Some Evidence on the Consistency of Banks' Internal Credit Ratings

Mark Carey

Federal Reserve Board

April 21, 2001

This paper was prepared in cooperation with Mark McCambley and Michael Lavin of Loan Pricing Corporation (LPC). For more information about LPC's Loan Loss Database (LLD) initiatives in North America and Europe, please contact Mark McCambley at 212-489-5455 or mmccambley@loanpricing.com. The lenders participating in the LLD generously shared the LLD with the Federal Reserve on a one-time basis as a contribution to efforts to advance risk-focused banking supervision and capital regulation. The views expressed herein are not necessarily those of the Board of Governors, other members of its staff, the Federal Reserve System, Loan Pricing Corporation, or any LLD participant institutions. Thanks to Irina Barakova for excellent research assistance. Correspondence about this paper should be addressed to Mark Carey at the Federal Reserve Board, Washington, DC 20551. (202) 452-2784 (voice), (202) 452-5295 (fax), mcarey@frb.gov.

Major banks and their regulators are increasingly interested in the accuracy and consistency of banks' internal ratings of the credit risk posed by borrowers and loans. In addition to the traditional influence of ratings on loan origination and monitoring, ratings are now used in pricing and capital allocation models and in setting reserves. Moreover, the BIS recently proposed revisions to the Basel Accord under which internal ratings would play a key role in setting regulatory capital requirements for banks.[1]

Especially in quantitative applications, borrower ratings are typically used as indicators of the probability a borrower will default. Grades such as 'E' or '5' are translated into an average probability of default (PD) for borrowers in the grade and the PDs are used as inputs to quantitative models or processes. As a practical matter, the PDs are often the single most important model input, so errors and inconsistencies in ratings and PDs can have substantial effects on pricing decisions, capital allocations, and business strategies. For example, if the PD for a borrower is taken to be 0.25 percent instead of 0.06 percent (used below as average values for BBB/Baa versus A-rated borrowers), the total capital requirement under the proposed Basel internal ratings based (IRB) approach for the borrower's loans would be about 2.5 percent versus 1 percent.[2] Loan interest rate spread hurdles from pricing models might be expected to differ by 15 basis points as a result of such a difference in capital allocations, nearly half the recent average drawn spread on loans to A-rated borrowers.[3]

Frequent inconsistencies between a lender's internal ratings and PDs and those of its competitors can affect origination decisions and competitive positions (by affecting capital allocations and pricing hurdles). Moreover, frequent errors in internal rating assignments can reduce a lender's earnings by reducing the quality and efficiency of

---

[1] Bank for International Settlements, "The New Basel Capital Accord" and associated documents, January 16, 2001 (www.bis.org).

[2] Assuming a 30 percent loss given default and no maturity adjustment.

[3] Suppose the IRB capital allocations are representative of those from banks' internal loan pricing models, and that the marginal cost of capital to a bank for equity and reserves is 10 percent (a 15 percent expected return less a 5 percent cost of debt funding). A 150 basis point difference in capital multiplied by a 10 percent difference in incremental cost of capital equates to a difference in required spread of 15 basis points. The broadly syndicated loan grid in Loan Pricing Corporation's December 25, 2000 issue of *Gold Sheets* reports the average spread for A-rated borrowers as 32.5 basis points.

borrower monitoring and intervention processes. Some inconsistencies across lenders in individual ratings and PDs are to be expected because reasonable people may differ about the credit quality of any given borrower, but biases in ratings or high frequencies of inconsistencies may indicate a need for internal reviews of rating criteria and procedures.

This paper presents evidence about the consistency of rating assignments across lenders using Loan Pricing Corporation's Loan Loss Database (LPC LLD). The data drawn from the LLD has information about individual commercial loans from the portfolios of more than two dozen lenders (all are banks) during 1994-98, including the internal rating assigned to each borrower, with such ratings mapped to a common ten-grade LPC scale.[4] Thus, it is possible to compare the ratings assigned by different lenders to the same borrower at the same time, and to produce summary statistics about the frequency and magnitude of disagreements across lenders.

For the pooled sample of pairs of internal ratings, the ratings are effectively the same in about 45 percent of cases, are within two grades in about 95 percent of cases, and are within three grades in 98 percent of cases. When grades are mapped to PDs using Moody's and S&P's long-run average default rates, and thence to capital allocations using the BIS proposed IRB capital formula, implied capital allocations differ by less than one percentage point for about half the borrowers, by less than two points for about 70 percent of borrowers, and by less than 10 percentage points for over 95 percent of borrowers.

Rating disagreements across lenders are not very predictable from borrower characteristics available in the LLD. Disagreements are somewhat less likely for large borrowers and for borrowers that have not drawn down much on their lines of credit, and perhaps a bit more likely for high-quality borrowers.

