

REFEREED PAPER

**THE SMRI INTERLABORATORY PROFICIENCY SCHEME -
REPEATABILITY, REPRODUCIBILITY AND MORE**

WALFORD SN AND PILLAY V

*Sugar Milling Research Institute NPC, c/o University of KwaZulu-Natal, Durban, 4041, South Africa
swalford@smri.org vpillay@smri.org*

Abstract

The Sugar Milling Research Institute NPC (SMRI) has for many years organised an inter-laboratory testing scheme, the results of which were used to determine the ‘Best Laboratory’ within the industry annually. Participating mill laboratories were supplied with Very High Polarisation (VHP) sugar and composite final molasses samples previously analysed at the SMRI and, in the case of VHP sugar, the South African Sugar Association Sugar Terminal laboratory. The samples were analysed at the mill laboratories and the results compared against the “assigned” values. Test scores of one or zero were assigned depending on whether the results were within or outside of industry accepted tolerances. The major drawback with this scheme is that binary scoring does not provide the participants with a system that can be used to monitor the ongoing performance of the laboratory.

The interlaboratory scheme was modified and a proficiency scheme was introduced in the 2015/16 season based on a z-score system. The system is based on a target standard deviation and the difference between the participant’s result (x) and an “assigned” analyte value (x_a). The z-score system was used to provide continuous quality feedback throughout the season. The overall performance of a laboratory for the season was monitored using the rescaled sum of z-scores (RSZ). The ultimate aim of the proficiency scheme is to improve and maintain the quality of analytical results. These results provide the SMRI with the necessary information to establish analytical precision values that can be fed back into the proficiency scheme.

Keywords: analysis, proficiency scheme, reproducibility, repeatability

Introduction

The determination of correct results for the analysis of Very High Pol (VHP) sugar and composite final molasses samples by the sugar mill laboratories is critical for process control and factory balance calculation. The use of these figures in penalty/bonus schemes also demands trust in the analysis results. Fundamental to the achieving of correct results is the use of a well-defined and written analysis method that is logical and easy to follow. The methods described in the South African Sugar Technologists’ Association (SASTA) Laboratory Manual (Anon, 2009) are followed by mill laboratories in the South African industry. The use of common methods allows the direct comparison of results from products of different mills. Consistent quality in result generation requires maintenance of minimum standards within the laboratory such as continuous staff training, maintenance of equipment and the use of quality chemicals and glassware. Quality also requires constant monitoring of laboratory processes to ensure that analytical results are correct.

Interlaboratory Scheme History

Emphasis on the quality of analytical results within the South African sugar industry has gained importance over the past 40 years. The critical relevance of a quality scheme at the South African Sugar Terminal (SAST) was recognised by Radford (1972) who introduced a statistical quality control system to improve the reliability of analyses. The SMRI, SAST and Hulett's Research and Development laboratories carried out a programme of analyses to investigate a system of quality control for raw sugar and final molasses analysis (Mellet *et al*, 1982). The results of this investigation led to the implementation of a scheme to compare the mill laboratory pol and brix results of weekly composite final molasses samples with those recorded by SMRI (Mellet *et al*, 1990). It was found that most mill laboratories showed large analysis discrepancies when compared with SMRI results and the authors concluded that the differences appreciably affected the overall season balance.

In an effort to improve the quality of the mill analytical performance, the SMRI offered laboratory audits and, since 1996, instituted an annual Best Laboratory award which included a floating trophy (Anon, 1996). The laboratory award was based on the differences of the mill results compared to SMRI analyte values for molasses and VHP sugar samples. Test scores of one or zero were assigned depending on whether the results were within or outside of industry accepted 'tolerances' based on the results of Mellet *et al* (1982). The major drawback with this scheme is that binary scoring only provides an indication that the result is acceptable or not. It does not provide the participants with a scoring system that can be used to monitor the laboratory performance on an ongoing basis and react to systematic trends and bias.

