NBER WORKING PAPER SERIES

THAT'S NEWS TO ME! INFORMATION REVELATION IN PROFESSIONAL CERTIFICATION MARKETS

Ginger Zhe Jin
Andrew Kato
John A. List

Working Paper 12390
http://www.nber.org/papers/w12390

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2006

Contact Information: Ginger Z. Jin, Department of Economics, University of Maryland, College Park, MD 20742; telephone: (301) 405-3484; email: jin@econ.umd.edu; website: http://www.glue.umd.edu/~ginger/. John A. List, Department of Economics, the University of Chicago, 1126 East 59th Street, Chicago, IL 60637; telephone: (773) 702-9811 email: Jlist@uchicago.edu; website: http://www.arec.umd.edu/jlist/. An earlier draft of the paper was distributed under the title "Evolution of Professional Certification Markets: Evidence from Field Experiments." We would like to thank the University of Maryland for providing funds to support this research and to three sportscard dealers who kindly participated in one of the field experiments. Gary Biglaiser, Rachel Croson, Glenn Harrison, Liesl Koch, Marc Nerlove, Tigran Melkonyan, Michael Riordan, Kyle Bagwell, Christopher Mayer, Raymond Fisman, Raphael Thomadson, Luis Cabral, John Rust, Dan Vincent, and Larry Ausubel provided useful remarks and discussion on an earlier version of this paper. Seminar participants at the University of Maryland, Columbia University, the ASSA meetings in San Diego, and the NBER Summer Institute also provided comments that helped shape the study. Any errors remain our own. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

That's News to Me! Information Revelation in Professional Certification Markets
Ginger Zhe Jin, Andrew Kato and John A. List
NBER Working Paper No. 12390
July 2006
JEL No. D8, C93

ABSTRACT

This study uses field experiments to investigate empirically the informational role of professional certifiers. We explore a certification market that has evolved in such a manner that provides a unique opportunity to measure the information provision of a monopolist certifier and that of subsequent entrants. Empirical results suggest that the certification industry plays a dual role: it reduces the information asymmetry between informed and uninformed parties and generates new information to all market players. Interestingly, the second role isn't conspicuous until the certification market becomes competitive, as the monopolist certifier credibly distinguishes lemons from non-lemons for the uninformed party, but adds little information to experienced agents. On the contrary, new entrants adopt more precise signals and use finer grading cutoffs to differentiate from the incumbent. Our measured differentiated grading cutoffs map consistently into prevailing market prices, suggesting that the market recognizes differences across multiple grading criteria.

Ginger Zhe Jin
Department of Economics
University of Maryland
College Park, MD 20742
and NBER
jin@econ.umd.edu

John A. List
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
jlist@uchicago.edu

I. Introduction

Market economies devote substantial resources to certify product quality—Educational Testing Services (ETS) offers SAT tests for college applicants, U.S. News & World Report ranks universities, Underwriters Laboratories certifies consumer and industrial products, Moody's reports bond ratings, and accounting companies audit financial reports for public corporations. In theory, a professional certificate may be valuable for at least two reasons. First, if one party of the trade possesses superior information about product quality, the certificate can alleviate the information asymmetry, and therefore attenuate the lemons problem and facilitate trade (Akerlof 1970). Second, professional certifiers might have the expertise to provide information to both sides of the market. Such information can significantly enhance allocative efficiency (Blackwell 1953).

Both roles have profound implications for markets, yet little is known empirically whether and when each arises. Indeed, while theories have advanced to making welfare comparisons across market structures (Lizzeri 1999, Franzoni 1999) and regulators express concerns about the market power of certifiers (SEC 2003), little is known about the primitive facts on market structure and certifier performance. For example, what information does a monopoly certifier provide? Who obtains useful information from such a certificate? How do subsequent entrants compete with the incumbent? And, whether, and to what extent, entrants provide market information are all fundamental questions to which we have limited insights. The lack of clean empirical evidence is not surprising since observational data alone might confound criteria differences and sorting effects, rendering field data suggestive, but not entirely compelling. Indeed, even when field data circumvent these problems, too many theoretically relevant factors change simultaneously to allow a clean comparative static test.

The goal of this paper is to use two controlled experiments to provide empirical insights on these basic questions. In doing so, we highlight an approach—field experiments—that might prove useful for future scholars studying related phenomena. For decades, a popular tool in the literature to answer such questions has been an event study. Event studies infer information content by comparing, for example, market prices before and after the release of bond ratings or analysts' earnings report. Assuming market price is a sufficient statistic of the information available to the market, the event study approach has two caveats: it is difficult to control simultaneous information flow; and it is difficult to pin down the exact timing of the arrival of the "certificate" (rumors may spread before the official announcement).

We overcome these difficulties by collecting data from controlled field experiments. Our field experiments are undertaken in naturally occurring settings where the key theoretical factors are identifiable and arise endogenously. Our chosen market—the sportscard grading industry—is attractive in this regard for several reasons. First, there is a generally agreed upon set of traits for grading sportscards, and quality is a major determinant of price. Second, the industry is relatively young, and thus far has been unregulated. Third, there has been little change in the grading technology but the industry has evolved dramatically over the last 15 years. Specifically, the first grading service, PSA (Professional Sports Authenticators), began operating in 1987 and now belongs to a publicly traded company. Due to institutional reasons detailed below, PSA has not changed its grading system since its inception. In 1999, the market expanded, and two competitors entered the market (Sportscard Guaranty LLC (SGC) entered in early 1999 and Beckett Grading Services (BGS) entered later in 1999). All three services continue operating today, and at least 14 other “fringe” grading companies have joined the market since 1999. In theory, these grading companies could compete in both price and grading criteria. Empirically, the "big three" graders (PSA, SGC and BGS) adopt similar price structures but differ in grading criteria.[1]

Based on this observation, our primary field experiment compares the information content of PSA grades to those of subsequent entrants, SGC and BGS. In particular, we submitted 212 sportscards to *all three* major certifiers for grading—PSA, SGC, and BGS—as well as to three professional dealers who differ by card-dealing experience. By making use of a random “round-robin” experimental design, we ensure proper inference about the relative information content across all graders. Data gathered in this field experiment are fit in a structural econometric model to recover two aspects of grading criteria: the grading cutoffs of each grader and the amount of noise in each grader’s signal. This approach allows us to conduct a direct comparison across certifiers and professional market traders. Furthermore, it allows us to compare the estimated grading criteria with actual market prices, and therefore detect whether the market understands the information conveyed in the certificates.

Several insights emerge. First, the monopolist, PSA, utilizes a signal that is as noisy as that of the experienced dealers. This finding is complemented by insights gained from a supplementary field experiment that was conducted in 1997, when PSA acted as the monopolist certifier: when the same card copy was auctioned with and without the PSA grade, non-dealers adjusted their bids in response to the publicized PSA grade, whereas dealers did not change their bidding distribution. Under our preferred

[1] PSA price has slightly increased over time, which is against the intuition that price should go down had newcomers intensified price competition. Moreover, among the big three, the price difference for the most commonly used grading service (grading a number of cards in 20-30 days turnover time) is no more than $1.

interpretation, the data suggest that PSA certificates were used to credibly distinguish lemons from non-lemons for the uninformed party, but added little information to the experienced market players.

In contrast, subsequent entrants—SGC and BGS—considerably sharpened the signal precision and adopted finer grading cutoffs in an attempt to differentiate from PSA. In doing so, they provided information to both dealers and non-dealers. Importantly, because SGC and BGS differentiated from PSA in grading cutoffs, the three certifiers provide a much finer signal than any individual certifier. This result suggests that although new entrants might capture market share from the incumbent, they do not entirely crowd out the information value of the incumbent's grading scheme. Rather, they add information value to the market.

These results highlight the dual role of certifiers: the certification industry, as a whole, not only reduces the information asymmetry between informed and uninformed parties, but also introduces finer information to the entire market. Finally, we find a consistent mapping between market prices and our empirically estimated grading cutoffs and signal precision, which provides a robustness check of our empirical methods and suggests that the market efficiently uses information on the differences across multiple grading standards.

The remainder of our study proceeds as follows. Section II reviews both theoretical and empirical literatures about professional certifiers. Section III provides a brief description of the sportscard certification market. Section IV discusses our experimental design and empirical results. Section V concludes.

III. Literature Review

Starting with Grossman (1981) and Milgrom (1981), many theorists have examined how intermediaries induce the market to reach a state of full information. For example, Biglaiser (1993) sets up a model of "middlemen" and presents some guidelines on which markets benefit from expert intermediaries. A related line of inquiry explores the theory of *independent* certifiers. Such certifiers do not trade the certified goods, rather they maximize profits by setting certification fee and grading criterion. Assuming certifiers can detect product quality with perfect accuracy and zero cost, Lizzeri (1999) shows that a monopoly certifier has incentives to provide a simple pass/fail certificate in order to extract information rents, but competition among intermediaries will lead to full information revelation. Franzoni (1999) examines a different setting where a third-party certificate of compliance is required for firms to engage in a regulated activity but detecting compliance involves unobserved efforts from the certifier. With certain liability imposed on certifiers, competition among certifiers will reduce certification fees but does not always improve social welfare.[2]

Guerra (2001) extends Lizzeri's model by allowing buyers to have a noisy estimate of product quality in the absence of quality certificate. This modeling innovation yields a disclosure of ordered ranks (say A, B, C) instead of the simple pass or fail. Hvide and Heifetz (2001) consider a free-entry model of certification, allowing each certifier to choose certification criterion and certification fee. They find that, in equilibrium, certifiers differentiate their grading criteria and the certification fee increases with the stringency of grading criterion.

