Collaborative development of near infrared calibrations for quality testing of wheat and barley breeding material: 2. The use of indicator variables to correct for interlaboratory biases in reference data

Brian G. Osborne
BRI Australia Ltd, PO Box 7, North Ryde, NSW 1670, Australia.

Ian J. Wesley
Grain Quality Research Laboratory, CSIRO Plant Industry, PO Box 7, North Ryde, NSW 1670, Australia.

Introduction

The philosophy adopted in the quest for more robust calibrations for wheat and barley quality testing was described in the first paper in this series. In essence, sample sets derived from six different laboratories are to be used to create a common spectral database. The problem of merging data measured on different instruments was solved by instrument standardisation. The next issue to be addressed was interlaboratory biases in the reference data. One option would be for all the reference analyses to be carried out in one laboratory. However, this would involve sample shipment which has been avoided by instrument standardisation. Furthermore, each laboratory will need to monitor calibrations against its own laboratory data. Therefore, the problem is to merge datasets involving reference data measured in different laboratories.

Indicator variables were first introduced into near infared (NIR) spectroscopy by Don Burns[1] and are included as an option in the calibration section of recent versions of ISI software. They are "dummy" constituents which are used when there is reason to believe that the accuracy of a calibration may be influenced by one or more systematic factors in the data, such as reference data generated in different laboratories.

The function of indicator variables is simply to activate or deactivate additional offset terms a_i in the calibration equation:

$$Y = a_0 + \sum_{i=1}^{j-1} a_i IV_i + \sum_{i=1}^{k} b_i x_i$$

where Y is the value to be predicted, a_0 and b_i are the calibration constants, x_i are the spectral data and IV_i are the indicator variables. The indicator variable IV tags a block of data and allows it to have its own offset value equal to $a_0 + a_i$. j separate blocks would require $j - 1$ indicator variables. It is impor-

tant to note that *IV* is a qualitative term which is assigned the value 0 or 1. Thus, it only contributes to a_0 when it has the value 1. Indicator variables play no role in the fitting of the *b* terms or the selection of the wavelengths or PLS factors.

In order to provide a demonstration of how to use indicator variables and to test the effect of indicator variables in correcting for interlaboratory bias, a typical set of calibration samples was selected and changes introduced into the data in such a way as to simulate plausible biases for the constituent in question.

Materials and methods

Samples

341 samples of whole Australian wheat were collected from each of the wheat-growing states during the 1996 harvest. Protein was determined by the Kjeldahl method ($N \times 5.7$).

NIR spectroscopy

All spectra were measured using an NIRystems model 6500 spectrophotometer (NIRystems Inc., Silver Spring, MD, USA) in reflectance mode. Spectra were measured using a coarse sample cell and sample transport mechanism. All spectral data were recorded from 400 nm to 2498 nm at 2 nm intervals and saved as the average of 32 scans for each sample. A 4-point Fourier smoothing was applied to the data during collection.

Data treatment and analysis

All data analysis was conducted within Infrasoft International (ISI) software WINISI (NIRystems Inc., Silver Spring, MD, USA). The raw log 1/*R* data were corrected for the effects of scatter using standard normal variate (SNV) and detrend and transformed into second derivative $d^2(\log 1/R) / d\lambda^2$ using an 8-point (16 nm) gap and an 8-point (16 nm) smoothing function. All subsequent analysis was performed on the scatter corrected, second derivative data over the range 800–2500 nm. Modified partial least squares (PLS) was used to develop calibration equations.

Simulation of interlaboratory biases

Three groups were selected systematically by sorting the sample set in order of increasing protein content then divided into three groups of 114, 114 and 113 samples by writing every second sample to the Group 2 file and every third sample to the Group 3 file. The protein contents of Group 1 were unchanged while 0.3% was added to all the values in Group 2 and 0.5% was deducted from all the values in Group 3.

Table 1. Values of indicator variables for three factors.

	IV1	*IV2*
Group 1	0	0
Group 2	0	1
Group 3	1	0

Application of indicator variables

Indicator variables were added to the calibration file by entering an "I" instead of a number to distinguish them from constituent data. In this simulation, there were three groups which required two indicator variables:

Since Group 1 had no bias applied, it was designated as the reference group and both its *IV*s given the value 0. The calibration derived using indicator variables included a separate term for each group of data.

