Grader Bias in Cattle Markets? Evidence from Iowa

Brent Hueth, John D. Lawrence and Philippe Marcoul

Working Paper 04-WP 355
March 2004

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011-1070
www.card.iastate.edu

Brent Hueth and Philippe Marcoul are assistant professors and John Lawrence is associate professor, in the Department of Economics, Iowa State University.

The authors gratefully acknowledge funding support from the Agricultural Marketing Resource Center, Center for Agricultural and Rural Development, Iowa State University. The authors also thank Joe Sellers, Diana Bodensteiner, and various producer members of the Chariton Valley Beef Alliance for generously agreeing to participate in this study. Maro Ibarburru provided valuable research assistance.

For questions or comments about the contents of this paper, please contact Brent Hueth, 371 Heady Hall, Iowa State University, Ames, IA 50011-1070; Ph: 515-294-1085; Fax: 515-294-0221; E-mail: bhueth@iastate.edu.

Abstract

Participants in U.S. markets for live cattle increasingly rely on federal grading standards to price slaughtered animals. This change is due to the growing prominence of "grid" pricing mechanisms that specifiy explicit premiums and discounts contingent on an animal's graded quality class. Although there have been recent changes in the way cattle are priced, the technology for sorting animals into quality classes has changed very little: human graders visually inspect each slaughtered carcass and call a "quality" and "yield" grade in a matter of seconds as the carcass passes on a moving trolley. There is anecdotal evidence of systematic bias in these called grades across time and regions within U.S. markets, and this paper empirically examines whether such claim is supported in a sample of loads delivered to three different Iowa packing plants during the years 2000-02.

Keywords: cattle markets, grader bias, quality measurement.

GRADER BIAS IN CATTLE MARKETS? EVIDENCE FROM IOWA

Introduction

The pricing of slaughter cattle in U.S. markets increasingly depends on carcass quality attributes assigned through so-called grid pricing mechanisms (e.g., McDonald and Schroeder, 2003; Whitley, 2002). The carcass attributes most widely employed for this purpose include U.S. Department of Agriculture (USDA) "quality" and "yield" grades. These two attributes are intended to reflect meat palatability and the quantity of usable meat on an animal in relation to total carcass weight, respectively. Although the adoption of grid-based pricing mechanisms is a relatively recent occurrence, very little has changed in the way the USDA grades carcasses. In the majority of commercially graded cattle, grading is accomplished through visual inspection of each animal, without the aid of physical measurement.

Naturally, there is grading error in this process. Modern slaughter and packing facilities are highly focused on efficiency and move carcasses through processing at a rapid rate. As a result, graders are typically asked to assess the quality and yield attributes of an individual carcass in less than 10 seconds, and grading error is an inevitable occurrence. However, so long as this grading process is "reasonably" unbiased, or at least consistently biased across time and location, it may be reasonable to expect little effect on efficiency or the distribution of returns in marketing cattle. The purpose of this paper is to examine whether bias exists in a sample of loads delivered to three different Iowa packing plants during the years 2000 to 2002, and, if so, to assess the direction and economic consequence of this bias.

The data that make this exercise possible (which we will describe in more detail) include carcass measures, in addition to quality and yield grade, that allow computation of the *true* yield grade. That is, we have data on individual animals delivered to various packing plants for which we observe the USDA called yield grade and the true yield grade. Although seemingly a simple matter (given this data), it turns out that measuring grader "bias" is somewhat difficult. An econometrician or statistician usually uses the term "bias" in reference to the difference between a sample estimate of some population *parameter*, and the true population parameter. In contrast, the grading bias that interests us is the difference between the *distribution* of called grades and true grades. Moreover, even this comparison is not quite what we want since at present no grading technology can costlessly measure true grade without error in the time frame demanded by current processing standards. Thus, it is also necessary to develop some reasonable model of how we think graders should perform and to compare the distribution of called grades with the distribution predicted by this model. If these two distributions differ significantly, then we will say that grading is "biased."

In what follows we briefly describe the grading procedure that is carried out in U.S. cattle markets and develop a formal model of bias estimation. We then describe our data in greater detail and present results. The final section discusses directions for future research and summarizes our results.

