AN ALGORITHM FOR THE COMPUTER EVALUATON OF QUANTITATIVE LABORATORY UNKNOWNS BASED ON ACCURACY AND PRECISION[1, 2]

EDWARD A. MOTTEL AND GILBERT GORDON, Departments of Chemistry, University of Iowa, Iowa City, Iowa 52242 and Miami University, Oxford, Ohio 45056

Abstract. An algorithm for evaluating student laboratory results is presented in which both proximity to the true result (accuracy) and the reproducibility of these results (precision) can be evaluated. The usefulness of such a scoring device and its pedagogical advantages are presented. The results are described in terms of a computer program LABGRADE which is available from the authors upon request.

OHIO J. SCI. 77(2): 63, 1977

Grading of experimental results will always be somewhat subjective in nature. In an attempt to make the evaluation of quantitative laboratory results less subjective, a general algorithm and computer program has been developed which can be used to evaluate and to assign a numerical grade for quantitative determinations. Unlike most evaluation procedures, the score is based both on the proximity of the results to the accepted value and the reproducibility of those results.

The results of student determinations of laboratory unknowns, like all scientific measurements, should fulfill the concepts of the scientific principle; that is, they should be close to the *true* or accepted value (accuracy), and reproducible (precision). In addition, it is also important that the person making the determinations, in this case a student, knows how to make any necessary numerical calculations.

In general, the student performs a specific experiment (e.g. a chemical acid-base titration) and records the raw data. In most laboratory situations the student can make the required numerical calculations, or alternatively, the raw data can be entered into a data reduction program, and the values appropriate to the student's unknown experimental result can be computer calculated. Many such programs have been reported in the chemical literature (Altenburg *et al*, 1968; Galyan and Ryan, 1972; Reiter and Budig, 1974). Comparison of the student calculated value to the computer calculated value has many pedagogical advantages and it helps to insure that the student has correctly made the required calculations. In some cases, authors have reported programs in which a statistical analysis was applied to the raw data (Rosenstein and Smith, 1962; Smith *et al*, 1965; Wellman, 1970; Wise, 1972), but only a few authors have noted the importance of precision to the scientific experiment (Wartell and Hurlbut, 1972; Jones and Lytle, 1973; Klatt and Sheafer, 1974, and none to our knowledge have included this as an integral part of the grading scheme. Either the student or computer calculated values can be used with the computer program reported here. This program called LABGRADE is written in FORTRAN IV and requires 40K bytes in addition to approximately 1000 bytes for every seven students. A class of 400 students requires 96K bytes. A sample printout and program listing is available from the authors. The input for LABGRADE consists of calculated values, rather than raw data, and thus LABGRADE is not meant to replace the computer calculation pro-

[1]Manuscript received April 5, 1976 and in revised form August 9, 1976 (#76-35).

[2]This paper was presented in part at the 9th Great Lakes ACS Regional Meeting, June 5, 1975, St. Paul, and at the 1st Chemical Congress of the North American Continent, December 2, 1975, Mexico City, Mexico.

grams, but rather to be used in conjunction with them. In view of the growing need for carefully documented programs (Hoffman, 1975) which meet specific teaching needs, we have developed both the basic algorithm and the LABGRADE program.

STATISTICAL ANALYSIS

LABGRADE is an iterative program which calculates the arithmetic mean and the standard deviation (σ) for all student generated results which are supplied as input data. Any results which lie outside of the 3σ range are excluded, since a result which has only normal uncertainties will occur within 3σ of the arithmetic mean 99.7% of the time, and a new arithmetic mean and standard deviation are computed. This procedure is repeated, until all remaining results fall within three standard deviations of the *true* value or the class mean. A graph of the reported concentrations of NaOH, which were obtained from a potassium acid phthalate titration, plotted as a function of the frequency with which those concentrations were reported (fig. 1) shows that the student-reported values for a large typical class have approximately a normal distribution. Graphically, in the output from the LABGRADE program, a normal curve with the same standard deviation as that calculated from the student results is plotted centering about the *true* value (if it is known) or the class average if the accepted value is not supplied as an input value.

