Acta Cryst. (1974). A30, 607

Systematic Intensity-Dependent Differences in Structure Factors Derived from the Single Crystal Intensity Measurement Project of the International Union of Crystallography

BY J. K. MACKENZIE

Division of Chemical Physics, CSIRO, P.O. Box 160, *Clayton, Victoria, Australia* 3168

(*Received* 19 *November* 1973; *accepted* 1 *April* 1974)

14 of the 17 sets of structure factor data for D(+)-tartaric acid collected on the initiative of the Commission on Crystallographic Apparatus of the IUCr have been compared without making explicit assumptions concerning either the nature of the interaction of X-rays with a crystal or of the statistical distributions of the measurement errors. It is shown that the greater part of the differences between the 14 sets of structure factors derived from the different single crystals depends systematically on the intensity and wavelength in a manner strongly suggesting extinction as the cause. These differences are apparent even for medium intensities and the largest structure factors may be underestimated by as much as a factor of two. When the major systematic intensity-dependent differences are removed the statistics of the residuals show an excess of both large and small values as compared with a normal (Gaussian) distribution. The effects of different weighting schemes on various R values are considered and it is shown that while the ordinary R value is fairly robust, weighted R values with weights that correspond to assuming $\sigma(F) \propto F$ or $\sigma(F) = \text{const.}$ grossly overemphasize respectively either the low or the high intensities.

1. Introduction

In most single crystal structure analyses, structure factors are derived from one set of intensity measurements made on one single crystal. Thus, there is no direct experimental basis for a proper estimate of accuracy, and hence of the physical significance of the measurement. Furthermore, the scientific literature contains very little numerical information concerning the measurement differences likely to be encountered in standard procedures of intensity data collection. This lack of information is commonly overcome either by guessing the functional dependence of the statistical errors (standard deviation proportional to structure factor, say) or by comparing observed structure factors with those calculated from some model. The latter process assumes that the calculated values are virtually error free, and is frequently circular in its application. Both procedures are unsatisfactory for the assessment of experimental accuracy.

For these reasons the publication of 17 sets of experimentally determined structure factors for D(+)-tartaric acid obtained from different crystals provides a unique opportunity to investigate the range of variation likely to be encountered in a practical situation involving a low-absorption, low-atomic-weight organic compound: a situation which is reasonably typical of many structure analyses. The Commission on Crystallographic Apparatus of the IUCr which initiated the collection of this data has already published a Report on the Project in two parts (Abrahams, Hamilton & Mathieson, 1970; Hamilton & Abrahams, 1970), only the first part of which (hereinafter called the Report) is of concern in the present paper. The Report examined the trends of differences between sets in relation to a number of variables (I, θ, h, k, l) and arrived at certain general conclusions, but did not give much indication of the magnitudes of the individual differences or of their functional form.

In the present paper the basic data of the Report has been re-examined and subjected to a closer study. For it is our belief, following earlier experience (Mackenzie & Maslen, 1968), that the extensive data derived in projects warrants study from a number of different viewpoints (Mathieson, 1969). In the present study the number of *a priori* assumptions has been reduced to a minimum and, in particular, there are no assumptions (other than those already implied by the data in the Report) concerning either the behaviour of a crystal when interacting with X-rays or the nature of the statistical distributions which describe the measurement errors. Although the conclusions are general rather than specific it is shown below that the greater part of the differences between sets of structure factors derived from the different single crystals used in the project depends systematically on intensity in a manner strongly suggesting extinction as the cause. Although some of the statistical analysis carried out in the Report indicated the presence of intensity-dependent differences, they were not regarded as important and there is no emphasis on them in its conclusions.

When confronted with data for analysis the first need is to subject it to 'primitive' tests which assume as little as possible about the nature of the underlying distribution of errors (Tukey, 1972). These tests include visual inspection, and the application of either 'distribution-free' statistical tests or ones which are 'robust' in the sense that the results are insensitive to a wide range of underlying statistical distributions of error. In the case of the Report visual inspection

first suggested the gross differences between sets 12 or 13 and the remainder of the data (Mathieson, 1969). The Report also used a simple R value to assess both the internal and the mutual consistency of the data; as will be seen later, the simple R value is relatively robust.

More sophisticated statistical tests usually make assumptions about the underlying statistical distributions. Thus, before such tests are applied there is a need to verify as far as possible that the data being analysed are not inconsistent with being drawn from the assumed class of distributions. The Report used both weighted R values and the technique of analysis of variance for which the weights and hence the standard deviations of the observations are assumed to be known apart from a scale factor. No attempt was made to estimate the standard deviations $\sigma(F)$ from the data, it being simply assumed* that $\sigma(F)=kF$. This assumption is inconsistent with the obvious requirement that low intensities (and small values of F) are associated with an irreducible minimum of error. As will be seen, it led to an overweighting of the low F values which was so gross that the 32 (out of 331) reflexions of lowest intensity contributed nearly 2/3 of the value of the weighted R value wR. Such an unsatisfactory state of affairs weakens somewhat the considerations in the Report relating both to wR and, to a lesser extent, the analysis of variance.

The investigations reported in the present paper began with a straightforward calculation of the standard deviations of the observations for each reflexion and then sought for evidence of dependence on F and other variables. The wealth of data available is such that these estimates should be quite reliable and so provide a reasonable basis either for assessment of accuracy or for weighting.

§2 of the present paper defines the raw data used. The primary calculations discussed in §3 are concerned with the exposure of systematic differences between sets and the allocation of their origin to the different crystals used. In these calculations the observations in the raw data were given equal weight and there was no rescaling. In subsequent calculations edited data were also manipulated. The edited data were obtained from the raw data by the removal of the 59 observations in brackets in Table 1 and the reasons for regarding these 59 observations as outliers are given in §4. §5 describes an *ad hoc* generalized scaling procedure which either removes most of the systematic differences or at least places all the crystals on a common basis. §6 deals briefly with the statistical distribution of the residuals, while §7 is concerned with the effects of using various weighting schemes on the values of some different R values. In both these sections a variety of weighting schemes are considered.

* If counting statistics were the main contributor to the errors this assumption corresponds to a constant total count for each reflexion.