The SMRI was awarded ISO 9001 laboratory accreditation in 1998, which was upgraded to ISO 17025 accreditation in 2006. A key pillar in the 17025 quality system is the requirement that the accredited laboratory actively participates in appropriate proficiency testing schemes. Benefits of participation include an independent assessment of performance, being able to monitor the quality of the measurements made and make appropriate improvements, demonstrate competency to customers and identify staff development and training needs. The SMRI has participated in the LGC SUPS (Sugar Proficiency Scheme) for five years. The scheme is run in association with the International Commission for Uniform Methods of Sugar Analysis (ICUMSA) and the results are used to continually monitor and improve the SMRI quality system.

Proficiency testing is an ideal mechanism to give mill laboratory staff the ability to both monitor and improve their analytical capabilities. A trial of a possible scheme was proposed in 2014 and reported by Walford and Pillay (2014) and was implemented in the 2015/16 season.

Repeatability and reproducibility

Fundamental to any analytical analysis is the need to have consistent terminology that is understood by all users. *Accuracy* is often used when *trueness* is meant, *error* used instead of *uncertainty*, and *precision* used when *repeatability or reproducibility* is meant. A brief explanation of these terms is given in Table 1.

Table 1. Explanation of quality terminology

Term	Explanation
Accuracy	The closeness of agreement between a single test result and the accepted reference value. When applied to a set of test results, it involves a combination of random components and a common systematic error or bias component.
Trueness	Closeness of agreement between the average values obtained from a large series of test results and an accepted reference value.
Error	The result of a measurement minus the true value of the measurand.
Uncertainty	An estimate attached to a test result which characterises the range of values within which the true value is asserted to lie. In routine analysis, the uncertainty associated with an individual result will usually be estimated from prior studies such as method validation studies.
Precision	Precision is a general term for the closeness of agreement between independent test results obtained under stipulated conditions. Two measures of precision are in common use: <i>repeatability standard deviation</i> and <i>reproducibility standard deviation</i> .

Precision is based on the dispersion of a set of data from its mean value, depends only on the distribution of random errors, and does not relate to the true value or the specified value. The precision result depends on the conditions of measurement and therefore these conditions must be specified when referring to an estimate of the precision reported. Lower precision is reflected by a larger standard deviation. The most commonly reported measure of precision is the standard deviation which is made up of two measures defined as follows:

- Repeatability standard deviation – determined under conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time (Anon, 2006a). This measure is often referred to as repeatability and abbreviated “r”.
- Reproducibility standard deviation - determined under conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment (Anon, 2006b). This measure is often referred to as reproducibility and abbreviated “R”.

The standard ISO 5725 additionally discusses an intermediate repeatability standard deviation where independent test results are obtained with the same method on identical test items in the same laboratory but by different operators using the same or different equipment over longer intervals of time (Anon, 1994). Repeatability and reproducibility conditions involve repeat execution of the **entire method** from the point at which the test portion is taken from the laboratory sample, and not just repeat instrumental determinations on prepared extracts. Based on these definitions, it is clear that the repeatability standard deviation will be smaller than the reproducibility standard deviation as the variance and randomness will increase when other laboratories are included. It was on this point that the original SMRI interlaboratory scheme was inadequate in that the repeatability values determined at the SMRI (*i.e.* within one laboratory) were used for the ‘tolerance values’ to bracket the SMRI target values for multiple laboratories. The variances in the analyses of the same sample in many different laboratories

will be greater than the variance when analysed in only one laboratory. It is for this reason that an interlaboratory proficiency scheme was introduced.

Proficiency testing

Proficiency testing comprises an interlaboratory system for the regular testing of the accuracy that the participant laboratories can achieve. The primary purpose of proficiency testing is to help laboratories detect and cure any unacceptably large inaccuracies in their reported results and is therefore designed as a self-help system to alert the participants whether they need to improve their analytical practice. These discrepancies could result from malfunctioning or poorly calibrated equipment, possible poor analyst performance or reagent deterioration. The format of proficiency schemes is described in several international guidelines and standards including the International Union of Pure and Applied Chemistry (IUPAC) protocol (Thompson *et al*, 2006), and the International Standard Organisation (ISO/IEC) guideline (Anon, 1997). The SMRI scheme follows the basic tenets of these guidelines.

