Journal of Agricultural Education
Volume 48, Number 4, pp. 11 – 19
DOI: 10.5032/jae.2007.04011

THE PSYCHOMETRIC PROPERTIES OF THE AGRICULTURAL HAZARDOUS OCCUPATIONS ORDER CERTIFICATION TRAINING PROGRAM WRITTEN EXAMINATIONS

Brian F. French, Assistant Professor
Daniel H. Breidenbach, Doctoral Candidate
William E. Field, Professor
Roger Tormoehlen, Professor and Department Head
Purdue University

Abstract

The written certification exam that accompanies the Gearing Up for Safety-Agricultural Production Safety Training for Youth curriculum was designed to partially meet the testing requirements of the Agricultural Hazardous Occupations Order (AgHOs) Certification Training Program. This curriculum and accompanying assessment tools are available for national implementation. Psychometric properties of the exam were examined based on a sample of high-school agricultural education students, who were consistent with the population of youth for which the AgHOs were designed to protect from certain workplace hazards. The analyses included evaluation of score reliability (i.e., internal consistency) and the theoretical structure of the exam via confirmatory factor analysis. Reliability estimates were satisfactory. Confirmatory factor analysis revealed that the preferred theoretical model had adequate fit. Implications and future directions are discussed.

Introduction

The purpose of the Agricultural Hazardous Occupations Order (*AgHOs*) training programs is to provide systematic and necessary training for persons, particularly youth under the age of 16, working in agricultural production. This training assists to ensure that minimum safety and health training requirements are met as prescribed under Subpart E-1 of Part 1500 of Title 29 of the Code of Federal Regulations (i.e., *AgHOs*), which is an amendment to the Fair Labor Standards Act (Exec. Order No. *507.71 and* 570.71, CFR 29, 1996). Under these regulations, certain tasks in agricultural workplaces have been identified as particularly hazardous for young workers and disallowed for persons under the age of 16. The Act, however, contains provisions to allow 14 and 15 year olds who meet certain criteria to be employed for specific tasks if certain training requirements are satisfied. Currently the Act does not apply to youth under the age of 16 who work on a farm owned by a parent or guardian.

The *AgHOs* regulations provide little guidance on how to assess an examinee's knowledge or skills on specific core competencies and provide no examination resources. The regulations require that an examinee must be able to demonstrate the knowledge of general agricultural safety practices and the ability to operate a tractor and two-wheeled trailer/implement over an obstacle course similar to those employed in the 4-H Tractor Operator Contest (U. S. Department of Labor, 1996). The Gearing Up for Safety-Agricultural Production Safety Training for Youth curriculum (Tormoehlen et al., 2003) is one recent effort to meet the *AgHOs* training requirements. Although not currently required by the *AgHOs*, the design team for the Gearing Up for Safety curriculum concluded that a consistent and objective method of assessing students prior to *AgHOs* certification was needed. Introduction of a clearly defined testing process would enable instructors to objectively, fairly, and

consistently assess the skills of each examinee seeking employment. A validated examination process would allow an examinee to be assessed by comparing performance to a set of standard criteria and identifying areas of weakness to be addressed prior to certification and employment.

The Gearing Up for Safety design team developed a three tier assessment process: a written exam, demonstration of tractor pre-operational safety inspection, and successful completion of tractor operation and driving exam over a standard course. The process and exam components were constructed to assess not only knowledge of the minimum core content areas specified by the *AgHOs* but also new agricultural workplace hazards not addressed by the 40-year-old law. The exam was designed such that it could be implemented with youth covered and not covered by the *AgHOs* exemptions as well as with workers entering the agricultural work force with little prior safety training. However, assessment instruments, regardless of the ability measured (e.g., farm safety), cannot be assumed to provide accurate information without proper psychometric evidence to support claims of what the instrument purports to measure. See the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) for a detailed discussion of the topic.

