Here $\mathrm{Ei}(x)$ denotes the exponential integral function defined as

$$x<0: \quad \mathrm{Ei}(x)=\int_{-\infty}^{x}(e^{t}/t)\,\mathrm{d}t$$

$$x>0: \quad \mathrm{Ei}(x)=-\lim_{\varepsilon\to 0}\left(\int_{-x}^{-\varepsilon}+\int_{\varepsilon}^{\infty}\right)(e^{-t}/t)\,\mathrm{d}t \qquad (22)$$

and $C=0{\cdot}577215\ldots$ is the Euler constant.

References

BIRD, D. M. & KING, Q. A. (1990). *Acta Cryst.* A**46**, 202-208.

BUXTON, B. F. & LOVELUCK, J. E. (1977). *J. Phys. C*, **10**, 3941-3958.

DOYLE, P. A. (1970). *Acta Cryst.* A**26**, 133-139.

DOYLE, P. A. & COWLEY, J. M. (1974). *International Tables for X-ray Crystallography*, Vol. IV, pp. 152-174. Birmingham: Kynoch Press. (Present distributor Kluwer Academic Publishers, Dordrecht.)

DOYLE, P. A. & TURNER, P. S. (1968). *Acta Cryst.* A**24**, 390-397.

FOX, A. G. & O'KEEFE, M. A. (1989). *Acta Cryst.* A**45**, 786-793.

GAUKLER, K. G. & GRAFF, K. (1970). *Z. Phys.* **232**, 190-204.

GOODMAN, P. & LEHMPFUHL, G. (1967). *Acta Cryst.* **22**, 14-24.

GORINGE, M. J. (1966). *Philos. Mag.* **14**, 93-97.

HALL, C. R. & HIRSCH, P. B. (1965). *Proc. R. Soc. London Ser. A*, **286**, 158-177.

HASHIMOTO, H. (1964). *J. Appl. Phys.* **35**, 277-290.

HASHIMOTO, H., HOWIE, A. & WHELAN, M. J. (1962). *Proc. R. Soc. London Ser. A*, **269**, 80-103.

MEYER-EHMSEN, G. (1969). *Z. Phys.* **218**, 352-377.

PENG, L.-M. & COWLEY, J. M. (1988). *Acta Cryst.* A**44**, 1-5.

RADI, G. (1970). *Acta Cryst.* A**26**, 41-56.

REIMER, L. (1984). *Transmission Electron Microscopy*, pp. 295-300. Berlin: Springer.

REIMER, L. & WÄCHTER, M. (1980). In *Electron Microscopy*, Vol. 3, edited by P. BREDEROO & G. BOOM, pp. 192-193. Leiden: Seventh European Congr. Electron Microscopy Foundation.

RENARD, D., CROCE, P., GANDAIS, M. & SAUVIN, M. (1971). *Phys. Status Solidi B*, **47**, 411-421.

ROSSOUW, C. J. & BURSILL, L. A. (1986). *Proc. R. Soc. London Ser. A*, **408**, 149-164.

SCHOBER, H. R. & DEDERICHS, P. H. (1981). In *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, edited by K.-H. HELLWEGE & J. L. OLSEN, Group III, Vol. 13a, pp. 1-191. Berlin: Springer.

STEEDS, J. W. (1983). In *Quantitative Electron Microscopy*, edited by J. N. CHAPMAN & A. J. CRAVEN, pp. 49-96. Glasgow: Scottish Univs. Summer School in Physics.

WEICKENMEIER, A. (1990). In preparation.

YOSHIOKA, H. (1957). *J. Phys. Soc. Jpn*, **12**, 618-628.

Acta Cryst. (1991). A**47**, 597-604

The Interpretation of Raw Diffractometer Data

BY A. T. H. LENSTRA, H. J. GEISE AND F. VANHOUTEGHEM†

University of Antwerp (UIA), Department of Chemistry, Universiteitsplein 1, *B*-2610 *Wilrijk, Belgium*

(*Received* 27 *July* 1990; *accepted* 26 *April* 1991)

Abstract

Statistical analysis reveals that the X-ray background rigorously follows a counting statistical distribution provided the measurements are made under truly fixed-time conditions in a small $(\sin\theta)/\lambda$ interval. The operational procedures to ensure this are not trivial. First, the design of, for example, the Enraf-Nonius CAD-4 diffractometer is such that measurements made at constant scan speed at different Bragg angles may have somewhat different measuring times. Failure to correct for this leads primarily to an increase in the variance of the data. Second, the use of a rapid prescan followed, when appropriate, by a slower main scan leads to a set of prescan data biased towards overestimates of background values and underestimates of raw intensities. The effort needed to extract from the data unbiased estimates of averages and variances of the X-ray background is rewarding. It can lead to a lowering of the standard deviation of a net intensity by up to one order of magnitude. This in turn means that many more intensities of weak reflections are reliably estimated and are hence worth including in the structure determination. This obviously leads to increased model accuracy. An example is given. A change in measuring procedure is recommended which will increase the efficiency of the standard background-peak-background procedure.

Introduction

The present standard data-reduction procedure operates on each reflection measurement separately and thus completely ignores any knowledge that might have been acquired prior to the current measurement. This is surprising considering the great impact ascribed to experience in all aspects of life. More specifically, statistical methods are available to separate the effects of random measurement errors from systematic factors, as well as to extract from data sets information which can be used to judge the quality

† Deceased 23 January 1991.

