

Near-Infrared Spectroscopy for Determination of Protein and Amylose in Rice Flour Through Use of Derivatives

Miryeong Sohn,^{1,2} Franklin E. Barton, II,¹ Anna M. McClung,³ and Elaine T. Champagne⁴

ABSTRACT

Cereal Chem. 81(3):341–344

The use of the derivative method for near-infrared (NIR) calibration was investigated to determine protein and amylose content in rice flour. Samples for two years, 1996 and 1999, were combined to give a wide range of the constituents for development of the calibration model. The NIR spectral data were transformed with Savitzky-Golay derivative with multiplicative scatter correction. To develop the best derivative models, the polynomial fits (quadratic, cubic, and quartic), convolution intervals (3–11 points for protein, 3–17 points for amylose), and derivative orders (1st derivative D1; 2nd derivative D2) were investigated. For the protein analysis, all polynomial fits with 3–11 points were acceptable to develop both the D1 and D2 models. However, the three-point quadratic and five-

point quartic fits were not acceptable for the D1 model, and the three-point quadratic fit was not acceptable for D2. For the amylose analysis, the D1 model produced generally better results than D2. Higher convolution intervals were required for the D2 model, whereas the D1 model was not affected by convolution intervals. A quadratic (or cubic) fit with 17-point convolution interval was acceptable for the amylose D2 model, and the quadratic fit with 5–11 points and cubic (or quartic) fit with 7–17 points were suitable for the D1 model. Based on the standard error of cross-validation (SECV), the calibration models developed using data for two years resulted in good precision with an SECV of 0.23% for protein using four factors and an SECV of 1.0% for amylose using 10 factors.

Some of the important factors related to the taste and texture of rice are degree of milling, moisture content, drying conditions, and chemical and physical properties. The property with the greatest effect on rice quality is the degree of milling, which significantly lowers protein content (Champagne et al 1997, 1998). The main variation in sensory texture attributes is related to amylose and protein content in rice (Windham et al 1997).

Near-infrared (NIR) spectroscopic techniques have been used as rapid and cost-effective analytical ways to assess rice quality. Recent studies have reported calibration models for determining amylose and protein content in rice. Delwiche et al (1995) developed NIR calibration models for amylose and protein analysis in ground milled rice. The results were good but a large number of factors was required, 18 for amylose and 16 for protein. Delwiche et al (1995) extended the application of NIR reflectance spectra to whole-grain milled rice samples. The results obtained were acceptable for both constituents, even though the standard error of the model was slightly higher compared with the ground samples (Delwiche et al 1996). Barton et al (1998, 2000) examined three types of spectroscopic techniques for rice quality and reported the optimal geometry for development of a NIR model. In these studies, the protein and amylose models had acceptable error levels but were developed using a sample set with a narrow range for protein and an uneven distribution of amylose content. Samples with a greater range and distribution for protein and better distribution of amylose content (particularly 0–10%) are required to adequately predict a more diverse population.

The derivative treatments have long been in use for analysis of spectroscopic data. First derivative and second derivative are both commonly used in modern spectroscopy, particularly in NIR spectroscopy. The digital derivative reduces baseline interference and increases the resolution of small absorbance bands. There are various methods of calculating derivative spectra: the point-difference method, gap method, Savitzky-Golay (SG) method (Savitzky et al 1964; Morrey 1968; Steiner et al 1972), and Norris method (Hopkins

2001a,b). For the SG derivative method, data points in the region around a central point are fit to a polynomial, and the analytical value of the derivative at the midpoint of the interval is taken as the value of the derivative at the wavelength of the central point. The SG method is the subject of the current study because it is widely available in many software packages, and the least-squares curve fitting and differentiation can be done in a single convolution operation, simply and elegantly (Hopkins 2001a). However, observation of convolution results is necessary to avoid over-smoothing and loss of resolution of underlying bands (Hopkins, 2001a).

The objective of this study was to update NIR calibration models for determining protein and amylose in rice flour using a sample set with greater range and distribution of sample values and to investigate the best derivative condition for the best model.

