THEORY OF INFORMATION AS APPLIED TO ANALYTICAL CHEMISTRY. II.*

LOSS OF INFORMATION

K. ECKSCHLAGER

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, Prague 6

Dedicated to honour the memory of Professor J. Hanuš on the occasion of the centenary of his birthday.

Received May 22nd, 1970

The case is considered in which a part of information contained in the results of analyses is lost by working under other than optimum conditions, when using dependent quantitative analytical methods. As loss of information, reduction of the amount of information, which occurs without a simultaneous change in redundance, is here defined.

In the first paper[1], fundamental relationships for the determination of the amount of information, which we obtain by analysis, were presented and the effect of the precision of results and of the number of parallel determinations on the amount of the information obtained as well as on redundance was discussed. As far as in paper[1] the change in the amount of information was considered, always the case was concerned in which simultaneously the redundance was changed. Here, the case will be mentioned, in which during evaluation of the dependent determinations under different experimental arrangement or for a different way of assessing the same data, but always with the same redundance, various amounts of information are obtained from the analytical results. Since we always aim at obtaining as largest amount of information as possible, the difference between the information which we can obtain as maximum, and that really achieved under given precision and redundance of the analytical method, is considered as the loss of information.

In analytical practice which involves use of the dependent method, *i.e.* method whose result is evaluated by comparison with the standard, the case may occur that part of the information is lost, if another than optimum experimental arrangement is chosen, or if the results are evaluated in another way than is the most suitable one. It is necessary to mention here, of course, that by the idea of another than optimum experimental arrangement or evaluation, unproper evaluation or such experimental

* Part I: This Journal *36*, 3016 (1971).

arrangement which leads to incorrect results is in no way meant. The aim of the present paper is to point out the effects that are responsible for the loss of information.

RESULTS AND DISCUSSION

When using dependent analytical method, the loss of information may occur by deviation from the optimum working procedure. By the optimum experimental arrangement such a working procedure is meant which makes it possible to gain maximum information on the composition of the sample to be analyzed and leads, even for a small deviation from the composition, to accurate results. At the same time, this working procedure shall in no way be more complicated than necessary. The deviation from the optimum procedure provides equally accurate results on the average, which we should obtain by maintaining exact procedure, but precision and the loss of information are impaired in contrast to the case in which the optimum working procedure is exactly maintained. If we speak here about non-maintaining the optimum working procedure, we must take into account also the cases in which the optimum working procedure is unknown and the procedure taken from the literature is different from the optimum one.

The impairment of the precision of results causes that the reliability interval of the results for a certain level of significance is increased in contrast to the case, when the optimum working procedure is exactly maintained. Since we defined[1] the uncertainty after the experiment as the reliability interval for the significance level $\alpha = 0{\cdot}039$, which resulted from the comparison with the uncertainty calculated according to the Shannon relationship[2], it is obvious that such a deviation from the optimum working procedure, which impairs precision of the results, is followed by the increase of the uncertainty as well as by decrease of the amount of information obtained, even though it may not lead to the distortion of the results. This will be shown by us on examples evaluating the results of the dependent analytical method by direct comparison with the standard, standard addition method and the calibration curve method.

When evaluating the results of the dependent analytical method, *i.e.* of the method whose result we determine by comparison with a standard, we can proceed either by comparing with one or more standards; we do that, as a rule, graphically by construction of the calibration curve. For the present, we consider the linear dependence only between the y quantity measured and the concentration x determined. When comparing with one standard, we use either direct comparison with external standard, or apply the standard addition method (internal standard). In case of the direct comparison, we calculate unknown concentration x, using relationship

$$x = (y_x/y_0)\, c_0 \,, \qquad (1)$$

where y_x and y_0 are the values of the quantities measured (signals) on the sample to be analyzed and on the standard, respectively, and c_0 is the concentration of the standard. In the standard addition method, we can proceed, for example, by pipetting two equal aliquots of the sample, and to one

of them standard amount is admixed; the whole analytical procedure is accomplished with both aliquots and the result is then calculated according to (*1*) from the values found, but here $y_0 = y_{x+0} - y_x$, where y_{x+0} is the value of the measured quantity for the aliquot with the standard addition.