Clearly, these models do not exactly match the structure of the sportscard grading industry. For example, most theories assume that sellers have perfect information about product quality, and therefore restrict the certifier's role to solving the lemons problem.

[2] The model restricts all certificates to pass/fail and asserts that in equilibrium all certifiers exert the same

In reality, there may be noise in all players' information set, which allows certifiers to provide information to both sides of the market. Despite the important differences, we believe the theoretical literature provides three insights that are useful benchmarks for our analysis. First, in the absence of competition, a monopoly certifier may not reveal full information. Second, competition in the certification industry should improve the information content of certificates. Third, if certifiers can choose grading criterion beyond the simple pass or fail, competition among certifiers is likely to lead to differentiation in grading criteria.

Interestingly, on the empirical side, the bulk of the literature focuses on certified goods rather than the certifier(s). A typical event study investigates how the market reacts to a change of certificate. For example, Ippolito and Mathios (1990) investigate how cereal consumers respond after the government lifted a ban of advertising on the health benefits of fiber cereal consumption (while the fiber content of ready-to-eat cereal is verifiable through independent sources). Jin and Leslie (2003) document how consumers and restaurants respond to the issue of restaurant hygiene grade cards. Numerous studies measure how the price of a financial asset reacts to bond rating, analyst report, or audited earnings report.[3] Aside from these event studies, researchers have documented price and/or quality differences between certified and uncertified goods in thoroughbred racehorses (Wimmer and Chezum 2003), collectible stamps (Dewan and Hsu 2005) and sports cards (Jin and Kato 2006). Chaney et al. (2004) examine how

---

effort in determining compliance.

[3]The evidence on bond ratings is inconclusive. Katz (1974), Grier and Katz (1976), and Hettenhouse and Sartoris (1976) report evidence that bond rating increases provided unanticipated information, but decreases did not. Hand et al. (1992), Ederington and Goh (1998), and others have found the opposite result—bond rating decreases provided new information but increases did not. Pinches and Singleton (1978), Wakeman (1981), and Weinstein (1977) found no evidence that bond rating changes provided new information in either direction. For financial analysts and auditors, the general conclusion is that stock prices are responsive to some of their reports, but not to all of them (Healy and Palepu 2001).

private firms select into different auditors and conclude that the fee-premium for the big-5 auditors disappears after controlling for selection.

Only a few studies draw direct comparisons across certifiers. For example, researchers have found that the market treats US bonds with split ratings differently from the bonds with equal ratings and the bonds with only one of the two ratings (Thompson and Vaz 1990, Cantor et al. 1997). These findings suggest that Moody's and S&P may differentiate in rating criteria. Yet because bond issuers can choose whether to obtain one or two ratings, these results are confounded with selection effects. To distinguish the two explanations, Cantor and Packer (1997) examine the factors driving the split ratings between Moody's, S&P, and two other rating agencies that accept voluntary request for bond rating. They find limited evidence of selection bias.

Berger et al. (2000) broaden the scope of professional certifiers to include both private certifiers and regulators. They use price and rating data to infer whether the government inspection and rating of a bank holding company Granger-cause a movement in Moody's rating of the same company, or vice versa. They find Granger-causality in both directions, which suggests that supervisors and bond rating agencies both acquire some information that aids the other group in forecasting changes in bank condition. Besides financial industries, differential ratings have also been documented in health plan report cards (Scanlon et al. 1998) and college rankings (Pike 2004).

As is clear, the existing empirical literature has cleverly used both price and multiple rating data to infer differences across certifiers. While econometric techniques are useful in identifying selection from the differentiation of grading scales, the evidence is indirect and does not reveal the full structure of grading differentiation. In comparison, the experimental approach used in this paper allows us to circumvent the selection issue and obtain direct estimates on grading criteria. Compared to the traditional event studies, field experiments enable us to focus on the informational content of professional certificate while controlling for numerous confounding factors that arise in an observational study.

III. Sportscard Grading

Each year, card companies design and print sets of cards depicting players and events from the previous season. Once the print run of a particular set has been completed, the supply of each distinct card in the set is fixed. The value of a particular card depends on its scarcity, the player depicted, and the physical condition of the card—i.e., condition of the edges, corners, surface, and centering of the printing. To track card condition, people often use a 10-point scale. For example, a card with flawless characteristics under microscopic inspection would rate a perfect "10" while obvious defects to the naked eye, including minor wear on the corners, would decrease the card's grade to a "7". The card's overall grade is computed via the aggregation of the various characteristics, and post-1980 sportscards that merit a grade below "7" are rarely traded.[4]

Card condition, especially at the high end, is hard to detect by the naked eye. Each collector may examine the card carefully (sometimes with the help of a magnifying glass) and obtain a noisy signal of the card condition. The noise of the signal decreases with experience, but most likely remains positive for even the most experienced dealers. In fact, it is not uncommon to observe two experienced dealers disagreeing on the condition of a specific card.

Professional grading offers an alternative channel to identify card condition. PSA

[4] Because grading is voluntary and costly, better quality cards are more likely to be graded. This is why very few post-1980 graded cards are ever observed in the 1 to 6 range, even though such grades exist and are given out when warranted. In practice, graded cards are usually "8" or above (Jin and Kato 2004a).

began offering grading services in 1987 and its parent company became publicly traded in 1999 (Collectors Universe, under Nasdaq ticker symbol CLCT). SGC entered the professional grading market in 1999, soon followed by BGS. As of 2002, PSA, BGS, and SGC remained the largest and most respected of the existing 15-20 grading services. We believe the breakdown of the PSA monopoly in 1999 is due partly to the onset of the Internet, as detailed in Jin and Kato (2005b). In 1998, eBay, the most popular auction site for sportscard transactions, went public. The Internet not only substantially reduces transaction cost, but also intensifies the information asymmetry between buyers and sellers. To overcome the information problem, the demand for professional grading services considerably increased after 1998. The demand shock, plus PSA's commitment to its initial grading criterion (as detailed below), opened profitable opportunities for potential entrants.

Professional grading is voluntary and costs \$6-\$20 per card, depending on package size and requested turnaround time; further, the fee is independent of the actual grade received. Graded cards are encased in plastic and sealed with a sonic procedure that makes it virtually impossible to open and reseal the case without evidence of tampering. The casing indicates the grading service, grade received, and a bar code with serial number that identifies the particular copy of the card. Anyone with Internet access can visit the grader's web site and verify the card's grade by serial number. Figure 1 provides an example of a PSA-graded 1985 Topps #401 Mark McGwire (*rookie*), an example of a BGS-graded 1993 Topps Traded #1T Barry Bonds, and an example of an SGC-graded 1991 Topps Tiffany #352 Ken Griffey Jr. *All Stars*.

PSA adopted integer grades from 1 to 10, whereas BGS adopted a slightly finer grading scheme, which included half grades from 1 to 10: 7.5, 8, 8.5, etc. SGC initially used a 100-point grading scale—e.g. 88, 92, 96—but soon provided equivalent conversion to a half-grade system similar to BGS, where 88 means 8, 92 means 8.5, 96 means 9 and 98 means 10. Interestingly, because SGC used only a limited number of grades in the original 100-point grading scale, the converted grades do not exhaust all possible half grades between 1 and 10. One curious omission is 9.5 – the converted SGC system has 7, 7.5, 8, 8.5, 9, and 10, but no 9.5. In comparison, the BGS scale includes all possible half grades, although BGS rarely gives a perfect grade of 10. Among the three certifiers, BGS is also the only one that offers sub-grades for centering, corner, edge and surface, in additional to the overall grade.

A casual comparison of grading scales suggests an interesting pattern: the first entrant, PSA, adopted a coarse grading scheme, the second entrant, SGC, adopted a finer scheme, and the third entrant, BGS, adopted an even finer grading scheme. Subsequent "fringe" entrants have generally followed this approach as well, adopting scales that are refinements of the existing certifiers' techniques.

We find it interesting that PSA has not changed its grading criteria since its inception. In theory, PSA could respond to the entries of SGC and BGS by changing its own grading criteria, but such a change is likely not optimal due to at least two important facts. First, because PSA never indicates date of certification, and thousands of previously and newly graded copies are traded daily in the same market, PSA is committed to one grading standard over time unless it wishes to upset the market. In this spirit, PSA has learned an important lesson from the coin market—one major coin certifier increased its grading upper bound from 60 to 64 in the 1970s, which generated a major market upset and was believed to contribute to the decline of coin trading (PSA also grades coins). Second, PSA remains the dominant player in the industry. Given the market expansion since 1998, PSA's grading business has grown rapidly (even though the growth could have been greater had entry not occurred). It would therefore be unwise to jeopardize a long-established reputation and a rapidly growing business to combat a relatively small market stealing pressure resulting from competitive entries. As a consistency check, we consulted a number of experienced sportscard dealers, who all confirmed the temporal stability of the PSA grading standard. As a whole, this represents convincing evidence, for any criterion change undetected by the market generates no benefit to PSA, and should have never been adopted in the first place.