0108-7673/91/050597-08$03.00

© 1991 International Union of Crystallography

of a single observation, provided the single observation and the data set are samples drawn from the same population. In this way relationships between measurements can be discovered which help to improve the accuracy of X-ray diffractometry. One example may prove our point. The net intensity I is routinely obtained by subtracting the local background B from the raw intensity R:

$$I = R - \gamma B.$$

The factor γ represents the ratio of the times used to measure R and B. Unfortunately, owing to the shape of the crystal and the design of the diffractometer, γ may vary slightly for reasons which are not immediately obvious. However, because R and B are pairwise related *via* the observed intensity profile, Lehmann & Larsen (1974) were able to design a procedure to optimize γ. Since measurements of background are numerous in diffractometric experiments, a statistical investigation of them seemed appropriate. By doing so we follow in the footsteps of French & Wilson (1978), who in their seminal paper on the application of Bayesian statistics on the problem of negative intensities state 'We believe that a complete structure determination based on a Bayesian treatment of the data will yield a significant improvement, especially when the available data are weak in intensity'. In this paper we will show that with the proper manipulation the distribution of B values, $P(B)$, can be well described by counting statistics, and that the average $\langle B\rangle$ and spread $\sigma^2(B)$ are only dependent upon $(\sin\theta)/\lambda$. Once $P(B)$ is known several possibilities arise for improving accuracy. For example, it will no longer be necessary to measure the background at each reflection. A much smaller number of observations suffices to produce $P(B)$ with a preset accuracy. This may save up to 30% of the time currently spent on single-crystal measurements. Furthermore, one can devise statistically correct criteria to detect outliers and design (software) repair procedures. Or one may calculate a statistically correct estimate of a data point, which for some reason escapes measurement. We will also demonstrate that knowledge of $P(B)$ allows the contribution of the background to $\sigma^2(I)$, the variance of the *net intensity*, to be eliminated almost completely. This translates into a reduction of detection limits by as much as a factor of ten. The effort to obtain the extra information is small: data reduction becomes a two-step procedure. In the first step one determines $P(B)$ from the recorded data and in the second step one uses this information in the evaluation of net intensities. It is appropriate to stipulate here what we understand by an intensity observation. We take the view that all measurements are made under correct diffractometer conditions (*e.g.* background measurements do not contain tails of reflection peaks, scan widths are correctly chosen, a crystal monochromator is used *etc.*). The procedures discussed in this work are strictly speaking only valid for measurements performed on an Enraf–Nonius CAD-4, because accuracy is obviously linked to the design and mode of operation of the diffractometer. However, similar arguments will apply to other diffractometers. Therefore, it seems useful to recall here some details of the data-collection process on an Enraf–Nonius CAD-4. After setting the crystal in the appropriate orientation to record the intensity profile of a particular reflection, the instrument performs a preliminary scan (prescan) in which the chosen scan angle α is scanned at high speed. The time needed to scan α is divided into 96 equal intervals and the number of counts collected during each interval is stored in the appropriate channel of a 96-channel recorder. Obviously, each channel number can be associated with a particular value of the Bragg angle θ, given the wavelength λ of the radiation and the value of α. Here, as usual, channels 1 to 16 are considered to contain the so-called left background BL, channels 81 to 96 the right background BR and channels 17 to 80 the raw intensity R. Table 1 gives *inter alia* the equations used in the CAD-4 logic while performing the background–peak–background (BPB) measurements. The ratio $I(\text{prescan})/\sigma[I(\text{prescan})]$, which is a program variable, is used to select one of several measuring strategies. In this work, if $I(\text{prescan}) < 0{\cdot}33\sigma(I)$ no final run is made and $I(\text{prescan})$ is used in the further analysis. We will refer to these data as prescan data. For the sake of completeness we mention that it is also possible for the prescan to be sufficient when the reflection is strong enough to give acceptable statistics even with such a fast scan. This never occurred in the data set used here. When, however, in this set $I(\text{prescan}) > 0{\cdot}33\sigma(I)$ the measuring time t for the final scan is adjusted to give an operator-chosen precision (say 1%) based on counting statistics. If $t < t_{\max}$, the maximum time allowed by the investigator, the fixed count strategy is chosen: the highest scan speed is selected which results in $I(\text{final})$ with the desired precision. Alternatively, if $t > t_{\max}$ the fixed-time strategy is chosen. That is the slowest scan speed is selected which gives an actual measuring time $\leq t_{\max}$, resulting in the highest number of counts attainable within the allowed time. We will refer to the latter data as fixed-time data. A complete description of the operational mode of the CAD-4 diffractometer is given by Schagen, Straver, Van Meurs & Williams (1988).

Preliminary analysis of background values

The data set used in this work is a routinely obtained set of measurements of a cyclic dipeptide. The structure is monoclinic, $P2_1$ (b axis unique), with four molecules per unit cell (Lenstra, Verbruggen, Bracke, Vanhouteghem, Reyniers & Borremans, 1991). The

Table 1. *Survey of definitions and equations used in this work*

v	scan speed (°min^{-1}) of measurement;
v_{max}	maximum scan speed (20·1° min^{-1}) on CAD-4, to which all data are scaled;
$C(i)$	number of counts in channel i;
BL	scaled left background;
BR	scaled right background;
R	scaled raw intensity;
σ^2(BL)	variance of BL, when counting statistics apply;
σ^2(BR)	variance of BR;
σ^2(R)	variance of R;
n	sample size.

$$\mathrm{BL} = (16g)^{-1} \sum_{i=1}^{16} C(i) \quad (1)$$

$$\mathrm{BR} = (16g)^{-1} \sum_{i=81}^{96} C(i) \quad (2)$$

$$R = (64g)^{-1} \sum_{i=17}^{80} C(i) \quad (3)$$

$$\sigma^2(\mathrm{BR}) = g^{-1}\mathrm{BL} \quad (4)$$

$$\sigma^2(\mathrm{BR}) = g^{-1}\mathrm{BR} \quad (5)$$

$$\sigma^2(R) = g^{-1} R \quad (6)$$

$$g = v_{max}/v \quad (7)$$

$$I = R - 2(\mathrm{BL} + \mathrm{BR}) \quad (8)$$

$$\sigma^2(I) = \sigma^2(R) + 4\{\sigma^2(\mathrm{BR}) + \sigma^2(\mathrm{BL})\} \quad (9)$$