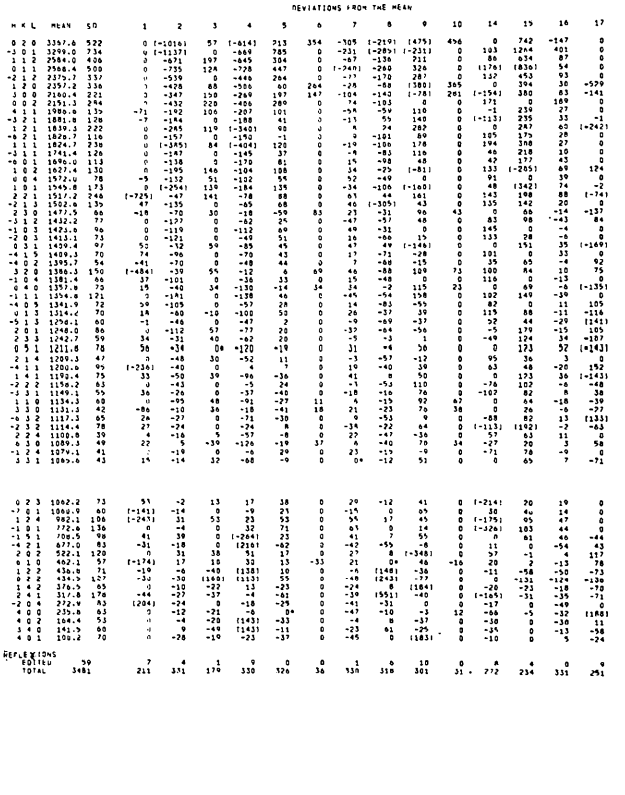
2. The raw data and its weighting

In this paper the term intensity is used interchangeably with $|F|^2$ and refers to the intensity actually measured modified by an Lp correction. It is concerned solely with a selection of observations called the raw data which are $|F|$ values extracted from Table 4 of the Report. The raw data are composed of all the data in the Report except those for sets 12, 13 and 11*b*; thus there were 3481 observations spread over 331 reflexions and made by 14 experimenters using different single crystals. The reasons for the exclusion of these three sets are given below. Sets 6 and 10, which comprised only 36 and 31 observations respectively were included in all calculations but are excluded from further discussion.

In the primary calculations the observations were all given equal weights. The reason for adopting this weighting scheme to begin with was that one of the objects of the present investigation was to discover how the standard deviations depend on either the reflexion being measured or the experimenter, and the initial hypothesis was that all variances (square of standard deviation) were equal. The discussion in §5 shows that, after the outliers and a systematic dif-

Table 1. *Mean values of* $|F|$, *standard deviations* (*SD*) *and residuals of the raw data for the* 50 *most intense reflexions*

Entries in brackets were subsequently edited out; the lower part of the table lists values only for those reflexions from which data was edited. All data were equally weighted and sets 11*b*, 12 and 13 excluded. The last column labelled 17 is for set 11*a*.



ference have been removed, this is not far from the truth (see Fig. 2 and Table 3).

There was no preliminary rescaling of the raw data. The Report had already scaled the data with $\sigma(F)=kF$ which gives the low intensities large weight and so optimizes the relative scale factors at the low-intensity end of the range of intensities. This is an advantage if it is believed that these low intensities are the least subject to errors such as extinction.

Sets 12 and 13 were excluded on the basis both of independent calculations and the evidence in the Report of a systematic l-dependent deviation of these sets from the others. Such deviations are consistent with maladjustment of the equi-inclination instruments used. In the independent calculations, separate scale factors were assumed for each value of $l=0,1\ldots 6$ for each of the 17 experimental sets and the method of Hamilton, Rollett & Sparks (1965) used to make the adjustment. The results showed that the scale factors for set 13 varied systematically with l over a 7:1 range (50:1 in intensity), those for set 12 varied over a 2:1 range (in the opposite sense), while in all other cases the greatest deviation was $\pm 11\%$ for sets 7 and 14. Similar results have been published previously by Mathieson (1969; Fig. 6). Experimenter 11 offered two sets of data and expressed a preference for 11*b*. However, this set was rejected on the rather arbitrary basis that it seemed from the Report to differ from the other sets slightly more than did the original set 11*a*.

3. Residuals of the raw data

The Report gives estimates of $|F|$ by a number of experimenters for each reflexion. Using the raw data for each reflexion in turn the mean and standard deviations of these estimates were computed together with the deviation of each estimate from the calculated mean. The (rounded) values for the 50 most intense reflexions are given in Table 1 while all the deviations are plotted against $|F|$ for each experimenter (except 6, 10) in Fig. 1.

With the exception of sets 9, 1 and 11*a*, these data are plotted in the following way. The reflexions were ordered by decreasing values of the mean $|F|$. Then, for each experimenter, the averages of ten consecutive reflexions are plotted against the mean value of $|F|$ with error bars indicating one standard deviation of these ten values on either side of the mean (this represents $\pm 1/10 \simeq \pm 3$ standard deviations of the mean). The individual results for the 20 most intense reflexions