Noise in rating assignments is important because of the adverse selection risks and internal inefficiencies that can result, but differences across lenders on average are especially important because they have implications for commercial lines of business as a whole and, under the IRB approaches, for portfolio regulatory capital requirements. Averaging each LLD lender's ratings relative to those of the rest of the pool, no lender was optimistic or pessimistic on average by a full grade, and only five were optimistic or

[4] As described further below, the LLD covers more years and has other variables as well.

pessimistic on average by more than half a grade. However, the relationship between grades, PDs and capital allocations is nonlinear, so small average disagreements about grades can translate into significant disagreements about PDs and capital. Three lenders have average PD differences larger than one percentage point. Ten lenders have average capital allocation differences larger than one percentage point.

Interpretation of these results is somewhat in the eye of the beholder. On the one hand, different lenders are arguably in substantial agreement about the credit quality of about half their borrowers, and rating disagreements appear relatively small on average. On the other hand, individual disagreements appear to be material to capital allocation and pricing decisions for a third or more of sample borrowers, and average disagreements in implied capital allocations are arguably economically significant for nearly half of the participating lenders. In my opinion, the results imply that 1) rating system inconsistencies are a source of material adverse selection and other business risks for many lenders; 2) a lender with a reliable rating system and that has confidence in its system may have a significant competitive advantage in the long run; 3) benchmarking studies like this one can aid understanding and tuning of rating systems; and 4) regulators interested in evaluating rating systems may be able to use results of benchmarking studies to efficiently allocate their examination and review resources across banks.

The PD and capital allocation estimates in this paper are only for illustrative purposes. This paper's evidence is about internal ratings, not "quantification" (the process by which ratings are converted into PDs) and not capital allocation systems, because internal estimates of PDs and capital allocations do not appear in the LLD, only ratings. Lenders' quantification and capital allocation systems are typically separate from their internal rating systems. A lender may have accurate and consistent internal ratings but still have PDs or capital allocations that differ from those of other lenders if the systems that produce them differ, but such differences would not be detected in this study.

Consistent ratings are not necessarily accurate ratings, and this paper's evidence is only about consistency. If borrower ratings are meant to bucket obligors according to their true, unobserved probability of default, then the banking industry as a whole might be consistently wrong about any given borrower or group of borrowers. However, long time series of loss experience data for unchanging rating systems are needed to reveal such industry-wide errors. Until such data become available, the best that any lender or the banking industry can do is make good use of available information.

To my knowledge, only one other benchmarking study has appeared, and it was done somewhat differently. Under the auspices of the Risk Management Association (RMA), a group of eight lenders pooled internal PD estimates as of March 31, 2000 for loan facilities having LPC Loan Identification Numbers (LINs, not part of the LLD), while excluding most loans to borrowers whose equity was known to be publicly traded.[5] The RMA study reports average differences in PD estimates for each lender relative to the others but no information about ratings or capital allocations. The average PD differences range from –0.62 percent to +0.71 percent, somewhat smaller than the range –1.18 percent to +0.79 percent found in this study. This paper and the RMA study are complementary, with one providing direct comparisons of PDs and the other of ratings. In addition, this study covers a longer time period, includes a larger number of lenders, and reports more information about the underlying borrowers and their relationship to differences in rating assignments.

Both studies offer cautionary notes for the conduct of benchmarking exercises in the future. A major barrier to benchmarking is the lack of a common identification system for loans and borrowers, making development of large samples difficult because the matching of observations across lenders is a laborious task. Moreover, differences in facility and obligor ratings, in the granularity of internal rating scales, and in the timing of rating assignments and reviews all complicate interpretation of results, as described further below.

Section 1 and Appendix A describe the LLD and the manner in which this paper's cross-lender comparisons are constructed. Section 2 presents results for the pooled sample, and Section 3 for individual lenders relative to the pool. Section 4 offers some evidence about the likelihood of rating disagreements for different types of borrowers. Section 5 discusses various practical barriers to the conduct of benchmarking studies such as this one and offers ideas about how such studies might be made easier, and Section 6 offers cautionary notes and concluding remarks.

[5] "EDF Estimation: A Test-Deck Exercise," in the RMA Journal, November 2000, pp. 54-61.

1.0 Data and Sample Construction

Lenders that participate in LPC's Loan Loss Database (LLD) contribute information annually about exposed loans from their commercial portfolios (all LLD participants are banks). Contributions do not include all portfolio loans. Instead, each year a stratified sample of newly originated loans is drawn randomly from each lender's portfolio. Information about these loans is collected both for the origination year and for subsequent years until the loans exit the portfolio. In drawing each year's sample, a dollar size cutoff (below which no loans are drawn) and a sample size (number of loans) is established for each lender in order to achieve a pooled sample large enough to support reasonably accurate estimation of default rates for that year's pool. The first samples of exposure data were collected in 1994 and the data used in this study extend through 1998 (recent collections extend the LLD's coverage into 2000).[6]

To support comparisons of rating assignments, a sample of pairs of ratings of the same borrower by different lenders is needed, with ratings on the same scale and current as of a common date. The LLD does not contain information that precisely satisfies these criteria, but it does contain ID fields for lenders, borrowers and loan facilities, and it has ratings and effective dates that can be used to form a serviceable sample of rating comparisons. Identification of multi-lender borrowers using the ID codes is straightforward.[7] The main issues involve the ratings, the dates, and the fact that data records are for loan facilities but ratings in the records are usually at the borrower level. Appendix A describes in detail how such matters are handled. To summarize:

- A given lender's internal grades may be mapped to fractional LPC grades, for example internal "E" to LPC "5.5." Fractional grades are preserved in

---

[6] The LLD includes somewhat separate samples of defaulted loans that extend back to the mid 1980s and which were collected for purposes of estimating recovery rates. Exposure information prior to default is often collected as well as recovery information for such loans. The LLD also includes exposure and default records constructed from LPC's *Dealscan* database. Because both subsamples are generated quite differently than the main LLD data, and because ratings in *Dealscan* (where available) are S&P or Moody's ratings rather than internal ratings, both subsamples are excluded from this study.