The purpose of this research was to examine the psychometric properties of the written exam, the first component of the assessment process. Specifically, this research examined score reliability (i.e., internal consistency), and evidence of construct validity (e.g., confirmatory factor analysis) of the written examination. The factor structure of the exam was examined with confirmatory factor analysis (CFA), which is a theory-driven analysis requiring specification of the relationship of the items to the underlying abilities or traits. Various indices can be used to measure the goodness of fit of the hypothesized factor structure. Thus, stronger evidence of construct validity can be provided about the scores within this framework compared to a more data driven approach (i.e., exploratory factor analysis). Additionally, CFA allows for comparison of the hypothesized model to rival hypotheses that may lead to stronger evidence of validity (Thompson & Daniel, 1996).

For purposes of the *AgHOs* certification program, the Gearing Up for Safety curriculum (Tormoehlen et al., 2003) assumes the construct of agricultural workplace safety includes knowledge about general farm and ranch safety, basic knowledge about tractors, implements, and machines, and knowledge and skills related to operating a tractor. The written exam was constructed with these three components of agricultural safety in mind. Therefore, one theoretical model to examine was a first-order three-factor model (Model C). However, rival theoretical models were possible and should be examined (Thompson, 2004). An example of a rival model was a single-factor model (Model A), where the previous mentioned components were simply facets of a single construct. Perhaps the exam scores simply reflect the influence of only one factor. Additionally, a two-factor model (Model B) was plausible that specifies farm safety knowledge as one factor and knowledge about operating machines and tractors as a second factor. Last, a higher-order model with three first-order factors (Model D) influenced by one second-order factor was theoretically defensible and was examined in comparison to the other models. Figure 1 shows Model D. The other models are nested within this model, in which one or more factors are collapsed into one another as described above.

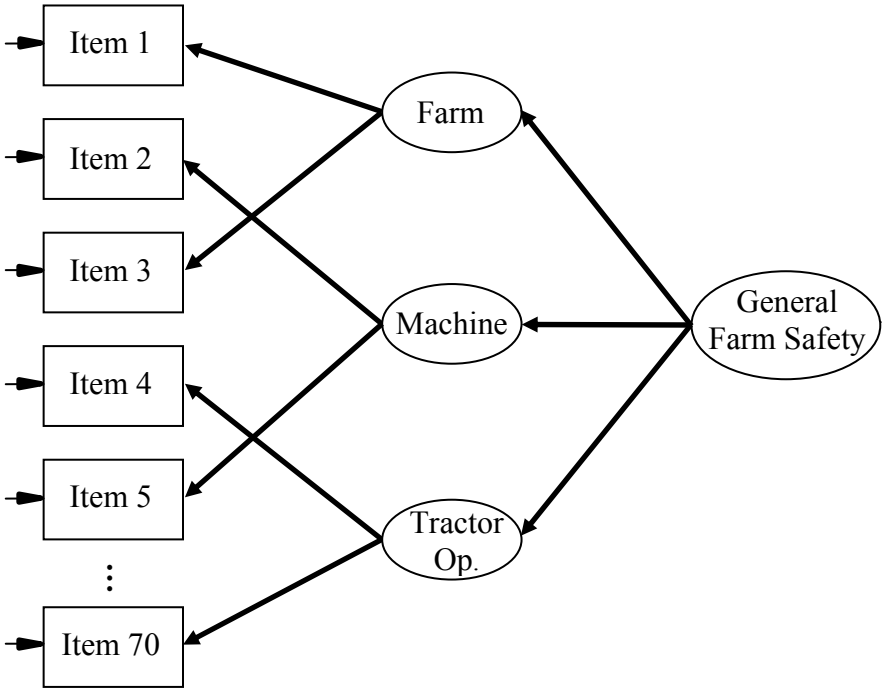

Figure 1. Model D: A three-factor model with one higher-order factor.