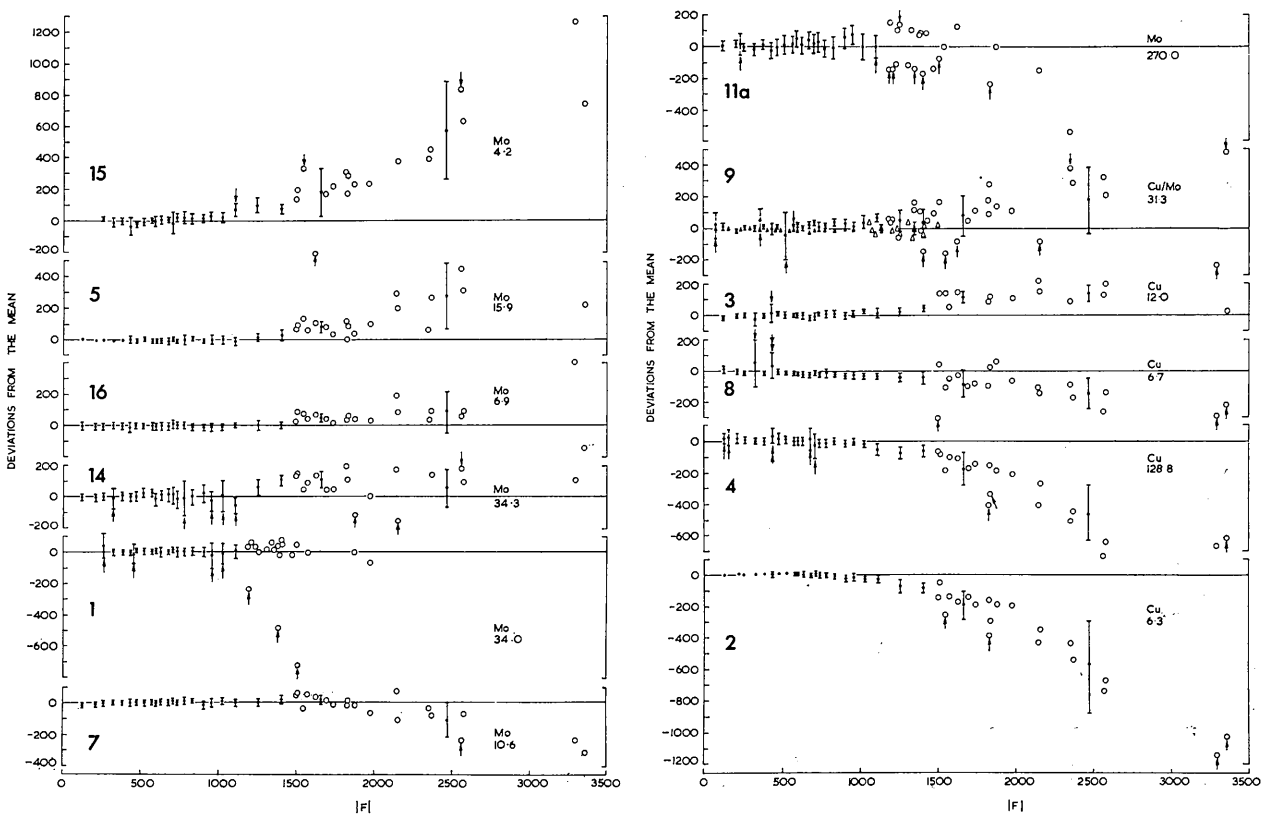


Fig. 1. Plot of residual deviations for the raw data and various experimental sets (number at left). The radiation used is indicated at the right of each plot and the crystal volume (in units of 10^{-12} m^3) immediately below it. The open symbols refer to individual values and the closed symbols (with error bars) to a mean and standard deviation for 10 reflexions of neighbouring intensity. The arrows mark each piece of data edited out. For set 9 the circles refer to data taken with Cu $K\alpha$ radiation and the triangles to data taken with Mo $K\alpha$ radiation.

($|F| > 1500$) are also plotted separately by open circles while for $|F| < 560$ only every second average is plotted except where there is an associated outlier when the error bar is broken. For set 9 the circles refer to results obtained with Cu $K\alpha$ radiation for $l=0,1,2$ and the triangles to results obtained with Mo $K\alpha$ radiation for $l=3,4,5$. The grouping of the observations has been changed slightly both in this case and also for sets 1 and 11*a*.

The standard deviations are similarly plotted in Fig. 2(*a*) against $|F|$ and show a clear tendency to increase with $|F|$. If the same data are plotted against angle of reflexion, the scatter of the points increases markedly at the lower angles of reflexion where the larger intensities occur and no clear trend with angle is discernible. It may therefore be concluded that the standard deviations depend mainly on the intensity.

Fig. 2(*a*) shows that the standard deviation increases with $|F|$ at more than a linear rate. Such a dependence could be due to (*a*) a statistically random error, (*b*) a systematic intensity-dependent difference, or (*c*) the sets being wrongly scaled, *i.e.* a non-physical systematic difference. Just which of these possibilities applies can be decided by reference to the actual deviations from the mean for each experimenter as shown by the plots in Fig. 1.

A statistically random error implies that for each experimenter positive and negative deviations should occur about equally often and at random. It is immediately obvious from Fig. 1 that while this may be so for middle to low values of $|F|$ it is certainly not true for a number of sets at high values of $|F|$. Thus there is clear evidence of a systematic error.

The scaling in the Report tended to equalize the data for low $|F|$ so that any change of scale would simply add a deviation proportional to $|F|$ to those already plotted. Since a number of the plots show deviations from a straight line through the origin and, what is more, curvatures in opposite senses, it is clear that no changes of scale can bring all the plots into reasonable agreement.

The inescapable conclusion is therefore that there are systematic differences between the sets of data which are intensity dependent and that these are most likely due to physical causes. It will now be shown that these differences are probably wavelength-dependent, though other factors can sometimes override this dependence.

The various plots in Fig. 1 are labelled with the set number, the type of radiation used and the approximate volume of the crystal in units of 10^{-12} m^3. Disregarding for the moment the two sets 11*a* and 9 at the top of the right-hand group, the remaining sets in this group all used Cu $K\alpha$ radiation while those in the left-hand group all used Mo $K\alpha$ radiation. It is clear that the magnitude of the systematic differences in the left-hand group decreases from large positive deviations to small negative ones for Mo $K\alpha$ radiation while for the right-hand group the differences decrease from a small positive deviation to large negative ones for Cu $K\alpha$ radiation. There does not seem to be any strong correlation with crystal volume.

Sets 11*a* and 9 do not seem to fall into the simple pattern. Set 11*a* used Mo $K\alpha$ radiation but has a negative trend as large as any of those using Cu $K\alpha$ radiation. However this set used by far the largest crystal. Set 9 is interesting because both radiations were used on the one crystal. While there was a reasonable coverage of $|F|$ values for Cu $K\alpha$ radiation (circles), only medium and low values of $|F|$ were measured with Mo $K\alpha$ radiation (triangles). If the above pattern of variation with wavelength were universal it would be expected that if reflexions with large $|F|$ values had been measured with Mo $K\alpha$ radiation, they would have been systematically above comparable values measured with Cu $K\alpha$ radiation. While there is no evidence that this is so, there is also no information concerning the way in which the two partial sets were brought to the same scale.