Grading Procedures in U.S. Cattle Markets

Grading services offered by the USDA are voluntary and provided at cost to meat packing firms upon request. Interestingly, meat grading was originally intended to be used soley for meat-market reporting but it subsequently developed into a trading standard (U.S. Department of Agriculture, 1997). At present, all meat packing firms employ USDA graders in their facilities and rely on federal grade standards for the majority of their trading.

The *quality grade* of a beef carcass can be one of (in decreasing order of meat palatability) the following: Prime, Choice, Select, Standard, Commercial, Utility, Cutter, and Canner. Technically, the quality grade of an animal is determined as a function of the degree of marbling (intramuscular fat) observed on the carcass, ranging from "practically devoid" to "slightly abundant," and the maturity (or age) class of the animal, the definition of which varies across animal type (e.g., steer versus cow). In practice, however, maturity is not known by graders, and marbling is not physically measured. Both traits are assessed by focusing on the visual state of certain key parts of the carcass. In the definition of the Prime quality grade, for example, an animal's maturity class is determined in part by observing "...slightly red and slightly soft chine bones and cartilages on the ends of the thoracic vertebrae that have some evidence of ossification." Similarly, for this particular maturity class, Prime requires "a minimum slightly abundant amount of marbling, and moderately firm ribeye muscle" (U.S. Department of Agriculture, 1997, p. 13). Thus, in addition to the subjectivity inherent in the definition of quality grade, the grading process itself is subject to additional subjectivity in the form of graders' sensory assessment of the relevant carcass attributes.

Yield grade is in principle defined more precisely than is quality grade. The USDA defines yield grades "1" (best) through "5" (worst) using a well-established prediction equation relating various carcass characteristics to a yield index (where "yield" refers to usable meat in relation to total carcass weight), together with a grading rule that specifies the intervals of the yield index that represent a distinct yield grade. The equation used by USDA is

$$\begin{aligned}\text{predicted yield} = 2.50 + 2.5 \times \text{fat thickness} + 0.20 \times \text{kph} \\ + 0.0038 \times \text{weight} - 0.32 \times \text{ribeye area}, \quad (1)\end{aligned}$$

where "kph" refers to kidney, pelvic, and heart fat. Yield grade is then 1 for predicted yield strictly less than 2.0; 2 for predicted yield between 2.0 and 2.99; 3 for predicted yield between 3.0 and 3.99; 4 for predicted yield between 4.0 and 4.99; and 5 otherwise.

In practice, however, this equation is never used during commercial grading operations. Instead, graders are asked (and trained) to assess visually key points of the carcass and to call a grade, 1 to 5, in a matter of seconds. For example, although there are slightly different criteria specified depending on the apparent size of the animal, a yield-grade 1 animal "...usually has only a thin layer of external fat over the ribs, loins, rumps, and clods and slight deposits of fat in the flanks and cod or udder. There is usually a very thin layer of fat over the outside of the rounds and over the tops of the shoulders and necks. Muscles are usually visible through the fat in many areas of the carcass" (U.S. Department of Agriculture, 1997, p. 11).