A further attractive feature of using the sportscard grading industry in our case study is that, whether buying or selling, all trading parties refer to a standard price guide for sportscards—*Beckett Baseball Cards Monthly* for baseball cards, *Beckett Football Cards Monthly* for football cards, etc. For each single type of ungraded card, Beckett collects pricing information from about 110 card dealers throughout the country and publishes a "high" and "low" price reflecting current selling ranges for Near Mint-Mint (8) copies. The high price represents the highest reported selling price and the low price represents the lowest price one could expect to find with extensive shopping. For graded cards, Beckett follows the same practice but lists price ranges for each grade level (usually 7 to 10) of frequently graded cards. When trading volume is high, Beckett reports separate prices for PSA, BGS, and SGC, and pools all other companies as "Others". Jin and Kato (2005a) report that market-clearing prices of graded cards closely track the "low" price listed in the Beckett price guide. This particular market feature allows us to treat Beckett "low" prices as a proxy of market-clearing prices and to map them with our empirically estimated grading cutoffs.

IV. Empirical Results

This section presents two field experiments and one price analysis. The first experiment identifies the grading criteria of the three professional certifiers. In complement, the price analysis detects whether the price structure prevailing in the trading market is consistent with the grading criteria discovered in the experiment. Further market examination is presented in the second experiment, where we investigate how different types of card traders react to the presence of a professional certificate.

IV.1 Experiment One

Experimental Design We began our field experiment by equally distributing 216 sportscards into 9 groups in February 2002. Upon the grouping, we randomly allocated the cards first to the three sportscard dealers (Kevin, Rick, and Rodney) and then to the three certifiers (PSA, SGC, and BGS). Specifically, Kevin received groups A, B, C; Rick received groups D, E, F; and Rodney received groups G, H, K. Once all three dealers finished grading, we mailed groups A, D, G to PSA; B, E, H to BGS, and C, F, K to SGC for official grading. All certifiers returned the cards by April 29, 2002, which marked the end of Round 1. In the next two rounds, we rotated the cards to be graded by one of the other graders until all 6 graders had graded *each* of the 216 cards. Table 1 presents the rotation details: each row represents a card group and each column represents one of the six graders.

The round-robin aspect of the experimental design is especially important for two reasons. First, each of the three professional certifiers places the graded card into a sonically sealed plastic casing upon certification and grading. To avoid confounding influences, when we received the graded cards from the certifiers, we recorded the card's grade and carefully chiseled off the plastic casing before re-sending the card to the other graders. Because the case is designed to prevent tampering, we may have inadvertently damaged the card. The round-robin rotation prevents one certifier from receiving systematically worse cards than another certifier. Indeed, we damaged 4 of the cards accidentally during the process; hence, our final data analysis uses 212 cards.

Second, for the three dealers who do not seal cards in plastic cases, grading entails physical handling. Although they are all experienced dealers and promised to handle the cards with care, there exists a chance that the grading process generated some minor damage to the cards. Such damage would upset future grades, but would not be easily detectable by even the trained eye. This fact represents the impetus for rotating the cards among dealers in such a way that even if the handling differed by dealer, each certifier on average faced the same distribution of card quality. Also note that in each round, dealer grading took place before certifier grading. In case dealers introduced an additional noise in card quality, we would capture it as part of a certifier's signal noise, thus *understating* the signal precision difference between certifiers and dealers. Since in the data we find that all certifiers are at least as precise as dealers, our conclusion is potentially strengthened.

Prior to moving to our empirical results, we should mention a few interesting aspects of our field design. First, none of the professional certifiers knew that we were running an experiment on the certification market and so they graded the cards under the assumption that they had been mailed to their company as "normal" cards to be graded. This was not a difficult task, as these three companies grade, on average, at least 10,000 cards per year. Nevertheless, when mailing the cards to each of the certifiers we took special precautions not to tip them off by using different consumer names and addresses in each round. Second, to ensure that this was a naturally occurring transaction, we paid the typical grading fee for PSA ($8), SGC ($6.5), and BGS ($9) to grade the cards, and we paid a flat-fee ($108) to our three dealers (whose requested fees were lower because they could grade the cards during slow times of the day at their retail shops). We were careful to choose professionals that had been shop owners in the sportscard market for at least five years and who had heterogeneous experience levels (Kevin: 8 years; Rick and Rodney: 14 years) to provide a demanding test of the professional certifiers.

Summary Statistics Different graders might adopt disparate grading cutoffs, hence it is important to highlight that the grades are ordinal and the raw grades are not readily comparable across graders (e.g., PSA 10 may not be equivalent to SGC 10). Moreover, because most grades are 8 or above and each grader has at most 5 possible grading categories at 8 or above (i.e., 8, 8.5, 9, 9.5, 10), a number of cards obtain identical grades from the same grader, thus creating ties. Inevitably, each grader has a lumpy distribution (see Table 2). Depending on how we order ties, the rank correlation of any two graders could be as low as 0.4 or as high as 0.9. For this reason, it is difficult to make sharp inferences from raw rank correlations.

To deal with these difficulties, we adopt an alternative approach. For any two cards randomly selected from the pool of 212 cards (call them A and B), we examine whether grader j and grader j' agree on their relative quality. If both j and j' agree that the quality of card A is superior to the quality of card B (i.e., $q_A > q_B$) or the two cards are of equal quality (i.e. $q_A = q_B$) , we define the two graders as *strongly consistent* for this card pair. If grader j rated $q_A > q_B$ but grader j' rated $q_A < q_B$, they are *strongly inconsistent*. If one grader rated $q_A > q_B$ but the other rated $q_A = q_B$, they are *weakly inconsistent*. After completing this comparison for all possible card pairs (22,366 in total), we compute the percentages in which grader j and grader j' are strongly consistent, strongly inconsistent, or weakly inconsistent. This exercise results in three matrices, which are provided in Table 3: panel A for strong consistency, panel B for strong inconsistency, and panel C for weak inconsistency. The three percentages, by definition, must sum to one in every cell.

Of particular interest is Panel B. The degree of strong inconsistency among professional certifiers is roughly 5%-7%, much lower than that among dealers (10%-13%), or that between professional certifiers and dealers (7%-13%). This suggests that professional certifiers, as a whole, are more compatible and more precise than dealers. Should all professional certifiers systematically miss some important component of card quality, the inconsistency between certifiers and dealers would have been much higher than that among dealers. The same logic applies if professional certifiers aim the main market but the three dealers were not representative of the mainstream. Short of this inconsistency, it is reasonable to assume independent evaluation noise among all six graders, rather than some systematic bias within professional certifiers or within dealers.

In the last row, we compute the average strong inconsistency for each grader as compared to the other five. Among professional certifiers, it is clear that BGS, the last entrant of our three certifiers, achieves the highest level of consistency with the other certifiers, and that PSA, which was once the monopolist certifier, is the least in accord. Panel A in Table 3 displays similar patterns: professional certifiers are more likely to be strongly consistent with each other than are certifiers with dealers, or dealers with dealers. Again, in terms of consistency, BGS is the sharpest and PSA is the least in accord[5].

While these summary statistics are suggestive, they do not provide explicit

[5] If we restrict attention to professional certifiers only, then PSA seems the best while a comparison between BGS and SGC produces the largest inconsistency. This holds because PSA adopts fewer grading cutoffs than the other two. For this reason, it is important to compare the three certifiers against a common comparison group (i.e. the three dealers).

estimates of grading cutoffs or grading precision, and therefore do not offer a strict comparison across all graders. We overcome these shortcomings by implementing a full structural model.

Structural Model Suppose card i has an unknown quality q_i, which is iid from a common distribution $F(q \mid \theta)$ where $\{\theta\}$ denotes the distributional parameters. Grader j observes an unbiased noisy signal $s_{ij} = q_i + \varepsilon_{ij}$, where the iid noise $\varepsilon_{ij} \sim N(0, \sigma_j)$ and σ_j denotes the degree of noise in grader j's grading system. Internally, grader j has a set of cutoffs, such as J_8, J_9, J_{10}, etc. Once grader j observes signal s_{ij}, she fits the signal within those cutoffs and assigns corresponding grade g_{ij}. For example, if $J_8 \le s_{ij} < J_{8.5}$, then $g_{ij} = 8$.

Of course, we observe only the final grade $\{g_{ij}\}$. According to the raw grade distribution in Table 3, g_{ij} could be one of (7, 8, 9, 10) if grader j is PSA, (7.5, 8, 8.5, 9) if j is BGS, (7.5, 8, 8.5, 9, 10) if j is SGC, (7.5, 8, 8.5, 9, 9.5) if j is Kevin or Rodney, or (6, 7, 7.5, 8, 8.5, 9, 9.5) if j is Rick. Note that we do not observe any card receiving a BGS 9.5 or BGS 10, implying that the cutoffs for BGS 9.5 and BGS 10 are higher than any cutoff we can estimate from our data.

We take $\{q_i\}$ as random effects (see below for a robustness check on this assumption). Thus, the unknown parameters are the quality distribution parameters $\{\theta\}$, grading cutoffs $\{J_g\}$, and signal precision $\{\sigma_j\}$. Defining $1_{i,j,g}$ equal to 1 if grader j gave card i a grade of g, we have the overall likelihood function

$$L = \prod_{i=1}^{212} \left\{ \int_{q_i} \left[\prod_{j=1}^{6} \sum_{g} 1_{i,j,g} \cdot \left[\Phi\left(\frac{J_{g+} - q_i}{\sigma_j} \right) - \Phi\left(\frac{J_g - q_i}{\sigma_j} \right) \right] \right] dF(q;\theta) \right\}$$

where Φ denotes the cdf of a standard normal and J_{g+} denotes grader j's cutoff that is immediately above grade g. Estimates are obtained via maximum likelihood.