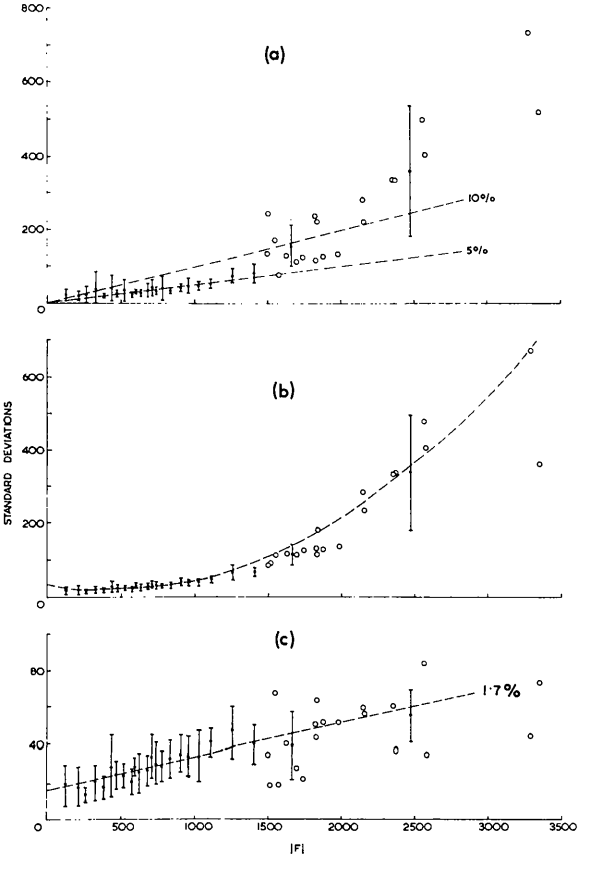


Fig. 2. Standard deviations for (*a*) the raw data, (*b*) the edited data and (*c*) the data corrected for the intensity-dependent systematic difference (note the 5× expansion of vertical scale). The curves drawn show for (*a*) deviations of 5 and 10%, (*b*) a quadratic fit to the edited data and (*c*) a linear fit (slope 1·7%) to the corrected data. Open circles refer to individual values and closed circles (with error bars) to a mean and standard deviation for 10 reflexions of neighbouring intensity.

The above assignment of the systematic differences to physical causes still leaves open the question of their origin. Is it associated with the individual crystal being measured or does it arise from the apparatus used in the measurement? It is difficult to answer this question unequivocally but a comparison of the present results with those of the American Crystallographic Association (ACA) single-crystal project is suggestive. In the ACA project the same single crystal of calcium fluoride was measured by seven experimenters so that the reported differences can probably be assigned to the measuring apparatus. The assessment of this data by Mackenzie & Maslen (1968) shows that the medium to high $|F|$ values were estimated with standard deviations of less than $1\frac{1}{2}\%$ while for low $|F|$ values the standard deviation was always less than $3\frac{1}{2}\%$.* On the other hand, in the present project it is apparent from Fig. 2(*a*) that the standard deviations for the medium to high $|F|$ values increase from 5 to over 15%. This comparison therefore suggests that the origin of the systematic differences lies mainly within the individual crystals used by the different experimenters. The Report draws a similar conclusion on the basis of a comparison of the R values in its Table 7.

4. Outlying observations

Inspection of the larger residuals indicates that there are some outlying or wrongly recorded observations. For example, several of the observational values corresponding to the bracketed entries at the bottom of Table 1 are two to three times the values estimated by any other experimenter for the same reflexion; this, in spite of the fact that the Report's scaling procedure (*cf.* §2) should have brought these low-intensity reflexions into near coincidence. It therefore seemed desirable to try to detect such outliers and find the effect of their removal.

Any process, such as the one below, which systematically rejects data is to some extent subjective and the results of subsequent calculations must be interpreted with some reservations. At worst the rejected data properly belong to the observations and so the results of the subsequent calculations are biased. At best these data are properly removed and a bias which would otherwise have arisen has been avoided. A more cautious approach is to repeat subsequent calculations using all the data for one set of calculations and rejecting certain data for the other 'edited' set. This establishes clearly how the conclusions are affected by the rejection of certain data. Another way of looking at the process is to regard it as one which divides the data into two or more homogeneous parts, each of which can be investigated separately; perhaps the suspected outliers can be referred to the original source of the observations. The initial rejection of sets 12 and 13 belongs to this latter way, the remaining raw data being treated separately here. The outliers in this remaining raw data are treated in the former manner by duplicating calculations.

The 59 observations considered to be outliers in the raw data are bracketed in Table 1 and marked with arrows in Fig. 1. After their removal every large error bar in Fig. 1 is reduced to about the same size as its neighbours and the resulting standard deviations, plotted in Fig. 2(*b*), behave much more regularly than the corresponding results for the raw data in Fig. 2(*a*). The reduction in the total residual sum of squares is given in Table 2.

Table 2. *Residual sum of squares for various calculations*

All data weighted equally. Sets 11*b*, 12, 13 excluded.

Data used	Number of scale factors/set	Residual sum of squares $\times 10^{-4}$	Degrees of freedom
Raw	None	2223·3	3150
Edited	None	1388·9	3091
Raw	1	1511·6	3136
Raw	2	723·3	3122
Raw	3	667·0	3108
Edited	3	289·2	3049

The decision to regard these 59 observations as outliers was reached in the following way.

The first problem was to remove as far as possible the systematic differences between sets which are associated with intensity. This was done by taking the observations ordered by mean intensity in small groups of 10 or 20 consecutive reflexions and rescaling them using the method of Hamilton, Rollett & Sparks (1965) with equal weights for all observations. Then, for each reflexion, the deviations of the observations from their mean were printed out together with two statistics suggested by Dixon (1962) as suitable for locating outliers. These statistics were C_1'=range/standard deviation and r_{10}=the difference between the two largest (or smallest) deviations/range; observations significant at the 1% level for either statistic were marked. A marked observation was put on the list of possible outliers only if it also stood out from other deviations for the same experimenter: Fig. 1 suggests that these deviations should all be of about the same magnitude. After removing the observations so found the calculation was repeated.