$$\langle \mathrm{BL} \rangle = n^{-1} \sum_n \mathrm{BL} \quad (10)$$

$$\langle \mathrm{BR} \rangle = n^{-1} \sum_n \mathrm{BR} \quad (11)$$

$$\langle \mathrm{R} \rangle = n^{-1} \sum_n R \quad (12)$$

$$\langle \sigma^2(\mathrm{BL}) \rangle = n^{-1} \sum_n \sigma^2(\mathrm{BL}) = n\langle \mathrm{BL} \rangle / \sum_n g \quad (13)$$

$$\langle \sigma^2(\mathrm{BR}) \rangle = n^{-1} \sum_n \sigma^2(\mathrm{BR}) = n\langle \mathrm{BR} \rangle / \sum_n g \quad (14)$$

$$\langle \sigma^2(\mathrm{R}) \rangle = n^{-1} \sum_n \sigma^2(R) = n\langle R \rangle / \sum_n g \quad (15)$$

$$s^2(\mathrm{BL}) = (n-1)^{-1} \sum_n (\mathrm{BL} - \langle \mathrm{BL} \rangle)^2 \quad (16)$$

$$s^2(\mathrm{BR}) = (n-1)^{-1} \sum_n (\mathrm{BR} - \langle \mathrm{BR} \rangle)^2 \quad (17)$$

$$s^2(R) = (n-1)^{-1} \sum_n (R - \langle R \rangle)^2 \quad (18)$$

molecules form two crystallographically independent chains in the same orientation and are separated by a vector (0·5; 0; 0·5), generating a pseudo-B centring. Consequently, most reflections with $h + l = 2n + 1$ are much weaker than those with $h + l = 2n$, a feature that renders the set very suitable for the present statistical analysis. The systematic weakness of the intensities follows also from the high fraction of the prescan data (27%, $n = 678$) which is about half the fraction of fixed-time data (59%, $n = 1474$). Moreover, the prescan as well as the fixed-time data are about uniformly distributed over the θ space. The number of fixed-count data, concentrated at low θ values, is rather small (14%, $n = 366$). We will neglect them in this study, thereby avoiding statistical complications linked to small samples sizes.

We scaled all data to the maximum scan speed and divided the sets (prescan and fixed time) into a suitable number of $(\sin\theta)/\lambda$ intervals to eliminate the complications arising from the slow decrease of background values with increasing Bragg angle (Keulen, 1969). The results are summarized in Table 2.

First, we will analyse the prescan and fixed-time data together. For each $(\sin\theta)/\lambda$ interval we investigate whether the 'true' background can be regarded as a constant (that is the background is not related locally to the nearby raw intensity value) and its measurement is only influenced by random errors. This proved to be true, because each series of background values exhibits a normal distribution {Gaussian of the form $n(B) \simeq \exp[-(\mu - B)^2/2s^2]$, where $n(B)$ signifies the frequency of B values, μ the sample average and s^2 the sample variance}. At this point we are able to describe a series of say 200 background measurements by their μ and s^2, but s^2 has no firm *physical* backbone yet.

Further progress requires that the distribution also obeys counting statistics, *i.e.* follows approximately a Gaussian with $\mu = s^2$. The time series of background values observed at the intensity control reflections showed this behaviour, but the distribution of B values over a series of *hkl* behaved erratically in most instances. An example of the latter can be seen in Table 3 in which the cumulative distribution functions of two samples (BL and BR, each containing 274 values) are compared to the standard properties of a normal distribution. A mere glance at the data shows that the sample distributions are only moderately related to a normal distribution. Application of the Kolmogorov-Smirnov test (see *e.g.* Lindgren, 1976) at the 5% significance level shows that the hypothesis that the sample distribution is equal to the Gaussian must be accepted for BR data and rejected for the BL data. Evidently, we have to identify and correct for additional yet anonymous complicating factors (systematic errors).

The fatal alternative would be that the background does not obey counting statistics because the values are related to the nearby intensity values, possibly by Rayleigh scattering. This would most probably show from Table 2 in correct behaviour [*i.e.* $\langle \mathrm{BL} \rangle \simeq \langle \mathrm{BR} \rangle$ and $\langle \sigma^2(\mathrm{BL}) \rangle = \langle \sigma^2(\mathrm{BR}) \rangle = s^2(\mathrm{BL}) = s^2(\mathrm{BR})$] of the prescan backgrounds (near weak or absent reflections) and deviating behaviour [*e.g.* $\langle \mathrm{BL} \rangle = \langle \mathrm{BR} \rangle$; $\langle \sigma^2(\mathrm{BL}) \rangle = \langle \sigma^2(\mathrm{BR}) \rangle \neq s^2(\mathrm{BL}) = s^2(\mathrm{BR})$] of the fixed-time backgrounds (near strong reflections). For the prescan backgrounds $\langle \sigma^2(\mathrm{BL}) \rangle$, $\langle \sigma^2(\mathrm{BR}) \rangle$, s^2(BL) and s^2(BR) appear indeed equal and their equality is confirmed by the generalized Wilcoxson-Mann-Whitney test (Lindgren, 1976) at 20% significance (two-sided). Such an equality is absent in the fixed-time backgrounds: $\langle \sigma^2(\mathrm{BL}) \rangle$ and $\langle \sigma^2(\mathrm{BR}) \rangle$ belong to one population and s^2(BR) and s^2(BL) to another. However, for almost every $(\sin\theta)/\lambda$ interval $\langle \mathrm{BR} \rangle$ and $\langle \mathrm{BL} \rangle$ of the prescan data are larger than $\langle \mathrm{BR} \rangle$ and $\langle \mathrm{BL} \rangle$ of the fixed-time data. This observation rules