The advantages of using a performance scoring system are:

- Results can be expressed in a form that is relatively easy to interpret and understand;
- Results can be summarised in graphical or tabular form to depict overall performance;
- A performance score allows participants to directly compare their own results with others; and
- Provided consistent statistical values are applied, the performance score can be used by participants to monitor and trend their own performances over time.

During the crushing season, the SMRI distributes up to seven different sets of identical subsamples of VHP raw sugar and final molasses to each participating mill laboratory. The mill laboratory personnel analyse the VHP raw sugar for colour, pol and moisture and the final molasses for pol and brix using SASTA Laboratory Methods. The SMRI Analytical laboratory also analyses the samples as unknown blind samples. Results are returned to the SMRI and used to evaluate the accuracy of the individual laboratories using a z-score system.

The majority of scoring systems in proficiency testing schemes compare the difference between the participants' result (x) and an assigned value (x_a) with a quality target (Ellison, *et al*, 2009; Thompson, *et al*, 2006). The quality target is usually expressed as a standard deviation (often called the standard deviation for proficiency assessment) and denoted ρ_p . Each scoring system has acceptability criteria to allow participants to evaluate their performance. The most common scoring system is the z-score. The z-score for a proficiency test result is calculated using equation 1.

$$z = \frac{x - x_a}{\rho_p} \quad (1)$$

The standard deviation (ρ_p) can be prescribed (e.g. legislative), by results from a reproducibility experiment, from a general model or the participants' results themselves and should be at a level that is accepted by the participants and can be regarded as 'fit-for-purpose'. The numerical value of the parameter should be such that the resultant z-scores can be interpreted by reference to the standard normal distribution. The primary idea of the z-score is to make all

proficiency test scores comparable so that the significance of a score is immediately apparent no matter what the identity of the test material or analyte concentration is.

The assigned value (x_a) can be determined by making use of a reference laboratory, directly using a certified reference material, by comparison of the proficiency test material with a certified reference material, by formulation using standards or by consensus of participants' results. The latter is the most widely used method for determining the assigned value x_a and was used in the 2015/16 SMRI scheme. All the test results were checked for outliers using the Grubbs test (Anon, 1995) and inappropriate test results were removed and the mean of the remaining results used as the assigned value x_a . The standard deviation of the range of results was used for the proficiency assessment p_p value. An example of a set of results based on this approach is shown in Table 2.

Table 2. Example of the results from the pol analysis of one sample of a raw sugar sent to participating laboratories (colour coded blocks explained in the text)

Lab ID	1st check		2nd check	
	Pol	Grubbs Z Score	Pol	z-score
1	99.47	0.50	99.47	0.28
2	99.44	0.28	99.44	0.07
3	99.00	2.84	99.00	-3.05
4	NR			
5	99.30	0.71	99.3	-0.92
6	99.58	1.28	99.58	1.06
7	99.58	1.28	99.58	1.06
8	99.31	0.64	99.31	-0.85
9	NR			
10	99.39	0.07	99.39	-0.28
11	99.44	0.28	99.44	0.07
12	99.43	0.21	99.43	0.00
13	99.54	0.99	99.54	0.78
14	99.48	0.57	99.48	0.35
15	99.31	0.64	99.31	-0.85
16	99.40	0.00	99.4	-0.21
17	99.41	0.07	99.41	-0.14
18	99.30	0.71	99.3	-0.92
19	NR			
Mean	99.40		99.43	
Std Dev	0.14		0.095	
Z critical	2.59		2.55	

NR = samples with no results

The mean and standard deviation of the 1st check results are determined and the Grubbs Z score calculated. The Grubbs Z score is compared to Z critical and if greater, the associated result is removed from the data for the determination of the recalculated mean and standard deviation.