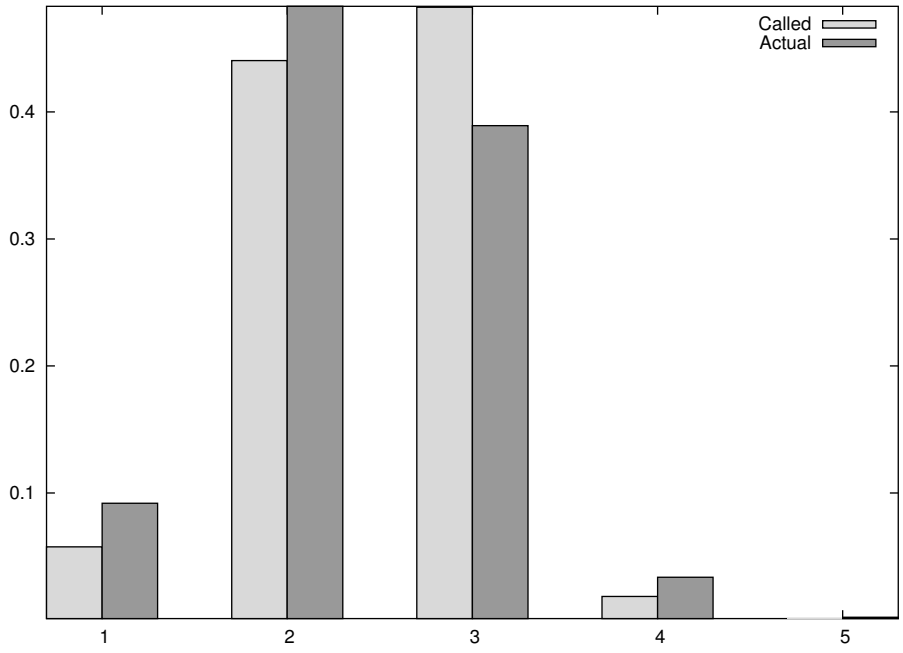

FIGURE 1. Unconditional called versus actual yield-grade distribution

Thus, as with quality grade, there is significant subjectivity in the call of an animal's yield grade by USDA graders. Nevertheless, an interesting difference between quality and yield grade is the ability, at least in principle, to measured precisely yield-grade given adequate time and resources and hence to compare graders' called yield with actual yield. Such a comparison is provided in Figure 1, where we plot histograms of called and actual

grades for the data in our sample. We describe the source and generation process for these data in a subsequent section, but for now it is sufficient to note the apparent difference between called and actual yield-grade distributions. USDA graders tend to call slightly fewer 1's, 2's, and 4's than actually occur, but also tend to call more 3's. Yield-grade 5's rarely occur, nor are they called often.

Of course, the histograms in Figure 1 mask important information regarding the nature of grading error. For example, it is impossible to tell from this figure if graders call more 3's than actually occur because they mistakenly call 3 when yield grade is actually 2, or if instead they mistakenly call 3 when yield grade is actually 1, 4, or 5. To better understand the nature of grading error in our sample, Table 1 presents the distribution of called grade conditional on actual grade. Later in the paper, we refer to this as the "conditional error distribution" of called grades.

Table 1 makes clear that graders error away from the endpoints of the distribution. When yield grade is actually 1 or 5, graders tend more often to call 2 and 4, respectively. Similarly, graders are far more likely to call 3 rather than 1 when yield grade is actually 2, and to call 3 rather than 5 when yield grade is actually 4. The unconditional mean yield of animals in our sample is 2.78, so another way to interpret the apparent "bias" of graders is to say that they error toward the mean.

TABLE 1. Distribution of called yield grade conditional on actual yield grade

	Called Yield				
Actual Yield	1	2	3	4	5
1	.41	.49	.10	0	0
2	.04	.68	.28	0	0
3	.01	.15	.83	.01	0
4	0	.01	.58	.40	.01
5	0	0	.25	.63	.13

This is not surprising given the task that graders perform. With limited time to call yield grade, and with no supporting physical measures of the relevant carcass attributes,

graders must *estimate* the yield index (and then perform the relevant calculation converting the yield index into yield grade). In carrying out such an estimation, and given graders' prior knowledge of the yield-grade distribution, it may be "efficient" to err in the manner observed in Table 1. Thus, without specifying some model of how we think graders *ought* to behave, it is impossible to conclude from the information in Figure 1 and Table 1 that graders are "biased," or to evaluate graders' efficacy in any other fashion. Moreover, thus far we have documented apparent differences in the distributions of called and actual yield grades without saying anything about the economic and statistical significance of these differences. In the next section, we develop a behavioral model of grading that will allow us to address these questions.

