smoker, certainly no further examination of the blood for carbon monoxide is necessary.

If a Van Slyke apparatus is available, and the analyst performs this technique, simultaneous operation of the diffusion reaction would verify results or detect error. This type of check analysis might find use in medicolegal considerations. However, the statistical evidence presented for the diffusion reaction and the estimation of percentage carboxyhemoglobin reveals that this technique stands alone on its own merit. This infers that dependence on the Van Slyke should be unnecessary for most clinical experiences.

REFERENCES

1. Van Slyke, D. D., and Neill, J. M., The Determination of Gases in Blood and Other Solutions by Vacuum Extraction and Manometric Measurement. J. Biol. Chem., Vol. 61, pp. 523-73, 1924.
2. Peters, J. P., and Van Slyke, D. D., Quantitative Clinical Chemistry Methods. Baltimore, Williams & Wilkins Company, p. 261, 1943.
3. Methods for Medical Laboratory Technician. TMS-227, p. 279, 1951.
4. Chinn, H., Pawel, N., and Redmond, R., A Simple Micro Method for Blood Carbon Monoxide Determination, J. Clin. & Lab. Med., Vol. 46, pp. 905-9, 1955.
5. Gray, C. H., and Sandiford, M., A Microdiffusion Method for the Estimation of Carbon Monoxide in Blood. The Analyst, Vol. 71, p. 107, 1946.
6. Conway, E. J., Microdiffusion Analysis and Volumetric Error. London, Crosby Lockwood and Son Ltd., pp. 230-3, 1947.
7. Gettler, A. O., and Freimuth, H. C., Carbon Monoxide in Blood: A Simple and Rapid Estimation. Am. J. Clin. Path., Vol. 13, pp. 79-82, 1943.
8. Fister, H. J., Manual of Standardized Procedures for Spectrophotometric Chemistry, New York, Standard Scientific Supply Corp., Method C-20.1, 1950.
9. Cochran, W. C., and Cox, C. M., Experimental Designs, New York, John Wiley and Sons, Inc., p. 103, 1950.
10. Ebel, R. L., Estimation of the Reliability of Ratings. Psychometrika, Vol. 16, pp. 407-24, 1951.
11. Fisher, R. A., Statistical Methods for Research Workers, (10th Ed.) Edinburgh, Oliver and Boyd, 1946.

APPENDIX

STATISTICAL EVALUATION OF THE PALLADIUM PRECIPITATION TEST OF THE PERCENT CARBON MONOXIDE SATURATION IN BLOOD

Richard L. Deininger, 2/Lt, USAF (MSC)
James E. Smithson

The palladium precipitation test appears well-suited to the problems posed by potential carbon monoxide poisoning. On the other hand, the fact that human judgments are involved raises the question about the reliability of the test. Human judgments are subject to error, and the question is whether these errors are so large as to make the test practically useless. Second, there is the question whether the relatively untrained technicians who would use the test in the field could reliably perform the reaction and make the judgment although both are apparently simple.

The reliability of the test is intimately connected with its usefulness. If the test is not reliable, then the technician cannot consistently separate the normal blood from the poisoned blood. However, if the test is reliable, it might be possible to distinguish a normal nonsmoker from a normal smoker and both of these from the slightly poisoned individual. Questions concerning the reliability of the test must be answered empirically through the scientific application of experimental and statistical techniques.

PROBLEM

Such scientific techniques were applied to the palladium precipitation test and the results of the statistical evaluation and the experiments by which it was accomplished are discussed. Four questions are answered. First, how reliable are the judgments alone? Second, how many usable categories are there in the judgments? Third, what sort of differences exist between judgments? Fourth, is the judgment per se or the combination of the performance of the reaction and the judgment more subject to error and unreliability?

THE METHOD OF ANALYSIS

Perhaps the most important characteristic of ratings is that they are distributed over a range of possible values. Many factors are responsible for the variation typical of a set of judgments. Undoubtedly the true saturations of the blood in the reacted samples is responsible for part of the variation. Other variation might be attributable to the judges themselves, to the state of the blood, or to the precision of the analysis reaction.