The proficiency z-score is then determined for each sample result (right hand column in Table 2). The z-score is interpreted as shown in Table 3. It is assumed that the z-scores follow a standard normal distribution and the following interpretation can be made:

- Approximately 95% of all z-scores will fall between -2 and $+2$;
- A score outside the range of -3 or $+3$ would be unusual and it is taken to indicate that the cause of the event should be investigated and remedied; and
- Scores in the range of -2 to -3 and $+2$ to $+3$ would be expected about one time in 20 so an isolated event of this kind is not a 'big deal'. However, successive scores in this range would require investigation.

Table 3. Interpretation of the z-score

$ z \leq 2$	Satisfactory performance
$2 < z \leq 3$	Questionable performance
$ z > 3$	Unsatisfactory performance

In the example shown in Table 2, only Laboratory 3 has an unsatisfactory performance (z-score greater than three) with the remainder having a satisfactory performance. Tabular information in Table 2 showing the results of each round is provided to all the participating laboratories by the SMRI at the end of each round.

The power of the proficiency scheme lies in the continuous monitoring of the z-score for any particular analyte analysis. The laboratory can record and interpret the accumulative z-scores for a particular analysis in the form of a Shewhart control chart (Ellison *et al*, 2009). The z-score is determined from the reproducibility estimates for the complete analytical method and thus represents a quantifiable statistical range of acceptable deviation. Thus, in most participant laboratories the run-to-run standard deviation for the analysis will normally be below the proficiency standard deviation with most of the scores falling within acceptable limits. A z-score outside the range ± 3 would occur very rarely and if it did occur, it would be more reasonable to suppose that the analytical system had produced a serious bias than a very unusual random error. Such an occurrence would demonstrate that the laboratory needed to take some kind of remedial action to eliminate the problem. Two successive z-scores falling between two and three (or between -2 and -3) could be interpreted in the same way. In fact all of the normal rules for interpreting the Shewhart chart such as the Westgard Rules (Thompson, *et al*, 2006) can apply to z-score charts. A typical z-score chart in the form of a Shewhart chart is shown in Figure 1. This shows an analysis method that is being followed correctly with the necessary instrumental and method checks in place. The individual analyte z-score charts are provided to each participating laboratory with results of each round of the scheme during the season.

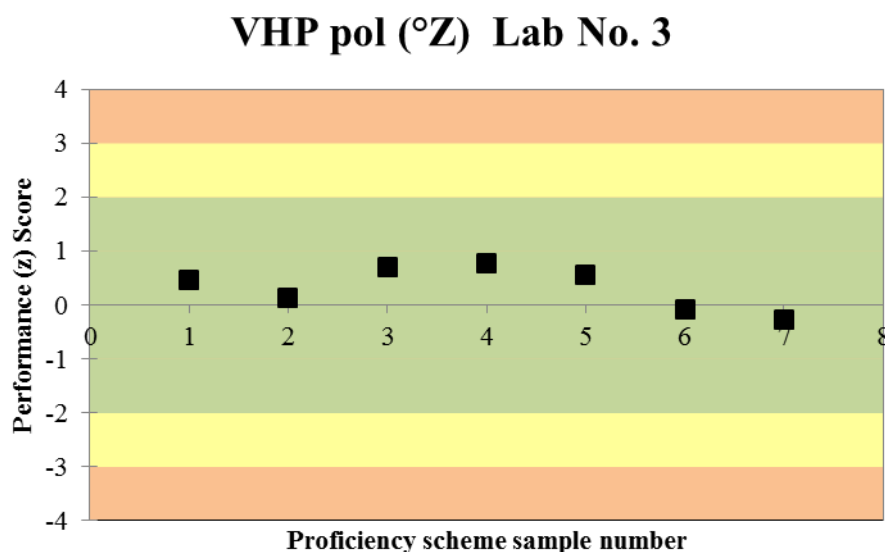


Figure 1. Example of a Shewhart control chart for pol analysis of VHP sugar across the 2015/16 season. Colours on the chart correspond to the interpretation in Table 3.