Methods

Participants

The participants were part of a larger study evaluating the effectiveness of a visually-based format of the Gearing Up for Safety instructional curriculum. High school students (grades 9–11, ages 14–19, mean age 15.8 years, N = 337, 247 male, 81 female, 9 unreported) enrolled in agriculture courses from three states (Indiana, Kentucky, Tennessee) were administered the exam within a three-week period during the 2004-2005 academic school year. Ethnicity of the participants included Caucasian (n = 317), African American (n = 1), Hispanic (n = 2), Asian (n = 2), and Native American (n = 2). Three participants reported being in more than one of these classifications, and the remaining 10 participants did not report their ethnicity.

Instrument

As the first assessment component of the Gearing Up For Safety curriculum, the written exam measures major content areas covering basic agriculture knowledge and competencies needed to safely operate agricultural tractors and machinery and to perform other general agricultural tasks allowable under the *AgHOs* exemptions. An item pool of 350 dichotomously scored (i.e., correct/incorrect) multiple-choice items was developed to cover the range of *AgHOs* regulations, including such content as (a) identifying tasks having the greatest risk of injury, (b) critical farm hazards, and (c) desired minimum competencies needed to safely operate agricultural tractors and equipment and perform other hazardous farm work. The items were developed and reviewed by content experts to establish content validity, as per recommendations of the *Standards for Educational and Psychological Testing* (AERA et al., 1999). This process not only ensured the items were appropriate in content but also covered the desired content areas and were in accord with the desired competencies identified by the content experts.

The exam administered in this study consisted of 70 items randomly selected within content constraints from the 350-item pool. The exam included roughly equal numbers of items from each of the 11 chapters in the Gearing Up for Safety

curriculum. Exam length was based on a balance between adequate representation of curricular content and keeping the exam to a reasonable length. The theoretical three-factor model for the exam included (a) 24 items to assess farm safety (*farm safety* factor), (b) 19 items to assess farm machines and implements (*machine safety* factor), and (c) 27 items to assess safe tractor driving practices and knowledge (*tractor operation safety* factor).

Procedures and Results

Psychometric analyses of the exam included internal consistency reliability (Cronbach's alpha), and a confirmatory factor analysis (CFA) to examine the instrument's internal structure (i.e., model fit). Alpha estimates were obtained for the total test score, as well as the three subscales. The full-scale score alpha was 0.874. The subscale values for *farm safety, machine safety,* and *tractor operation safety* were 0.644, 0.647, and 0.783, respectively. The full-scale score reliability estimate was adequate for the purpose of the exam whereas the estimates of reliability for the subscale scores were somewhat lower than usually desired (Nunnally & Bernstein, 1994). Selection of other items from the item bank to improve reliability is possible but was not implemented, as it was not possible given the testing situation. The remaining items will be evaluated on future tests.

All items were submitted to an empirical item analysis to examine item difficulty (i.e., proportion correct on each item) and discrimination (i.e., item-total correlation) with the intent of evaluating potentially problematic items. For instance, lower discriminating items may not predict total scores as well as higher discriminating items. With lower discriminating items, item performance will not differ systematically for students with low and high total scores. Ten items had low correlation with total score (less than 0.1). One of these ten items had a negative discrimination value indicating that more examinees in the high scoring group responded incorrectly compared to the low scoring group. Items such as these will need to be reviewed and possibly revised. When this item was deleted, alpha estimates increased slightly for the total score (r = 0.876) and for the *farm safety* subscale score (r = 0.653). This item was excluded from further analysis.

Confirmatory Factor Analysis

The item level CFA was conducted with M*plus* 3.11 (Muthén & Muthén, 1998–2004). M*plus* was used to apply robust weighted least squares (RWLS) estimation to dichotomous data (i.e., item level data scored correct/incorrect), per the recommendation of Finney and DiStefano (2006). A common approach in practice is to use methods designed for continuous data when analyzing such variables. However, treatment of categorical data (e.g., dichotomous data) as continuous in CFA (a) violates the assumption of multivariate normality, (b) may distort the factor structure, and (c) may result in biased parameter and fit index estimates (Finney & DiStefano, 2006; Lubke & Muthén, 2004). RWLS estimation is based on work by Muthén, du Toit, and Spisic (1997), among others, which has been shown to work well in certain conditions for the CFA context using categorical data. Specifically, RWLS uses an asymptotic distribution-free covariance matrix to represent the covariance between item responses and the trait underlying the dichotomous responses. RWLS estimation does not require the inversion of the weight matrix used in the standard WLS approach, which in turn leads to greater stability with small samples and with various factor models, number of variables, and number of response options (Beauducel & Herzberg, 2006; Flora & Curran, 2004).