In determining the reliability, the total variation of the ratings or judgments is separated into two or more parts. First, the variation attributable to the true saturations is always separated. Second, there is always some variation that is attributable to error. In addition, the variation due to one or more additional factors (such as the judges) may be isolated. The important point is as follows: when one of these additional components is systematically analyzed from the ratings, it is no longer part of the error. Conversely, if one of these factors is randomized throughout the ratings, then it is included in the error term.

The procedures for systematically collecting the ratings are called experimental designs (1), and have been developed to a high degree both in theory and practice. The methods for analyzing these ratings and for estimating the variability attributable to the various factors come under the topic of the analysis of variance (4), and the reliability of the ratings may be estimated by using an elementary form of the analysis of variance. The procedure for estimating the reliability has been discussed by Ebel (3), and will be abstracted here.

In the simplest case, the total variance of the ratings is separated into one component due to the true saturations, and another due to errors. That is:

$$\text{Var}(r) = \text{Var}(t) + \text{Var}(e) \quad (1)$$

where $\text{Var}(r)$ is the total variance, $\text{Var}(t)$ that due to the saturations, and $\text{Var}(e)$ that attributable to errors. If the variance due to differences between judges, $\text{Var}(j)$, is systematically removed, then we write:

$$\text{Var}(r) = \text{Var}(t) + \text{Var}(j) + \text{Var}(e') \quad (2)$$

where $\text{Var}(r)$ and $\text{Var}(t)$ are the same as before, but where $\text{Var}(e')$ is less than $\text{Var}(e)$ if there are any true differences between the judges.

The ratings are reliable to the extent that the variance attributable to the true saturations exceeds that due to random errors (3). The reliability coefficient, r , is the index commonly used to express this relation and is defined as:

$$r = \frac{\text{Var}(t)}{\text{Var}(t) + \text{Var}(e)} \quad (3)$$

If additional components of the total variance are isolated, then $\text{Var}(e')$ is used in formula 3 rather than $\text{Var}(e)$. When the variation of the ratings depends only on the true saturations, then:

$$\text{Var}(e) = 0$$

and

$$r = \text{Var}(t)/\text{Var}(t) = 1.00$$

However, if the variance of the ratings is attributable to errors alone, then:

$$\text{Var}(t) = 0$$

and

$$r = 0/\text{Var}(e) = 0.00$$

Thus, the reliability coefficient behaves as an index of efficiency where the error variance is viewed as lost or wasted.

The next question is whether it is possible to distinguish between 40 different carbon monoxide saturations (varying only slightly), or whether only two or so, broad saturations can be determined. To estimate the number of usable categories in the ratings, the mathematical theory of communication (5) is applied. According to this theory, the usable information is called the information transmitted (I_t), and is equal to the logarithm to the base two of the number of usable categories, N :

$$I_t = \log_2 N \quad (4)$$

The rationale for selecting this particular function is explained in the original report of the theory (5, p. 3).

Actually, the judgments under discussion need not fall into precise categories, for the ratings could cover an infinitely large number of values. In a case such as this, the information transmitted can still be calculated. However, the number of usable categories becomes an abstraction so that the discussion is in terms of a number that is functionally equivalent to N usable, discrete categories.

The ratings have a given variance, $\text{Var}(r)$, and from this variance, a maximum potential amount of information transmitted, $H(r)$, is calculated as follows:

$$H(r) = 1/2 \log_2 (2\pi e \text{Var}(r)) \quad (5)$$

Because of random errors, all this potential information is not transmitted; and the actual information transmitted is less than the maximum by the equivocation or ambiguity due to the errors. The variance attributable to the errors is known and equals $\text{Var}(e)$. For this given variance the maximum information lost due to errors will be:

$$H_t(r) = 1/2 \log_2 (2\pi e \text{Var}(e)) \quad (6)$$

Thus, a conservative estimate of the maximum information transmitted would be the difference between these two values, or:

$$I_t = H(r) - H_t(r) \quad (7)$$

$$I_t = \log_2 \left[1 + \frac{\text{Var}(t)}{\text{Var}(e)} \right]^{1/2} \quad (8)$$

By virtue of equation 4, the equivalent number of usable categories may be introduced into equation 8 in place of I_t . Taking the antilogarithm of both sides of the equation produced by this substitution, the number of usable categories is equal to:

$$N = \left[1 + \frac{\text{Var}(t)}{\text{Var}(e)} \right]^{1/2} \quad (9)$$

The number of usable categories and the reliability coefficient are related to one another, as would be expected. The exact relation is:

$$N = 1/(1 - r)^{1/2}$$

The final step is to estimate the variance found in equations 3 and 9. These estimates are obtained by subjecting a set of ratings to an analysis of variance, that in turn produces a series of mean squares (one for each variance component) from which the variances are estimated. The reliability coefficient or the number of usable categories could be calculated by estimating the variances from the mean squares and then substituting these estimates into the proper formulas. However, it is simpler to express formulas 3 and 9 in terms of the proper mean square values.