Measuring Grader "Bias"

Suppose that graders, upon observing a carcass, receive an unbiased signal s of the carcass's true yield y. Without loss of generality we can write $s = y + \varepsilon$ with $E[\varepsilon] = 0$ and $E[\varepsilon^2] = \sigma^2$. To makes things empirically tractable, we will assume that y is distributed independent of ε and that both variables are normally distributed with the mean and variance of y given by μ and σ_y^2, respectively. Then s and y are jointly normally distributed with correlation coefficient $\rho = \sigma_y/(\sigma + \sigma_y)^{1/2}$.

There are a variety of ways in which we might expect graders to behave, but from a normative perspective they *ought* to make use of their signals in forming an estimate of y. Perhaps the simplest strategy a grader might pursue is to form an estimate of the true yield index y, conditional on the signal he observes, and then base his called yield grade on this estimate. Under the maintained assumption that graders know the true distribution of the yield index,[1] and given our assumption on the joint distribution of y and s, a grader can compute the expected value of y conditional on s using $\mathrm{E}[y|s] = \mu + \rho^2(s - \mu)$. Having performed this computation, the grader can then substitute $E[y|s]$ for y into the USDA yield-grade standard $g_s(y) = 5 - \sum_{j=2}^{5} 1\{y < j\}$ to yield the behavioral rule

$g_b(s) = 5 - \sum_{j=2}^{5} 1\{E[y|s] < j\}$. In what follows, we will assume that graders use this strategy.[2]

If $g_b(s)$ (combined with the estimation of y conditional on s) is a reasonable description of how graders behave, then it should yield predictions similar to what we observe. As noted in the previous section, we are concerned both with the aggregate distribution of called-grade outcomes and with the error distribution of called grades. To see what is implied by this model of behavior, first note that we can rewrite $g_b(s)$ as

$$g_b(s) = 5 - \sum_{j=2}^{5} 1\{s < \tilde{j}(j, \rho, \mu)\}, \tag{2}$$

where $\tilde{j}(j, \rho, \mu) = (j - (1 - \rho^2)\mu)/\rho^2$. Thus, an immediate consequence of $g_b(s)$ is that graders will tend away from calling extreme grades. In particular, for all $j < \mu$, expression (2) implies $\tilde{j}(j, \rho, \mu) < j$. In other words, graders using the rule $g_b(s)$ shade the grade standard for a "good" yield grade (low y) by making the USDA standard more strict. An analogous result holds for $j > \mu$ but in the opposite direction, in which graders shade the grade standard for a "bad" yield grade (high y) by making the USDA standard more lax ($\tilde{j}(j, \rho, \mu) > j$).

Next, we would like to evaluate how this shading affects the likelihood of observing a given called grade, in relation to the likelihood of observing the same true grade. Whether or not this shading makes high- and low-yield grades more or less likely than what we should expect given the population distribution for y and the use of the USDA grading standards $g_s(y)$ also depends on the distribution for s. To answer this question, first consider $\Pr[g_b(s) = 1]$. For this comparison, note that under the USDA grading rule, and for the given population distribution of y,

$$\Pr[g_s(y) = 1] = \Pr[y < 2] = \Phi\left(\frac{2 - \mu}{\sigma_y}\right), \tag{3}$$

where Φ is the standard normal cumulative distribution function. In contrast, under the behavioral rule $g_b(s)$,

$$\Pr[g_b(y) = 1] = \Pr[s < \tilde{j}(2, \rho, \mu)] = \Phi\left(\frac{2 - \mu}{\rho\sigma_y}\right), \tag{4}$$

so that the behavioral rule is likely to yield fewer yield-grade 1's than is the USDA rule, provided $\mu > 2$. A similar computation for yield grade 5 reveals that $\Pr[g_b(5) = 5] = 1 - \Phi((5-\mu)/\rho\sigma_y)$, which is lower than the likelihood of observing yield grade 5 under the USDA rule provided that $\mu < 5$. Thus, under the reasonable assumption the mean of the population distribution for y lies somewhere in the model of the scale defining the USDA yield grade, graders will tend to call too few 1's and 5's, relative to the expected actual frequency of 1's and 5's. Moreover, if for some k, $k < \mu < k+1$, then

$$\Pr[g_b(s) = k] = \Phi\left(\frac{k+1-\mu}{\rho\sigma_y}\right) - \Phi\left(\frac{k-\mu}{\rho\sigma_y}\right) > \Pr[g_s(y) = k] = \Phi\left(\frac{k+1-\mu}{\sigma_y}\right) - \Phi\left(\frac{k-\mu}{\sigma_y}\right), \quad (5)$$

so that graders will tend to call grade k more frequently than it will actually occur.