MATERIALS AND METHODS

Sample Preparation

Two sets of rice samples were prepared from harvests in 1996 ($n = 90$) and 1999 ($n = 128$). Samples from 1996 were described by Barton et al (2000); 78 samples were grown in United States (California, Texas, Louisiana, and Arkansas), three came from Taiwan, three from Korea, and six from Australia. Samples from 1999 were provided by the USDA-ARS, Rice Research Unit, Beaumont, TX, from the plant breeder nursery. The sample sets were combined to give one robust sample set with better range and distribution for amylose and protein content. The samples were shelled using a rice machine (model SB, Satake Engineering, Toyota, Japan) and then immediately milled. Milling conditions were as reported for the previous study by Barton et al (1998). Samples of ground flour were obtained with a Satake cyclone mill that was heated to 40°C and equipped with a vibrating trough to admit a steady and uniform supply of rice to the mill.

Reference Analyses

Protein ($N \times 5.95$) was determined by the method of combustion using a Leco model FP-2000 nitrogen analyzer in duplicate assays on a 0.5-g assay of ground rice (Approved Method 46-30, AACC 2000). Apparent amylose was determined by the method of Juliano (1971) at the Rice Research Unit. All replicates were averaged for use in the data analysis.

Near-Infrared Spectroscopy

Near-infrared spectroscopic analyses were performed using a monochromator (model 6500, NIRSystems, Silver Springs, MD).

¹ USDA-Agricultural Research Service, Richard B. Russell Agricultural Research Center, Athens, GA 30605.

² Corresponding author. E-mail: msohn@qaru.ars.usda.gov

³ USDA-Agricultural Research Service, Rice Research Unit, Beaumont, TX 77713.

⁴ USDA-Agricultural Research Service, Southern Regional Research Center, New Orleans, LA 70179.

The instrument was operated by the software package WINISI v. 2.01 (Infrasoft International, Port Matilda, PA). The reflectance spectral data were scanned over the range of 400–2,498 nm at 2-nm intervals and then truncated to 1,100–2,498 nm (700 data points). Samples were packed in a spinning cup. Triplicate spectra were collected on each sample and stored as log (1/R).

Data Processing and Chemometrics

All spectra were averaged to produce a single spectrum for each sample and then converted to JCAMP format to import into the Unscrambler software (v. 7.6, CAMO, Trondheim, Norway) for chemometric analysis. The spectral data were preprocessed by multiplicative scatter correction (MSC) (Isaksson and Næs 1988) to remove scatter effects and transformed with the Savitzky-Golay (SG) method. The first derivative (D1) and second derivative (D2) were used to develop the PLS model because derivatives of orders higher than two are more sensitive to noise and have not been shown to have an advantage in calibration (Hruschka 2001). Three different polynomial fits of second (quadratic), third (cubic), and fourth (quartic) orders, and 3–17 points in convolution intervals were used as the derivative conditions.

Partial least squares (PLS) regression (Martens and Næs 1989) was performed and a model was developed for predicting dependent variables from spectra. Random cross-validation of the model was used to evaluate the performance with 20 segments and 13 samples per each segment. Each cross-validation subset was mean-centered. Performance statistics were accumulated on each group of removed samples. Model performance was reported as the correlation coefficients (R^2), the standard error of cross-validation (SECV), and the average of the residuals (bias).

RESULTS AND DISCUSSION

Figures 1 and 2 show the distribution of the protein and amylose content for rice samples. The 1996 samples represent the rice found in commercial use; the gaps in protein and amylose content

arise because of selection for specific end uses. The 1999 samples were obtained from plant breeder stock and used to fill the gaps and extend the range somewhat. As shown in Fig. 1, the range of protein in the 1996 samples was 4.89–11.3% with very few samples having <8% protein and with some gaps. The distribution of the protein content was improved after adding the 1999 samples, and the total sample set has a range of 4.89–12.48% for protein content. Distributions of amylose content are shown in Fig. 2, where most of the data gaps in the 1996 sample set were filled by 1999 sample set, even though there were still very few samples with 1–7% amylose. When one analyzes for protein and amylose, one must realize that an increase in one component results in a relative decrease in the other. This is complicated for amylose because the starch fraction contains amylose and amylopectin, whose proportions can vary considerably. Thus, it is possible to have a rice cultivar in which both protein and amylose increase or decrease. For this reason, it was necessary to obtain samples that had the needed protein contents and varying amylose-to-amylopectin ratios. A total range of the amylose content was 0.2–25.7% for the combined sample set. The combination of data for two years gave better range and distribution of sample values compared with the data for one year.