[7] LPC of course knows the identity of each lender but records it in the data as an anonymous ID code. The author did not have access to any lender or borrower names. Names of borrowers are also collected and converted into anonymous ID codes, and LPC reportedly tries hard to match borrowers by name across lenders (taking minor differences in spelling and other characteristics into account) so that common borrowers have common ID codes. Loan facilities for each borrower are assigned ID codes as well, but these are specific to a lender. LPC does not attempt to match facilities across banks because of the lack of reliable identifying information.

computing PDs and capital allocations, but for convenience of exposition grade differences are rounded to the full grade level in reporting relative frequencies. For example, a difference from 0 to 0.5 is reported as no difference.

- The "effective dates" of observations for a given borrower in the LLD typically are not exactly the same across lenders. For example, lender A's information about borrower X might be as of May 1, 1996 and lender B's information as of August 18, 1996. In the results reported in the main body of this paper, any dates falling in the same calendar quarter are treated as being "at the same time." Results from two other methods of lining up observations on the timeline are similar and are reported in Appendix A.
- Some LLD observations contain facility grades rather than borrower grades, but this study is intended to be about borrower grades. Where multiple facilities exist for a borrower at a given lender, the riskiest assigned rating is used. Results are robust to elimination of borrowers where ratings differ across facilities.

1.1 Conversion of grades into PDs and capital allocations

The economic importance of differences in grade assignments depends on where on the credit quality spectrum the differences fall. A two grade difference at the safe end, say between AAA and A on S&P's scale, has moderate implications in an absolute sense for interest rate spreads and capital allocations, but two-grade differences between BBB and B, or BB and CCC, have major absolute implications. Because LLD borrowers span the credit quality spectrum and because statistics on average differences across lenders are of interest, it is helpful to convert grades into PDs and capital allocations and to report average cross-lender differences for those items.

To do so, I convert individual grades into average PD values using LPC's mapping from its grades to rating agency grades. Each agency grade is converted to PDs using rounded-off or interpolated long-run average default rates from recent Moody's and S&P studies, as shown in Table 1. Where a borrower carries a fractional grade in the

LLD, I compute its corresponding PD by linear interpolation. (For LPC grades 7 through 10, I set PD values somewhat arbitrarily.)

Given a PD for a borrower, I compute a capital allocation using the Basel Committee's recently proposed IRB formula.[8] The right column of Table 1 reports such allocations for each grade when the mapped PD values are used. As noted previously, PDs and capital allocations are purely to aid economic interpretation of differences of opinions about grades, and are not necessarily indicative of the values used by any particular lender.

Table 1. Illustrative PDs and capital allocations for LPC grades

LPC Grade*	Mapped Equivalent Agency Grade (S&P/Moody's)*	Moody's default rate (percent)	S&P default rate (percent)	Average PD used here (percent)	Implied total capital allocation (percent)
1	AAA,AA/Aaa,Aa	0.04	0.01	0.02	0.53
2	A/A	0.08	0.04	0.06	1.03
3	BBB/Baa	0.30	0.21	0.25	2.49
4	BB+,BB/Ba1,Ba2	0.60	0.75	0.67	4.66
5	BB-/Ba3	2.50	1.14	1.85	8.80
6	B/B	4.50	5.16	4.83	15.60
7	CCC/Caa (oaem)	10	10	10	23.15
8	CC,C/Ca,C (substandard)	20	20	20	30
9	D (doubtful)	100	100	100	30
10	D (loss)	100	100	100	30

* LPC grades 7-10 are designed to correspond to the classifications used by U.S. bank regulators, as shown in the parentheses. The correspondence between agency grades and regulatory classifications is likely to be inexact; I therefore set PDs for grades 7-10 somewhat arbitrarily. Moody's default rates are from "Historical Default Rates of Corporate Bond Issuers, 1920-1999," Moody's Investors Service Special Comment, January 2000, Exhibits 30-31. S&P's are from "Ratings Performance 1999: Stability and Transition, Standard & Poor's, February 2000, Tables 3 and 11. Some numbers from both studies were rounded.