Thus, at least in terms of aggregate frequencies of occurrence for the various grades, the behavioral rule $g_b(s)$ yields predictions that are broadly consistent with observed called grades. We are also interested in assessing the extent to which $g_b(s)$ results in a conditional error distribution qualitatively similar to that in Table 1. To do so, we need to choose a particular parameterization for the joint distribution of y and s. In the next section, we first present a brief description of our data and then describe a strategy for choosing this parameterization using maximum likelihood. To the extent that we are able to reproduce the error distribution in Table 1, we will conclude that graders are "unbiased" or at least that they seem to be using a reasonable behavioral rule.

Empirics

Data

The data for our analysis come from a sample of loads delivered to three different packing plants in Iowa during the years 2000 to 2002. During this period, a small group of southern Iowa cattle producers payed a third party to collect carcass attributes of their slaughtered animals beyond those reported in grid-pricing closeout sheets. In particular, for each animal delivered by participating producers, the third party measured and

reported (among other things) the characteristics needed to compute the actual yield index according to equation (1), and USDA quality and yield grades.

This sample is not randomly selected and thus potentially not representative of all loads delivered to the relevant plants. Results subsequently reported should be interpreted with this caveat in mind. Another potential caveat is the possibility that the presence of third parties (payed by cattle producers) altered graders' behavior. However, discussions with plant and USDA representatives suggest this is not a concern. The total number of animals delivered by the producers in our sample represents a small fraction of the total animals graded on any given day, and moreover it is not uncommon for producers to request measurement of carcass attributes beyond yield and quality grade.

Table 2 summarizes the data in our sample. Overall, we have observations on 3,841 graded carcasses, including 501 animals delivered to Plant A, 1,431 delivered to Plant B, and 1,909 delivered to Plant C. With the exception of deliveries to Plant A in 2001, yield-grade 2 accounts for roughly 40 percent of graded carcasses. There is slighly more variation across years and plants for yield-grade 3 calls, but in over half of the cases reported in Table 2, yield-grade 3 accounts for at least 50 percent of graded carcasses. Yield-grade 1 is called on between 1 and 15 percent of graded carcasses, and yield-grade 4 is called on 2 percent or less of graded carcasses, with notable exceptions in Plant B during years the 2000 and 2002. Yield-grade 5 is called on an almost negligible number of graded carcasses.

TABLE 2. Data summary

Animals with	Plant A			Plant B			Plant C		
Yield Grade	2000	2001	2002	2000	2001	2002	2000	2001	2002
1	0.01	0.07	0.05	0.07	0.15	0.08	0.03	0.09	0.05
2	0.42	0.66	0.36	0.47	0.46	0.49	0.40	0.39	0.40
3	0.55	0.27	0.57	0.43	0.38	0.41	0.55	0.52	0.54
4	0.02	0.00	0.01	0.23	0.00	0.17	0.02	0.00	0.00
5	0.00	0.00	0.00	0.01	0.00	0.00	0.03	0.00	0.00
Total Animals	238	183	80	546	143	742	1119	221	569

In the following section, we evaluate whether the called grades summarized in Tables 1 and 2 can be plausibly generated by graders' employing the behavioral rule $g_b(s)$ discussed in the previous section.

Estimation

Our basic unit of observation is a graded carcass, and we index these $n(=3841)$ observations by i. Of course, the signal (s) observed by the grader is an unobserved latent variable from the perspective of an econometrician. We denote the carcass yield (estimated via equation (1)) associated with each observation by y_i, and the USDA called grade by g_i. In addition, associated with the ith observation is a K–vector of characteristics x_i that conditions the distribution of y_i. For tractability, we assume these characteristics affect only the location of the distribution for y_i.