An analysis of variance produces one mean square for the errors, M_e or $M_{e'}$, and at least a second mean square for the true saturations, M_t . Either of the error mean squares is a direct estimate of the error variance; however, the saturations mean square contains the variance due to the saturations plus a certain amount of the error variance. Thus, formulas 3 and 9 appear more complicated when expressed in terms of the mean squares:

$$r = \frac{M_t - M_e}{M_t + (k - 1)M_e} \quad (10)$$

$$N = \left[1 + \frac{M_t + M_e}{kM_e} \right]^{1/2} \quad (11)$$

In the above formulas, k is the number of ratings obtained for each reacted sample. It is not necessary in every case that k be the same for each of the n reacted samples rated. In the event that k can and does vary from reacted sample to reacted sample, k_o is used:

$$k_o = \frac{1}{(n - 1)} \left[\sum k + \frac{\sum k^2}{\sum k} \right] \quad (12)$$

Formulas 10 and 11 may be used with the more complicated analyses of variance by substituting $M_{e'}$ for M_e .

EXPERIMENT 1

(Phase I)

The palladium precipitation test can be separated into two phases: (1) the performance of the reaction using a particular sample of blood; and (2) the judgment of the resulting reacted sample. The first study dealt with the second step, judgments of the reacted samples. If the judgments proved reliable enough, a subsequent study would have the technician do both steps in order to evaluate the entire production.

This initial experiment involved the 40 integral percents saturation from 1 to 40%. Forty judges with widely varied backgrounds made the judgments in this experiment. The experimental design had each judge rate four samples and each sample rated by four judges. However, no judge rated all 40 samples, but rather a single set of four. No judge rated a set that no other judge saw, and some judges never rated samples that some other judges saw.

Procedure

The 40 saturations were randomly arranged into a sequence of 40 numbers that in turn was divided into 10 parts, each containing four saturations. Then, in order of appearance, four judges rated the four samples in one and only one of these sets. When a set had been rated by four judges, it was discarded and a new set of four samples was reacted and used.

The experimenter carried out all the reactions. Highly saturated transfusion blood was used in differing volumes in order to produce the 40 different apparent saturations. The actual saturation of the transfusion blood was determined by the Van Slyke volumetric method, but with the exception of this test no further Van Slyke determinations were made.

The 40 judges were military and civilian employees of the Aero Medical Laboratory, Wright Air Development Center, whose occupations covered a wide range (clerical, administrative, medical, psychological, and so forth).

The experimenter worked with one judge at a time, and the judge rated one sample at a time by comparing it with the chart of standards. This chart consisted of six artist's representations of similar reacted samples placed at the 0, 5, 10, 20, 30 and 40% marks along a scale ranging from 0 to 40.

In general, the design and analysis of the experiment were such that random errors in the dilution of the blood and the performance of the reaction would lead to low reliability in the ratings. Also, the analysis did not require the same number of ratings of each sample, nor that ratings be obtained for all 40 samples. Thus, the seven ratings lost during the experiment did not damage the results. The four ratings of the 20% saturation were lost, as were one rating from each of three other percents saturation (15, 16 and 17%).

Results

The simple analysis of variance produced two components (see Table A). The mean square attributable to the saturations was 44 times as large as that due to errors. From this it is concluded that the saturations have a larger effect on the ratings than errors.

Because the mean square due to saturations so overshadowed that due to errors, the reliability was very high. The value of the coefficient was 0.92, which is exceptionally good and very near the maximum of 1.00. To calculate the value, the two mean squares contained in Table A and the average number of ratings for each sample calculated from formula 12 were substituted into equation 10.