Finally, the listed possible outliers were carefully considered individually and the observations of largest intensity were counted as outliers only after the systematic difference had been removed by the method described in the next section.

In any individual experiment outliers are hard to detect unless repeated observations are made. Almost 2% of the present observations have been classified as outliers and it is tempting to ask if this is the rate of temporary and undetected malfunction of the (automated) equipment used. A colleague who operates an automated equipment for a very different experimental

* These figures were obtained after disregarding two outlying sets and the most intense reflexion.

purpose reports a similar rate of errors due to supposed equipment malfunction but it is to be noted that a re-examination of a large mass of astronomical observations [Ash, Shapiro & Smith, 1971, see item (6), col. 2, p. 552] suggests a rate of 1 % for transcription errors.

5. Eliminating intensity-dependent differences

In this section a purely phenomenological attempt is made to estimate and remove the major intensity-dependent differences between sets. While no physical significance should be attached to the particular numerical values of the constants evaluated, the split of the deviation into systematic and random parts is probably valid. On the assumption that the systematic difference has a smooth dependence on intensity the discussion in the last section suggests trying to represent it by means of variable scale factors with a simple polynomial dependence on intensity. This assumption is supported by the results of Fig. 4 of Denne (1972).

Writing F rather than $|F|$ for the magnitude of the structure factor it is assumed that the value F_{hi} estimated by the ith experimenter for the reflexion with index h is given by

$$F_{hi}=g_1F_h+g_2F_h^2+g_3F_h^3+e_{hi}\,, \qquad (1)$$

where e_{hi} is the random error, $g_r\equiv g_r(i)$ for $r=1,2,3$ are scaling constants and F_h is a (true) value of the structure amplitude which is free of systematic error and hopefully independent of the particular crystal used for measurement. The values of F_h obtained below are 'extrapolated means' obtained from all the observations taken together. For this purpose the usual method of adjusting the scale factor needs some generalization.

A least-squares method which generalizes that of Hamilton, Rollett & Sparks (1965) was used to evaluate all the parameters. The weighted sum of squares

$$S=\sum_{hi}\sum w_{ij}\{F_{hi}-g_1(i)F_h-g_2(i)F_h^2-g_3(i)F_h^3\}^2 \qquad (2)$$

is minimized by an iterative two-stage calculation. In the first stage values of the F_h are regarded as given and minimization of S with respect to variation of the $g_r(i)$ leads to a number of independent least-squares adjustments of standard type which determine values of g_1,g_2,g_3 and a residual sum of squares for each set of experimental data. In the second stage the current values of $g_r(i)$ are used and S is minimized, for each h, for variation of F_h; this requires the solution of a fifth-degree polynomial. The values of $g_r(i)$ and the associated values of F_h are normalized by the condition $\sum_i[g_1(i)]^2=14$, the number of experimental sets. Finally, the whole calculation is repeated until S is judged to have reached its final minimum value.

These calculations were carried out for various degrees of complexity in the polynomial (1) ranging from no adjustment through one scale factor per set with $g_2(i)\equiv g_3(i)\equiv 0$ to three scale factors per set. Some numerical values are given in Tables 2, 3 and 4.

Table 2 shows the marked decrease in the total residual sum of squares S which arises both from increasing the number of scale factors per set and from the editing of the data. Standard statistical F-tests show significance but their detailed validity is open to question since the errors are not normally distributed (see §6).

In Table 3 the order of the sets is the same as in Fig. 1. The upper group used Mo radiation and the lower Cu radiation; sets 6 and 10 are also listed. Comparison of the scale factors g_r for the raw and

Table 3. *Values of three constant scale factors with standard deviations in units of the last figure*

All data equally weighted and sets 11*b*, 12, 13 excluded. Starred entries for raw data differ appreciably from those for edited data.

Set	Edited data g_1	$g_2\times10^4$	$g_3\times10^9$	Residual sums of squares $\times10^{-4}$	Raw data g_1	$g_2\times10^4$	$g_3\times10^9$	Residual sums of squares $\times10^{-4}$
15	0·9841± 47	−0·966± 29	5·27± 31	25·6	0·9766± 69	−0·946± 42	5·26± 46	55·6
5	0·9721± 28	−1·145± 17	6·74± 18	16·3	0·9680± 36	−1·154± 22	7·02± 25	25·6
16	0·9905± 25	−1·327± 15	7·96± 16	12·7	0·9852± 35	−1·325± 22	8·11± 24	24·5
14	1·0025± 58	−1·250± 37	6·41± 39	42·5	0·9920± 76	−1·288± 47	7·32± 53*	78·4
1	1·0078± 71	−1·281± 98	1·71±285	8·9	1·0603±228*	−2·310±304*	28·50±879*	99·6
7	1·0055± 26	−1·413± 16	7·92± 17	13·6	1·0099± 28	−1·493± 17*	8·99± 19*	14·9
11*a*	1·0044±148	−1·004±172	−8·02±372	74·4	1·0222±159	−1·386±174*	0·31±373*	102·3
9	0·9920± 57	−1·134± 52	5·59± 87	25·0	0·9762± 78	−1·089± 48	5·18± 54	100·9
3	1·0037± 36	−1·296± 22	7·65± 27	7·5	0·9980± 46	−1·259± 27	7·00± 30	12·7
8	0·9899± 40	−1·497± 35	9·25± 60	15·4	0·9889± 60	−1·528± 37	9·84± 41	66·9
4	1·0446± 38	−1·982± 22	13·24± 24	27·7	1·0481± 54	−2·080± 33*	15·10± 37*	57·2
2	1·0512± 38	−2·034± 34	13·67± 59	14·9	1·0415± 35	−2·001± 21	13·30± 24	24·1
6	0·9805±100	−1·124± 58	6·51± 64	2·6	0·9695± 87	−1·059± 49	5·57± 52	2·1
10	0·9674±107	−0·941± 60	4·53± 66	2·3	0·9569±101	−0·883± 56	3·72± 59	2·1
				289·4				666·9

edited data shows that in the majority of cases there are no appreciable differences. Detailed consideration of the scale factors (for the edited data) shows that sets 2 and 4 are distinctly different from the remainder and that sets 1 and 11*a* appear anomalous. The values for the lower set using Cu radiation are more variable than those for the upper set (Mo) and are ordered in the same way as suggested by Fig. 1. The upper set is not so well ordered in relation to Fig. 1 but the numerical values of g_2 and g_3 increase together and because of their opposite sign tend to offset one another. This is potentially dangerous situation in curve fitting and suggests that the use of higher-degree polynomials would be undesirable.