Given these assumptions, our endogenous variable g_i is related to y_i and x_i according to the relationship

$$g_i = 5 - \sum_{j=2}^{5} 1\{x_i\beta + \rho^2(y_i + \varepsilon_i - x_i\beta) < j\}, \tag{6}$$

where β is a vector of parameters that affect the mean value of y_i. Given this relation, we can compute the probability of observing any given grade, conditional on the unknown parameters $\theta = (\rho, \beta, \sigma_\varepsilon)$. In particular, defining $\hat{\varepsilon}_i(j,\theta) = (j - x_i\beta)/\rho^2 + x_i\beta - y_i$, we can write the log likelihood for observation i as

$$\ell(\theta|g_i,y_i,x_i) = 1\{g_i = 1\}\log\Phi\left(\frac{\hat{\varepsilon}_i(2,\theta)}{\sigma_\varepsilon}\right) + 1\{g_i = 5\}\log\Phi\left(1 - \frac{\hat{\varepsilon}_i(5,\theta)}{\sigma_\varepsilon}\right) + \sum_{k=2}^{4} 1\{g_i = k\}\log\left(\Phi\left(\frac{\hat{\varepsilon}_i(k+1,\theta)}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\hat{\varepsilon}_i(k,\theta)}{\sigma_\varepsilon}\right)\right). \tag{7}$$

We will estimate the unknown parameters θ through maximization of this likelihood and will compare the conditional error distribution of the rule $g_b(s)$ under these parameters with that in Table 1.

Results

To estimate the parameter vector θ, we ignore heterogeneity across plants and years and concentrate the likelihood in (7) with the sample mean for y as our estimate of μ. We

further (somewhat arbitrarily) choose σ_ε to be 1. Because the value of ρ is not of interest in itself, we whether we get an unbiased estimate of this parameter is unimportant. Instead, we are interested in knowing how well the behavioral rule $g_b(s)$ (conditional on observed yield indices y_i) can replicate the distribution of called grades g_i.

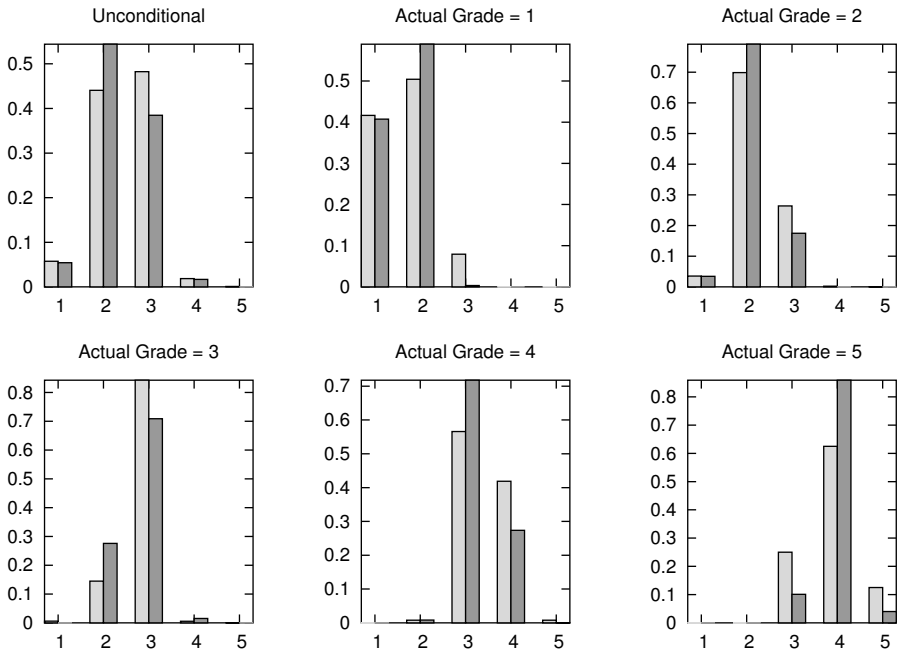

FIGURE 2. Unconditional and conditional called versus predicted grades (lightly shaded bars correspond to called values; darkly shaded bars correspond to predicted values)

Maximizing the expected value of the likelihood in (7) yields an estimate for ρ of 0.82. The predicted distribution of grades at this value of ρ (and for $\mu = 2.78$ and $\sigma_\varepsilon = 1$) is presented in Figure 2. The first box in this figure, labeled "Unconditional," contains called frequencies and predicted probabilities for each grade. Overall, the model does a remarkably good job of predicting the aggregate frequency of 1's and 4's called by USDA graders, but it predicts too many 2's and too few 3's.