Table 2. *Summary of data set; see Table* 1 *for definitions of symbols*

$(\sin\theta)/\lambda$	n	⟨BL⟩	⟨R⟩	⟨BR⟩	⟨σ^2(BL)⟩	⟨$\sigma^2(R)$⟩	⟨σ^2(BR)⟩	s^2(BL)	$s^2(R)$	s^2(BR)	$\sum g$
Prescan data											
0 -0·16	1	56·7	162·0	43·3	18·9	54·0	14·4	-	-	-	3
0·16-0·31	22	18·0	62·5	16·6	6·0	20·8	5·5	11·6	64·8	7·5	66
0·31-0·39	32	15·7	57·7	15·3	5·2	19·2	5·1	4·4	45·7	8·6	96
0·39-0·45	38	14·5	50·7	12·6	4·9	16·9	4·2	6·7	56·4	4·8	114
0·45-0·49	72	11·7	39·0	9·3	3·9	13·0	3·1	5·0	40·7	3·8	216
0·49-0·53	88	10·4	35·5	8·8	3·5	11·8	2·9	4·4	40·5	4·2	264
0·53-0·56	111	9·1	31·9	7·9	3·0	10·6	2·6	2·7	21·9	2·7	333
0·56-0·59	110	8·8	30·5	8·1	2·9	10·2	2·7	3·4	26·8	3·4	330
0·59-0·62	120	8·2	29·1	7·8	2·8	9·7	2·6	2·4	18·7	2·6	360
0·62-0·65	84	8·2	29·3	7·8	2·8	8·9	2·6	2·9	14·4	2·5	252
Fixed-time data											
0 -0·16	1	37·4	152·8	32·4	2·0	8·0	1·7	-	-	-	19
0·16-0·31	112	16·9	231·1	16·4	0·9	12·4	0·9	3·8	39025	4·0	2128
0·31-0·39	232	14·9	121·0	14·3	0·8	6·4	0·8	1·8	5486	3·0	4408
0·39-0·45	178	13·2	110·8	12·0	0·7	5·8	0·6	1·6	6626	3·8	3382
0·45-0·49	163	11·3	85·4	10·0	0·6	4·6	0·5	1·5	3686	2·3	2987
0·49-0·53	168	9·7	54·1	8·4	0·5	3·0	0·5	1·0	565	1·0	3024
0·53-0·56	179	8·7	42·0	7·8	0·5	2·3	0·4	0·8	216	0·9	3222
0·56-0·59	180	8·1	37·4	7·6	0·4	2·1	0·4	0·9	81	0·8	3240
0·59-0·62	154	7·6	34·8	7·3	0·4	1·9	0·4	0·6	80	0·6	2772
0·62-0·65	107	7·6	34·5	7·3	0·4	1·9	0·4	0·7	68	0·9	1926

out Rayleigh scattering of the diffracted beam as an important factor. It suggests that a systematic bias affects the prescan background rather than the fixed-time background and that the cause is connected with the measuring strategy. Further inspection of Table 2 shows that the ratio s^2/σ^2 is about 1 for the prescan background, but larger than 2 for the fixed-time background. Thus, an error source produces in the latter a variance as important as the variance due to counting statistics.

Finally, we note that s^2/σ^2 for the fixed-time data increases with decreasing Bragg angle. These indications point to a cause connected to the diffractometer.

In the following sections we will show how ⟨BR⟩ and ⟨BL⟩ of the prescan data are biased estimates of the 'true' background and how all s^2 values are biased estimates of the 'true' background variance σ^2.

Bias as a result of measuring strategy

Prescan data exist as a separate subset because the investigator wishes to save measuring time and chooses a strategy to omit those reflections from the final scan for which, say, $I/\sigma(I) < 0{\cdot}33$. This criterion allows through reflections for which by coincidence the raw intensity is underestimated and/or the background is overestimated in preference to reflections for which by coincidence the raw intensity is overestimated and/or the background is underestimated. Therefore, underestimates of the background tend to be selectively missing from the set and prescan backgrounds are overestimates of the 'true' background. The sample means ⟨BL⟩ and ⟨BR⟩ will be too high, while the sample variances: s^2(BL) and s^2(BR) will be lower than the variances due to counting statistics. The fixed-time set is not limited by a selective criterion and thus this bias does not exist. Hence, in this case ⟨BR⟩ and ⟨BL⟩ are more faithful estimates of the 'true' background.

Table 3. *Cumulative distribution function of a normal distribution also obeying counting statistics* ($\mu = \sigma^2 = 7{\cdot}69$), *compared to a sample of* 274 *BL values and to a sample of* 274 *BR values*

Both samples are drawn from the set of Table 2 in the $(\sin\theta)/\lambda$ interval 0·59-0·62 Å^{-1}.

	$<-2\sigma$	$<-1\sigma$	$<-0\sigma$	$<+1\sigma$	$<+2\sigma$	$<+3\sigma$
Normal distribution	0·02	0·16	0·50	0·84	0·98	1·00
Left-side sample	0·03	0·25	0·60	0·84	0·97	1·00
Right-side sample	0·03	0·15	0·47	0·78	0·96	1·00

To quantify the impact of the selection criterion $I/\sigma < 0{\cdot}33$ we performed a Monte Carlo simulation of a prescan measurement of a net intensity I of zero counts. For a reflection in the interval $0{\cdot}59 \le (\sin\theta)/\lambda \le 0{\cdot}62$ Å^{-1} the best estimate of the background is ⟨BR⟩ = ⟨BL⟩ = 7·5 counts (Table 2) with ⟨σ^2(BR)⟩ = ⟨σ^2(BL)⟩ = 2·5 counts. The actual prescan is made at 1/3 of the maximum speed. So the local backgrounds BR and BL are equal to 22·5 counts and R is 90 counts. A series of 1000 (index j) such measurements were simulated. Counting statistical error margins were included by setting a count rate $\mathrm{BL}(j) \vee \mathrm{BR}(j) = 22{\cdot}5 + k \times 22{\cdot}5^{1/2}$. The factor k is calculated by a Gaussian random generator. As generating function we used $(-2 \ln u_1)^{1/2} \cos(\pi u_2)$, where u_1 and u_2 are random numbers. If by doing so a negative BL(j) or BR(j) value occurred, it was set equal to zero. When the criterion $I < 0{\cdot}33\sigma(I)$ was applied to the 1000 reflections 36% would trigger a final scan. The remaining 64% are thus typical for prescan data. They gave averages ⟨BR⟩ and ⟨BL⟩ = 8·4 counts (at maximum speed) with a variance of 1·7 counts. As the model requires, the average (8·4