Model fit was evaluated by several criteria: chi-square significance test, comparative fit index (CFI), Tucker-Lewis fit index (TLI), root mean square error of approximation (RMSEA), and the Weighted Root Mean Square Residual (WRMR). WRMR applies only to RWLS estimation; values less than 1.0 indicate good fit (Yu & Muthén, 2002). The chi-square index and RMSEA are measures of how well the observed covariance matrix is reproduced by the parameter estimates. CFI and TLI both indicate the relative fit of a given model as compared to a null model, but the TLI

adjusts for parsimony. Hu and Bentler (1999) recommend examining combinations of fit indices when evaluating model fit. In particular, a TLI or CFI of at least 0.95 in conjunction with an RMSEA less than 0.06 is suggested as evidence of fit.

Models A, B, and C (i.e., first-order one-, two-, and three-factor models, respectively) were estimated. Recall Model C was considered the primary model, as this is how the exam was constructed. Model D is a higher-order model with three first-order factors (*farm safety, machine safety*, and *tractor operation safety*) and a single second-order factor that influences the three first-order factors. The higher-order model is not statistically more parsimonious compared to the three-factor model but should be considered if first-order factors are correlated. Additionally, the higher-order model may be more consistent with the manner the scores are used in practice. That is, passing the exam is based on a single total score, yet for interpretation of examinee skills for further evaluation, all three subscale scores may be viewed.

The item-level CFA using M*plus* (Muthén & Muthén, 1998-2004) resulted in acceptable fit for all four models tested. As can be seen in Table 1, fit was nearly identical across models. In the absence of a clearly superior model based on statistical criteria, one must rely on theory to guide model selection. Thus, Model D (i.e., higher-order model) was selected as the best fitting model, even though it did not meet strict model fit guidelines (e.g., CFI > 0.95), as it is not clear how these guidelines function with RWLS (Beauducel & Herzberg, 2006). Since Models C and D are statistically equivalent in terms of parameters estimated, fit for the two models are identical. However, Model D accounts for the first-order interfactor correlations and is consistent with exam score use, thus giving it the advantage as the selected model. That is, this model would be preferred to use in practice as it is supported theoretically and empirically.

Table 1
Fit Measures for Four Factor Models Examined

Fit Statistic	Model A	Model B	Model C	Model D
χ^2 (*df*)	315.571 (220)	315.407 (220)	313.906 (220)	304.667 (220)
p-value	< 0.001	< 0.001	< 0.001	< 0.001
CFI	0.899	0.899	0.901	0.901
TLI	0.915	0.915	0.916	0.916
RMSEA	0.036	0.036	0.036	0.036
WRMR	1.06	1.06	1.06	1.06

Note. The degrees of freedom for RWLS are estimated according to a formula given in the M*plus* Technical Appendices.

As seen in Table 2, each of the first-order factors loaded strongly on the second-order factor (range of loadings = 0.93 - 0.99). The higher-order factor accounted for 86% to 98% of the variance in the first order-factors. The first-order factor loadings were generally low to moderate and quite variable. The *tractor operation safety* items had slightly higher loadings than the other two factors. Several (i.e., n = 24) factor pattern coefficients were not significant. Additionally, some (n = 8) standard errors appeared to be somewhat inflated (M = .94, range = 0.00 – 3.57).

These issues may indicate necessary revision of some of these items and/or problems associated with a small sample size, a large number of items, and the use RWLS estimation. Specifically, Flora and Curran (2004) do caution that correctly specified models may be incorrectly rejected (high Type I error rate) with the combination of a small sample and a complex model.