Results and discussion

The calibration obtained using the original data is shown in Figure 1 and the calibration statistics are given in Table 2. The effect of introduction of the simulated interlaboratory biases is to increase the *SECV* and decrease the r^2 (Figure 2, Table 2). The dotted lines in Figure 2 show the lines fitted to each

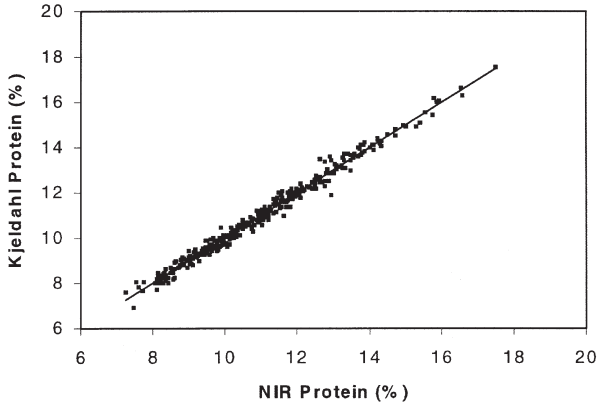

Figure 1. Plot of Kjeldahl vs NIR protein without interlaboratory biases.

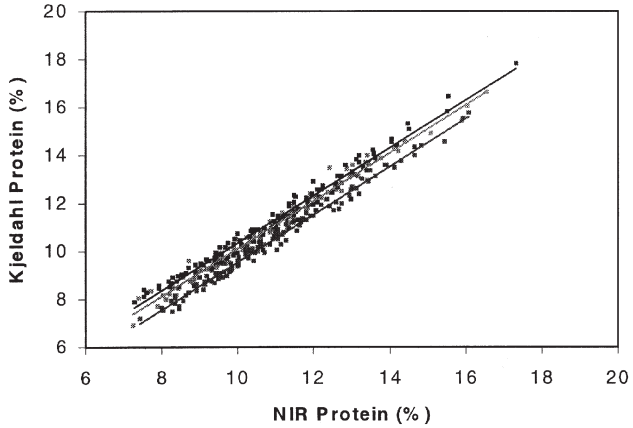

Figure 2. Kjeldahl vs NIR protein with interlaboratory biases.

Table 2. Calibration statistics for whole wheat protein.

Experiment	*SECV* %	R^2
No biases	0.22	0.99
Simulated biases	0.42	0.95
Biases corrected	0.22	0.99

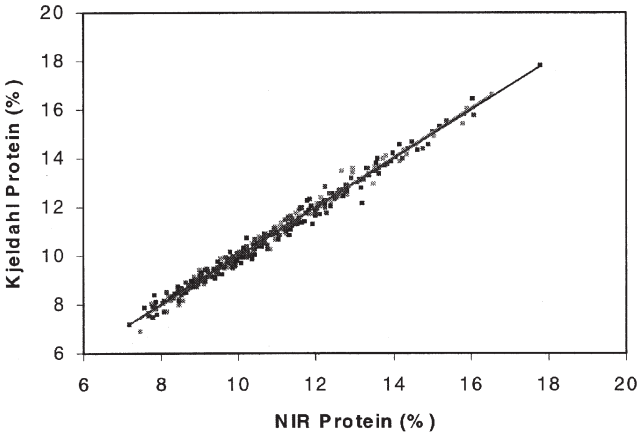

Figure 3. Kjeldahl vs NIR protein after correction of interlaboratory biases using indicator variables.

of the separate groups. Application of indicator variables resulted in complete removal of the effect of the biases (Figure 3, Table 2). Furthermore, the biases calculated by subtraction of each pair of a_i values were in close agreement with the known biases (Table 3). Therefore, the indicator variables exerted their anticipated effect in detecting and removing biases of known magnitude.

A recent example of the successful application of indicator variables in a practical situation has been in the NIR measurement of digestible energy content of cereals for growing pigs.[2] The high cost of the reference assay necessitated sourcing calibration samples, together with their associated reference data, from four different laboratories. In this application, the prediction sum of squares was al-

Table 3. Interlaboratory biases calculated using indicator variables with added biases in parentheses.

	Group 1	Group 2	Group 3
Group 1	0	—	—
Group 2	0.25 (0.3)	0	—
Group 3	–0.53 (–0.5)	–0.78 (–0.8)	0

most halved by using indicator variables to fit different offsets to the data from each laboratory. In addition, the values of the different offsets provided proof of the existence of interlaboratory biases and an estimate of their magnitudes.

Conclusion

Indicator variables enable calibrations to be developed using samples analysed in different laboratories. They both compensate for the effect of interlaboratory biases on the accuracy of the calibration and allow the biases to quantified without the need for a common set of samples to be analysed in the different laboratories.

References

1. D.A. Burns, in *Handbook of Near-Infrared Analysis*, Ed by D.A. Burns and E.W. Ciurczak. Marcel Dekker, New York, pp. 317–328 (1992).
2. R.J. van Barneveld, J.D. Nuttall, P.C. Flinn and B.G. Osborne, *J. Near Infrared Spectrosc.* **7,** 1 (1999).