[8] The formula is RW_C= (LGD/50) x BRW_C (PD), or 12.50 x LGD, whichever is smaller. IRB capital requirements cover both expected and unexpected losses and thus correspond roughly to equity and reserves needed per dollar of loan equivalent exposure at default. Loss given default (LGD) is assumed to be 30 percent, no maturity adjustment is included in calculations, PDs shown in tables are in percentage form but in calculations are converted to the decimal form wanted by the formula, and $BRW_C(PD) = 976.5\ x\ N(1.118\ \ x\ \ G(PD)+1.288)\ x\,(1+.0470\ \ x\ \ (1-PD)/PD^{0.44})$, where N() is the standard normal cdf and G() is the inverse of N(). See page 49 of The New Basel Capital Accord, January 16, 2001.

2.0 Frequency and Extent of Rating Disagreements in Pooled Data

Table 2 reports marginal and cumulative frequencies of rating disagreements using the common-quarter method of producing comparable pairs of ratings (described in Appendix A), which yields 2588 pairs. A difference of zero grades means lenders agree. Lenders are in agreement about rating assignments in a bit less than half of the comparable cases, differing by one grade for about another 40 percent of cases, and by more than one grade in about 17 percent of cases. The maximum observed disagreement for any single borrower is 5.5 grades.

Tables 3 and 4 report frequencies of various levels of disagreement across lenders about implied PDs and capital allocations, respectively.[9] Lenders are in exact agreement about grades and thus about implied PDs and capital allocations in about one-quarter of cases.[10] However, cumulative frequencies run up toward 100 percent more slowly than with respect to grade differences. For about one-quarter of sample observations, PD differences are more than 50 basis points and capital allocation differences are more than 2 percentage points. Such differences can be economically substantial, and their existence for so many cases indicates the potential for significant pricing differences across lenders and thus adverse selection risks. Capital allocation differences above 8 percentage points exist for about 8 percent of the comparable observations.

Dramatic disagreements about PDs and capital allocations are somewhat more frequent than dramatic grade disagreements because even modest grade disagreements among the riskier grades have significant implications for PDs and capital. Looking back at the last two columns of Table 1, it appears that single-grade disagreements around the border between investment and speculative grades, and especially among the speculative grades, are most likely to result in economically large implied disagreements about PD and capital. On the one hand, such rating disagreements might be less worrisome because loans to speculative borrowers carry higher spreads and thus in some respects there may be more pricing slack to absorb noise in pricing model hurdle rates. On the

[9] Cross-lender differences are computed by converting each lender's grade into a PD and thence into a capital allocation and taking differences of those values.

[10] As noted previously, the 44 percent of cases with "zero" disagreement reported in Table 1 include disagreements of up to half a grade, and such fractional disagreements are preserved in calculations of PD and capital differences.

Table 2. Frequency of Rating Disagreements Across Lenders for the Same Borrower (percent)

Number of grades of difference*	Common-quarter method N = 2588	
	Marginal frequency	Cumulative frequency
0	44	44
1	39	83
2	12.1	95.1
3	3.3	98.4
4 or more	1.6	100

* '0' means disagreements of half a grade or less, '1' disagreements of more than half a grade but less than or equal to 1.5 grades, etc.

Table 3. Frequency of Implied Probability of Default (PD) Differences Across Lenders for the Same Borrower (percent)

PD Difference (basis points)	Common-quarter method N = 2588	
	Marginal frequency	Cumulative frequency
0	23.8	23.8
>0, ≤10	23.2	47.0
>10, ≤25	20.6	67.6
>25, ≤50	9.1	76.7
>50, ≤100	4.4	81.1
>100, ≤200	9.2	90.3
>200, ≤500	8.7	99.0
>500	1.0	100

Table 4. Frequency of Implied Total Capital Allocation Differences Across Lenders for the Same Borrower (percent)

Capital difference (percentage points)	Common-quarter method N = 2588	
	Marginal frequency	Cumulative frequency
0	23.8	23.8
>0, ≤0.5	19.4	43.2
>0.5, ≤1	7.3	50.5
>1, ≤2	19.2	69.7
>2, ≤4	12.1	81.8
>4, ≤8	10.7	92.5
>8, ≤15	6.9	99.4
>15	0.6	100

other hand, the commercial loan market is highly competitive and thus pricing disagreements tend to influence loan origination volumes. Moreover, because levels of default risk are higher for riskier borrowers, the consequences of pricing-related adverse selection can also be more serious for long-run portfolio performance.

When average rating, PD and capital allocation differences are broken out by year (not shown in tables), no trend toward greater agreement among lenders is evident. The mean rating difference in each year is generally not far from one grade; the mean PD difference is usually not far from 0.75 percentage points; and the mean difference in capital allocations is usually close to 2 percentage points.

Some observers are most interested in the performance of the "pass" grades (1 through 6 on the LPC scale). The "regulatory" grades (7-10) are largely meant to capture credits in or near default or at very high risk of default. When the analysis is restricted to borrowers rated 6 or better, which eliminates 113 observations, the distribution of disagreements is generally similar, but the incidence of very large disagreements across lenders is reduced somewhat. The percentage of PD differences over 200 bps drops from 9.7 to 8.6, and none remain that are larger than 500 bps. The percentage of capital allocation differences larger than 4 percentage points drops from 7.5 to 6.3 percent, and only 0.1 percent are larger than 15 percentage points.