Estimation Results To allow flexibility, we assume $F(q;\theta)$ to be a beta distribution with two free parameters $0 < a \leq 10$ and $0 < b \leq 10$. Beta is a general type of distribution on the support of (0,1), and importantly, it includes the uniform distribution, as well as PDFs that increase or decrease with various concavity/convexity. Our empirical results presented below are qualitatively similar to those under different bounds of $\{a,b\}$.

Empirical results are reported in three panels. Table 4 Panel A presents the estimated grading cutoffs and precisions $\{J_g, \sigma_j\}$ for all six graders. Panel B conducts Wald tests for statistical significance in grading cutoffs of the three professional graders. Panel C tests the statistical significance in grading precision among all six graders. We omit cutoff comparisons for individual dealers because they do not offer grading service for regular business. We ask them to grade by the most detailed scales, however, including all half grades and applying their own grading criteria to ensure that we obtain the most conservative estimation of their grading precision.

All grading noises are strictly positive. Consistent with Table 3, the latest entrant in the professional grading industry – BGS – has the smallest grading noise and is most agreeable with the other graders. For the other two certifiers, the second entrant, SGC, is less noisy than the first entrant PSA ($\sigma_{SGC} < \sigma_{PSA}$), though the difference is not statistically significant. The amount of grading noise is very close between PSA and the most experienced dealers (Rick and Rodney), while the least experienced dealer (Kevin) is noisier than all the other five, especially BGS and SGC.

Note that the first certifier, PSA, utilizes a signal that is statistically as noisy as those of the experienced dealers. Unlike PSA, the second entrant—SGC—sharpens its signal precision beyond the least experienced dealer in our sample, while the third entrant—BGS—adopts a signal that is statistically more precise than all three dealers. This result suggests that later entrants, especially BGS, provide more precise information than PSA.

Full estimation results also shed light on grading cutoffs. The first two certifiers, PSA and SGC, adopt similar cutoffs in their common grade categories: SGC 10 is not distinguishable from PSA 10, SGC 9 is not distinguishable from PSA 9, and SGC 7.5 is very close to PSA 8. The finer categories that SGC tends to add – SGC 8 and SGC 8.5 – are between PSA 8 and PSA 9. In contrast, the third entrant, BGS, adopts a rather different strategy: it defines finer categories on the high end – BGS 9 is between PSA 9 and PSA 10, but not close to either end; while BGS 9.5 and BGS 10 are certainly above PSA 10.

It is worth mentioning that, although SGC and BGS use finer scales than PSA, the whole system encompassing all three certifiers is much finer than any certifier or dealer alone. This result suggests that, although new entrants might capture market share from the incumbent, they do not replace the existing grading system. Rather, by improving grading precision and adopting differentiated grading cutoffs, they add information value to the whole industry.[6] In response, facing multiple (noisy) certification systems, a seller can strategically maximize the grade of a specific card quite easily. For example, he

[6] It is difficult to directly test whether the three professional grades (PSA, BGS, SGC) together provide significant new information to individual collectors. Because we must destroy the previous professional grade before obtaining a grade from the next certifier and many ungraded copies appear identical in front of naked eyes, it is impossible to present the three grades at the same time and convince collectors that the three grades apply to the same card copy. This difficulty motivates us to infer the informational value of professional grades by testing graders in our main field experiment.

could send the card first to BGS, crack it open and resend it to PSA if the BGS grade is lower than 9.5, crack open the PSA case if the PSA grade is less than 10, and try it again with SGC. Of course, this practice will stop at some point when the cost of repeated grading becomes too high. Although we do not have enough data to empirically test for this phenomenon, it is commonly observed in the field. This phenomenon is also non-unique to sportscard grading: at least 15 MBA programs claim in the top 10, and multiple producers within the same industry claim to have the single best quality.

The procedure described above assumes the underlying card quality conforms to a beta distribution. Although the beta distribution encompasses a number of specific distributions (such as uniform), it remains an arbitrary assumption. Instead of trying other distributions that are equally arbitrary, we conducted a robustness check by allowing card-specific fixed effects. Specifically, we treat all card qualities $\{q_i\}$ as free parameters. This is the least constrained model and can accommodate any empirical distribution of the underlying card quality. The relevant estimation details are contained in Appendix. The identifiable parameters from the fixed effects approach generate qualitatively similar results as the random effects approach: cutoffs are ranked in the same order, and relative magnitudes are similar. This consistency provides confidence that the main results of our paper are robust to the distributional assumption for the underlying card quality.

To summarize, the first experiment has two main findings: (1) the incumbent certifier produces a signal that is as noisy as individual dealers, but later entrants improve in signal precision; (2) later entrants also differentiate in grading cutoffs, as a result the whole system encompassing all three certifiers is much finer than any certifier alone.

These findings are consistent with the theoretical literature about certifiers, but they raise two economic questions: first, if a certifier has a better signal than anybody else in the market, does the market understand the information conveyed in the certificate? If the answer is no, certifiers may lack the incentives to gather and release such information. We address this question by analyzing the relationship between trading price and grading cutoffs. The second question pertains to the information role of professional certifiers. In theory, if a certifier's signal noise is independent of the noise in a trader's self evaluation, the certificate will always help the trader improve his knowledge on the underlying quality of the card. However, to what degree a professional certificate provides new information to various card traders is an empirical question. The second field experiment intends to shed light on this question.

IV.2 Mapping grading criteria with price data

There are two reasons to believe that understanding multiple grading standards is not a trivial task. As shown in the first experiment, even experienced dealers do not have a more precise signal than any of the three professional certifiers. This implies that individual traders face a challenge of separating grading noise from grading criteria. While the numerical grades adopted within each grading standard imply an obvious ordinal rank, the grades across certifiers are not directly comparable. Without an experiment like ours, it is difficult to conclude whether BGS 9 is above or below SGC 10. These difficulties raise a natural concern that a market that lacks the ability to understand multiple grading scales may motivate certifiers to shirk in grading efforts thus undermining the fundamental role of professional certification.

Because our first field experiment identifies the certifiers' grading criteria independent of market price, we can contrast the estimated grading criteria with the perceived criteria as revealed by the market price. If our experimental approach provides meaningful estimates and the market understands the fundamental differences across multiple grading standards, then we should observe a consistent mapping.

To implement our approach, we take the Beckett "low" book price as a proxy of market-clearing price (Jin and Kato (2006) have shown a close relationship between market transaction price and the Beckett "low" price). Our price sample consists of 32 baseball cards that were similar to our experimental cards (i.e., identical technologies), and have detailed book prices by grade and certifier.[7] We use Beckett guides dated February 2002–October 2003 to maximize sample size. Defining the unit of observation as card-certifier-grade, we have 2,022 observations in total, and all available grades are 8 or above. To deal with demand changes across cards and over time, we deflate each price by the PSA 8 price of the same card in the same month. So a deflated price of 2 should be interpreted as 200 percent of its benchmark price. We then compute the average of deflated prices by grade and certifier.[8]

Figure 2 plots grading cutoffs in the upper panel and contrasts them with the average deflated prices in the lower panel. In the upper panel, the horizontal axis is the grading cutoffs estimated in the full model, and the vertical axis is the grading scale ranging from 7 to 10. Each vertical line in the graph denotes the grading cutoff for a specific grade and a specific certifier. To distinguish among certifiers, we use blue lines

[7] The card identities are 1989 Upper Deck #1 Ken Griffey Jr., 1989 Upper Deck #25 Randy Johnson, 1990 Leaf #220 Sammy Sosa, 1990 Leaf #300 Frank Thomas, 1990 Upper Deck #17 Sammy Sosa, 1991 Bowman #569 Chipper, 1991 Upper Deck Final Edition 2F Pedro Martinez, 1992 Bowman #82 Pedro Martinez, 1992 Bowman #461 Mike Piazza, 1992 Bowman #532 M. Ramirez, 1993 Bowman #511 Derek Jeter, 1994 Upper Deck #24 Alex Rodriguez, 1995 Bowman's Best #B2 Vlad Guerrero, 1995 Bowman's Best #B7 A. Jones, 1998 Fleer Tradition Update #U87 T. Glaus, 1998 Fleer Tradition Update #U100 Drew, 1999 Bowman #350 A. Soriano, 1999 Fleer Tradition Update U5 A. Soriano, 1999 Topps Traded T65 A. Soriano, 1991 Upper Deck Final #17F Thome, 1999 Upper Deck Ultimate Victory #136 A. Soriano, 2001 SP Authentic #211 Prior, 2001 SP Authentic #212 Teixeira, 2001 SP Authentic #91 Ichiro Isuzu, 2001 SP Authentic #126 Pujols, 2001 Upper Deck Victory #564 Ichiro, 2001 Bowman #254 Pujols, 2001 SPx #206 Pujols, 2001 Upper Deck #295 Pujols, 2001 Upper Deck Sw Spt #121 Pujols, and 2001 Upper Deck Sw Spt #139 Prior.

[8] Regression analysis controlling for card type and time trend yields the same rank of prices; hence our discussion focuses on the raw averages rather than on regression coefficients.

for PSA, black lines for SGC, and pink lines for BGS. In the lower panel, the horizontal axis is the deflated prices (interpreted as multiples of PSA 8 price) and the vertical axis is the grading scale from 7 to 10. The observed price schedule is a convex, increasing function of grade within each certifier – BGS 9.5 is priced as high as 12.26 times the benchmark price, while that number drops to 2.79 for BGS 9, 1.336 for BGS 8.5, and 1.022 for BGS 8. This confirms the industry understanding that the main action in card grading is to seek a grade at the very high end.