A poor z-score is indicative of a problem but is not diagnostic so further information is required to determine the origin of a poor result. Examples of possible actions include re-examination of the records for the analysis of the proficiency testing material, checking for systematic or sporadic mistakes in calculations or incorrect weights or volumes being used.

As mentioned, the original SMRI interlaboratory ‘Best Laboratory’ award was based on a total binary scoring system which is not ideal. Furthermore, the system did not give the laboratory staff feedback about the quality of their particular results. The proficiency testing scheme control chart provides the staff with an ideal feedback system regarding the quality of the results they produce. A number of monitoring systems can be used to assess the z-score over longer periods. Performance of a laboratory for a single test over time can be monitored using parameters such as the rescaled sum of z-scores (RSZ) or the sum of squared z-scores (SSZ) (Thompson, *et al*, 2006; Ellison *et al*, 2009).

The RSZ is determined using equation 2, and can be interpreted on the same basis as a single z-score.

$$RSZ = \sum_{i=1,n} \frac{z_i}{\sqrt{n}} \quad (2)$$

This statistic has the useful property of demonstrating a persistent bias or trend so that the sequence of z-score results [1.5, 1.5, 1.5, 1.5] provides a statistically significant RSZ of 3.0 even though no one single result is significant at the 95 % confidence level. However, RSZ could conceal two large z-scores of opposite sign that approximately cancel.

The SSZ statistic is determined using equation 3.

$$SSZ = \sum_{i=1,n} z_i^2 \quad (3)$$

This statistic has the advantage of avoiding the cancellation of large z-scores of opposite sign but is less sensitive to small biases. SSZ is especially sensitive to outliers such as single high z-scores.

A modified RSZ was used as the basis for reviewing the performance of each laboratory over the season. The absolute RSZ score for each analyte for each participating laboratory was determined and summed at the end of the season to provide a 'score' that could be used to rank the participating laboratories (Table 4).

Table 4. Example of the final RSZ scores achieved by the individual mills

Mill	Raw sugar			Molasses		Total
	Pol	Moisture	Colour	Pol	Brix	
1	1.14	1.64	0.80	0.78	0.90	5.26
2	0.94	1.64	1.14	1.19	0.57	5.48
3	1.77	1.93	1.11	1.50	1.34	7.65
4	1.34	1.56	1.71	2.18	1.98	8.77
5	2.96	2.12	2.37	1.30	1.60	10.36
6	2.57	2.03	2.20	2.21	1.39	10.39
7	1.65	4.06	1.09	1.64	2.04	10.49
8	2.21	2.89	3.04	1.69	1.19	11.01
9	2.12	3.07	2.07	3.34	0.96	11.57
10	2.24	3.44	2.47	2.36	1.46	11.97
11	5.60	2.02	4.90	0.77	2.61	15.90
12	2.89	1.96	3.20	3.24	5.89	17.17
19	2.97	3.07	5.13	4.33	3.04	18.54

Scheme improvements

In an ideal proficiency testing scheme the value used for ρ_p is determined by fitness for purpose and represents the amount of uncertainty in the result that is tolerable in relation to the purpose of the analysis. The system was tested over a two-year period and run as a scheme for the 2015/16 milling season. A fundamental shortcoming of the SMRI Proficiency Interlaboratory Scheme as described in this paper has been the use of the calculated standard deviation being used as the standard deviation for proficiency assessment (ρ_p). As was shown in equation 1, this value is used in the calculation of the z-score but varies from round to round (Table 5).