counts) is larger than the unbiased average (7·5 counts) and moreover it is in excellent agreement with the observed values ⟨BR⟩ = 8·2 counts and ⟨BL⟩ = 7·8 counts (see Table 2). Since the biased background average can be directly enumerated from the unbiased value, the statement 'background is a local constant' still holds. Turning to the variances one notes that in agreement with the model the Monte Carlo variance [σ^2(counting statistics) + s^2(selection criterion) = 1·7] is smaller than σ^2(counting statistics) = 2·5 and that s^2(selection criterion) = $-0{\cdot}3\sigma^2$ (counting statistics). However, in contrast to expectations, Table 2 reveals that for most $(\sin\theta)/\lambda$ intervals the observed variances are larger than the counting statistical ones. A similar discrepancy ($s^2 > \sigma^2$) is present in the 'unbiased' fixed-time data. Thus, both subsets suggest a second external error source, preferably common to both sets, affecting the spreads.

Bias as a result of diffractometer design

The Enraf–Nonius CAD-4 has scan speeds between 0·4 and 20·1° min^{-1}. However, only those speeds are actually accessible for which the reduction factor ($g = v_{max}/v_{actual}$) is an integer ($1 \le g \le 50$). On the other hand, the angle to be scanned is enumerated as $\alpha = a_0 + \delta \tan\theta$. Thus a continuously variable α is combined with a set of discrete scan speeds. This combination forces the scan time to become a continuously adjustable parameter. Fig. 1 gives as a function of θ the time spent in reality to measure reflections in the so-called fixed-time set as they were measured in the particular scan mode used for this crystal [mode in which short-term intensity fluctuations are determined by a non-equal test (Schagen *et al.*, 1988)]. The relevant diffractometer settings are: signal profile scan width $\alpha = 2 + 0{\cdot}5 \tan\theta$; total scan width performed = $1{\cdot}5\alpha$ and maximum scan time (t_{max}) spent on one scan = 90 s. It follows that real fixed-time measurements only occur if α is constant

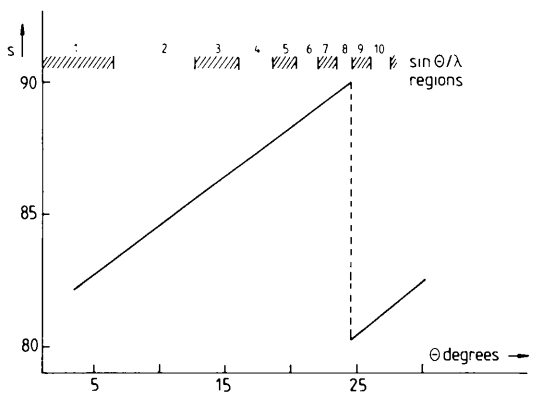

Fig. 1. Scan time as a function of θ for so-called fixed-time measurements (see text).

because either θ is constant or δ is zero. The latter is an unlikely event because δ is used in the measurement strategy to take care of wavelength dispersion (De Wolf & Lenstra, 1984). In a series of intensity-control measurements, however, θ is constant. For this series the background was found to follow counting statistics ($s^2 = \sigma^2 = \mu$). Averaging, as we did in Table 2, over different values of $\theta(hkl)$ we automatically randomized the scan-time differences within each $(\sin\theta)/\lambda$ interval. Table 4 shows the time difference (Δt) spanning each interval and the spread (s_3) resulting from the time randomization under the assumption that the individual reflections are uniformly distributed over the $(\sin\theta)/\lambda$ interval. This assumption is better justified for small intervals at high θ than for the larger interval at small θ.

In the model outlined so far, the total variance is the sum of the variances of three individual error sources: counting statistics (s_1), selection criterion (s_2) and time randomization (s_3). From the previous section we take $s_2^2 = 0$ for the fixed-time data and $s_2^2 = -0{\cdot}3s_1^2$ for the prescan data. The latter will be a good estimate for the prescan set in the interval $0{\cdot}59 \le (\sin\theta)/\lambda \le 0{\cdot}62$ Å^{-1}, but a crude one for the other intervals. In fact the mean and variance of a prescan set can vary considerably as a function of the original distribution (*i.e.* before the selection criterion is applied). Nevertheless, a good agreement is found (Table 4, last two columns) when we compare the estimates of $(\sum s_i^2)/s_1^2$ with the values of the sample variance: counting statistical variance ratio.

We conclude that the model has linked the observed variances (s^2) to the counting statistical variances (σ^2) sufficiently well to explain why s^2 is closer to σ^2 in prescan data than in fixed-time data and to show that at low θ values the error due to time randomization is more important than the error due to counting statistics.