The remaining boxes in Figure 2 contain the conditional error distributions for called and predicted grades. When actual grade is 1 or 2, USDA graders call more 3's than are predicted by our behavioral model. This tends to benefit packers who receive relatively

high-quality animals at a low-quality price. Interestingly, when true yield grade is 3 or 4, USDA graders do a *better* job of prediction than does our behavioral model. Graders seem to be able to recognize yield grade 3 and 4 animals. When true yield grade is 5, USDA graders call more 5's and fewer 4's than predicted by our behavioral model, but they also tend to call more 3's. Of course, this benefits growers who potentially receive a relatively high price for low-quality animals. However, there are very few observations corresponding to an actual grade of 5, so comparison of called and predicted error distribution conditional on this grade is potentially misleading.

Given these results, it seems reasonable to conclude that USDA graders are not obviously biased. There is weak evidence of bias when actual grades are 1 or 2. An obvious next step for in our analysis is to examine whether the error distributions observed in Figure 2 are statistically different. If so, we can then examine the extent to which the grading bias observed for actual yield grades 1 and 2 are also economically significant.

Conclusion

This paper uses data from a sample of loads delivered to three different packing plants in Iowa to examine empirically the extent to which USDA graders are biased in their assessment of yield grade. Our data contain carcass characteristics sufficient to compute the true yield grade, which we then compare with USDA called grade for each of 3,841 graded carcasses. In our analysis, we define "bias" in reference to a behavioral grading rule that specifies how graders call grades, given imperfect observation of true yield.

Overall, we find that our behavioral rule is capable of explaining important qualitative features of the observed distribution of called grades. Most importantly, our model allows for grading error. When the actual grade is 1, for example, graders often call a grade of 2 or 3. Our behavioral model captures this feature, though in this particular case it predicts too many 3's and not enough 2's. When the actual yield grade is 3 or 4, USDA graders actually do a better job of calling grades than our model predicts. To the extent that USDA

graders err, they seem to do so in a way that benefits packers, though we have not yet evaluated the economic or statistical significance of this effect.

Endnotes

1. Graders typically observe hundreds of carcasses per day and presumably have a good sense of the distribution of quality, though the exact quality of any given carcass cannot be observed.

2. Alternatively, the grader might instead choose $\hat{k}$ to maximize the probability that the actual grade is $\hat{k}$, conditional on the observed signal s. That is, for each $k \in [1, \ldots, 5]$, the grader might compute $\Pr[\text{actual grade} = k|s]$ and then choose the argmax from this set of probabilities. This is perhaps a more sensible strategy, though it is also a more computionally burdensome strategy for grader to employ. The probability of each possible actual grade must be computed and then the largest of these selected. In contrast, the strategy described previously requires computation of a single expectation, and this uniquely determines the called grade. In any case, the purpose of the present paper is to examine whether called grades are consistent with *some* reasonable behavioral model. We leave exploration of alternative behavioral rules that might better explain USDA grader behavior for future research.

References

McDonald, R. A., and T. C. Schroeder. 2003. "Fed Cattle Profit Determinants under Grid Pricing." *Journal of Agricultural and Applied Economics 35*(1): 97–106.

U.S. Department of Agriculture. 1997. *United States Standards for Grades of Carcass Beef.* Washington, D.C.: U.S. Department of Agriculture, Agricultural Marketing Service, Livestock and Seed Division.

Whitley, J. E. 2002. "The Political Economy of Quality Measurement: A Case Study of the USDA Slaughter Cattle Market." *Australian Journal of Agricultural and Resource Economics 46*(4): 515–38.