The informational analysis reveals that there were nearly four usable categories in the ratings. The exact value, 3.46, was estimated by substituting the two mean squares in Table A and the average number of ratings per sample into equation 11.

In addition to being reliable, the average rating of each sample appeared fairly accurate. The average scattered about the 45° line indicating perfect accuracy, except at the higher saturations where there was a definite tendency to underestimate the saturations and at the lower saturations where some overestimation occurred. The ratings appeared more variable as the true percent saturation and the average rating

TABLE A

A SUMMARY OF THE ANALYSIS OF VARIANCE
USING THE RATINGS OBTAINED IN EXPERIMENT 1

SOURCE	SUM OF SQUARES	d.f.*	MEAN SQUARE	F RATIO
Percent CO Saturation	16 894	38	444.6	44
Error	1 152	114	10.1	
Total	19 046	152		

* Note: This is an abbreviation for degrees of freedom. The mean square is the quotient of the sum of squares divided by the degrees of freedom.

increased. A finer analysis suggested either (1) the variance increases linearly with the mean rating, or (2) the square root of the variance increases linearly with the mean rating. However, this tendency was not consistent. Statistically, we cannot reject the hypothesis that the variances corresponding to each of the 39 saturations are from the same population. It is important to answer this question, since the power of the analytic techniques applied to these ratings is reduced by true differences between the variances.

Conclusions

The judgments are highly reliable even though differences between judges should lower the error term and increase the reliability. Similarly, removing such differences should increase the number of usable categories in the ratings above the three or four now apparent. The possible relation between the average rating and the variance of the rating warrants further study.

EXPERIMENT 2

(Phase II)

The second experiment concerned judgments made under more nearly operational conditions. Although the initial study was very helpful, it gave no indication how the personnel of a hospital clinical laboratory performed as judges. Second, it did not tell what differences existed between judges. And third, it provided no data concerning the effect of clinically obtained blood samples.

The second study supplemented the first and had the following characteristics: (1) the judges were personnel assigned to a hospital clinical laboratory, (2) each judge rated each reacted sample, and (3) blood samples from a nonsmoker, a normal smoker, and an autopsy (cause of death - carbon monoxide poisoning) were used.

Procedure

The experimental design had 10 judges rate each of four reacted samples covering the range from 0 to 35% saturation.

The three sources of blood provided four different saturations. The Van Slyke volumetric determination of the nonsmoker's blood failed to show any carbon monoxide present. The Van Slyke determination of the normal smoker's blood gave a reading of 8% saturation, while the same analysis of the autopsy blood showed it to be 75% saturated. One sample of the autopsy blood was diluted 50% so that the effective saturation was 35%, and a second sample was diluted 75% resulting in an effective saturation of 18%. Thus, the samples judged had 0, 8, 18 and 35% carbon monoxide saturation.

As in the first study, the experimenter carried out all the reactions. He did the volumetric Van Slykes, diluted the blood where necessary, and performed the palladium precipitation test. Only one sample of each saturation was reacted, and all judges rated the same four reacted samples.

Ten of the officer, enlisted and civilian personnel at the hospital clinical laboratory, Wright-Patterson Air Force Base, volunteered to make judgments. Some specialized in clinical laboratory work; however, others specialized in clinical chemistry, hematology, and the like.

The judgments were made in the same manner as in the first study. Forty ratings, plus four ratings by an 11th, color blind judge, were gathered.

Results

The first analysis concerned the possible relation between the average rating and the variability of the ratings. For this analysis, the variance of the 10 ratings for each reacted sample was computed. The ratings for the 35% saturation sample were much more variable than those for the other three samples. The statistical test, a Bartlett's test, indicated such results would happen less than one time in 100 by chance alone. This finding supports the hypothesis that the variability of the ratings increases with their average. However, the common transformations suggested by the first study failed to remedy the situation, and this implies that some other factor may be responsible. Inspection of the ratings for the 35% sample reveal that most of the variability is due to two judges who rated the sample as 20 and 25%.