The histograms in Fig. 3 present the SECV values of PLSR model for protein developed using the D1 and D2 methods. The quadratic polynomial fit used 3–11 points in the convolution intervals, whereas the cubic and quartic fits used 5–11 points because five points is the minimum required. For the D2 models, all polynomial fits with 3–11 points produced similar results with an SECV of $0.24 \pm 0.01\%$ except the three-point quadratic and five-point quartic fits. Both models produced high errors (>0.3%). Performance of the models developed using D1 was better than those using D2 in all polynomial fits and convolution intervals. A three-point quadratic fit produced slightly higher SECV values compared with the other D1 models. The best protein model was from the five-point quartic fit of the D1; R^2 and SECV were 0.982 and 0.230%, respectively. The fact that the three-point quadratic fit for the D1 and

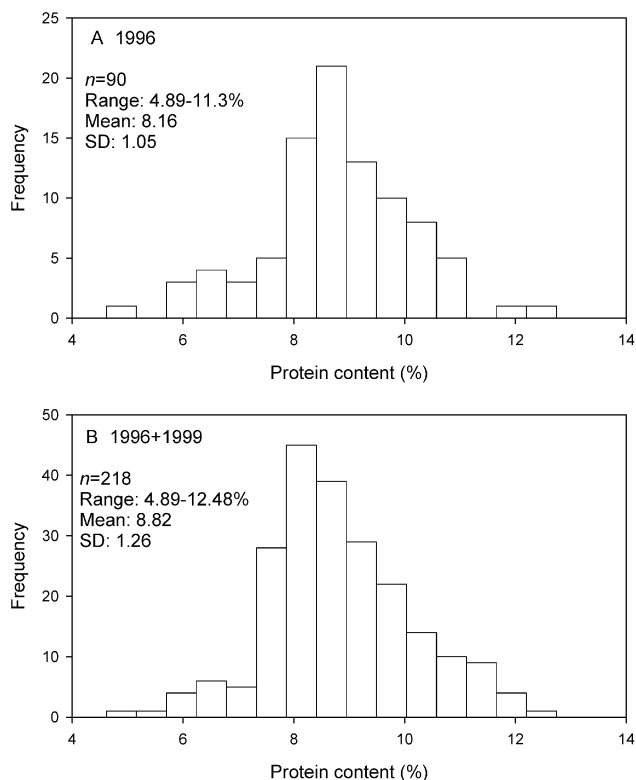


Fig. 1. Distributions of protein contents of rice samples (A) 1996 sample set and (B) 1996 plus 1999 sample set.

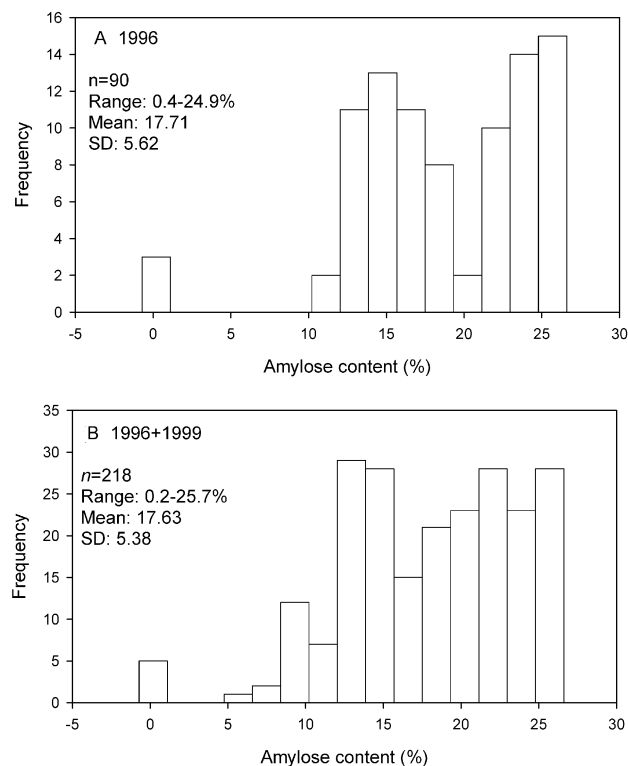


Fig. 2. Distributions of amylose contents of rice samples (A) 1996 sample set and (B) 1996 plus 1999 sample set.

D2, and five-point quartic fit for the D1 resulted in higher errors is probably related to poor noise rejection of the spectra. Hopkins (2001a) reported in his work analyzing polystyrene that a five-point quartic fit produced effective noise for both the D1 and D2 model, even though D2 was significantly higher (RSSK/Norm = 3.13) than D1 (RSSK/Norm of 0.9), here the RSSK/Norm was interpreted as an index of the noise reduction factor and calculated as the square root of the sum of the squares of the convolution coefficients, divided by the normalization constant. However, in our study the five-point quartic fit was suitable for the D1 model and produced the best results for protein analysis.