Improvement of background values

In the previous sections we concluded that in each $(\sin\theta)/\lambda$ interval the observed backgrounds can be adequately described given the moments of the counting statistical distribution. This conclusion has far-reaching repercussions on the enumeration of net intensities. In this section we discuss the improvement of accuracy in I values that follows from the improvement of B values alone. Conditional probability theory is a particularly useful tool to this end and therefore we recall here some aspects. The likelihood function $P(B|b)$ expresses the probability that the value B is experimentally observed under the condition that the ideal value is b. The posterior function $P(b|B)$ expresses the probability density of b under the condition that B has been observed. Having measured B, one is more interested in b and $\sigma^2(b)$ or, in mathematical language, in the first and second

Table 4. *Background variances influenced by the use of fixed scan speeds in connection with variable scan widths*

No.	$(\sin\theta)/\lambda$ interval ($Å^{-1}$)	Δt* (%)	Background† (counts)	s_1‡ (counts)	s_3‡ (counts)	$\sum s_i^2/s_1^2$§ Estimated	Observed
Fixed-time data							
2	0·16-0·31	2·68	35431	188	950	26·5	4·3
3	0·31-0·39	1·47	64357	254	946	14·9	3·0
4	0·39-0·45	1·13	42613	206	482	6·5	4·2
5	0·45-0·49	0·78	31812	178	248	2·9	3·5
6	0·49-0·53	0·80	27367	165	219	2·8	2·0
7	0·53-0·56	0·63	26582	163	167	2·1	1·9
8	0·56-0·59	0·62	25434	159	158	2·0	2·1
9	0·59-0·62	0·66	20651	144	136	1·9	1·5
10	0·62-0·65	0·67	14349	120	96	1·6	2·0
Prescan data							
2	0·16-0·31	2·68	1142	34	31	1·5	1·7
3	0·31-0·39	1·47	1488	39	22	1·0	1·3
4	0·39-0·45	1·13	1545	39	17	0·9	1·3
5	0·45-0·49	0·78	2268	48	18	0·8	1·3
6	0·49-0·53	0·80	2534	50	20	0·9	1·3
7	0·53-0·56	0·63	2831	53	18	0·8	1·0
8	0·56-0·59	0·62	2789	53	17	0·8	1·1
9	0·59-0·62	0·66	2880	54	19	0·8	0·9
10	0·62-0·65	0·67	2016	45	14	0·8	1·0

* Difference (in %) in scanning times spanning the interval $\Delta t = 200\,(t_{max} - t_{min})/(t_{max} + t_{min})$, with t_{max} and t_{min} as the maximum and minimum scan times, respectively.

† Background measurement (in counts) = 0·5 $(\sum g)$ (⟨BR⟩+⟨BL⟩).

‡ Variance due to counting statistics: s_1 = (background)$^{1/2}$; variance due to selection criterion: $s_2^2 = 0$ for fixed-time data; $s_2^2 = -0{\cdot}3s_1^2$ for prescan data; variance due to time randomization: $s_3 = 0{\cdot}01\Delta t$ (background).

§ Ratio of total variance to variance (counting statistics). Estimated: $\sum_i s_i^2/s_1^2$; observed: $[s^2(\mathrm{BL}) + s^2(\mathrm{BR})]/[\langle\sigma^2(\mathrm{BL})\rangle + \langle\sigma^2(\mathrm{BR})\rangle]$.

moments of $P(b\,|\,B)$:

1st moment $$b = \langle b\,|\,B\rangle = \int bP(b\,|\,B)\,db \qquad (19)$$

2nd moment $$\sigma^2(b) = \langle b^2\,|\,B\rangle = \int b^2P(b\,|\,B)\,db. \qquad (20)$$

$P(B\,|\,b)$ and $P(b\,|\,B)$ are related through the theorem of Bayes:

$$P(b\,|\,B) = P(B\,|\,b)P(b)/P(B), \qquad (21)$$

where $P(b)$ and $P(B)$ are the unconstrained probabilities of b and B, respectively. The function $P(b)$ is often called the 'prior'. We now set out to find the proper analytical expressions for the various probability functions in (19)-(21).

Since B follows counting statistics, B is linked to b through

$$P(B\,|\,b) \simeq \exp\,[-(B-b)^2/2b] \qquad (22)$$

provided B is not too small. If B represents a very small number of counts a Poisson distribution should replace the Gaussian form.

In the situation at hand $P(B)$ concerns one particular experimental observation, having one particular value, *i.e.* the one observed. Thus $P(B)$ is a delta function and serves in (21) merely as a normalizing factor. $P(b)$ must express the prior knowledge ('experience') one can add to the evaluation of the data at hand. When a background value B is observed near a particular reflection in a $(\sin\theta)/\lambda$ interval with $N+1$ observations, the total background experience prior to the $(N+1)$th observation summarizes the available N data. Thus, it must account for $N\langle B\rangle$ counts and for a standard deviation $\{N\langle B\rangle\}^{1/2}$. For a single observation one has to divide by N and the prior distribution becomes

$$P(b) = \exp\,[-(b-\langle B\rangle)^2N/2\langle B\rangle]. \qquad (23)$$

Since prior and likelihood are normal distribution functions denoted by $P(b) = N(m_p, \sigma_p^2)$ and $P(B\,|\,b) = N(m_1, \sigma_1^2)$, the posterior $P(b\,|\,B) = N[b, \sigma^2(b)]$ is also a normal distribution function. The relevant posterior moments are

$$\frac{1}{\sigma^2(b)} = \frac{1}{\sigma_p^2} - \frac{1}{\sigma_1^2} \quad \text{and} \quad b = \frac{\sigma_p^2 m_1 + \sigma_1^2 m_p}{\sigma_p^2 + \sigma_1^2}.$$

When the number of observations N becomes large, *i.e.* when experience really counts, these moments reduce to

$$b = \langle b\,|\,B\rangle = \langle B\rangle \qquad (24)$$

$$\sigma^2(b) = \langle b^2\,|\,B\rangle = \langle B\rangle/N. \qquad (25)$$

Although (24) and (25) are the mathematical expressions of the well known worldly wisdom *experientia vincit*, they have not yet been implemented in X-ray diffractometry. On the simplest level one may do so by realising that a single observation of B counts is obviously a single draw out of the distribution of potential (allowed) observations $P(b\,|\,B)$. To improve the 'hit-and-run' quality of the single observation one may replace B and $\sigma^2(B)$ by the first and second moments of $P(b\,|\,B)$, *i.e.* by b and $\sigma^2(b)$. On a somewhat deeper level one may realise that the average over N observations depends only on a fraction $1/N$ of the last observation. Thus, the benefit of the last experiment decreases rapidly with N and