According to the law of accumulation of errors the variation coefficient of the result, which we obtain by comparison with one standard, is given by

$$v = \sqrt{(v_x^2 + v_0^2)} = \sqrt{[(s_y/y_x)^2 + (s_y/y_0)]^2} . \qquad (2)$$

Then the standard deviation of the result is given by

$$s = x \, . \, s_y \sqrt{[(1/y_x^2) + (1/y_0^2)]} . \qquad (3)$$

It is then possible to determine for various y_x and y_0 values as well as standard deviation s, the amount of information from the fundamental relation

$$I = \log_e (c_2 - c_1) \sqrt{n}/2st , \qquad (4)$$

where c_1 and c_2 are the limits presumed, in which content of the component to be determined is involved, n is the number of determinations, and t is the critical value of the Student distribution for $\alpha = 0{\cdot}039$. Provided that the standard deviation s_y was determined from a great number of the determinations, *i.e.* that $n_s \to \infty$ and y_x or y_0, or both, are equal or smaller than 10 (*i.e.* that full scale of the apparatus

TABLE I

Amount of Information in Comparison with Standard

Values measured		$x \, . \, s_y$			
y_x	y_0	0·01	0·10	0·20	1·00
10	5	9·29	6·98	6·29	4·69
10	8	9·62	7·32	6·63	5·02
10	9	9·69	7·39	6·70	5·09
10	10	9·74	7·44	6·74	5·14
9	10	9·69	7·39	6·70	5·09
8	10	9·62	7·32	6·63	5·02
5	10	9·29	6·98	6·29	4·69
9	9	9·64	7·34	6·64	5·04
8	8	9·52	7·22	6·53	4·92
5	5	9·05	6·75	6·06	4·63
8	5	9·24	6·95	6·24	4·63

to measure y equals $y_{max} = 10$) if $(c_2 - c_1) = 100$ and $n = 1$ and the amount of information I was calculated in natural units "nit" for various values $x \,.\, s_y$ and various y_x and y_0 values and their mutual ratio, then is the amount of information summarized in Table I. It is evident from this Table that the largest amount of information is obtained from such measurements in which $s_y x$ is the smallest one and $y_x = y_0$, and only when sensitivity of the apparatus is fully utilized, *i.e.* when $y_x - y_0 = y_{max}$. The loss of information is higher, the greater is the deviation from these optimum conditions, and is relatively higher, the greater is the product $x \,.\, s_y$.

Similarly, the determination of the loss of information for the deviation from normal conditions using the standard addition method may be accomplished. Since $s_{y_{x+0}} - y = s^2_{y_{x+0}} + s^2_{y_x}$, it holds then for $s_{y_x} = s_{y_0} = s_y$, *i.e.* for $s_{y_{x+0}} = 2 \,.\, s_y$ that

$$v = v^2_{y_x} + v^2_{y_{x+0}} - y_x = s_y \sqrt{\left[\left(\frac{s_y}{y_x}\right)^2 + \left(\frac{2s_y}{y_{x+0} - y_x}\right)^2\right]}. \qquad (5)$$

The standard deviation is then given by

$$s = x \,.\, s_y \sqrt{[1/y^2_x + 2/(y_{x+0} - y_x)^2]}\,. \qquad (6)$$

Several values of the amount of information in "nit" units are listed in Table II. Assuming that $y_{x+0} = 10$ and equal conditions as in the previous example are involved, it follows for the amount of information that the loss of information is higher, the greater is the deviation from optimum conditions. For lower $s_y \,.\, x$ value, however, the loss in question is relatively smaller. In contrast to the method of direct comparison with the standard, both absolute and particularly relative loss of information is greater in the standard addition method, because of the fact that with the latter method, determination of y_0 is accomplished from difference $y_{x+0} - y_x$, and is thus subject to a greater uncertainty. The standard addition method, however, is from practical point of view of great advantage in that it actually fully compensates the possible systematic error, due to different composition of the sample and standard to be analyzed, and to a great extent also the possible drift, *i.e.* time instability of the values found by the measurements.