Each determination the student reported is assigned a variable number of points for accuracy and, if the student has made multiple determinations, a separate number of points is assigned for precision. The number of points assigned to accuracy for each determination corresponds to the height of the normal curve at that given abscissa as is shown in figure 1. A linear scale is established with the maximum point value occurring at the *true* or accepted value (i.e. 0.0σ) and a zero point value is established at 3σ. For example, the first determination reported by Student A,

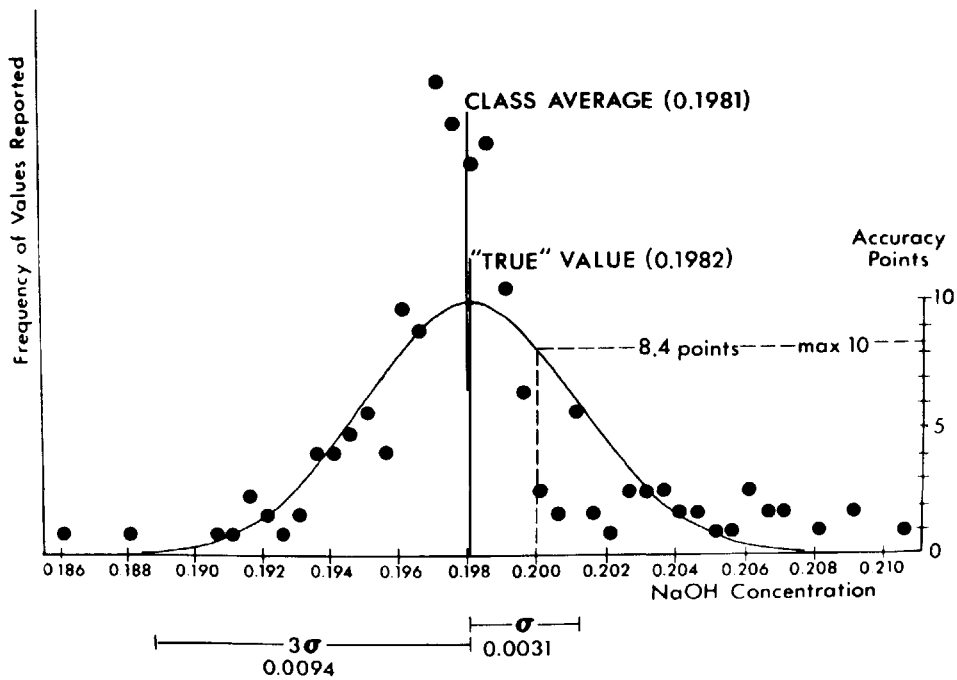

FIGURE 1. Assignment of Accuracy Points (Frequency of Values Reported as Function of Values Reported). Solid line corresponds to a normal distribution centered on the *true* value. Circles denote student results.

corresponding to the result noted by the dotted line in figure 1 was a NaOH concentration of 0.2000 M. Based on the class results this determination is worth 8.4 accuracy points out of a maximum score of 10 points (chosen for this example) as can be seen in figure 1. The same student also reported two additional determinations as shown in table 1.

TABLE 1

Reported results and evaluation of student A

Reported Conc. ± SD	Accuracy Points*	Precision Points**	Total Points
0.2000 ± 0.0010	8.4	4.8	13.2
0.2024 ± 0.0014	4.1	4.5	8.6
0.2007 ± 0.0003	7.3	5.0	12.3
Mean 0.2010			
Skew Pts.			5.0
Total Score (maximum = 50)			39.1

*10 points maximum per determination for accuracy.
**5 points maximum per determination for precision.

Precision points can only be assigned if the student has made multiple determinations, since the term precision implies reproducibility of results. If the student has made multiple determinations on the same system, the average value for that student is calculated and the deviation of each result from the individual student's arithmetic mean is determined. The number of precision points assigned to a given result is obtained by comparing the deviation of that result with the standard deviation for the whole class. Graphically, this is equivalent to centering the normal curve about the individual student's arithmetic mean, and obtaining the precision point value from the height of the curve corresponding to each determination as illustrated in figure 2. Again the maximum point value will occur when the deviation of an individual determination is zero, and no point value will result if the deviation for that determination is larger than 3σ.