As would be expected, editing the data has quite an appreciable effect on the residual sums of squares given in Table 3. However, even after editing, the ratio of largest to the smallest is not improved and remains about 10:1 (sets 6, 10 omitted). Since there are 200–300 individual contributions to all these values (see Table 1) there seem to remain real differences in the residual random error for the various experimental sets; whether this is due to technique or to the crystal is not entirely clear, but it is more likely due to technique. The standard deviations of the residuals for each reflexion after removal of systematic error are shown in Fig. 2(*c*); for the higher values these standard deviations are about 2% and very similar to those attained in the ACA project mentioned previously. Further, if these residuals are displayed in the same order as the angle of the corresponding intensity, no systematic dependence is apparent. This suggests that these residuals are probably due to instrumental effects and that a major systematic difference which depends only on the intensity has been effectively removed.

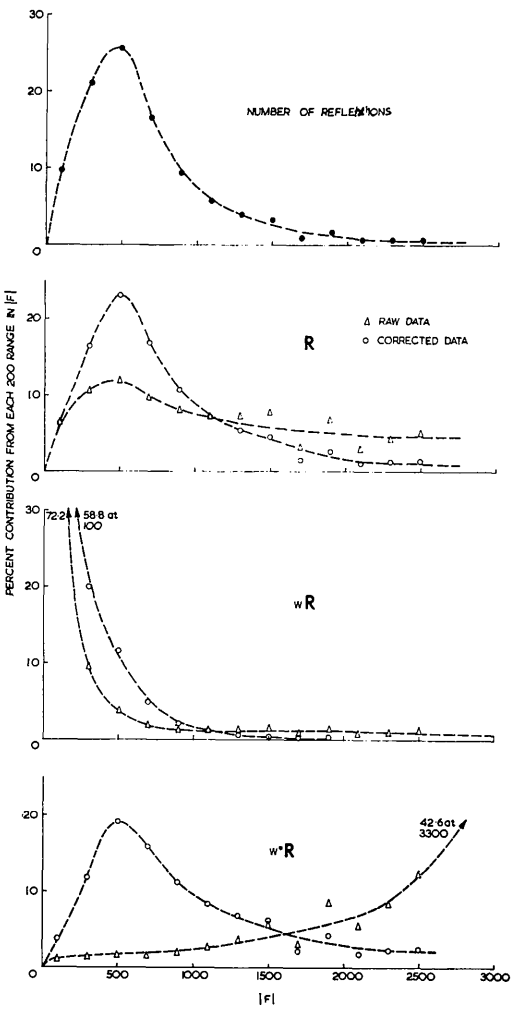


Fig. 3. Graphs showing for each 200 range of $|F|$ the number of reflexions and the contributions to various R values from each of these ranges.

The extrapolated mean values of F_h given in Table 4 are interesting because for the larger F values they are nearly a factor of two greater than the original mean values and more than 50% greater than the largest observed value. This is in the expected direction if the systematic difference is due to extinction. It might be argued that with the formal flexibility of variable scaling implied by equation (1) only a change in g_r is required and whether F_h is above or below any observed value is irrelevant. However, test calculations on artificial data with systematic *and random* errors as specified in equation (1) showed that the method of calculation in fact reproduced the original (true) F_h values reasonably well in spite of the fact that these values were all systematically above the artificial data. Thus given the correctness of the representation (1), the calculations give the correct answers. The extrapolated values in Table 4 may be too large but if the systematic differences are due to extinction there is a very real possibility that its influence may be much greater and extend to much lower intensities than has been previously realized.

Table 4. *Some extrapolated mean values derived from three constant scale factors*

Reflexion	Original mean	Extrapolated values for edited data	raw data
112	2584·0	5116·3	5158·0
230	1477·5	1915·8	1946·9
140	1010·8	1186·7	1200·5
540	502·3	539·8	542·9
$\bar{4}01$	203·3	208·4	209·3

6. Distribution of residuals

Throughout all the preceding calculations statistics of the residuals were printed out. For each experimenter's set the squares of the residuals were sorted into 11 ranges so chosen that if the deviations were normally distributed with a variance independent of $|F|$ there would be equal numbers of deviations in each of the

first 10 ranges and none in the final outlier range in which only $\frac{1}{1000}$ of the values would be expected to lie; the appropriate statistical distribution is that of χ^2. The variances (square of the standard deviation) for each reflexion were similarly sorted and counted.

Of course both the raw and edited data showed gross departures from normality due to the presence of systematic differences. But even when these differences had been removed as described above departures from normality were still apparent. These departures took the form of an excess both of small deviations and of large deviations and of outliers. Such behaviour is consistent with the steady increase in standard deviation with F shown in Fig. 2(*c*) since the large deviations tend to be associated with large F and the small deviations with the large number (see Fig. 3) of reflexions with small F. On the other hand, the small increase in the standard deviation of the residuals with $|F|$, which is only 2%, may indicate a further systematic effect. But, since it could also be an artifact of the method of removing differences, it did not seem profitable to pursue the matter further.

When the variation of standard deviation with F was allowed for by division of the deviations for each reflexion by values obtained from the straight-line approximation in Fig. 2(*c*), the departures from normality were still noticeable in both the statistics for each individual experimenter and in the statistics for the variances of the reflexions. Again the departures took the form of an excess of both small and large deviations and outliers. (Were all the outliers removed by the method of §4?) Even when the experimental sets were weighted approximately inversely as the squares of their overall standard deviations, departures from normality, although evident, were less pronounced and perhaps even acceptable. The same excess of both large and small residuals is reported for astronomical data by Ash, Shapiro & Smith (1971, see footnote Col. 1, p. 555). Assessment of the statistics is difficult because every change of scale and of weighting scheme changes the deviations. Since the considerations of the present paper do not rest on establishing the normality of the distribution of the errors the matter has not been followed any further.