When comparing with more than one standards, we construct, as a rule, the calibration curve for dependence $y = f(x)$. This dependence is very often linear, *i.e.*,

$$y = a + b \,.\, x\,. \qquad (7)$$

If we construct the calibration curve from m points, each of them corresponding to another concentration, x, of the component to be determined, and the actual determination is repeated n-times, the standard deviation is according to[3] given by

$$s = \frac{s_y}{b} \sqrt{\left(\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_A - \bar{y})^2}{b^2\, \Sigma(x_i - \bar{x})^2}\right)}, \qquad (8)$$

where $\bar{x}$ and $\bar{y}$ are mean values of the range of concentrations or of the quantities measured, in which the calibration curve is valid and the mean value found by the measurements is $\bar{y}_A = (1/n)\Sigma y_i$. The calibration curve slope according to (7), *i.e.* value b, is simultaneously a measure of the sensitivity of the determinations. It is obvious from relation (8) that for the case, for which $n = 1$, and $\bar{y}_A = \bar{y}$,

$$s = s_y/b \sqrt{[(m + 1)/m]} . \qquad (9)$$

For some values b, for the number of points from which the calibration curve had been constructed ($m = 2$, 3, 4, and 10), and for several values, s_y, of the standard deviation measurements, the amount of information for $y = 10$ is given in Table III. From here it follows that the larger amount of information is obtained, the greater is the number of points from which the curve had been constructed, even though the effect of m, starting from a higher value, *e.g.* from $m > 5$, is no more too distinct. Furthermore, larger amount of information will be gained, if the determination is more precise and sensitive, *i.e.* the greater is tangent b. According to relation (8), however, it depends also on the number, n, of parallel determinations and on difference $\bar{y}_A - \bar{y}$ which increases value s, lowering thus also the amount of information, the greater is the difference itself.

In order to be able to calculate the amount of information also in this case, we must choose a certain model in advance, according to which the calibration curve is constructed. Assuming that $\bar{y}_{max} = 10$ and for various m values, the model, for example, given in Table IV can be involved. The x_i values can be then easily calculated from relation (7).

If we assume that the calibration curve passes through origin of the co-ordinate system, $a = 0$, then $x_i = y_i/b$. Thereby all the values are known that are necessary to introduce relation (8), and the amount of information for some values n, $\bar{y}_A - \bar{y}$,

TABLE II

Amount of Information in the Standard Addition Method

y_{x+0}	y_x	$(y_{x+0} - x_x)$	$x \,.\, s$			
			0·1	0·10	0·20	1·00
10	2·5	7·5	8·61	6·31	5·62	4·01
10	3·0	7·0	8·73	6·43	5·74	4·13
10	3·5	6·5	8·82	6·55	5·82	4·25
10	4·0	6·0	8·86	6·55	5·86	4·25
10	4·5	5·5	8·87	6·57	5·87	4·27
10	5·0	5·0	8·85	6·55	5·86	4·25
10	5·5	4·5	8·80	6·50	5·81	4·20
10	6·0	4·0	8·73	6·43	5·73	4·12
8	3·6	4·4	8·64	6·34	5·65	4·04
8	2·0	6·0	8·09	5·78	5·09	3·48