Focusing on ranks, we find that the ordering of grading cutoffs is consistent with the price order. Comparing PSA versus BGS, we find that both cutoffs and prices have $BGS9.5 > PSA10 > BGS9 > PSA9 > BGS8.5 > BGS8 > PSA8$. The relative position of SGC grades at the high end is also consistent: the cutoff (and price) of SGC 10 is less than PSA 10 but higher than BGS 9. The only inconsistency between the two panels is that SGC is usually priced significantly lower than PSA at the same grade, even if their cutoffs are not statistically different. This result could be due to our small sample sizes, or due to a first mover advantage of PSA. BGS is better able to overcome this disadvantage, likely because it is more precise and strategically differentiates at the high end.

IV.3 Experiment Two

The first experiment allows us to compare the three professional certifiers while using three dealers as a common comparison group. Because it focuses on grading criteria and the number of dealers is small, the experiment does not lead to a convincing conclusion of how a professional certificate changes a trader's information set and how such change differs across different types of card traders. Insights in this regard can be obtained from another field experiment we carried out in 1997. At that time, PSA was the only professional certifier.

Experimental Design The goal of the experiment is to detect whether the PSA grade of sportscard quality delivers information to dealers and non-dealers. The experiment was carried out on the floor of a sportscard show located in a major Southern city in 1997. It consisted of four steps: (1) we auctioned 4 ungraded sportscards and determined the winner, (2) we purchased the cards back from the auction winners,[9] (3) we immediately had PSA grade the cards via their 1-hour, $50 per card, on-site grading system, and (4) we auctioned the same card as a graded variant. The entire procedure took place at the same card show in the morning or afternoon, allowing us to match the cards identically across the ungraded/graded treatment, and to control whatever factors might affect the demand for sportscards over time or across locations.[10]

Each participant's auction experience typically followed three steps: (1) inspecting the good, (2) learning the rules, and (3) concluding the transaction. In Step 1, a potential subject approached the experimenter's table and inquired about the sale of the sportscard displayed on the table. The experimenter then invited the potential subject to take about five minutes to participate in an auction for the sportscard displayed on the table. In Step 2, the subject learned the allocation rules. To perform the simplest possible test of the effect of information on bids, we chose an allocation mechanism–William Vickrey's (1961) second-price auction–which has proven straightforward in other field experiments (List 2001). To ensure that the graded and ungraded auctions could be run in the same few hours, we limited the number of participants to 30 in each auction, 15 dealers and 15

[9] We were able to re-purchase all four of the ungraded cards from the auction winners at, or just above, the winner's bid.

[10] We also considered reversing the order (i.e., auctioning off graded cards, buying them back, cracking the seal, auctioning off the identical ungraded cards), but we wished to avoid inadvertently damaging the cards when cracking the seals, which would lead to incorrectly rejecting the null of a treatment effect because the ungraded card would not be the "identical" card of the graded card.

non-dealers.

Finally, in Step 3 the subject filled out a survey (the survey and auction instructions are in the spirit of List (2001; 2002)), after which the experimenter explained that the subject should return at the top of the hour to find out the results of the auction (in some cases the auction did not "clear" until the top of the next hour). If a subject did not return for the specified transaction time, she would be contacted and would receive her cards in the mail (postage paid by the experimenter) within three days of receipt of her payment. For each ungraded auction, we also asked the participating subject what PSA grade she thought the auctioned card would receive if it were graded.

We followed several steps to maintain experimental control. First, no subjects participated in more than one treatment. Second, if the individual agreed to participate, she could pick up and visually examine each card (in sealed cardholders, with the graded card condition clearly marked if they were participating in the graded auction). The experimenter worked one-on-one with the participant, and imposed no time limit on her inspection of the cards. Third, treatment type was changed at the top of each hour, so subjects' treatment type was determined based on the time they visited the table at the card show. To further control for temporal selection effects, the ungraded/graded auctions were paired so the bidding in any ungraded/graded pair took place in either the morning or the afternoon. Further, our dealer table was situated at the front of the card show and thus consumers entering the market were the auction participants. Finally, the sportscard market naturally includes subjects of varying experience. Thus, we can capture the distinction between those consumers that have intense market experience (dealers) and those that have less market experience (nondealers). Limiting each auction to 15 dealers and 15 non-dealers, we could not find any significant demographic difference between bidders in the ungraded session and bidders in the graded session. This guarantees that each ungraded/graded pair highlights the change in information rather than any selection by the grading status.

Results Table 5 summarizes the 4x2 experimental design. In total, we observed data from 240 subjects: 120 bids and expected grades for ungraded cards, and 120 bids for graded cards. The table can be read as follows: row 1, column 1 shows that 15 dealers and 15 non-dealers placed bids for the ungraded Ripken Jr. 1982 *Topps* card. The median non-dealer believed the card would grade at PSA 7 if it were graded (s.d. = 3.3), and bid on average \$27.9 (s.d. = \$40.9). The median dealer believed the card would grade at PSA 8 if it were graded (s.d. = 0.6), and bid on average \$41.0 (s.d. = \$20.6).

Data suggest two differences between dealers and non-dealers: first, dealers predicted the PSA grade much better than the non-dealers. Dealers are not only more likely to expect the actual PSA grade at the median, but also exhibit much smaller variance in the expected grade. Second, while the mean and variance of nondealers' bids are considerably influenced by the PSA certificate, dealers are largely unaffected. For nondealers, both parametric and non-parametric Mann-Whitney tests suggest that the bid distributions observed across the graded and ungraded auctions are statistically different at the $p < .05$ level for the Ripken, Thomas, and Griffey card. No statistical significance is achieved for the Sanders card, probably because the non-dealers expected the PSA grade correctly at the median. Furthermore, the bid variances in all four of the graded auctions are significantly less than the bid variances in each of the ungraded auctions at the $p < .05$ level. Alternatively, neither the bid mean nor variance is significantly different across the graded and ungraded cards in the dealer data at conventional levels.

Based on Table 5, we reach two conclusions: first, dealers know more about card quality than non-dealers; second, the information revealed by the PSA certificate results in significant changes in the non-dealers' bidding distribution, but no significant changes in the dealers' bidding distribution.

Changes in the bidding distribution are subject to many possibilities. To give a sense of what settings we view our results as most relevant, consider the ungraded auction as an auction where every bidder receives one private value signal and one common value signal. The private value signal is independent across bidders. But the common value signal is equal to the sum of the unknown true quality plus noise. Though the noise is independent across bidders, the common value signals are associated by the true quality. Some bidders (say dealers) know more about the common value because their common value signals are less noisier. When the professional grade is made available, it releases a piece of public information on top of each bidder's private signals. We take the professional grade as another noisy proxy of the true quality. Though the auction literature has devoted enormous effort to examining the impact of public information on auction revenue (e.g., Milgrom and Weber 1982), it does not provide any specific prediction on the bidding strategy, especially in the presence of asymmetric bidders in a sealed second-price auction.

Under this framework, the publicized PSA grade potentially yields two changes in the bidding strategy: first, it provides new information about card quality, resulting in an update in the bidder's private evaluation of the card (unconditional on winning or losing the auction). Because the submitted bid is always an increasing function of the underlying evaluation, the change in evaluation in turn leads to a change in the submitted bid. If this is the primary reason driving the bidding difference between dealers and non-dealers, then the results suggest that non-dealers re-evaluate the card to a significant extent after observing the PSA grade, but dealers do not.

The second possibility is that the PSA grade reduces the uncertainty the bidder faces, thus allowing the bidder to bid more aggressively. In other words, the public information leads to a reduction in the winner's curse. If this is the main reason for the bidding difference between dealers and non-dealers, then this effect must be more prevalent for the non-dealers than for the dealers, suggesting the information is more useful for the non-dealers.

We cannot distinguish between the two explanations without a mapping of a specific bidding function (which depends on model assumptions and often involves multiple equilibria). Since the dealers' bidding distribution changes little (in both mean and variance) upon the release of the PSA grade, however, we conclude that neither effect occurs for dealers and therefore the PSA certificate adds little new information to dealers. Alternatively, regardless of the exact mechanism underlying the bidding function, the PSA grade must provide a significant amount of new information to non-dealers, as their distribution has significant changes in both the mean and variance.

The insignificant dealer response to the PSA grade revelation seems inconsistent with the strong theoretical notion that any signal that contains independent noise should help a card trader to improve his information on card quality. Such inconsistency can be attributed to at least two reasons: first, dealers' bids have a much tighter distribution than non-dealers' bids, and the sample size may be too small to detect statistical changes in a tight distribution. Second, sportscards may have both private and common value to collectors. If the private value is iid across collectors, it is statistically indistinguishable from the evaluation noise.[11] But private value, by definition, is unaffected by the

[11] The structural model as described for the first experiment remains valid in this new framework. If we

publication of the PSA grade. If most variation across dealers is due to their difference in private value, this variation remains regardless of how each dealer makes use of the PSA grade to update his view on the common value. This potentially explains the lack of dealers' response to the PSA grade. Unfortunately, data limitations prohibit us from separating these two explanations. Under either interpretation, however, our findings suggest that the PSA grade is more informative to non-dealers than to professional dealers, thus reducing the information asymmetry between the two types of card traders.

V. Concluding Comments

This paper uses two field experiments to explore the information content of professional certifiers in an evolving certification market. Our findings indicate that the actual role of professional certificates goes beyond solving the lemons problem: when neither party of the trade possesses perfect information about the product quality, professional certificates may provide valuable information to both sides of the market. In our case, such a result hinges critically on the role of competition in the certification market. The first certifier provided certificates that credibly distinguished lemons from non-lemons for the uninformed party, but added little information to experienced players in the market. Since the first certifier is committed to maintaining consistency in its grading criteria, new entrants compete by utilizing more precise signals and differentiated grading cutoffs. In doing so, the subsequent entrants enrich the overall grading scale used in the market and therefore provide information that is potentially useful to *all* trading participants, including well-informed sellers.