Table 5. *Comparison of* $\langle R\rangle$ *and* $\langle\sigma^2(R)\rangle$ *of prescan data with* $2\{\langle BR\rangle+\langle BL\rangle\}$ *and* $2q\{\langle\sigma^2(BR)\rangle+\langle\sigma^2(BL)\rangle\}$ *of fixed-time data*

$(\sin\theta)/\lambda$ (Å^{-1})	$\langle R\rangle$	$2\{\langle BR\rangle+\langle BL\rangle\}$	$\langle\sigma^2(R)\rangle$	$2q\{\langle\sigma^2(BR)\rangle+\langle\sigma^2(BL)\rangle\}$
0·16–0·31	62·5	66·6	20·8	22·8
0·31–0·39	57·7	58·4	19·2	20·3
0·39–0·45	50·7	50·4	16·9	16·5
0·45–0·49	39·0	42·6	13·0	13·4
0·49–0·53	35·5	36·2	11·8	12·0
0·53–0·56	31·9	33·0	10·6	10·8
0·56–0·59	30·5	31·4	10·2	9·6
0·59–0·62	29·1	29·8	9·7	9·6
0·62–0·65	29·3	29·8	8·9	9·6

becomes marginal at large N. The measuring strategy should be adapted to this if one wants to maintain a decent cost-to-benefit ratio. A successful strategy, however, combines caution with care. In our opinion, the standard preliminary scan, which includes the measurement of $R(hkl)$ and $B(hkl)$, should be the initial step. Then the local background observation $B(hkl)$ should be tested against the prior distribution $P(b)$, the knowledge of which could be obtained, for example, in the test phase when cell parameters, diffractometer settings *etc.* are determined. If $B-b<3(b)^{1/2}$ we have 99·7% probability that B is compatible with the expected b, *i.e.* B is not an outlier [(22)]. If so, we only have to remeasure $R(hkl)$ in the final scan and may omit the (re)measurement of $B(hkl)$. This may save up to one third of the final scan time. Or, alternatively, the full scan time may be used to increase the value of $R(hkl)$ by 30%. If B is identified as an outlier, the final scan should remeasure both R and B, which is today's default mode. In this connection we recall that background values measured during a prescan show an average bias of 10% due to the strategy, geared to save measuring time. This strategy-introduced bias is easily eliminated by replacing B and $\sigma^2(B)$ by b and $\sigma^2(b)$. The latter are strategy independent. Thus a time-saving strategy can be applied for its own purpose without distorting in any way the final result.

The splitting of $R(hkl)$ into its components

In the classic background–peak–background procedure the net intensity I and $\sigma^2(I)$ are calculated according to (8)–(9) (Table 1). One should realise that only R and $\sigma^2(R)$ are directly accessible from experiment and that BR, BL, σ^2(BR) and σ^2(BL) are merely estimates of the true B and $\sigma^2(B)$ which are needed. However, since we now accept the background everywhere within a certain $(\sin\theta)/\lambda$ interval follows counting statistics, we can say that an experimental value of B – if it were possible to perform the experiment – must lead to a result that follows a distribution centred around $\langle B\rangle$ and with a variance $\sigma^2(B)=\langle B\rangle$. The validity of this statement can be checked against the information available in Table 2. The prescan data form a select group of observations in which $I\simeq 0$. Thus, the actual averages $\langle R\rangle$ and variances $\langle\sigma^2(R)\rangle$ should be equal to $2\langle B\rangle$ and $2\langle\sigma^2(B)\rangle$, respectively, with values coming from outside the prescan set. Unbiased estimates of $2\langle B\rangle$ are $2\{\langle BL\rangle+\langle BR\rangle\}$, where $\langle$BR$\rangle$ and $\langle$BL$\rangle$ are taken from the fixed-time data. Similarly, unbiased estimates of $2\langle\sigma^2(B)\rangle$ are $2q\{\langle\sigma^2(BL)\rangle+\langle\sigma^2(BR)\rangle\}$, with $\langle\sigma^2$(BR) and $\langle\sigma^2$(BL)$\rangle$ taken from the fixed-time set and q representing the average scan speed of the fixed-time data divided by the average scan speed of the prescan data. Table 5 gives the comparison, showing an excellent agreement. In passing we note that $\langle R\rangle$ is always slightly smaller than $2\{\langle BR\rangle+\langle BL\rangle\}$, reflecting the fact that the prescan data set favours data with an underestimate of R.

Having established that $\sigma^2(B)=2\langle B\rangle$ and that R follows counting statistics [*i.e.* $\sigma^2(R)=R$], it follows that $\sigma^2(I)=I$. If we incorporate the uncertainty on the background average itself, $\sigma^2(I)$ increases slightly to $\sigma^2(I)=I+2\langle B\rangle/N$, a change which is only important for very small I. The approach allows us to infer values for $\sigma^2(R)$, $\sigma^2(I)$ and $\sigma^2(B)$ as if R, I and B could be separately measured.

Just as we were looking for the 'ideal' background b after having measured one particular B value, we will now look for r and i, the ideal values of R and I, respectively. In other words we are looking for the first moment of $P(r|R)$ and $P(i|I)$. Starting as before from Bayes's theorem we have

$$P(r|R)=P(R|r)P(r)/P(R). \qquad (26)$$

In contrast to the previous section $P(r)$ is now problematic. Under ordinary measuring conditions R is only measured once per reflection, so that the prior $P(r)$ cannot be constructed in a similar way to $P(b)$. For lack of better knowledge we represent $P(r)$ by a uniform density distribution and introduce the constraint that r should always be positive, *i.e.* $P(r)=0$ for $r<0$. In its present form [(26)] the properties of $P(r|R)$ are dictated by $P(R|r)$ because the single experiment should be and is more decisive than the (lack of) prior knowledge.