Of the 2588 usable observations of at-least-two-lenders-for-the-same-borrower-at-the-same-time, 1476 of them are part of a sequence, meaning that observations in different quarters are available for the same borrower and pair of lenders.[11] Such sequential observations shed light on the persistence of disagreements about ratings. Table 5 displays the percentage of rating disagreements that stay the same from one period to the next, both overall and for different levels of the early-period disagreement. Overall, disagreements are rather persistent, with about 60 percent of cases showing the same rating disagreement (and individual rating values) from one period to the next. However, large disagreements are less persistent. For example, in only 32 percent of the cases where first-period disagreements are three grades or more do both lenders' ratings remain the same in the next period. These statistics are consistent with some rating

[11] The time between periods is not held constant. Sequential observations for the same borrower and lenders are treated the same whether they are separated in time by one quarter or several years.

disagreements arising from genuine differences of view about borrower credit quality and rating criteria, and with some arising because rating errors are not caught (or not caught quickly) by lenders' credit review activity. [12]

3.0 Are Some Lenders Consistently More Optimistic or Pessimistic Than Average?

Noise in rating assignments and implied PDs and capital allocations is important, but perhaps even more important are differences across lenders on average. Lenders that are substantially more optimistic or pessimistic about credit quality than their peers may or may not turn out to be correct in the long run, but knowing that such differences of view exist is helpful in forming loan pricing and origination strategies and in understanding competitive positions in the commercial lending marketplace.

The left panel of Table 6 (the first six columns) presents unweighted average differences in grade assignments, implied PDs and capital allocations of each lender relative to the rest of the pool. Results are shown only for the twenty lenders having more than 100 comparable cases relative to the pool. Individual lenders are identified by an arbitrary ID number which is different than the internal LLD ID number. They are sorted according to their mean difference in capital allocations, from relatively most optimistic (negative numbers) to relatively most pessimistic. Average differences in capital allocation shown in the left panel range from –1.63 percentage points to +1.77. For PDs the range is –1.18 to +0.79 percentage points, and for grades the range is an average rating difference of from –0.80 to +0.57 grades.

Interestingly, the rank ordering of average differences in grade assignments is not one-for-one with the ordering of average differences in capital allocations. For example, Lender 2, with an average capital allocation difference of –1.50 percent, has an average difference in grade assignments of -0.03 grades. That is, lender 2 is in agreement with the pool about grade assignments on average. However, the *identity* of the borrowers about which Lender 2 disagrees with other lenders has a substantial impact on its relative

[12] Only two cases of comparable ratings at different lenders exist for borrowers that defaulted. In both cases, only one of the lenders records a default and a significant worsening of the rating. The other lender's rating stays the same or worsens slightly.

Table 5. Persistence of Rating Disagreements Over Time

Size of early-period disagreement	Percentage of sequential cases where disagreement is the same as in the next period
Full sample	61
0 grades	65
1 grade	62
2 grades	52
3 grades	32
4 or more	25

Table 6. Average Differences of Grades, PDs and Capital Allocations of Each Lender Relative to the Pool

	Unweighted Average Difference From Pool			P-value for Statistical Test of Hypothesis That Difference in Capital Allocation Is 0		Memo: Exposure-Weighted Average Difference From Pool	
Lender	Capital allocation (percent)	PD (percent)	Grades	Parametric t-test	Sign-rank test	Lender	Capital allocation
1	-1.63	-1.18	-0.57	0.00	0.00	3	-1.32
2	-1.50	-1.08	-0.03	0.00	0.01	2	-1.21
3	-1.31	-0.62	-0.44	0.00	0.00	4	-1.19
4	-1.25	-0.29	-0.67	0.00	0.00	6	-1.18
5	-1.22	-0.23	-0.80	0.00	0.00	1	-1.02
6	-1.03	-0.26	-0.38	0.00	0.00	5	-0.92
7	-0.93	-0.19	-0.64	0.00	0.00	9	-0.67
8	-0.62	0.17	-0.36	0.00	0.00	7	-0.66
9	-0.51	-1.09	0.13	0.35	0.85	8	-0.57
10	-0.39	-0.15	0.08	0.00	0.53	11	-0.40
11	0.04	-0.06	0.24	0.92	0.54	10	-0.12
12	0.17	-0.02	0.17	0.26	0.00	15	-0.06
13	0.21	0.05	0.26	0.47	0.03	16	-0.02
14	0.27	-0.20	0.12	0.07	0.01	13	0.06
15	0.65	0.09	0.42	0.05	0.00	14	0.07
16	0.88	0.79	-0.04	0.04	0.61	12	0.11
17	1.04	0.35	0.36	0.00	0.00	18	0.38
18	1.05	0.65	0.29	0.00	0.00	17	0.85
19	1.58	0.62	0.57	0.00	0.00	20	1.20
20	1.77	0.50	0.48	0.00	0.00	19	1.91

capital allocations and average PD differences. As noted previously, a difference of one grade has much larger absolute consequences for PDs and capital allocations if it occurs among the riskier grades.