The fact that new entrants improve the information content of professional

---

allow iid private value in addition to evaluation error, the only interpretation change is that the sum of private value and evaluation noise has about the same variance between PSA and dealers. If we assume zero private value for professional graders and some private value for dealers, our results suggest that the certificates depends on two industrial features: first, there has been an unexpected demand shock that increased the demand for professional certificates. Second, the incumbent certifier is committed to maintaining one grading standard over time. In the absence of either, the incumbent certifier could have adopted or adjusted its standard to meet the new demand. While the two conditions restrict our ability to extending the findings to other certification industries, they facilitate the empirical account of grading differentiation. As shown in Hvide and Heifetz (2001), grading differentiation could arise in a general model of certifier competition. Empirically, grading differentiation is common in almost every certification industry, and the differentiation could be vertical along one dimension (such as sportscard quality and bond default risk) or horizontal across many dimensions (like in restaurants, colleges and health plans).

An important normative consideration is that new entrants in a professional certification market might provide both benefits and costs, and therefore may not unequivocally be welfare-improving. The benefits arise from better information content embedded in the entrants' grading scales that are often finer and differentiated. Given that there is a fair amount of noise in the new and old grading systems, however, the increased competition in the certification industry might generate incentives for repeated grading, which possibly results in duplicate and excessive certification. Another cost lies in learning the market positioning of the new grader—for every new certifier, the market not only needs to learn its grading criteria, but also must determine the relative position of the newcomer's grading scale to that of all existing certifiers. Since each individual often has less information than any one certifier, this learning process could be long and costly. On this front, any normative model would require more formal theoretical

---

evaluation error of PSA is at least as noisy as that of the dealers.

structure.

References

Akerlof, George (1970): “The Market for ‘Lemons’: Qualitative Uncertainty and the Market Mechanism”, *Quarterly Journal of Economics* 84: 488-500.

Berger, Allen N.; Sally M. Davies and Mark J. Flannery (2000): "Comparing Market and Supervisory Assessments of Bank Performance: Who Knows What When?," *Journal of Money, Credit and Banking*, 32(3) Part 2: What Should Central Banks Do? August 2000, pp. 641-667.

Biglaiser, Gary (1993): “Middlemen as Experts,” *The RAND Journal of Economics*, 24: 212-223.

Blackwell, D. (1953), “Equivalent Comparison of Experiments," *Annals of Mathematic Statistics*, 24:265-272.

Cantor, Richard and Frank Packer (1997): “Differences of Opinion and Selection Bias in the Credit Rating Industry,” *The Journal of Banking and Finance*, 21: 1395-1417.

Cantor, Richard, Frank Packer, and Kevin Cole (1997): “Split Ratings and the Pricing of Credit Risk,” *The Journal of Fixed Income*, December 1997:72-82.

Chaney, Paul K.; Debra Jeter and Lakshmanan Shivakumar (2004) "Self-Selection of Auditors and Audit Pricing in Private Firms." *Accounting Review*, Jan2004, Vol. 79 Issue 1, p51-72.

Dewan, S. and Vernon Ning Hsu, (2005) “Adverse Selection in Reputations-Based Electronic Markets: Evidence from Online Stamp Auctions,” Working Paper.

Ederington, Louis H. and Jeremy C. Goh (1998): “Bond Rating Agencies and Stock Analysts: Who Knows What?,” *The Journal of Financial and Quantitative Analysis*,” 33(4): 569-585.

Franzoni, L.A. (1999) "Imperfect Competition in Certification Markets," in Bortolotti, B. and Fiorentini, G. (eds.), *Organized Interests and Self Regulation: An Economic Approach*, Oxford University Press, pp. 158-176.

Grier, Paul and Steven Katz (1976): “The Differential Effects of Bond Rating Changes Among Industrial Public Utility Bonds by Maturity,” *Journal of Business* 49: 226-239.

Grossman, Sanford (1981): “The Informational Role of Warranties and Private Disclosure about Product Quality.” *Journal of Law and Economics*, 24: 461-489.

Guerra, Gerardo A. (2001) "Certification Disclosure and Informational Efficiency: A Case for Ordered Rankings of Levels" *University of Oxford Department of Economics Discussion Paper Series.*

Hand, J., R. Holthausen, and R. Leftwich (1992): “The Effect of Bond Rating Agency Announcements on Bond and Stock Prices,” *The Journal of Finance* 57: 733-752.

Healy, Paul M., and Krishna G. Palepu (2001): “Informational Asymmetry, Corporate Disclosure, and the Capital Markets: A Review of the Empirical Disclosure Literature,” *Journal of Accounting and Economics*, 31: 405-440.

Hettenhouse, George W. and William L. Sartoris (1976): “An Analysis of the Informational Value of Bond-Rating Changes,” *Quarterly Review of Economics & Business*, Summer 1976, Vol. 16: 65-78.

Hsiao, Cheng (1991) "Identification and Estimation of Dichotomous Latent Variables Models Using Panel Data" *The Review of Economic Studies*, 58(4): 717-731.

Hsiao, Cheng (1986), *Analysis of Panel Data*, Cambridge University Press, 1986.

Hvide, Hans K. and Aviad Heifetz (2001): “Free-Entry Equilibrium in a Market for Certifiers" Norwegian School of Economics, *working paper*.

Ippolito, Pauline M. and Alan D. Mathios (1990): "Information, Advertising and Health Choices: A Study of the Cereal Market" *RAND Journal of Economics*, 21(3): 459-480.

Jin, Ginger Z. and Phillip Leslie (2003): "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards" The Quarterly Journal of Economics, 118(2): pp. 409-452.

Jin, Ginger Z. and Andrew Kato (2006): “Price, Quality and Reputation: Evidence From an Online Experiment,” *forthcoming RAND Journal of Economics*.

Katz, Steven (1974): “The Price and Adjustment of Bonds to Rating Reclassifications: A Test of Bond Market Efficiency,” *The Journal of Finance*, 29(2): 551-559.

List, John A. (2001): “Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards,” *American Economic Review* (2001), 91[5]: 1498-1507.

List, John A. (2002): "Preference Reversals of a Different Kind: The More is Less Phenomenon," *American Economic Review*, 92 (5): pp. 1636-1643.

Lizzeri, Alessandro (1999): “Information Revelation and Certification Intermediaries,” *The RAND Journal of Economics*, Summer 1999.

Milgrom, Paul R. (1981): “Good News and Bad News: Representation Theorems and Applications”, *The Bell Journal of Economics*, 12: 380-391.

Neyman, J. and Elizabeth L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations" *Econometrica*, 16(1):1-32.

Pike, Gary (2004): ”Measuring Quality: A Comparison of US News Rankings and NSSE Benchmarks” Research in Higher Education 45(2), March 2004.

Pinches, George E. and J. Clay Singleton (1978): “The Adjustment of Stock Prices to Bond Rating Changes,” *The Journal of Finance*, 33(1): 29-44.

Scanlon, Dennis P., M. Chernew, S. Sheffler, and A. M. Fendrick (1998). “Health Plan Report Cards: Exploring Differences in Plan Ratings.” *Journal on Quality Improvement*, 24(1), pp. 5-20, January.

Thompson, G. Rodney and Peter Vaz (1990): "Dual Bond Ratings: A Test of the Certification Function of Rating Agencies," *The Financial Review*, 25(3): pp. 457-471.

U.S. Securities and Exchange Commission (SEC) (2003): "Concept Release: Rating Agencies and the Use of Credit Ratings under the Federal Securities Laws" available at http://www.sec.gov/rules/concept/33-8236.htm.

Vickrey, William (1961): “Counterspeculation, Auctions, and Competitive Sealed Tenders.” *Journal of Finance*, 1961, 16(1): 8-37.

Wakeman, L.M., (1981), "The Real Function of Bond Rating Agencies," Chase Financial Quarterly, 18-26.

Weinstein, Mark (1977): “The Effect of A Rating Change Announcement on Bond Price” *Journal of Financial Economics* 5: 29-44.

Wimmer, Bradley S., and Brian Chezum (2003): “An Empirical Examination of Quality Certification in a `Lemons Market’,” *Economic Inquiry*, April 2003, 41(2): 279-91.

Appendix: Fixed Effects Robustness Check

Under the fixed effects approach, the likelihood function is:

$$L = \prod_{i=1}^{212}\prod_{j=1}^{6}\left\{\sum_{g} 1_{i,j,g} \cdot \left[\Phi\left(\frac{J_{g+} - q_i}{\sigma_j}\right) - \Phi\left(\frac{J_g - q_i}{\sigma_j}\right)\right]\right\}$$

This introduces a renormalization problem. Should the grades be continuous, $\{q_i\}$ would have been identified as card fixed effects. When grades are ordinal with unknown cutoffs and unknown noise, however, it is possible to renormalize the structure. Specifically, we can take one grader (j') as a benchmark, redefine the true card quality as $\tilde{q}_i = q_i + \varepsilon_{ij'}$, and transform the signal error as $\tilde{\varepsilon}_{ij'} = 0$ for grader j' and $\tilde{\varepsilon}_{ij} = \varepsilon_{ij} - \varepsilon_{ij'}$ for grader $j \neq j'$. This renormalization treats grader j' to be as precise as observing the truth, which results in perfect prediction for grader j' (i.e. $\tilde{\sigma}^2_{j'} = 0$), and an increase of grading noise for the other graders (from σ^2_j to $\tilde{\sigma}^2_j = \sigma^2_j + \sigma^2_{j'}$). The optimal strategy in terms of maximum likelihood is to choose the least noisy grader as the benchmark.