and for standard deviation, s_y, of the measurement can be thus calculated. The values in natural units are presented in Table V. It is evident from Tables III and V that the greatest loss of information occurs when working in the extreme region of the calibration curve, with low sensitivity and low precision, in particular if the calibration curve had been constructed from a small number of points and if a small number of parallel determinations is carried out. After all, this follows from what has been reported by Doerffel and Hildebrand[3]. Moreover, we will note here, to what extent the appropriate effect is applied. The relative loss of information is small when making use of fairly high sensitivity and if the determination is rather precise. Then the fact that we do not work in the middle part of the calibration curve does not nearly become evident. Only during evaluation at the ends of the calibration curve, some loss of information takes place. Of no great effect is either a number of points from which the calibration curve is constructed, if $m \geqq 10$. Common analytical practice regards the number of points, $m = 5$, to construct the calibration curve as adequate; it must

TABLE III

Amount of Information in the Calibration Curve Method
$\bar{y}_A - \bar{y} = 0$, $n = 1$.

Sensitivity		S			
b	m	0·01	0·1	0·2	1·0
0·5	2	6·89	4·59	3·90	2·29
	3	6·96	4·65	3·96	2·35
	4	6·98	4·68	3·99	2·38
	5	7·01	4·70	4·01	2·40
	10	7·04	4·74	4·05	2·44
1·0	2	7·59	5·29	4·59	2·99
	3	7·65	5·35	4·65	3·04
	4	7·68	5·37	4·68	3·07
	5	7·70	5·40	4·70	3·10
	10	7·74	5·44	4·74	3·14
2·0	2	8·28	5·98	5·29	3·68
	3	8·34	6·04	5·35	3·73
	4	8·37	6·07	5·37	3·77
	5	8·39	6·09	5·40	3·79
	10	8·44	6·13	5·44	3·83
10	2	9·89	7·59	6·90	5·29
	3	9·95	7·65	6·96	5·35
	4	9·98	7·38	6·98	5·37
	5	10·00	7·70	7·00	5·40
	10	10·04	7·74	7·05	5·44

never be smaller than $m = 3$. Of somewhat greater effect is the number of parallel determinations; this is eventually apparent from Table VI in which for $b = 0{\cdot}5$, $\bar{y}_A - \bar{y} = 0$, $s_y = 0{\cdot}01$, the amount of information for various m and n values

TABLE IV
Calibration Curve Model for Various m

m	y_i				
2	1·0	9·0	—	—	—
3	1·0	5·0	9·0	—	—
4	1·0	4·0	6·0	9·0	—
5	1·0	2·5	5·0	7·5	9·0

TABLE VI
Amount of Information for Various m and n Values
$b = 0{\cdot}5, \bar{y}_A - \bar{y} = 0, s_y = 0{\cdot}01$.

n	m				
	2	3	4	5	10
1	6·80	6·96	6·98	7·01	7·05
2	7·44	7·53	7·59	7·62	7·70
3	7·74	7·85	7·91	7·96	8·06
4	7·94	8·03	8·14	8·19	8·32

TABLE V
Amount of Information in the Calibration Curve Method

$(\bar{y}_A-\bar{y})$	b	m	n	s 0·01	s 1·00	$(\bar{y}_A-\bar{y})$	b	m	n	s 0·01	s 1·00
0	0·5	3	1	6·96	2·35	1	0·5	3	1	6·94	2·24
			2	7·54	2·93				2	7·47	2·86
			3	7·85	3·25				3	7·77	3·16
		5	1	7·01	2·40			5	1	6·91	2·30
			2	7·63	3·02				2	7·67	2·96
			3	7·96	3·36				3	7·89	3·28
	2·0	3	1	8·34	3·73		2·0	5	1	8·39	3·79
			2	8·92	4·32				2	9·01	4·41
			3	9·24	4·63				3	9·35	4·74
		5	1	8·39	3·79	4	0·5	5	1	6·61	2·01
			2	9·01	4·41				2	7·07	2·46
			3	9·35	4·74				3	7·31	2·70
							2·0	5	1	8·12	3·51
									2	8·60	4·00
									3	8·86	4·25

