

Report on Reproducibility of Methods of Chemical Analysis used in the Iron and Steel Industry

BS 4237 : 1967

THIS REPORT, having been approved by the Iron and Steel Industry Standards Committee and endorsed by the Chairman of the Engineering Divisional Council, was published under the authority of the General Council on 29 December, 1967.

The Institution desires to call attention to the fact that this report does not purport to include all the necessary provisions of a contract.

In order to keep abreast of progress in the industries concerned, British Standards are subject to periodical review. Suggestions for improvements will be recorded and in due course brought to the notice of the committees charged with the revision of the standards to which they refer.

A complete list of British Standards, numbering over 4000, fully indexed and with a note of the contents of each, will be found in the British Standards Yearbook, price 15s. The BS Yearbook may be consulted in many public libraries and similar institutions.

This report makes reference to the following British Standard:

BS 1121. Methods for the analysis of iron and steel.

British Standards are revised, when necessary, by the issue either of amendment slips or of revised editions. It is important that users of British Standards should ascertain that they are in possession of the latest amendments or editions.

The following BSI references relate to the work on this report:
Committee reference ISE/18 Draft for comment D64/307

CO-OPERATING ORGANIZATIONS

The Iron and Steel Industry Standards Committee, under whose supervision this report was prepared, consists of representatives from the following Government departments and scientific and industrial organizations:

Alloy Steels Association
 Board of Trade
 *British Cast Iron Research Association
 British Constructional Steelwork Association
 British Electrical and Allied Manufacturers' Association
 *British Iron and Steel Federation
 British Ironfounders' Association
 British Mechanical Engineering Federation
 British Railways Board
 *British Steel Castings Research Association
 Council of Ironfoundry Associations
 Council of Iron Producers
 Crown Agents for Oversea Governments and Administrations
 Engineering Equipment Users' Association
 Federation of Civil Engineering Contractors
 Institute of British Foundrymen
 Institute of Iron and Steel Wire Manufacturers
 Institute of Marine Engineers
 Institution of Civil Engineers
 Institution of Mechanical Engineers (Automobile Division)
 Institution of Production Engineers
 Institution of Structural Engineers
 Iron and Steel Board
 Iron and Steel Institute
 Joint Iron Council
 Lloyd's Register of Shipping
 *Ministry of Defence, Army Department
 *Ministry of Defence, Navy Department
 Ministry of Labour, H.M. Factory Inspectorate
 *Ministry of Technology, National Physical Laboratory
 *National Association of Drop Forgers and Stampers
 Oil Companies Materials Association
 Royal Institute of British Architects
 Shipbuilding Employers' Federation
 Society of British Aerospace Companies Ltd.
 Society of Motor Manufacturers and Traders Ltd.
 Stainless Steel Development Association

The Government department and scientific and industrial organizations marked with an asterisk in the above list, together with the following, were directly represented on the committee entrusted with the preparation of this report:

British Iron and Steel Research Association
 Ministry of Technology, Laboratory of the Government Chemists
 Permanent Magnet Association
 Royal Institute of Chemistry
 Individual analysts and statisticians

REPORT ON REPRODUCIBILITY OF METHODS OF CHEMICAL ANALYSIS USED IN THE IRON AND STEEL INDUSTRY

FOREWORD

This British Standard is based on a report prepared by an ad hoc panel of the British Iron and Steel Research Association Methods of Analysis Committee*. This panel was set up to examine the difficulties which were experienced in obtaining and expressing adequately the reproducibility of procedures used in metallurgical analysis and to make recommendations as to the procedure to be followed when preparing any future methods for the analysis of iron and steel.

The recommendations made were approved by BISRA Chemical Analysis Committee, and are embodied in this report. While they were formulated with particular reference to the methods of analysis for iron and steel, published as parts of BS 1121, and the examples quoted relate to such methods, the recommendations are of more general application, and may be used by other committees preparing British Standard methods for the analysis of metallurgical materials.

An appendix gives specific recommendations as to the form in which the statement of reproducibility should be included in the British Standard 1121 series of methods of analysis for iron and steel.

1. REPRODUCIBILITY AND ACCURACY

A glossary of terms used in this report is given in Appendix C.