We maximize (1) by choosing the true quality of every single card $\{q_i\}$, the grading cutoffs $\{J_g\}$, and the grading precision $\{\sigma_j\}$. The computation converges to selecting BGS as the zero-noise benchmark. This is not surprising given the fact that both Tables 3 and 4 suggest BGS to be the most agreeable grader. When we exclude BGS from the data set, the algorithm converges to picking the second least noisy grader – SGC – as the benchmark. Such a pattern confirms our intuition: with no knowledge of the true quality, it is difficult to measure how noisy an expert grader is relative to the truth. Rather, we learn which grader is more precise than the others.

Setting one grader as the benchmark introduces another identification problem, however. By definition, the benchmark grader has zero noise and therefore his ordinal grades would be perfectly predicted conditional on the true card quality. If the benchmark grader assigns grade g to all cards with $\tilde{q} \leq q_0$ and grade $g+1$ to all cards with $\tilde{q} \geq q_0 + x$, his grading cutoff for grade $g+1$ could be anywhere between q_0 and $q_0 + x$. In other words, the overall likelihood function has a flat area at the maximum and cannot find a unique solution for the benchmark grader's grading cutoffs. The under-identification will prevent us from comparing the grading criteria across graders.

The random effects approach avoids the renormalization problem because the quality distribution is set different from the noise distribution.[12] Random effects also avoid the incidental parameter problem that exists for most fixed effects estimation with short panels (Neyman and Scott 1948; Hsiao 1986; 1991). Adopting an arbitrary rule to determine the benchmark grader's cutoffs,[13] we can obtain the fixed effects results.

[12] In practice, we set $F(.)$ as beta, and the noise distribution as normal.

[13] We adopt a sequential procedure. First, taking a set of true card quality as given, we identify grading cutoffs and grading precisions by ordered probit. Second, given the estimated grading cutoffs and precisions, we choose the true card qualities to maximize the likelihood and iterate the two steps until all parameters converge. When the algorithm identifies the benchmark grader and sets its grading noise to zero, we compute the benchmark graders' cutoff J_g as the average between the highest card quality with grade g-1 and the lowest card quality with grade g. Standard errors are bootstrapped under the same rule. Detailed algorithm description and estimation results are available at http://www.glue.umd.edu/~ginger/research/.

Table 1. Field experiment: the round-robin design

Total 216 Cards	PSA	SGC	BGS	Kevin	Rick	Rodney
Card Group A	Round 1 Step 2	Round 2 Step 2	Round 3 Step 2	Round 1 Step 1	Round 3 Step 1	Round 2 Step 1
Card Group B	Round 2 Step 2	Round 3 Step 2	Round 1 Step 2	Round 1 Step 1	Round 3 Step 1	Round 2 Step 1
Card Group C	Round 3 Step 2	Round 1 Step 2	Round 2 Step 2	Round 1 Step 1	Round 3 Step 1	Round 2 Step 1
Card Group D	Round 1 Step 2	Round 2 Step 2	Round 3 Step 2	Round 2 Step 1	Round 1 Step 1	Round 3 Step 1
Card Group E	Round 2 Step 2	Round 3 Step 2	Round 1 Step 2	Round 2 Step 1	Round 1 Step 1	Round 3 Step 1
Card Group F	Round 3 Step 2	Round 1 Step 2	Round 2 Step 2	Round 2 Step 1	Round 1 Step 1	Round 3 Step 1
Card Group G	Round 1 Step 2	Round 2 Step 2	Round 3 Step 2	Round 3 Step 1	Round 2 Step 1	Round 1 Step 1
Card Group H	Round 2 Step 2	Round 3 Step 2	Round 1 Step 2	Round 3 Step 1	Round 2 Step 1	Round 1 Step 1
Card Group K	Round 3 Step 2	Round 1 Step 2	Round 2 Step 2	Round 3 Step 1	Round 2 Step 1	Round 1 Step 1

Notes: Round 1 in blue, Round 2 in black, and Round 3 in pink. The total number of cards in use is 216. Four of them were damaged, so the final sample size is 212.

Table 2. Field Experiment: Grade Distribution by Grader

	PSA	BGS	SGC	KEVIN	RICK	RODNEY
4	0	0	0	0	1	0
4.5		0		0	0	0
5	0	0	0	0	0	0
5.5		0	0	0	0	0
6	0	0	0	0	1	2
6.5		0		0	0	0
7	1	2	2	1	2	0
7.5		3	3	4	3	2
8	66	43	11	37	45	25
8.5		124	49	129	92	62
9	134	40	134	40	57	120
9.5		0		1	11	1
10	11	0	13	0	0	0
Total	212	212	212	212	212	212

Notes: Each cell represents frequency. Blank means the grade is not applicable to the grader.

Table 3. Summary Statistics by Degree of Consistency

Panel A: % strongly consistent (both graders said A>B, A=B or A<B)

	psa	bgs	sgc	kevin	rick	rodney
PSA	1.000					
BGS	0.491	1.000				
SGC	0.537	0.465	1.000			
Kevin	0.409	0.399	0.418	1.000		
Rick	0.377	0.492	0.414	0.402	1.000	
Rodney	0.408	0.492	0.475	0.428	0.429	1.000
sum (except self)	2.223	2.339	2.308	2.057	2.114	2.232
average (except self)	0.445	0.468	0.462	0.411	0.423	0.446
Ranks by average	4	1	2	6	5	3

Panel B: % strongly inconsistent (one grader said A>B, and the other said A<B)

	psa	bgs	sgc	kevin	rick	rodney
PSA	0.000					
BGS	0.059	0.000				
SGC	0.053	0.070	0.000			
Kevin	0.111	0.109	0.100	0.000		
Rick	0.130	0.089	0.109	0.131	0.000	
Rodney	0.111	0.069	0.091	0.103	0.118	0.000
sum (except self)	0.463	0.396	0.423	0.554	0.577	0.492
average (except self)	0.093	0.079	0.085	0.111	0.115	0.098
Ranks by average	3	1	2	5	6	4

Panel C: % weakly inconsistent (one grader said A=B and the other said A>B or A<B)

	psa	bgs	sgc	kevin	rick	rodney
PSA	0.000					
BGS	0.450	0.000				
SGC	0.411	0.465	0.000			
Kevin	0.480	0.492	0.482	0.000		
Rick	0.493	0.419	0.478	0.467	0.000	
Rodney	0.481	0.438	0.435	0.469	0.453	0.000
sum (except self)	2.314	2.265	2.269	2.389	2.309	2.276
average (except self)	0.463	0.453	0.454	0.478	0.462	0.455
Ranks by average	5	1	2	6	4	3

Table 4. Full Model Estimation

Panel A: Estimates

	PSA		SGC		BGS		KEVIN		RICK		RODNEY	
	coeff.	std.err.	coeff.	std.err.	coeff.	std.err.	coeff.	std.err.	coeff.	std.err.	coeff.	std.err.
σ	0.1553	0.0287	0.1218	0.0212	0.0909	0.0165	0.2518	0.056	0.1624	0.0268	0.1505	0.0256
cutoff 6									0.1401	0.1376		
cutoff 7									0.1841	0.1300		
cutoff 7.5			0.2489	0.1227	0.3103	0.1141	-0.0623	0.1963	0.2412	0.1243	0.2014	0.1341
cutoff 8	0.1481	0.1404	0.3118	0.1185	0.3616	0.1121	0.1038	0.1585	0.2908	0.1209	0.2532	0.1282
cutoff 8.5			0.4145	0.1164	0.5497	0.1142	0.4255	0.1217	0.5228	0.1143	0.4502	0.1184
cutoff 9	0.5691	0.1146	0.5778	0.1147	0.7924	0.1129	0.8995	0.126	0.7545	0.1148	0.6317	0.1144
cutoff 9.5							1.3810	0.2047	0.9824	0.1216	1.1315	0.1308
cutoff 10	0.9732	0.1201	0.9149	0.1132								

Note: Assume the true card quality conforms to an iid Beta distribution on the support of (0,1) with two free parameters $0 < a \leq 10$ and $0 < b \leq 10$. Maximum likelihood identifies the cutoffs, the grading precisions, and the beta distribution parameters simultaneously. Blank cells indicate non-applicable.