With background and net intensity as independent elements and omitting the normalizing δ function

Table 6. *Comparison of results of analyses performed on intensity data after a classic treatment and after a Bayesian treatment*

	Classical data	Bayesian data
Number of observations	1155	2518
Number of variables	270	270
$\sum \lvert F_{obs} - F_{calc} \rvert$	569	1483
$\sum \lvert F_{obs} \rvert$	12 398	15 186
$\sum w \lvert F_{obs} - F_{calc} \rvert^2$	3767	9410
$\sum w \lvert F_{obs} \rvert^2$	2 132 080	2 751 960
$\sum \lvert F_{obs} - F_{calc} \rvert / \sum \lvert F_{obs} \rvert$	0·046	0·102
$\{\sum w \lvert F_{obs} - F_{calc} \rvert^2 / \sum \lvert F_{obs} \rvert^2\}^{1/2}$	0·042	0·058
Maximum density in difference Fourier map (e Å^{-3})	0·22	0·28

$P(R)$, we can rewrite (26) as

$$P(r|R) \underset{R}{\simeq} P(2B|2b)P(2b)P(R-2B|i)P(i). \quad (27)$$

$P(2b)$ is the background prior while $P(2B|2b)$ and $P(R-2B|i)$ are counting statistical distributions. $P(i)$ is taken as a uniform density distribution for $i \geq 0$, while $P(i) = 0$ for $i < 0$ and $r = i + 2b$. Using these distributions in (27) and using only those potential observations for which $I = R - 2B$ one can calculate numerically $\langle i \rangle$ and $\sigma^2(i)$ as the best estimate of a net intensity and its variance. The calculations were executed using a preliminary version of the program *BAYES*. In the program the triple integration is performed by summing over the permitted values of r, b and B and takes into account the above-mentioned boundary conditions. As expected, values of $\langle i \rangle$ did not differ much from those of I, but $\sigma^2(i)$ values were considerably smaller than $\sigma^2(I)$ values predicted by counting statistics [(9), Table 1].

To check the consequences of the ideas outlined above, two independent least-squares analyses were performed. In the first we used the classic values B, I and $\sigma(I)$ of 1155 reflections for which $I \geq 3\sigma(I)$. In the second we used the Bayesian corrected values b, i and $\sigma(i)$ for 2518 reflections for which $i \geq 3\sigma(i)$. In both analyses the same full-matrix least-squares strategy (concerning decisions which atoms should be refined anisotropically *etc.*) was used throughout (for details see Lenstra *et al.*, 1991). Experimental structure factors were weighted individually according to $\sigma(I)$ or $\sigma(i)$, respectively. Since $\sigma(i) \ll \sigma(I)$, an additional 1363 reflections with small net intensities are included in the second analysis; the weighting scheme also changed drastically, putting more emphasis on the lower-intensity reflections. The results of these analyses are summarized in Table 6. One notes only a small increase in R_w (from 0·042 to 0·058) upon introduction of 1363 weak reflections, usually considered poorly determined. Unweighted R doubles (from 0·046 to 0·102), as can be expected. Addition of 1363 small $F(hkl)$ values (average 0·2 on an absolute scale) hardly changed $\sum F(\text{obs.})$, but affected $\sum [F(\text{obs.}) - F(\text{calc.})]$. The fact that $\sum [F(\text{obs.}) - F(\text{calc.})]$ changes almost linearly with the number of contributing reflections is another indication that the Bayesian treatment produces rather good intensity estimates. Furthermore, the maximum peak height in the difference Fourier map based on 1155 reflections is 0·22 e Å^{-3}, whereas it is 0·28 e Å^{-3} with 2518 reflections. Thus, the selective addition of weak intensities hardly influenced the final noise level. The doubling of the number of observations in the least-squares calculations produced a decrease in the e.s.d.'s of the atomic parameters. As expected, the new e.s.d. values were about a factor $2^{1/2}$ smaller. The maximum observed parameter (positional as well as anisotropic thermal parameters) difference between the two converged models amounted to three times the e.s.d. of the first model. Most differences were found to be smaller than 1 e.s.d. Using i and $\sigma(i)$ we found the maximum Δ/σ ratio to be 5 for the added weak data. With a background of about 150 counts ($\langle BR \rangle = \langle BL \rangle = 75$ counts) a reflection with $I = 5$ counts has in our philosophy a variance of $5 + 150/200 = 6$ counts, whereas in the standard BPB analysis $\sigma^2(I) = 150 + 5 + 4 \times 75 = 455$. With the latter estimate of variance *all* values Δ/σ would be smaller than 1, a statistically very unlikely event. We conclude that the new estimate $\langle \sigma(i) \rangle = 2{\cdot}4$ counts is more realistic than the previous $\langle \sigma(I) \rangle = 21$ counts. The use of individual $\sigma(i)$ and i thus leads to many more reflections meeting the criterion $i > 3\sigma(i)$ and thereby reduces the e.s.d.'s of both positional and vibrational parameters without producing any detectable artefacts in the analysis.

The authors gratefully acknowledge financial support by the Belgian National Science Foundation (KFGO) and National Lottery. This text presents in part research results of the Belgian Program on Interuniversity Attraction Poles initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. Scientific responsibility, however, remains with the authors.

References

De Wolf, M. & Lenstra, A. T. H. (1984). *Bull. Soc. Chim. Belg.* **93**, 437-444.

French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517-525.

Keulen, E. (1969). PhD thesis, Univ. Groningen, The Netherlands. (In Dutch.)

Lehmann, M. S. & Larsen, F. K. (1974). *Acta Cryst.* A**30**, 580-584.

Lenstra, A. T. H., Verbruggen, M., Bracke, B., Vanhouteghem, F., Reyniers, F. & Borremans, F. (1991). *Acta Cryst.* B**47**, 92-97.

Lindgren, B. W. (1976). *Statistical Theory*, 3rd. ed., Ch. 11. New York: MacMillan.

Schagen, J. D., Straver, L., Van Meurs, F. & Williams, G. (1988). *Operators Guide to Enraf-Nonius CAD-4 Diffractometer.* Enraf-Nonius, Delft, The Netherlands.