The variance of the untransformed ratings was analyzed and the results are summarized in Table B. Three important results appear: (1) the error mean square is comparable to that from the first study, (2) no gross differences between judges appear, and (3) the saturations mean square is nearly 195 times the error mean square. Apparently, the laboratory people are as accurate as the non-laboratory personnel used in the first study. Also, one laboratory judge is about as accurate as the next in spite of the two low ratings made for the 35% sample. In view of the relative magnitude of the saturations mean square, it is not surprising to find the reliability very high (0.95). Finally, there are four usable categories in the ratings — slightly more than found in the initial study (4.51 compared with 3.46). The ratings by the color blind judge are indistinguishable from those of the normal judges.

TABLE B

A SUMMARY OF THE ANALYSIS OF VARIANCE
USING THE RATINGS OBTAINED IN EXPERIMENT 2

SOURCE	SUM OF SQUARES	d.f.*	MEAN SQUARE	F RATIO
Percent CO Saturation	5 815.5	3	1 938.5	194.6
Differences Between Judges	127.4	9	14.2	1.42
Error	269.0	27	9.96	
Total	6 211.9	39		

* Note: This is an abbreviation for degrees of freedom. The mean square is the quotient of the sum of squares divided by the degrees of freedom.

Conclusions

The second experiment offers no conclusive evidence that the ratings are more variable with samples of higher percents saturation. The lack of homogeneity of the variances seems due to one or so values rather than to a consistent trend. On the other hand, there is no definite evidence that such a relationship does not exist.

In general, the judgments per se seem little influenced by the type of judge, the individual judge, or the type of blood sample used. Also, the fact that there are four usable categories in the ratings is evidenced in the ratings themselves. Not one judge confused the smoker's blood with the nonsmoker's, nor any one sample with any other sample.

EXPERIMENT 3 (Phase III)

In view of the fine results of the first two experiments, a third was undertaken which investigated the ability of clinical laboratory personnel to perform the reactions and to make the judgments. This study tested whether laboratory personnel differed from one another when performing both phases of the test, and whether these people became more skillful during the course of performing four reactions and judgments. These features required a more complicated experimental design, but seemed important enough to justify the effort.

In addition, the third experiment studied the possible relation between the average rating and the variability of the ratings. The four saturations used in the experiment were selected on the hypothesis that the variance and the mean were linearly related (that is, the saturations were equally spaced along a logarithmic scale of saturations). Although not a full test of the hypothesis that the ratings were more variable with the higher saturations, the selection of the four saturations did permit a prediction concerning the reliability of the ratings. The set of saturations selected should yield a reliability of 0.92 to 0.95 if (a) the errors introduced by the performance of the reaction are small in comparison with those due to the judgments, and (b) the variances of the ratings increase proportionally with the average rating. Condition (a) was assumed during the previous studies, and condition (b) was still in doubt.

Procedure

Two different four-by-four Latin Square designs were used in the third study. The Latin Square design is used widely in agricultural and biological research where practical considerations restrict the experimenter (1, p. 103). The total variance is broken into four parts: one attributable to the random errors, another to the "treatments," a third to the "rows," and the last to the "columns." The design requires that each treatment appear once and only once in each row and each column, so that there are as many rows as columns as treatments.

In the present study, the rows of the Latin Squares represented the different laboratory personnel serving as subjects. There were four rows in each square, hence a total of eight persons served in the experiment. The columns represented the order in which the samples were reacted and judged. There were four columns and each saturation was reacted and judged first by one judge in each square, second by another judge in each square, and so forth. Four different percents saturation were the treatments. Since each saturation was reacted and judged once by each judge in each square, there were eight different ratings of each. The particular squares used were selected randomly from the many ones possible (1).

The blood samples were produced by diluting saturated transfusion blood either with water or less saturated transfusion blood. The actual percents saturation were not made known to the experimenter. The four saturations were to be 1.6, 4.0, 10.1 and 25.3%; however, errors undoubtedly occurred in the dilution process, since the Van Slyke volumetric determinations subsequently done by the experimenter found the following saturations: negligible, 3.0, 10.5 and 24.8%. One sample of each saturation was given to the subject in a clean, stoppered test tube identified only by a small label containing a code number.

The subjects in this study were at the same clinical laboratory that assisted in the second experiment and most of the people in the third also served in the second. The rows of the Latin Squares were numbered from 1 to 8 and arranged in a random permutation. Subjects were assigned a member of this permutation in order of appearance. In addition to having made judgments previously, most of the subjects had seen the experimenter perform a sample reaction at the time of the second experiment. Some of the subjects watched the experimenter perform a second reaction just before they themselves started on their first reaction. Finally, all subjects were given a typed series of instructions that outlined the steps in the experiment.