The p-values shown in Table 6 for t-tests and sign-rank tests of the hypothesis that the average difference in lenders' capital allocations from those of the pool is zero should not be taken too seriously. The p-values generally imply that the hypothesis is strongly rejected (a value of 0.01 means a rejection at the 99 percent confidence level). However, the probability distribution of the differences is not normal and the differences are not independent, so the test statistics are not well suited to the task and are shown only for illustrative purposes.

The two rightmost columns of Table 6 display average differences in capital allocations when individual differences are weighted by the dollar amount of the lender's exposure to the borrower (the differences shown in the left panel of the table are based on equally weighted observations).[13] Although the results are roughly similar regardless of the weighting, the rank ordering of lenders changes somewhat, and the range of average differences is not as large as in the equally-weighted case. These results imply that lenders are somewhat more likely to agree about the credit quality of borrowers that take out larger loans (more evidence supporting such an interpreation appears below).

I inspected average differences for each lender by year and found that individual lenders' optimism or pessimism relative to the pool was roughly the same across years for 11 of the 20 lenders. For three others, only a couple of years of data are available, too few to make even rough judgments about stability. Of the remaining six lenders, three tended toward greater agreement with the pool over time, two toward greater pessimism, and one toward greater optimism.

Overall, the lender-by-lender statistics imply that differences of opinion about credit quality can be economically material on average for some lenders, but such average differences are not huge. If these results generalize to the banking industry as a

[13] Exposure is approximated as the dollar amount of outstandings plus three-quarters of the dollar amount of unused commitments. Where commitment amounts are missing or zero and an amount is available for the same facility in a previous period, that amount is used.

whole, it would be particularly good news for regulators, who may soon need to validate IRB capital allocations. Many of the average differences in average capital allocations shown in the rightmost column of Table 6 seem to me to be bearable from the standpoint of competitive equity and safety and soundness. By focusing on the outliers, regulators may be able to achieve fair outcomes without enormous effort or burdens on banks.

4.0 For Which Borrowers are Disagreements More Likely?

Intuition suggests that the frequency and magnitude of disagreements should be larger for small borrowers and for private firms because information is less easily available about such firms. To summarize results, rating disagreements are somewhat less likely the larger the borrower, but disagreements are no less likely for public firms or those with an agency rating. Disagreements are somewhat more likely the greater the fraction of the commitment that is used and if at least one lender evaluates the borrower as posing very low risk (the equivalent of A or safer). However, these predictive relationships are not strong. Other characteristics, like industry, location, time since origination, and seasoning of the loans, have little predictive power.

Data available in the LLD that can conveniently be used at the borrower level include an indicator for whether the obligor is a private entity or one that has issued securities publicly, the borrower's agency rating (if any exists), the maximum number of years since origination of any of the borrower's facilities, an SIC code, and a location code. I judgmentally converted SIC codes into a 14-industry coding, and made indicator variables for industry, location, and the time since origination of the borrower's oldest loan (0-1 years, 2-5, >5). I used two types of proxies for borrower size: 1) the natural log of the amount of total commitments or total outstandings for the borrower in the year of the observation (aggregating amounts across lenders), and 2) the natural log of the borrower's total assets. In principle, the latter is a less noisy measure of size, but total assets is missing for about 15 percent of observations.

Table 7 presents results for logistic regressions predicting whether rating disagreements are less or more than half a grade ("agreement" vs. "any disagreement") and whether disagreements are more than 2.5 grades (existence of a large disagreement). Included in the regressions, but not shown in the table to save space, are dummy variables identifying each lender, industry dummies, dummies for the number of observations in the usable dataset for the given borrower, and location dummies. Coefficients on all dummies except those for lenders are not robustly statistically significantly different from zero, and only seven to nine of the lender coefficients are significant, consistent with the evidence in Table 6 that rating disagreements by lender are generally not large on average.

Table 7. Results of Fixed Effects Logistic Regressions Predicting Rating Disagreement

Variable	Any Rating Disagreement		Rating Disagreement > 2.5 Grades	
	Coefficient	P-value	Coefficient	P-value
Log Assets	-0.171	0.0001	-0.107	0.0555
Drawdown rate	0.493	0.0533	1.303	0.0001
Seasoned 2 to 5	0.008	0.9389	0.032	0.8349
Seasoned >5 yr	0.278	0.2654	0.548	0.0908
Agency-rated	0.034	0.7744	-0.011	0.9430
Private firm	-0.398	0.0891	0.273	0.3529
One safe rating	0.595	0.0001	-0.188	0.2468

As shown in the left columns of Table 7, the size of the borrower is a statistically significant predictor of the existence of any rating disagreement, with disagreement about ratings less likely for larger borrowers. Loan seasoning, the existence of an agency rating, and whether the borrower is private are not statistically significant predictors. Other things equal, rating disagreements are more likely for borrowers with one safe rating (at least one lender rated the borrower the equivalent of A or better). A higher drawdown rate is positively related to the probability of rating disagreements.