It is necessary first to discuss the relationship between reproducibility and accuracy.

For the purposes of this standard the term 'reproducibility'[†] means the consistency of replicate tests; it indicates the measure of agreement between values obtained when using a particular method.

'Reproducibility' where only one operator is involved using only one set of apparatus is more specifically defined as 'repeatability' (i.e. within operator reproducibility).

* Title changed in 1961 to Chemical Analysis Committee.

[†] It is recognized that certain other industries use the term 'precision' for the concept termed 'reproducibility' in this standard and restrict the term 'reproducibility' to the concept described in this standard as 'between operator reproducibility'. The term 'precision' is not used in this standard since within the iron and steel industry it has been considered that it might be misinterpreted as meaning 'accuracy'.

Should the consensus of opinion in industry in general decide on definitions which differ from those employed in this standard the Technical Committee will review the position.

It should be stressed that reproducibility alone tells us nothing about the accuracy of values obtained. Thus a method of analysis may have a bias towards either high or low results and still be reproducible. Alternatively, the presence of some interfering element in a particular sample may also result in a bias giving rise to values that, whilst reproducible, are inaccurate. There is no mathematical means of assessing whether values obtained are biased. This must be done by analysing samples or synthetic solutions in which the amount of the element to be determined is accurately known. This should always be done during the development of a new method where the primary aim is accuracy.

In practice both accuracy and reproducibility must be taken into account, and ideally both high accuracy and good reproducibility are sought. Once it has been established that there is no detectable bias in a method, then the better the reproducibility the narrower will be the range which is expected to include the true value.

2. MEASUREMENT OF REPRODUCIBILITY

Before it is possible to consider the best method of measuring reproducibility a brief account must be given of the underlying statistical theory. Every analyst knows that when replicate determinations are made on a sample, the results are frequently not identical even if the sample is homogeneous. Errors occur during weighing, estimation of solution volumes, etc. and in the absence of bias, these are equally likely to be positive or negative. It is the resultant of these errors that appears in the final analytical figure. In general there will be a tendency for some errors to neutralize the effects of others by operating in the reverse direction and it is usually found that in replicate determinations, those results in which the above errors almost neutralize one another occur more frequently than those with a higher nett error. An illustration of this typical condition is given in Fig. 1 in which the results of 57 determinations of nickel in a 3.19% nickel steel are recorded to 0.01% and the frequency of results at each level shown as ordinates.

It is important to consider the change that would occur to Fig. 1 if the intervals were decreased in size and at the same time the number of determinations increased. Under these conditions the steps between ordinates would become smaller until ultimately an infinitesimally small interval size coupled with an infinity of results would yield a smooth frequency distribution curve shown in Fig. 2. This curve (the Normal curve) has certain properties which are of immediate interest. Firstly it should be noticed that the 'tails' are asymptotic to the x -axis; there are no discontinuities which might help in defining reproducibility and all that can be inferred is that results become less frequent as their nett positive or negative errors increase. There is a chance, albeit a slight one, that nett errors of considerable magnitude will occasionally arise and therefore the range (the difference between the maximum and minimum values) of a set of replicate determinations will in general increase with the number of replicates.

A second important property of the Normal curve is that the proportion of results occurring within any interval on the x -axis is equal to the area bounded by the x -axis, the interval lines and the Normal curve, (see Fig. 3) divided by the total area under the curve. The Normal curve therefore permits the calculation of the probability of a single determination lying within any chosen limits. Since most analytical replicates with which this report is concerned approximate to a normal distribution the fact that the tails of the Normal curve are asymptotic signifies that any reproducibility limits given for an analytical method must be on a probability basis. Thus as the reproducibility limits are increased the probability of falling outside them decreases but never vanishes. Some agreed probability level (usually 95 %) must therefore be clearly stated with any quoted reproducibility level, and since the probability of any result falling within the reproducibility levels is not 100 % an occasional analysis must be expected to lie outside these levels. The standard deviation (see below) of a Normal curve facilitates the calculation of the reproducibility of results at any chosen probability level. Thus approximately 95 % of results will lie within ± 2 standard deviations from the mean or 99.7 % within ± 3 standard deviations from the mean.