Table 5. Differences in the calculated standard deviation of the pol results for the VHP sugar for each round of the proficiency testing scheme

Proficiency Round	Standard Deviation
1	0.099
2	0.093
3	0.073
4	0.102
5	0.057
6	0.105
7	0.100

The values currently being used thus describes the end-user's requirements in each round, not the reproducibility and uncertainty of the method being used. The primary purpose of the proficiency test is to allow the participants to confirm that they are complying with already defined external criteria. To overcome this, the results collected during the 2015/16 proficiency scheme were used to determine suitable 'fit-for-purpose' ρ_p values which will remain constant throughout the season. Analyte standard deviation values were determined for each method based on the procedures of Youden and Steiner (1975) and are shown in Table 6. These values will be used throughout the 2016/17 season.

Table 6. Test method standard deviation for proficiency assessment values (ρ_p) determined from the 2015/16 season proficiency scheme results

Analyte	Standard deviation for proficiency assessment value
VHP pol	0.10
VHP moisture	0.02
VHP colour	85
Final molasses brix	0.5
Final molasses pol	0.3

Conclusion

The SMRI Interlaboratory Proficiency Scheme was successfully introduced in the 2015/16 crushing season. The scheme provides the mill laboratory staff with a method for regularly assessing the quality and accuracy of the results in their laboratories using z-scores. The results of the scheme can furthermore help laboratories detect and cure any unacceptably large inaccuracies in their reported results by monitoring the z-score charts for each of the analysis methods (VHP sugar pol, moisture and colour and composite final molasses pol and brix). The current VHP sugar and molasses proficiency testing scheme has been extended to include laboratories outside South Africa. A refined sugar proficiency scheme was developed and introduced during the 2016/17 season to include South African and African refineries that are members of the SMRI.

REFERENCES

- Anon (1994). ISO 5725-6. *Accuracy (trueness and precision) of measurement methods and results -- Part 6: Use in practice of accuracy values*. International Organization for Standardisation. Geneva.
- Anon (1995). IUPAC. Protocol for the design, conduct and interpretation of method performance studies. *Pure Appl Chem*. 67:331-343.
- Anon (1996). Sugar Milling Research Institute Annual Report 1995-1996. p18.
- Anon (1997). ISO/IEC Guide 43-1:1997. *Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes*. International Organization for Standardisation. Geneva.
- Anon (2006a). ISO 3534-2. *Statistics -- Vocabulary and symbols -- Part 2: Applied statistics. Section 3.15* International Organization for Standardisation. Geneva.
- Anon (2006b). ISO 3534-2. *Statistics -- Vocabulary and symbols -- Part 2: Applied statistics. Section 3.20* International Organization for Standardisation. Geneva.
- Anon (2009). SASTA Laboratory Manual including the Official Methods. 5th edition. Mount Edgecombe.
- Ellison SLR, Barwick VJ and Farrent JD (2009). *Practical Statistics for the Analytical Scientist. A Bench Guide*. RSC Publishing. 2nd edition, 268pp. Cambridge.
- Mellet P, Lionnet, GRE, Kimmerling ZJ and Bennett PJ (1982). Standards for the analytical precision of sugar and molasses analysis. *Proc S Afr Sug Technol Ass* 56:55-57.
- Mellet P, Ronaldson MA, Excell, TL and Koenig MFS (1990). Some questions and answers concerning analysis in factory laboratories. *Proc S Afr Sug Technol Ass* 64:189-194.
- Radford AJ (1972). Use of statistical quality control at the Durban Sugar Terminal. *Proc S Afr Sug Technol Ass* 46:133-137.
- Thompson M and Wood R (1995). Harmonized guidelines for internal quality control in analytical chemistry laboratories. *Pure Appl Chem* 67:649-666.
- Thompson M, Ellison SLR and Wood R (2006). International Harmonized protocol for the proficiency testing of analytical chemistry laboratories. *Pure Appl Chem* 78:145-196.
- Walford SN and Pillay V (2014). The z-score as a possible alternative measure for the SMRI inter-laboratory scoring system. *Proc S Afr Sug Technol Ass* 87:Poster.available on request.
- Youden WJ and Steiner EH (1975). *Statistical manual of the Association of Official Analytical Chemists*. AOAC, Gaithersburg, MD, USA.