By evaluating the most effective PLS model for each principal component, it was concluded that four factors were sufficient to validate the protein contents (Fig. 4). As expected, the three models that produced higher SECV values showed higher y -variances in all factors when compared with the other models.

Derivative models for the amylose are shown in Fig. 5. The convolution intervals used were extended to 17 points. The quadratic and cubic fits produced almost the same results for the D2 model, whereas the cubic and quartic fits have the same pattern for the D1 model. For the D2 models, the SECV values were decreased gradually by increasing the number of points in the convolution interval. The quadratic and cubic fit models produced better results than the quartic fit model. Again, the three-point quadratic and five-point quartic fits produced unacceptable models with an SECV of $3.0 \pm 0.1\%$. The best D2 model for amylose was

achieved from the quadratic (or cubic) fit with 17 points with an R^2 of 0.979 and an SECV of 1.002%. The D1 method produced a stable model performance regardless of changes in the polynomial fits or convolution intervals, even though the SECV of the quadratic fits tended to increase slightly after 13 points. The quadratic fit with 5–11 points and cubic (or quartic) fit with 7–17 points were suitable derivative conditions for the amylose D1 model.

Ten factors were the optimal choice for the amylose modeling (Fig. 6). In the three-point quadratic and five-point quartic fit, the y -variances decreased by up to six factors. However, beyond this point, no more improvement was observed despite increasing the number of factors. This result confirms that the three-point quadratic and five-point quartic fits are unacceptable for amylose modeling. The best calibration models for both constituents were developed with good precision (Fig. 7).

Actually the NIR method tends to perform best when the constituent range is not extreme, so the model with data for two years has a somewhat lower precision than the model with data for one year. The previous protein models using data for one year developed by Delwiche et al (1995) and Barton et al (2000) resulted in an SEP of 0.107% using 16 factors (range of 5–10%) and SECV of 0.14% (range of 7.03–10%), respectively. Compared with the two models, the new protein model with data for two years has slightly less precision. However, it is definitely acceptable and robust if one considers better distribution of sample values, better range, and the fewer factors. The amylose model with

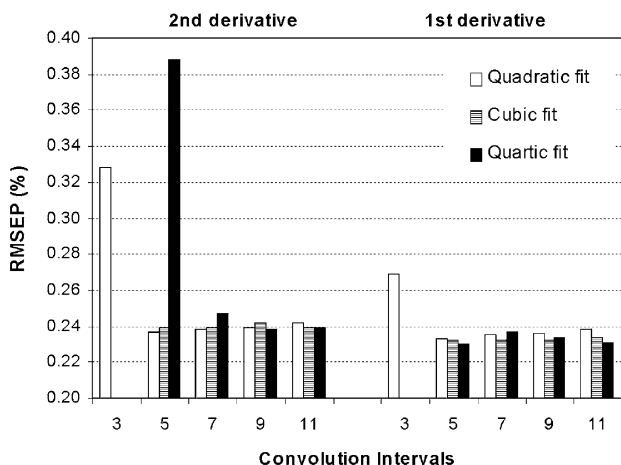


Fig. 3. Histogram of standard error of cross-validation of protein models developed using Savitzky-Golay derivative method.

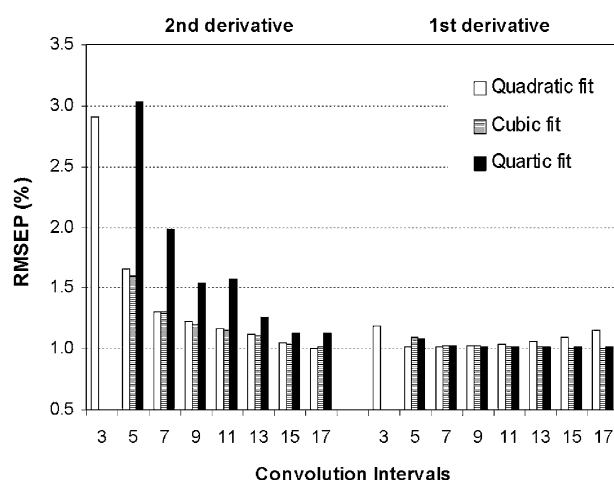


Fig. 5. Histogram of standard error of cross-validation of amylose models developed using Savitzky-Golay derivative method