It will be appreciated that the Normal curve is an ultimate which is never attained in practice since the number of replicate determinations is finite. The problem is always present, of inferring the characteristics of the Normal curve which represents an infinity of results from a finite number of replicate determinations. Thus means and standard deviations from experimental data are estimates of the true mean and standard deviation (i.e. of the Normal curve); as the number of data increase so these estimates approach more closely the true values. It follows that if only a few results are available the mean and standard deviation estimates can be very far from the true values.

The uncertainty associated with using the estimated standard deviation as a measure of the true standard deviation must reflect in the probability level assigned to any reproducibility limits. Thus if only 5 replicates are available, the estimated standard deviation must be multiplied by 2.57 instead of 2 to obtain the reproducibility at a 95 % probability level. Table 1 shows that when approximately 20 or more results are available a multiplication factor of 2 may be used without involving appreciable error.

TABLE 1

No. of results	Multiplication factor for 95 % confidence limits
5	2.57
10	2.23
15	2.13
20	2.09
30	2.04
60	2.00
> 100	1.96

3. DISCUSSION OF THE RELATIVE MERITS OF RANGE AND STANDARD DEVIATION

Hitherto the expression for reproducibility of methods of analysis for iron and steel, included in the various parts of BS 1121* has been based on the range of values obtained by a number of co-operating analysts. The possible advantages and disadvantages of this system compared with a more refined statistical approach are considered below.

Both range and standard deviation are statistical measures of the spread of replicate chemical analyses. Range is the difference between the greatest and least values observed. Its only real advantages are the ease with which it can be calculated and its obvious interpretation. It has two serious disadvantages:

- (1) There are no real limits to the values a variable can take; all that can be said is that high and low values are less frequent than the intermediate ones. Range is consequently very sensitive to the occurrence of one of these extreme values.
- (2) Range is liable to fluctuate considerably from batch to batch and tends to increase with the number of tests.

Standard deviation is the square root of the mean of the squares of the deviations of all values from the mean. Since the true mean for an infinite number of values is unknown, for a group of N values an estimate of the standard deviation is made using the formula:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$$

where $\bar{x}$ is the mean of N values and x is any individual value.

The main advantages of standard deviation are:

- (1) It is possible to estimate the proportion of values which lie within certain limits.

* B.S. 1121, 'Methods for the analysis of iron and steel'.

- (2) It is relatively stable, regardless of the number of values considered (above a certain minimum) or the incidence of extreme values.
- (3) It is possible to express standard deviation as a quantity made up of components due to laboratory error, operator error and random error, by the use of a good experimental plan. Conversely, since it is possible to separate these individual sources of error the major weaknesses of a test procedure can be readily identified.

To illustrate the comparison between range and standard deviation a block of 1000 values was taken from a table of random normal deviates*. Fig. 4 shows the histogram of these values.

The first 5 values were taken and the range and standard deviation computed. A further 5 values were added and the range and standard deviation re-computed. This process was repeated until results were obtained for 5, 10, 20, 50, 100, 500 and 1000 values as shown in the following table:

TABLE 2

No. of values	Mean	Range	Standard deviation <i>s</i>
5	3.012	0.23	0.094
10	3.013	0.24	0.083
20	3.010	0.28	0.092
50	3.002	0.39	0.095
100	2.988	0.40	0.091
500	2.995	0.59	0.094
1000	2.998	0.71	0.099

The true values for the whole population of 25 000 are a mean of 3.0 and a standard deviation of 0.1. It can be seen that as more figures become available the mean and standard deviation remain fairly stable and close to their true values, whilst the range steadily increases as the probability of exceptionally large or small values increases. Thus the inherent stability of standard deviation, in spite of outlying values, can be seen.

In addition, we can make probability statements regarding the distribution in terms of standard deviation. We would expect:

68.3% of the results to lie within $\pm 1s$ of the mean

95.5% of the results to lie within $\pm 2s$ of the mean

99.7% of the results to lie within $\pm 3s$ of the mean

The actual state of affairs for the 1000 results is summarized in Table 3.