It is only when the errors are normally distributed that the usual statistical tests can be relied on to give valid results. By the nature of the investigation this cannot be done until this very late stage. For this reason all statements concerning explicit levels of significance have been carefully avoided in the present work. Unfortunately one can only record that sums of squares either increase or decrease in various circumstances and make subjective but nevertheless informed statements of opinion concerning them

7. Some effects of weighting schemes

The purpose of the present section is to comment on some single numbers of the R-value type which are sometimes used to measure the agreement between sets of data. It will be shown that the weighting scheme used in the Report may have undesirable consequences.

The Report defines two measures of agreement between the ith experiment and the set of mean values

$$R_{i\mu}=\sum_h |F_{ih}-\bar{F}_h| / \sum_h \bar{F}_h \qquad (3)$$

and (g for general)

$$w^g R_{i\mu}=\{\sum_h [(F_{ih}-\bar{F}_h)/\sigma]^2 / \sum_h [(\bar{F}_h/\sigma)^2]\}^{1/2}. \qquad (4)$$

The Report assumed $\sigma(F)=kF$ so that

$$wR_{i\mu}=\{\sum_h [(F_{ih}-\bar{F}_h)/\bar{F}_h]^2 / \sum_h 1\}^{1/2}, \qquad (5)$$

while the other extreme assumption $\sigma(F)=k$ gives

$$w^* R_{i\mu}=\{\sum_h [F_{ih}-\bar{F}_h]^2 / \sum_h \bar{F}_h^2\}^{1/2}. \qquad (6)$$

Thus, $wR_{i\mu}$ is the root-mean-square fractional deviation, $w^*R_{i\mu}$ a root-mean-square deviation and $R_{i\mu}$ a mean of the magnitudes of the deviations.

In order to assess the contributions from various ranges of F to these R values it is assumed that the deviations of F_{ih} from their mean value $\bar{F}_h$ are normally distributed with standard deviations s_h. Then, if an average over n experimental values is taken, the expected value of

$$\sum_i (F_{ih}-\bar{F}_h)^2/n \quad \text{is} \quad s_h^2(n-1)/n\,,$$

and of

$$\sum_i |F_{ih}-\bar{F}_h|/n \quad \text{is} \quad s_h[2(n-1)/n\pi]^{1/2}.$$

Two expressions for s_h have been used. One derived from the smooth curve in Fig. 2(*b*) for the edited raw data (triangles in Fig. 3)

$$s_h=35-7\times 10^{-2}F+0{\cdot}8\times 10^{-4}F^2 \qquad (7)$$

and the other from the straight line in Fig. 2(*c*) for the edited raw data corrected for systematic differences (open circles in Fig. 3)

$$s_h=17+0{\cdot}017F\,. \qquad (8)$$

The relative numbers of reflexions in each 200 range of F are shown at the top of Fig. 3 while the other graphs show the relative contribution from each range of F to the various R values. The absolute values of R and wR calculated for the edited raw data are in substantial agreement with an appropriate mean of values given in the Report for the raw data while values calculated from the corrected data are compatible (they are not strictly comparable) with values given in the Report for the internal consistency.

In making statistical comparisons it is usually considered desirable to have each deviation contributing

as nearly equally as possible to the test statistic. Thus from the theoretical point of view the standard deviation σ in w^*R should be chosen as nearly equal to the true value as possible while from the practical point of view the contribution for various ranges of F in Fig. 3 should follow the curve for the number of reflexions. From this point of view the statistic wR used in the Report disastrously overweights the low F values since 60% or more of the total is contributed by the 32 reflexions of lowest intensity. The disaster occurs because $\sigma(F)$ goes to zero with F, a situation which is not physically plausible. The statistic w^*R is almost as bad for the edited raw data but by contrast overweights the high intensities; however, it is satisfactory for the corrected data. The statistic R stands out as being rather less sensitive to maladjustment of the weighting scheme.

In spite of these deficiencies, as pointed out earlier in connexion with scaling, there may be good reason to prefer a scheme which overweights some F values at the expense of other ranges. The important thing is to be clear what one wants to do and the extent to which it has been done.

8. Conclusions

The main overall conclusion is that distinct crystals of tartaric acid as used for X-ray structure factor determination in the IUCr project differ in the intensities they give for a particular reflexion in a manner which systematically increases with intensity until it is generally about 30% (15% in $|F|$) for the strongest reflexions. These differences are more extensive than previously realized and begin to be apparent for reflexions of medium intensity. They are consistent with being due to extinction and the largest structure factors may be underestimated by as much as a factor of two.

The more detailed conclusions which can be drawn from the preceding investigation are summarily listed below in order of decreasing certainty under each of the five headings.

1. *Analysis of the raw data*

(*a*) There are systematic intensity-dependent differences in the structure factors derived from the different single crystals. These differences begin to be apparent at $|F|=1100$ and increase to be about 15% in $|F|$ for the largest structure factors of $|F|=3300$.

(*b*) These differences are not due to improper scaling and any dependence on angle of reflexion is minor [see 3(*b*) below].

(*c*) The differences are probably wavelength-dependent, though other factors such as crystal volume can over-ride this dependence.

(*d*) They are due to physical differences between the crystals themselves rather than the apparatus or measuring technique used.

(*e*) The differences are consistent with the phenomenon of extinction.

2. *Outliers*

(*a*) There are certainly some outliers.

(*b*) They may constitute as much as 2% of the total number of observations.

(*c*) This may be the rate either of temporary undetected malfunctions of the automated equipment used or of human transcription errors.

3. *Removal of intensity-dependent differences*

(*a*) The systematic differences can be approximately represented by means of a polynomial cubic in the true structure factor $|F_h|$.

(*b*) After removal of the systematic differences the residuals show no obvious dependence on angle of reflexion.

(*c*) The standard deviation of these residuals is finite for small $|F|$ and rises linearly with $|F|$ to about 2% for the higher intensities.