Table 2
Standardized Pattern Coefficients and Uniqueness Estimates for the Higher Order Model

Farm Safety			*Machine Safety*			*Tractor Operation Safety*		
Item	Pattern	Uniqueness	Item	Pattern	Uniqueness	Item	Pattern	Uniqueness
27	0.663	0.561	50	0.874	0.233	11	0.825	0.316
53	0.655	0.566	38	0.669	0.551	47	0.782	0.390
68	0.609	0.626	48	0.651	0.576	28	0.654	0.570
56	0.569	0.669	51	0.636	0.598	19	0.630	0.602
62	0.564	0.681	46	0.582	0.662	21	0.624	0.613
61	0.561	0.681	36	0.573	0.673	29	0.620	0.616
67	0.551	0.692	42	0.490	0.759	60	0.610	0.625
63	0.550	0.690	44	0.471	0.778	9	0.576	0.671
49	0.504	0.744	41	0.463	0.785	26	0.555	0.696
65	0.495	0.754	24	0.429	0.818	32	0.521	0.727
66	0.494	0.750	35	0.366	0.866	30	0.512	0.740
57	0.388	0.850	15	0.352	0.876	54	0.507	0.744
70	0.356	0.873	52	0.327	0.894	3	0.486	0.775
59	0.355	0.872	45	0.309	0.903	17	0.486	0.762
55	0.325	0.892	18	0.181	0.967	31	0.480	0.767
2	0.318	0.899	14	0.146	0.978	25	0.478	0.770
4	0.229	0.947	5	0.139	0.981	43	0.462	0.785
58	0.204	0.958	39	0.128	0.984	37	0.459	0.789
64	0.201	0.959	8	0.006	0.999	20	0.448	0.797
34	0.157	0.976				40	0.363	0.868
6	0.110	0.988				12	0.320	0.897
22	0.102	0.989				33	0.306	0.907
69	0.023	0.999				23	0.288	0.917
						10	0.258	0.934

Farm Safety			*Machine Safety*			*Tractor Operation Safety*		
Item	Pattern	Uniqueness	Item	Pattern	Uniqueness	Item	Pattern	Uniqueness
						1	0.219	0.953
						13	0.117	0.987
						16	0.108	0.988

General Safety

Factor	Pattern	Uniqueness
Farm Safety	0.993	0.014
Machine Safety	0.992	0.017
Tractor Operation Safety	0.930	0.136

Note. Items are in order from largest to smallest pattern coefficient. Structure coefficients are available upon request from the first author.

Discussion

The reliability and validity evidence from this study provide information about the usefulness of the written exam that accompanies the Gearing Up for Safety Agricultural Production Safety Training for Youth curriculum. Although some items are in need of review, internal consistency reliability for the total score scale was acceptable. Additionally, the results of the confirmatory factor analysis for the first-order three-factor model and corresponding higher-order model had acceptable fit to the data. The latter was selected as the most useful for practice.

The CFA results did not reveal an obviously "best" model. All of the hypothesized models show virtually identical fit. In the absence of evidence against any model, we can with confidence interpret the exam in light of the most useful model. Model D best represents how the exam is used: A single score is used to determine whether an examinee may proceed to the other two steps in the certification process. This single score, rather than three subscale scores, can be viewed as an indicator of the second-order factor, *general farm safety*. But since the model has three first-order factors (i.e., *farm safety, machine safety, tractor operation safety*), there is the possibility of examining subscale scores for diagnostic reasons, as some examinees will not pass the exam.

Allowing a young person to progress through the certification process and potentially placing that young person in a hazardous worksite exposes the individual to considerable risk of injury or death. The *AgHOs* training programs and certification process, such as those similar to the Gearing Up for Safety curriculum, are focused on minimizing this risk. For instance, by examining the subscale scores, targeted solutions for those examinees not passing may include such activities as additional studying or training to target areas that appeared deficient based on exam performance. In any event, the examinee would be required to acquire the necessary knowledge and demonstrate the appropriate competencies before proceeding through the certification process.