* See 'Tracts for computers No. 25, Random normal deviates', by Herman Wold.

TABLE 3

No. of s.d.'s	Lower limit	Upper limit	Expected percentage	Actual percentage
1	2.9	3.1	68.3	67.8
2	2.8	3.2	95.5	95.7
3	2.7	3.3	99.7	99.7

Since the range increases with the number of tests, it cannot be utilized in the same manner.

To illustrate this principle on some actual chemical analysis data, the results of 57 determinations of nickel, on a standard steel sample containing 3.19%, shown in Table 4 were considered.

TABLE 4

NICKEL DETERMINATIONS (ACCEPTED VALUE 3.19 %)

3.20	3.21	3.16	3.16	3.19	3.18	3.17
3.19	3.18	3.21	3.20	3.21	3.14	3.21
3.19	3.20	3.21	3.13	3.16	3.21	3.17
3.20	3.22	3.22	3.19	3.23	3.25	3.22
3.20	3.18	3.19	3.19	3.21	3.18	3.18
3.17	3.17	3.18	3.18	3.18	3.19	3.18
3.20	3.17	3.18	3.20	3.20	3.20	3.19
3.18	3.20	3.10	3.20	3.19	3.20	3.19
3.19						

These were all individual determinations made at different times. The mean value of the 57 results was 3.189 and the standard deviation 0.0239.

81 % of the results lie within $\pm s$ and 93 % within $\pm 2s$, and all but one result within $\pm 3s$. The distribution of the results is shown in Fig. 1.

4. WAY OF EXPRESSING REPRODUCIBILITY

In general, it is required to express reproducibility in a simple way and this can be quoted as a multiple of the standard deviation. Reasons for choosing standard deviation have been given above. The recommended figure is 2 standard deviations, since the 95 % confidence limit (Appendix C) is of more practical value than a confidence limit of 68 %.

Repeatability is also expressed in similar terms.

In cases where reproducibility varies with concentration level or type of sample, it will be necessary to quote the different confidence limit for each level or sample. Where the standard deviation is constant it will be possible to use a single figure.

5. EXPERIMENTAL DESIGN

Much effort is expended in the search for new and improved analytical methods, and this will culminate in a final experimental plan carried out to confirm the suitability of the method and assess its reproducibility. A method which survives the laboratory examinations to the stage of this final plan will be known to be free from bias or, at least, it will be thought more accurate on theoretical grounds than any alternative method. A great deal of effort will be wasted if attempts are prematurely made to define reproducibility before a method is known to measure what it sets out to measure.

The object of the experimental design is to assess the several sources of variation which together cause the differences in results obtained from test to test. If tests on the same sample are carried out by different operators, in general they will not obtain identical results. The mean for each operator will vary about the mean of all operators' results (variation between operators). In the same way, other causes of variation can be assigned and isolated in the experimental design. Even when all the known sources of error are taken into account, and the same operator in the same laboratory carries out repeat tests, there will be differences in the results and this is known as the within operator variance.

When the separate sources of variation have been estimated, it is possible to assess the readiness of the test for use, and to make precise statements about its reproducibility. It is then possible to calculate with any desired degree of confidence the number of tests required to make an accurate estimate within pre-set limits. The method is elaborated in Appendix A.

This information is of great value and it is of the utmost importance that sufficient work be done in these final stages of method development to ensure a reliable estimate of reproducibility.

Ideally more than one operator in each laboratory should participate in order to separate the laboratory variation, and the operator variation. This is however, considered to be impracticable and in general one highly experienced operator only from each laboratory will participate. It should be noted that in these circumstances the within operator and within laboratory variations are synonymous.

Each operator will be required to carry out at least 3 determinations on each sample and it is important also that within a given project each operator shall undertake the same number of tests on each sample and that the same number of operators shall examine each sample. Providing at least 10 laboratories will co-operate in a given series, 3 determinations by each laboratory on each sample will be adequate. Alternatively with only 5 or 6 co-operating laboratories, 5 tests per laboratory will be required. The following procedure should be strictly adhered to in all cases.

Each chemist will first satisfy himself that he is familiar with and skilled in the particular technique. If there is any reason to suspect that the procedure is