(*d*) Not all the experimental sets have equal precision.

(*e*) The extrapolated values $|F_h|$ may lie above the measured values by as much as a factor of two.

4. *Distribution of residuals*

(*a*) The presence of the systematic differences in the raw data leads to gross departures of the residuals from being statistically distributed in accordance with a normal distribution.

(*b*) Even when the systematic differences are removed, departures from normality remain, there being an excess of both large and of small deviations.

(*c*) When each experimental set is further weighted inversely as the square of its overall standard deviation, the above departures from normality are reduced to what may be acceptable limits for some purposes.

(*d*) Some traditional tests of statistical significance (such as an F test) which rely on the assumption of a normal distribution of errors and which were applied in the Report may be invalid.

5. *Effects of weighting schemes on R values*

(*a*) The root-mean-square fractional deviation wR proposed in the Report is such that out of 331 reflexions only those 32 reflexions of lowest intensity contribute 60% of the value of wR. The index wR, which assumes that $\sigma(F)=kF$, is unrealistic in that it requires $\sigma(F)$ to tend to zero with F and so results in an overweighting of low-intensity reflexions.

(*b*) The other extreme assumption $\sigma(F)=$constant almost as disastrously overweights the high intensities. The corresponding index w^*R is just a root-mean-square deviation.

(*c*) The more usual R index which is a mean of the magnitudes of the deviations is more robust.

I wish to thank A. McL. Mathieson both for many discussions concerning the philosophy of comparing data and for his persistent encouragment without which this paper may not have been written. I am also indebted to other colleagues L. D. Calvert (NRC, Canada)

and W. A. Denne for a critical reading of the manuscript.

References

ABRAHAMS, S. C., HAMILTON, W. C. & MATHIESON, A. MCL. (1970). *Acta Cryst.* A**26**, 1–18.
ASH, M. E., SHAPIRO, I. I. & SMITH, W. B. (1971). *Science*, **174**, 551–556.
DENNE, W. A. (1972). *Acta Cryst.* A**28**, 192–201.
DIXON, W. J. (1962). *Contributions to Order Statistics*, chap. 10. Edited by A. E. SARHAN and B. G. GREENBERG. New York: John Wiley.
HAMILTON, W. C. & ABRAHAMS, S. C. (1970). *Acta Cryst.* A**26**, 18–24.
HAMILTON, W. C., ROLLETT, J. S. & SPARKS, R. A. (1965). *Acta Cryst.* **18**, 129–130.
MACKENZIE, J. K. & MASLEN, V. W. (1968). *Acta Cryst.* A**24**, 628–639.
MATHIESON, A. MCL. (1969). *Acta Cryst.* A**25**, 264–275.
TUKEY, J. W. (1972). *Quart. Appl. Math.* **30**, 51–65.

Acta Cryst. (1974). A**30**, 616

Etude Expérimentale des Susceptibilités Diamagnétiques Moléculaires. I. Méthode Générale

PAR G. VAN DEN BOSSCHE ET R. SOBRY

Laboratoire de Cristallographie et de Physique de l'Etat Solide, Université de Liège au Sart Tilman, B4000 Liège, Belgique

(*Reçu le* 21 *janvier* 1974, *accepté le* 8 *avril* 1974)

Lonsdale and Krishnan used a method for relating crystal and molecular susceptibilities in which they calculated the molecular tensor from the principal crystal susceptibilities assuming that the directions of the principal axes of the molecular tensor are known. Their method is only applicable when the molecular symmetry permits. To avoid this inconvenience and to allow an interpretation of molecular anisotropies in terms of chemical groups, we assume that each molecule consists of a skeleton on which different substitutions are made. The molecular tensor is thus the sum of the different elementary tensors associated with the molecule and related to an orthogonal and common set of axes. Several molecules with common skeletons and substitutions were studied to obtain the different elementary tensors; the set of equations which relate elementary tensor components to crystal susceptibilities in the approximation of weak intermolecular interactions was solved by a least-squares method. Our method was applied to benzene, hexachlorobenzene and hexamethylbenzene separately at first and then all together. The results are given for the substitution of a hydrogen atom by a chlorine atom or a methyl on a benzene ring. Several possible applications of this method are discussed.

Introduction

Au cours des 20 dernières années, le nombre d'études consacrées aux mesures de susceptibilités magnétiques à l'état solide n'a cessé de décroître. La raison de ce désintérêt réside dans les difficultés inhérentes à l'interprétation des résultats expérimentaux. La détermination de la susceptibilité moyenne ne permet le plus souvent qu'une simple comparaison de la valeur mesurée et de plusieurs valeurs calculées déduites des différentes systématiques théoriques ou expérimentales. Quant aux mesures de l'anisotropie magnétique du cristal, elles ne permettent de déduire le tenseur moléculaire que dans certains cas extrêmement favorables (Lonsdale & Krishnan, 1936). Ce dernier tenseur s'avère pourtant essentiel lorsque l'on veut interpréter les mesures en termes des différents groupements qui constituent la molécule et ainsi contribuer à l'élaboration de modèles théoriques satisfaisants pour rendre compte des propriétés magnétiques des molécules, encore inexpliquées pour des molécules aussi usuelles que le benzène, le naphtalène et l'anthracène (Caralp & Hoarau, 1968, 1969, 1972). Les développements récents de la résonance magnétique nucléaire permettent de relier les déplacements chimiques observés aux susceptibilités magnétiques moléculaires (Memory, 1968) et confèrent un intérêt certain à l'interprétation des mesures de susceptibilités cristallines. En fait, le problème posé est double. Il importe d'abord de déduire le tenseur moléculaire et ensuite d'expliquer les variations observées d'un composé à l'autre en termes de forces inter ou intramoléculaires. Il sera alors possible de déterminer l'influence des divers types d'interactions et d'orienter la recherche future de modèles théoriques qui rendent compte de l'importance relative des différentes contributions.

Pour déterminer les valeurs principales du tenseur magnétique moléculaire, Lonsdale & Krishnan (1936) ont écrit les équations qui relient les valeurs principales aux susceptibilités cristallines mesurées et à