Progress toward reducing the probability of injury or death of youth working in agricultural environments can be facilitated by ongoing research in this area. For instance, related to the written exam, continued evaluation, both quantitatively (e.g., item analysis) and qualitatively (e.g., item content review), of the item pool will help to ensure that the measurement of examinee knowledge is conducted with an

instrument with the appropriate reliability and validity evidence. For instance, there are many items in the item bank that need evaluation and may function better compared to the items used here or may require revision. These remaining items will be evaluated on future tests, and better functioning items could lead to improved internal consistency reliability, for example. Although not a focus of this paper, similar types of evidence should be gathered for the other components of the certification process (i.e., preoperational inspection, tractor operation/driving exam). Without such evidence, scores and certification, which depends on these scores, lack meaning.

Application to Field / Research

Based on the available evidence, the written certification exam is a useful tool for assessing safety knowledge related to hazardous work that may be performed on a farm (e.g., servicing various machines, operating a tractor) by youth workers, including those under 16 years of age. However, the importance of the factor structure is in need of further evaluation because validation is limited with only a single study. For instance, cross-validation with an independent sample would strengthen validity evidence. Thus, these results provide the first set of empirical evidence for test score validation for the written exam. This information is essential for users to have confidence in the use of the scores received by participants in the Gearing Up for Safety Training program. Test score validation is a judgment based on an integration of empirical results and theoretical rationales (Messick, 1989). Future research will need to assess other forms of score reliability and validity. For instance, estimates of the (a) stability of scores over time (e.g., test-retest reliability), (b) measurement invariance (e.g., measuring the same trait in the same manner) across groups (e.g., sex, race/ethnicity) and geographic regions, and (c) predictive power (i.e., predictive validity) of the scores for certain outcomes (e.g., lower death and injury rates, fewer close calls, or job success) would enhance the evidence for determining the usefulness of the exam scores and assist with continued exam development (e.g., item pool maintenance).

In closing, given that the overall goal of the *AgHOs* is to enhance the health and safety of youth working in agricultural production, the predictive validity of this exam, and the overall certification program, is perhaps the most important validity evidence for the certification program. To that end, long-term data collection regarding death and injury rates for youth who are certified and not certified must be continually examined.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186-203.

Exec. Order No. *507.71 and* 570.71, CFR 29 (1996).

Finney, S. J. & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course.* (pp. 269-314).

Flora, D. B. & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Lubke, G. H. & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11,* 514-534.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Muthén, B., du Toit, S.H.C. & Spisic, D. (in press). Robust inference using weighted least squaresvand quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*.

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide* (3rd ed.). Los Angeles: Author.

Nunnally, I. H., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, D.C.: American Psychological Association.

Thompson, B. & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Tormoehlen, R., Field, W., Fox, R., Personette, C., Vollmer, W., & Ortega, R (2003). Gearing up for safety: Production agricultural safety for youth. [Computer software]. West Lafayette, IN: Purdue University.

Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

BRIAN F. FRENCH is an Assistant Professor in the Department of Educational Studies at Purdue University, Purdue University, 100 North University St., West Lafayette, IN 47907-2098. E-mail: dbreiden@purdue.edu. E-mail: frenchb@purdue.edu.

DANIEL H. BREIDENBACH is a doctoral candidate in the Department of Educational Studies at Purdue University, 100 North University St., West Lafayette, IN 47907-2098. E-mail: dbreiden@purdue.edu.

WILLIAM E. FIELD is a Professor in the Department of Agriculture and Biological Engineering at Purdue University, 225 S University Street, West Lafayette IN 47907-2093. E-mail: field@purdu.edu.

ROGER TORMOEHLEN is a Professor and Head of the Department of Youth Development and Agricultural Education at Purdue University, 615 W. State Street
W. Lafayette, IN 47907. E-mail: torm@purdue.edu.

Acknowledgements

The research was sponsored by the grant Development, Implementation and Evaluation of a Model Administrative Management System for the HOSTA Program, Cooperative State and Research Service- USDA, AWARD NO: 04-41521-03019.