These results are somewhat surprising in that the ease of availability of information about the borrower (whether it is a public firm or has an agency rating) does not seem strongly related to the likelihood of rating disagreements. Access to information may be easier for larger borrowers, but larger borrowers also take out larger loans, which at most banks receive closer scrutiny by senior credit staff, and which therefore are less likely to be misrated. More disagreements for borrowers with one safe rating may arise because distinctions of credit quality become increasingly finicky the safer the borrower.

Patterns are somewhat different in the rightmost two columns of Table 7, which report predictors of large rating disagreements. Such disagreements are somewhat less likely for large borrowers, the existence of one safe rating is not predictive, and a higher drawdown rate by the borrower significantly increases the likelihood of large rating disagreements.[14] It is not clear why drawdown rates are predictive. Perhaps drawdowns are sometimes indicative of changing credit quality, and some lenders are quicker to change ratings than others.

Results are qualitatively similar when dummy variables are dropped from the logistic regressions, when ordinary least squares regressions for the size of rating disagreements are estimated, and when the amount of aggregate commitments to the borrower replaces total assets as a measure of borrower size (not shown in the table).

Some insight into the relative importance of different sources of rating disagreements can be obtained by inspecting pseudo-R^2 values with and without the lender dummies and with and without the other explanatory variables.[15] In the "Any rating disagreement" regressions, when lender dummies are dropped the pseudo-R^2 drops from 0.091 to 0.048, whereas when all predictors but the lender dummies are dropped the pseudo- R^2 is 0.032. The numbers are higher in the large-rating-disagreement regressions (0.159, 0.086, and 0.059), but the relative effectiveness of the different sets of predictors is similar.

Overall, it appears that differences of opinions about ratings are hard to predict, with the existence of agreement or disagreement largely determined by idiosyncratic factors and not by lender policy or borrower characteristics. Somewhat more than half of the predictable portion of agreement/disagreement may be described as being due to implications of common banking industry practices, as captured by the variables other than the lender dummies (for example, known patterns in the intensity of internal reviews

[14] Inspection of drawdown rates revealed them to be zero or near zero for half the usable sample. Replacement of the continuous drawdown variable with indicators for any outstanding balance and for outstandings over 75 percent of the credit limit revealed that only the former is significant in the any-rating-disagreement regression, but both are significant in the large-rating-disagreement regression (not shown in the table).

[15] In conventional regressions, the R^2 is a measure of the fraction of variation in the left-hand-side variable that is predictable by the regressors. A rough idea of the share of predictable variation due to different regressors may be gained by computing the R^2 with and without the given regressors. In logistic regressions, the conventional R^2 measure cannot be calculated, but a commonly used alternative pseudo- R^2 is 1-(log-likelihood-with-all-specified-regressors/log-likelihood-with-only-the-intercept).

of rating assignments according to borrower size). Somewhat less than half of the predictable portion is due to differences in individual lenders' views or policies, as captured by the lender dummies. It is important to note that such "policy" differences across lenders are some combination of formal policies about credit quality evaluation as set by lender management as well as consistent biases in how lender staff implements the evaluations. Such biases may not be as intended by a bank's management.

5.0 Some Observations About Challenges to Benchmarking Studies

Financial institutions or regulators wishing to conduct benchmarking studies like this one face a number of challenges, and careful attention to data design is a practical necessity. Perhaps the biggest challenge is generating large samples of comparable pairs of ratings, and the biggest barrier to that is matching borrowers and loans across lenders and time. This study benefits from use of the LLD and thus from LPC's efforts to match borrowers across lenders, but even with such efforts the common-quarter method of aggregating observations across time yields only 2588 comparable pairs of ratings from a database of over 40,000 individual rating observations.

Other studies may face even greater barriers to achieving large samples. The previously cited Risk Management Association (RMA) study involved machine matching using loan ID numbers assigned by LPC, but such ID numbers are currently assigned to only a small fraction of loans, and thus the resulting sample included only a few hundred comparable observations. RMA's attempt to use KMV CreditMonitor borrower ID numbers reportedly foundered because many loans involve obligors that, while perhaps related to an entity in the KMV database, are somewhat different and pose different risks. Taking that finding at face value, the use of CUSIP identifiers seems likely to involve similar problems. Perhaps the most promising near-term alternative is the U.S. bank regulators' Shared National Credit (SNC) database, which includes borrower IDs assigned by personnel maintaining the database. However, internal ratings currently are collected only from a single lead bank for each SNC loan, so benchmarking would require an expansion of the SNC data collection effort. In the longer term, an investment by the industry in common borrower and loan ID numbering systems would make it possible for any group of financial institutions to conduct benchmarking studies at relatively low cost.

Synchronization of data collection across lenders, and consistency of rating types (borrower vs. facility) are also important. As noted previously, differences in rating assignments that have different as-of dates or that result from comparing different types of ratings are difficult to interpret, especially for borrowers whose credit quality is changing rapidly.

Also challenging is achieving consistent mappings of each lender's internal grades to a common scale. It is tempting to simply map grades to average PDs and collect and benchmark only those, but such a strategy may be less useful in the long run than one that collects both mapped grades and PDs. If one lender's PDs differ on average from those of the pool, is that because of differences in rating practices or because the lender's quantification of its grades differs from the norm? Without benchmarking information about both grades and PDs, it is difficult for a lender to interpret benchmarking results and thus to draw conclusions about how it should change its systems.