can be seen. This is well apparent in the determination of a certain redundance, *i.e.* for $(m + n) = \text{const.}$, but for different ratios $m:n$. Even though the number of the points used to construct the calibration curve is $m \geqq 5$, more parallel determinations may still rather distinctly enhance the amount of information obtained. Here. of course, no loss of information can be spoken of, because also the redundance simultaneously changes with the change in the number of determinations and points used for the construction of the calibration curve. Of considerable effect is, of course, sensitivity characterized by value b of the tangent. The increase of sensitivity cannot be, of course, made without any limits; it is surely advantageous to use, for example, in photometry as long a measuring cell as possible, but in that case, the determinations over only a limited range of concentrations can be made, and in the measurement near the limit of determinability, an increase of variability of the blank experiment can appear unfavourable. When using the calibration curve, we must, of course, always be conscious of the possibility that a systematic error may be produced, due to the fact that the standards as well as samples are of different composition. When using certain calibration curve for a longer time, we should eliminate the drift, or at least limit its effect, by carrying out the calibration as often as possible. In the next example, we will note the effect of variability of the blank in terms of a source of the loss of information, in spite of the fact that problem of reading off the blank was discussed earlier[4] also from the viewpoint of the amount of information. As far as the non-zero value of the blank must be taken into account the standard deviation of the determination is given by

$$s_x = \sqrt{(s_{x_i}^2 + s_{x_0}^2)}, \qquad (10)$$

where s_{x_i} is the standard deviation of the actual determination and s_{x_0} that of the blank experiment. The amount of information for some different s_{x_i} and s_{x_0} values and for a number of parallel determinations of $n = 1$ to 5, is listed in Table VII.

TABLE VII

Amount of Information when Reading Off the Blank
$c_2 - c_1 = 100$, $n_s \to \infty$.

n	$s_{x_i} = 0{\cdot}01$				$s_{x_i} = 1{\cdot}0$			
	$s_{x_0} = 0$	0·005	0·01	0·05	$s_{x_0} = 0$	0·5	1·0	5·0
1	7·79	7·68	7·44	6·16	3·18	3·07	2·83	1·56
2	8·14	8·03	7·79	6·51	3·53	3·42	3·18	1·91
3	8·34	8·23	7·99	6·71	3·74	3·62	3·38	2·11
4	8·48	8·37	8·13	6·85	3·88	3·76	3·52	2·25
5	8·60	8·49	8·25	6·97	3·99	3·88	3·64	2·37

We assume here that $(c_2 - c_1) = 100$ and both standard deviations are determined from a sufficiently great number of determinations, *i.e.* that $n_s \to \infty$. Here, the loss of information can always be compared for equal number, n, of parallel determinations only, because the change in n is connected with the change in redundance. When reading off the blank, the relative loss of information is the smaller, the more precise are the results; here, it is not essential, whether higher precision was achieved because of the variability of the results being small or because of the fact that a greater number of parallel determinations is carried out. It is evident at the same time that the number of determinations is of no considerable effect. Maximum amount of information will be achieved only, if we are working with reagents of such a purity that reading off the blank is unnecessary[4].

In the conclusion, it may be quite generally outlined that the loss of information always occurs, when uncertainty of the results is inadequately increased. With dependent methods, the possibility for the loss of information is, as a rule, greater than in case of the independent methods in which the working procedure offers no such possibility for the deviation from the optimum procedure, which would lead to the precision impairment of the results.

REFERENCES

1. Eckschlager K.: This Journal *36*, 3016 (1971).
2. Shannon C. E.: *A Mathematical Theory of Communication.* The Bell System Technical Journal 1948.
3. Doerffel K., Hildebrand W.: Wiss. Z. der TH für Chemie „Carl Schorlemmer", Leuna-Merseburg *11*, 30 (1969).
4. Eckschlager K.: Chem. listy *61*, 272 (1967).

Translated by J. Hejduk.