Table 4 Panel B: Test of significant difference across grading cutoffs

Null hypothesis for cell (ij) : cutoff in row i = cutoff in column j

PSA vs. SGC

	SGC 7.5	SGC 8	SGC 8.5	SGC 9	SGC 10
PSA 8	-0.1008	-0.1637 *	-0.2663 ***	-0.4296 ***	-0.7668***
	(0.1037)	(0.0980)	(0.0938)	(0.0927)	(0.1031)
PSA 9	0.3202 ***	0.2572 ***	0.1546 ***	-0.0087	-0.3458***
	(0.0615)	(0.0491)	(0.0360)	(0.0241)	(0.0411)
PSA 10	0.7243 ***	0.6614 ***	0.5588 ***	0.3955 ***	0.0583
	(0.0820)	(0.0725)	(0.0627)	(0.0530)	(0.0549)

PSA vs. BGS

	BGS 7.5	BGS 8	BGS 8.5	BGS 9
PSA 8	-0.1621	-0.2135 ***	-0.4016 ***	-0.6443 ***
	(0.1000)	(0.0958)	(0.0931)	(0.0954)
PSA 9	0.2588 ***	0.2074 ***	0.0194	-0.2234 ***
	(0.0485)	(0.0385)	(0.0237)	(0.0262)
PSA 10	0.663 ***	0.6116 ***	0.4236 ***	0.1818 ***
	(0.0689)	(0.0626)	(0.0526)	(0.0498)

SGC vs. BGS

	BGS 7.5	BGS 8	BGS 8.5	BGS 9
SGC 7.5	-0.0614	-0.1127 *	-0.3008 ***	-0.5436 ***
	(0.0740)	(0.0679)	(0.0620)	(0.0620)
SGC 8	0.0016	-0.0498	-0.2378 ***	-0.4806 ***
	(0.0638)	(0.0566)	(0.0492)	(0.0498)
SGC 8.5	0.1042 *	0.0529	-0.1352 ***	-0.378 ***
	(0.0546)	(0.0459)	(0.0352)	(0.0363)
SGC 9	0.2675 ***	0.216 ***	0.0281	-0.2147 ***
	(0.0479)	(0.0378)	(0.0213)	(0.0221)
SGC 10	0.6046 ***	0.5533 ***	0.3652 ***	0.1224 ***
	(0.0563)	(0.0483)	(0.0369)	(0.0371)

Note: For row i column j, we report (the cutoff in row i - the cutoff in column j) with standard error in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All the tests use the estimates reported in Table 4A.

Table 4 Panel C: Test of significant difference across grading precisions

	σ of SGC	σ of BGS	σ of Kevin	σ of Rick	σ of Rodney
σ of PSA	0.0336	0.0644 **	-0.0965	-0.0071	0.0048
	(0.0359)	(0.0325)	(0.0627)	(0.0401)	(0.0398)
σ of SGC		0.0309	-0.13 **	-0.0407	-0.0287
		(0.0299)	(0.0587)	(0.0339)	(0.0325)
σ of BGS			-0.1609 ***	-0.0715 **	-0.0596 *
			(0.0593)	(0.0307)	(0.0305)
σ of Kevin				0.0894	0.1013 *
				(0.0600)	(0.0596)
σ of Rick					0.0119
					(0.0361)

Note: For row i column j, we report (σ in row i - σin column j) with standard error in parentheses. ***p<0.01, ** p<0.05, * p<0.1. All the tests use the estimates reported in Table 4A.

Table 5: Results from the 1997 Auction Field Experiment

Card Type	Ungraded	Graded
Ripken Jr. 1982 *Topps*	n=30 (PSA 7; 2.5) Bid = $34.7 (32.2) Non-dealer bid = $27.9 (40.9) (PSA 7; 3.3) Dealer bid = $41.0 (20.6) (PSA 8; 0.6)	n=30 (PSA 8) Bid= $48.0 (17.2) Non-dealer bid = $51.7 (13.0) Dealer bid = $44.3 (20.3)
Sanders 1989 *Score*	n=30 (PSA 7; 2.2) Bid = $34.3 (32.3) Non-dealer bid = $44.3 (40.8) (PSA 8; 3.0) Dealer bid = $22.0 (15.2) (PSA 7; 1.1)	n=30 (PSA 7) Bid= $30.7 (22.5) Non-dealer bid = $40.2 (24.5) Dealer bid = $21.1 (15.9)
Thomas 1990 *Leaf*	n=30 (PSA 8; 2.3) Bid = $70.8 (43.4) Non-dealer bid = $66.3 (53.5) (PSA 7; 3.2) Dealer bid = $75.3 (31.4) (PSA 8; 0.8)	n=30 (PSA 9) Bid= $90.0 (22.3) Non-dealer bid = $96.9 (21.4) Dealer bid = $83.0 (21.7)
Griffey Jr. 1989 *Upper Deck*	n=30 (PSA 7.5; 2.8) Bid = $41.0 (35.9) Non-dealer bid = $36.7 (47.8) (PSA 5.5; 3.5) Dealer bid = $45.3 (18.7) (PSA 8; 0.8)	n=30 (PSA 8) Bid= $56.3 (22.3) Non-dealer bid = $65.0 (24.6) Dealer bid = $47.6 (16.2)

Notes: Row 1, column 1 shows that 30 bidders placed bids for the ungraded Ripken Jr. 1982 *Topps* card. The median bidder believed the card would grade at PSA 7 if it was graded (s.d. = 2.5). Mean bid was $34.7 (s.d. = 32.2). Non-dealers bid on average $27.9 (s.d. = $40.9) and the median non-dealer believed the card would grade at PSA 7 if it was graded (s.d. = 3.3). Dealers bid on average $41.0 (s.d. = $20.6) and the median dealer believed the card would grade at PSA 8 if it was graded (s.d. = 0.6). Each auction had 15 non-dealers and 15 dealers.

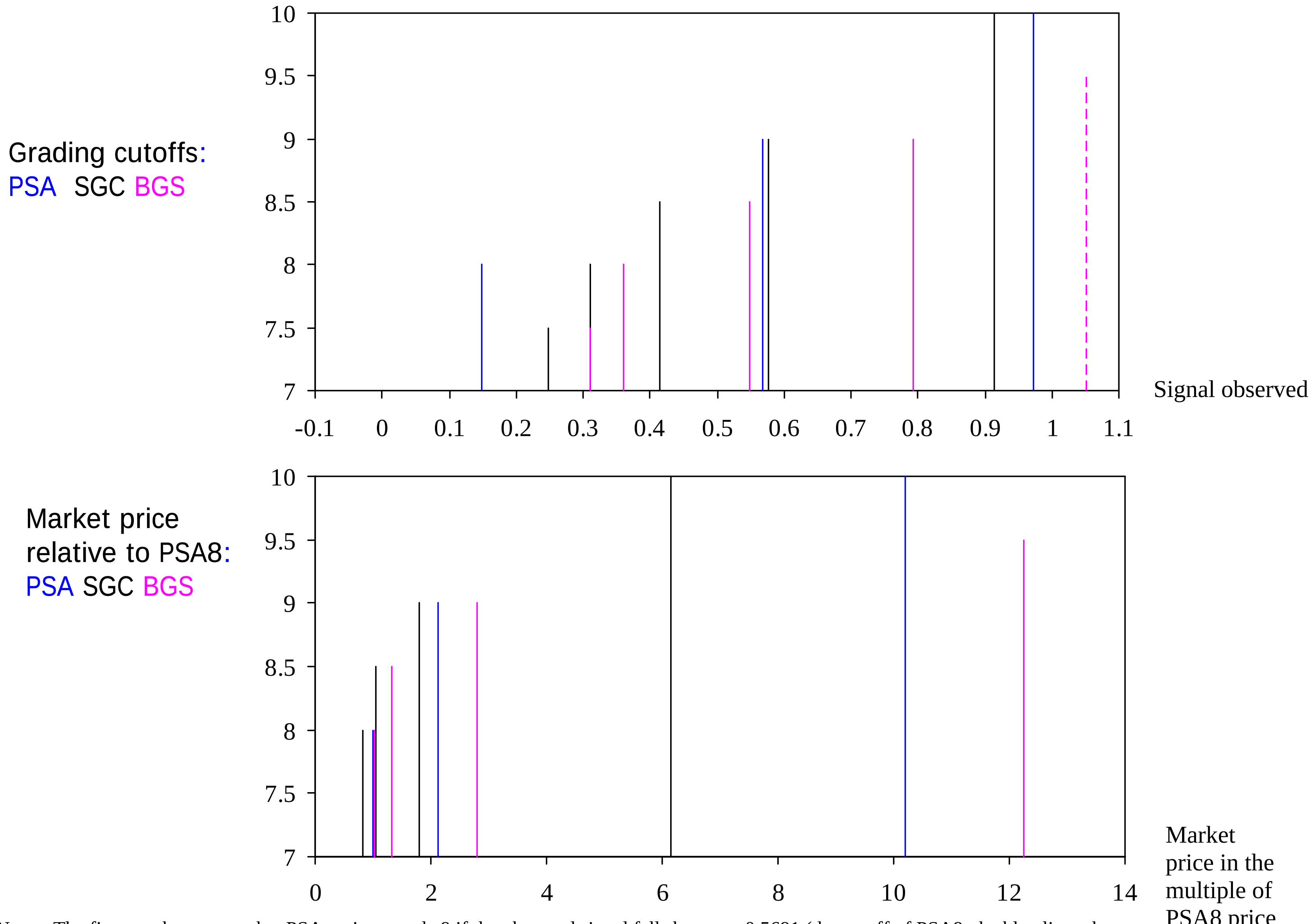
Figure 1. Examples of Graded Cards

BGS (serial number at the back)

SGC (96 is equivalent to 9 in a 1-10 scale)

PSA

Figure 2. Contrast of grading cutoffs and deflated price by grade and grader



Notes: The first graph suggests that PSA assigns grade 9 if the observed signal falls between 0.5691 (the cutoff of PSA9, the blue line whose height equals 9) and 0.9732 (the cutoff of PSA10, the blue line whose height equals 10). The second graph shows that on average the market price of a PSA9 card is 2.137 times of the PSA8 price conditional on the same card type. The magnitude of BGS9.5 cutoff is constructed because we do not observe a BGS9.5. However, the deflated price of BGS9.5 is precisely estimated based on Beckett low book price.