An optional base or skewing score, which corresponds to the minimum number of points credit the student receives for just doing the experiment may also be employed. Thus, for three titrations the previous student would obtain a score of 39.1, if 5.0 points were assigned for simply doing a minimum of three titrations. In other words, the skew score is used to differentiate between the student that carried out the experiment and obtained

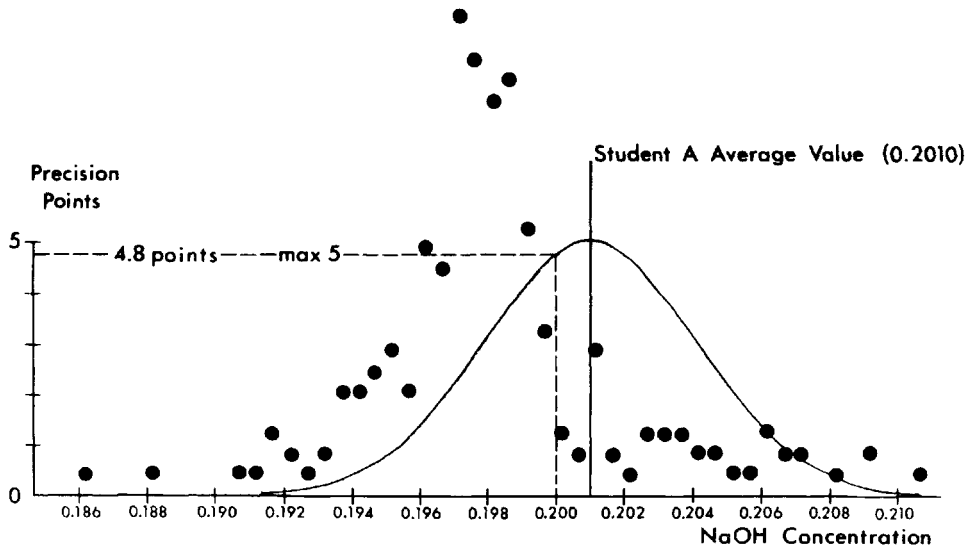

FIGURE 2. Assignment of Precision Points (Frequency of Values Reported as a Function of Values Reported). Solid line corresponds to a normal distribution centered on Student A's average value. Circles denote student results.

poor results and the student who failed to report the required determinations.

Prior to each computer run, the instructor determines the maximum number of accuracy points and the maximum number of precision points (assuming more than one result is submitted by the student) for each determination reported by the student. At this point, the instructor also decides if the student should be assigned any base points and the maximum and/or minimum number of individual student results to be used in computation of the student's total score.

DIAGNOSTIC CAPABILITIES

Perhaps one of the most helpful and often overlooked advantages of a program such as LABGRADE is its diagnostic capabilities. For example, if the class average differs significantly from the *true* or accepted value, then it is likely there is: an error in the instructions, the procedure or some systematic error; a contaminated reagent; or an incorrectly determined *true* value. In any of these cases it would be important to be able to reconsider the details of the grading procedure used in this experiment.

A second diagnostic aid might correspond to the observation of a bimodal distribution of values as is shown in figure 3. This could arise from a contaminated reagent being used in one or more of the laboratories. A possible solution to this problem would be to grade each of the laboratories separately, using the class average for the *true* value in those laboratories suspected of having contaminated reagents.

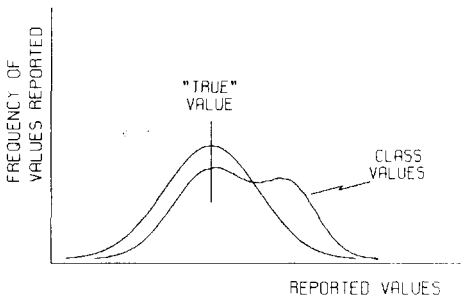

FIGURE 3. Expected Class Distribution and Example of Bimodal Class Distribution.

Finally, it should be pointed out that, when the class is considered as a group, the class average comes remarkably close to the accepted value once the statistically insignificant determinations are discarded. As can be seen in table 2, after seven iterations the class average becomes virtually identical to the *true* value. After 10 iterations, all remaining determinations fall within 3σ of the accepted values and the class average differs from the accepted value by only 0.0001 molar units, or one part per two thousand (0.05%). Thus a large number of class determinations, even from a class that is just learning to titrate, can be a sensitive method to check the *true* value which is often blindly accepted without any additional evaluation. For titration experiments, our *true* values were taken as the average of three determinations made by one of the laboratory teaching staff immediately preceding each laboratory. For other experiments such as the determination of the equivalent weight of magnesium, the accepted values were obtained from handbooks or other reference materials.