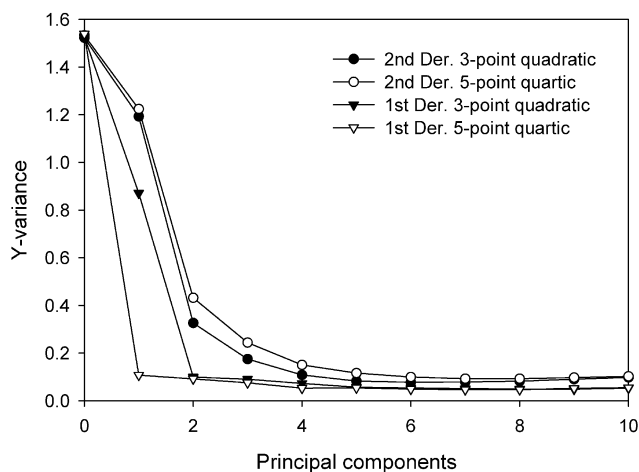


Fig. 4. Principal components vs. y -variances for protein PLS models.

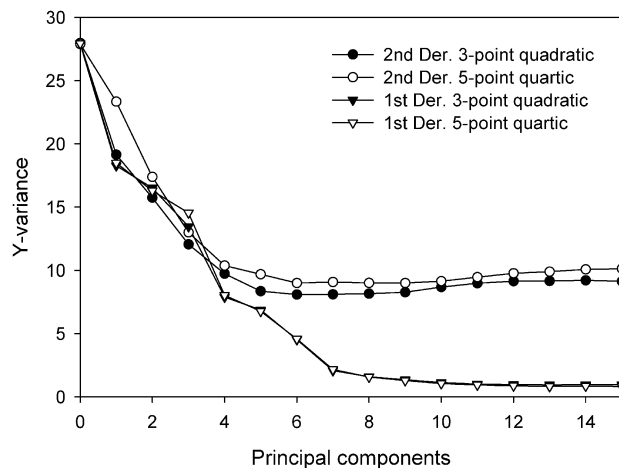


Fig. 6. Principal components vs. y -variances for amylose PLS models.

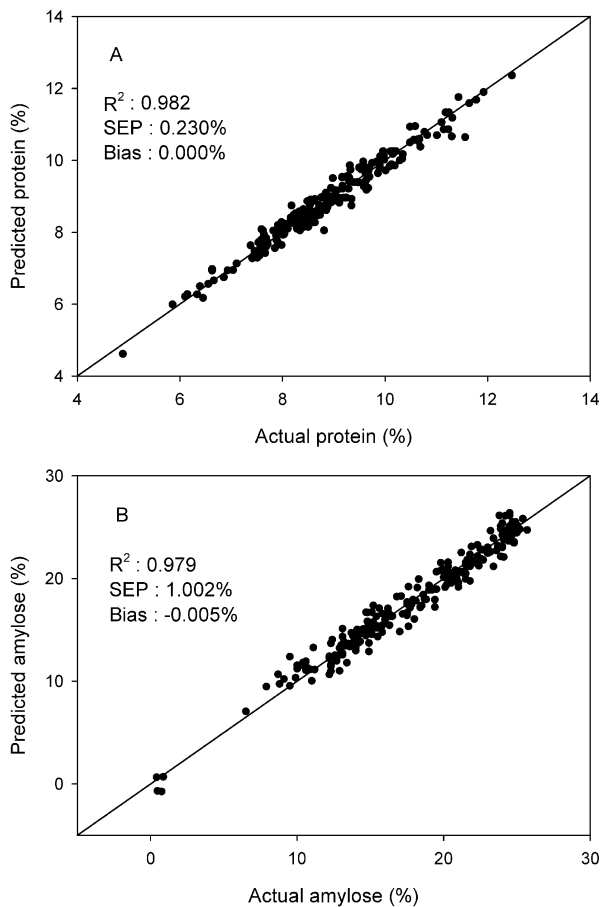


Fig. 7. Scatter plots of reference values vs. NIR validated values of the best PLS models for protein (**A**) and amylose (**B**). Protein model was generated from 5-point quartic fit of D1; amylose model was generated from the 17-point quadratic fit of D2 derivative.

data for two years is more robust than that built using data for one year (SEP = 1.04% using 18 factors, SECV = 1.25%) because fewer factors were required for the model. The sample set used has better distribution of sample values and greater diversity of amylose-to-amylopectin ratio.