Immediately after demonstrating the reaction, the experimenter turned the samples, instruction sheets, and report slips over to the officer in charge of the laboratory, who then supervised the study without further aid from the experimenter. Each subject reacted and judged one sample at a time. Upon reporting his or her results to the officer in charge, the next sample was given out. The judgments were the same as in the previous study.

Results

The third study indicates that there is no relation between the variability of the ratings and the saturation of the sample. The trend in the first two studies appears slightly reversed, and the statistical test comparing the variance of the ratings for each saturation is not significant.

Next, the variance of the 32 ratings obtained was analyzed by a complicated, but standard procedure that combined the results of the two Latin Squares and provided more stable estimates of the several variances. Table C summarizes the results: (1) the subjects did not change in skill as they proceeded from one sample to the next (the order mean square is less than the error mean square); (2) no true differences between judges appear (the subjects mean square is barely twice the error mean square); and (3) the errors appear smaller in variability than those found in the first two studies.

TABLE C

A SUMMARY OF THE ANALYSIS OF VARIANCE
USING THE RATINGS OBTAINED IN EXPERIMENT 3

SOURCE	SUM OF SQUARES	d.f.*	MEAN SQUARE	F RATIO
<u>Order of Performance</u>	<u>1.4</u>	<u>3</u>	<u>.46</u>	-----
<u>Subjects (total)</u>	<u>44.5</u>	<u>7</u>	<u>6.36</u>	1.96
Subjects	43.4	6		
Squares	1.1	1		
<u>Percents CO Saturation</u>	<u>2 752.6</u>	<u>3</u>	<u>917.53</u>	282.3
<u>Error (total)</u>	<u>58.5</u>	<u>18</u>	<u>3.25</u>	
Deviations	39.5	12		
Treat x Sq.	17.6	3		
Col. x Sq.	1.4	3		
<u>Total</u>	<u>2 856.9</u>	<u>31</u>		

* This is an abbreviation for degrees of freedom. The mean square is the quotient of the sum of squares divided by the degrees of freedom.

The reliability coefficient was somewhat higher than anticipated, for it turned out to be 0.97. There were six usable categories (6.03 to be exact) in the ratings. This is an interesting finding, for the saturations in the third study ranged from 1 to 25% while those in the other studies ranged from 1 to 40%. As before, the average rating agreed well with the Van Slyke determination, for the averages were 1.4, 4.0, 13.4 and 25.1%.

Conclusions

The greater reliability in the final study appears due to a decrease in errors at all saturations rather than to the selection of particular saturations. In other words, the hypothesis that the means and variances are related is not acceptable even though the reliability coefficient is very high.

The higher reliability in this last study could be explained several ways. It is possible that the judgments are more accurate when the personnel do the reactions themselves. However, it is more probable that the particular judges selected in this study were more skillful at the beginning of the last study than at the start of the second because of the experience they received in the second experiment. Nevertheless, one fact is apparent: the amount of training obtained in this way certainly was not very extensive.

EVALUATION

The three experiments show that the judgments are exceptionally reliable and are more influenced by the saturations than by errors. Second, there are enough usable categories in the test to warrant its use as a screening device. Third, the judges do not differ appreciably from one another: the typical clinical laboratory personnel are about as accurate as one another in performing the reactions and in making the judgments.

APPENDIX
BIBLIOGRAPHY

1. Cochran, W. G. and Cox, G.M., Experimental Designs, New York: John Wiley and Sons, Inc, 1950.
2. Conway, E. J., Microdiffusion Analysis and Volumetric Error, London: Crosby Lockwood and Son, Ltd, 1947
3. Ebel, R. L. Estimation of the Reliability of Ratings, Psychometrika, 1951, Vol. 16, pp. 407-24.
4. Fisher, R. A., Statistical Methods for Research Workers (10th ed.), Edinburgh, Oliver and Boyd, 1946.
5. Shannon, C. E. and Weaver, W., The Mathematical Theory of Communications, Urbana: The University of Illinois Press, 1949.