TABLE 2

Determinations used in successive iterations for a typical set of student determinations

Standard Deviation	*True* or Accepted Value	Class Average	No. Detn. Statistically Retained
0.6973	0.1982	0.2955	249
0.0386	0.1982	0.2083	244
0.0216	0.1982	0.2040	236
0.0127	0.1982	0.2009	228
0.0088	0.1982	0.1992	219
0.0058	0.1982	0.1985	211
0.0040	0.1982	0.1983	207
0.0035	0.1982	0.1982	204
0.0033	0.1982	0.1981	202
0.0031	0.1982	0.1981	202

In comparing the results from the same experiment among different laboratories, we have observed that the magnitude of the standard deviation is constant if no changes in procedure, reactants or scoring formula are made. For example, a comparison of three different laboratories doing the same experiment gave standard deviations of 2.711×10^{-3},

2.013×10^{-3} and 3.134×10^{-3}. In all cases, the class average refined to within 0.0001 molar units of the accepted value, even when the accepted value was not input. For small laboratories (less than 100 total determinations) it is preferable to input the *true* value, because the sample size does not seem to be large enough. For large laboratories (>200 determinations) the sample size appears to be large enough and either the accepted value or class average properly refine to indistinguishable values. As a result it should be possible to provide instantaneous grading for schools with on-line computer facilities, in addition to the possibility of comparing classes from year to year. Class comparison may aid those laboratories which have relatively low enrollments, and which, as a result, may suffer from poor statistics. A detailed comparison of the standard deviations and the proximity of the class average to the accepted value should also provide a mechanism to measure the effects of changing the wording of a procedure, the procedure itself, or some other part of the experiment such as the descriptions, generally improved procedures or better indicators which frequently result in smaller standard deviations and more accurate results.

Acknowledgments. We would like to thank Kathy Dittemore for her aid in the initial development of our series of computer programs and Kenneth Sando, Dwight C. Tardy and John R. Doyle for their helpful comments. We would also like to acknowledge The University of Iowa Graduate College for computer funds to develop this program, and The University of Iowa Computer Center staff for their assistance.

LITERATURE CITED

Altenburg, J. F., L. A. King and C. Campbell. 1968. Computer grading of general chemistry laboratory reports. J. Chem. Educ. 45: 615–616.

Galyan, R. H. and V. A. Ryan. 1972. Computer aided undergraduate radio-chemistry instruction utilizing time sharing terminals. J. Chem. Educ. 49: 591.

Hoffman, A. A. J. (ed.) 1975. Proceedings of the 1975 Conference on computers in the undergraduate curricula (CCUC/6), Fort Worth, Texas.

Jones, D. E. and F. E. Lytle. 1973. Computer aided grading of quantitative unknowns. J. Chem. Educ. 50: 285–286.

Klatt, L. N. and J. C. Sheafer. 1974. Changing student attitudes about quantitative analysis laboratory. J. Chem. Educ. 51: 239–242.

Reiter, R. C. and J. F. Budig. 1974. A time-sharing computer application to general chemistry laboratory work. J. Chem. Educ. 51: 43.

Rosenstein, R. D. and S. R. Smith. 1962. Using computing machines to grade student analysis reports. J. Chem. Educ. 39: 620–621.

Smith, S. R., R. Schor and P. C. Donohue. 1965. Computer grading of small numbers of laboratory unknowns. J. Chem. Educ. 42: 224–225.

Wartell, M. A. and J. A. Hurlbut. 1972. A Fortran IV program for grading quantitative analysis unknowns. J. Chem. Educ. 49: 508.

Wellman, K. M. 1970. Computer grading of introductory organic chemistry laboratory results. J. Chem. Educ. 47: 142.

Wise, G. 1972. Computer programs for undergraduate physical chemistry students. J. Chem. Educ. 49: 559–560.