CONCLUSIONS

It is important to carefully select the derivative conditions such as polynomial order, convolution interval, and order of derivative to generate the best model for determining protein and amylose in rice flour. The quadratic and cubic fits produced almost the same results in D2 models, and the cubic and quartic fits were similar for the D1 models. For protein, any polynomial fit with 5–11 points was acceptable for both the D1 and D2 models, except for the five-point quartic fit D2 model. For amylose, the 17-point quadratic (or cubic) fit was acceptable for the D2 models, whereas the 5–11

points quadratic fit or 7–17 points cubic (or quartic) fit were suitable to develop the D1 amylose models. The PLS models developed using a data set for two years resulted in good precision and were robust for both protein and amylose analysis.

ACKNOWLEDGEMENTS

We gratefully acknowledge the technical assistance of Judy Davis.

LITERATURE CITED

- American Association of Cereal Chemists. 2000. Approved Methods of the AACC, 10th Ed. Methods 46-11A and 46-30. The Association: St. Paul, MN.
- Barton, F. E., II, Windham, W. R., Champagne, E. T., and Lyon, B. G. 1998. Optimal geometries for the development of rice quality spectroscopic chemometric models. *Cereal Chem.* 75:315-319.
- Barton, F. E., II, Himmelsbach, D. S., McClung, A. M., and Champagne, E. T. 2000. Rice quality by spectroscopic analysis: Precision of three spectral regions. *Cereal Chem.* 77:669-672.
- Champagne, E. T., Bett, K. L., Vinyard, B. T., Webb, B. D., McClung, A. M., Barton, F. E., II, Lyon, B. G., Moldenhauer, K., Linscombe, S., and Kohlwey, D. E. 1997. Effects of drying conditions, final moisture content, and degree-of-milling on rice flavor. *Cereal Chem.* 74:566-570.
- Champagne, E. T., Lyon, B. G., Min, B. K., Vinyard, B. T., Bett, K. L., Barton, F. E., II, Webb, B. D., McClung, A. M., Moldenhauer, K. A., Linscombe, S., McKenzie, K. S., and Kohlwey, D. E. 1998. Effects of postharvest processing on rice texture profile analysis. *Cereal Chem.* 75:181-186.
- Delwiche, S. R., Bean, M. M., Miller, R. E., Webb, B. D., and Williams, P. C. 1995. Apparent amylose content of milled rice by near-infrared reflectance spectrophotometry. *Cereal Chem.* 72:182-187.
- Delwiche, S. R., McKenzie, K. S., and Webb, B. D. 1996. Quality characteristics in rice by near-infrared reflectance analysis of whole-grain milled samples. *Cereal Chem.* 73:257-263.
- Hopkins, D. W. 2001a. Derivatives in spectroscopy. *Near Infrared Analysis* 2:1-13.
- Hopkins, D. W. 2001b. What is a Norris derivative. *NIR News* 12:3-5.
- Hruschka, W. R. 2001. Data analysis: Wavelength selection methods. In: *Near-Infrared Technology in the Agricultural and Food Industries*. P. Williams and K. Norris, eds. Am. Assoc. Cereal Chem.: St. Paul, MN.
- Isaksson, T., and Næs, T. 1988. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *Appl. Spectrosc.* 42:1273-1284.
- Juliano, B. O. 1971. A simplified assay for milled rice amylose. *Cereal Sci. Today* 16:334-340, 360.
- Martens, H., and Næs, T. 1989. *Multivariate Calibration*. J. Wiley and Sons: Chichester, UK.
- Morrey, J. R. 1968. On determining spectral peak position from composite spectra with a digital computer. *Anal. Chem.* 40:905-914.
- Savitzky, A., and Golay, M. J. E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36:1627-1639.
- Steiner, J., Termonia, Y., and Deltour, J. 1972. Comments on smoothing and differentiation of data by simplified least square procedure. *Anal. Chem.* 44:1906-1909.
- Windham, W. R., Lyon, B. G., Champagne, E. T., Barton, F. E., II, Webb, B. D., McClung, A. M., Moldenhauer, K. A., Linscombe, S., and McKenzie, K. S. 1997. Prediction of cooked rice texture quality using near-infrared reflectance analysis of whole grain milled samples. *Cereal Chem.* 74:626-632.

[Received June 30, 2003. Accepted November 14, 2003.]