Table 8 illustrates the potential effect of differences in details of quantification, both for benchmarking purposes and for levels of capital allocation, by contrasting results using the average agency-grade default rates shown in Table 1 with results when PDs are based on an alternative set of default rates for each grade which are smaller than those in Table 1.[16] The second column of Table 8 reproduces the average capital allocation differences from Table 6, while the third column of Table 8 shows the average capital allocation differences when the alternative PDs are used. Average differences are generally smaller with the alternative PDs, and the rank-ordering of banks is changed somewhat.[17]

[16] The alternative PD values are approximate (not exact) observed default rates computed from the LLD, and are as follows: For LPC grade 1: 0.00%; for 2: 0.02%; for 3: 0.04%; for 4: 0.35%; for 5: 0.65%; for 6: 3.25%; for 7: 6.45%; for 8: 20%; for 9: 100%; for 10: 100%.

[17] Average differences are not always smaller in absolute value because the relative size of differences across grades is different in the two quantifications. Even though levels of PDs are smaller on average in the alternative quantification, some differences of PDs across grades are larger and for any given lender the mix of sizes of positive and negative PD differences changes, so a lender with many rating disagreements involving those grades may have average PD and capital allocation differences that are larger in absolute value under the alternative quantification.

The last two columns of Table 8 illustrate the effect that different quantifications can have on the level of commercial portfolio capital requirements. It is important to note that the capital ratios shown are not representative of those that would be calculated for the given banks' actual portfolios because LPC's sampling strategy draws disproportionate numbers of low-risk borrowers. Thus, the sample portfolios in the LLD

Table 8. Average Differences of Capital Allocations of Each Lender Relative to the Pool, And Average Capital Allocations for Each Lender, for Different Quantifications

	Average Capital Allocation Difference From Pool (percent)		Level of Allocated Capital (percent)*	
Lender	Agency-based quantification	Alternative quantifcation	Agency-based quantification	Alternative quantification
1	-1.63	-1.09	-	-
2	-1.50	-1.56	-	-
3	-1.31	-1.06	3.68	2.17
4	-1.25	-0.74	4.17	2.71
5	-1.22	-0.65	2.34	1.55
6	-1.03	-0.65	2.94	1.67
7	-0.93	-0.54	-	-
8	-0.62	-0.21	3.53	2.33
9	-0.51	-0.53	5.81	3.69
10	-0.39	-0.73	2.76	1.47
11	0.04	-0.20	3.93	2.35
12	0.17	0.14	3.22	1.84
13	0.21	0.00	-	-
14	0.27	0.13	3.90	2.58
15	0.65	0.24	6.93	4.54
16	0.88	0.01	4.66	3.32
17	1.04	1.02	-	-
18	1.05	0.76	2.44	1.63
19	1.58	1.05	5.41	3.51
20	1.77	1.41	6.42	4.33

*** Important note:** The levels of IRB capital ratios shown for each sample bank are not representative of those that would obtain for the banks' actual portfolios because LPC's sampling strategy draws higher proportions of low-risk assets than of high-risk assets, and low-risk assets require less capital. Cells with a "-" are those where sample portfolio sizes were relatively small.

are safer and attract less allocated capital than would the underlying portfolios. However, proportional differences across banks within a column and across columns are indicative of the fact that different quantifications of the same grades can result in substantial differences in allocated capital, even for identical portfolios and rating assignments.

6.0 Cautionary Notes and Concluding Comments

This study's results imply that most lenders are in rough agreement on average about the credit quality of common borrowers, and the unpredictability of individual rating disagreements is consistent with a conclusion that lenders' rating and review systems generally use easily available information. However, some lenders are economically significantly more optimistic or pessimistic than others on average. Moreover, the relative frequency of disagreements about individual borrowers is fairly high. Such disagreements appear to be primarily idiosyncratic and not matters of bank policy. They probably represent some combination of simple differences of opinion about individual credit quality and rating errors by individuals assigning ratings. Thus, further attention by a bank to its rating assignment and review procedures can reduce adverse selection risks and enhance the earnings efficiency of commercial lending business lines. Moreover, differences in ratings can translate to economically significant differences in average capital allocations, as can differences in quantifications (the default probabilities that are associated with each grade for modeling purposes).

Such conclusions may be too mild. There are reasons to believe that this study's data and methods are likely to understate the size and incidence of rating disagreements across lenders. Thus, this study's results should be generalized with caution to the banking industry or even to the total portfolios of individual LLD participating lenders. Four reasons to be cautious stand out:

- The lenders that participate in the LLD are likely to be those that are particularly attentive to the quality of their rating systems and risk management, and thus they may be less likely than the average commercial lender to make large volumes of rating assignment errors or to have unusually optimistic or pessimistic ratings.

- In order to obtain reasonably accurate measurements of default rates for each grade while avoiding the costs of collecting total-portfolio data, LPC samples proportionately more borrowers from the safer grades than from the riskier grades, resulting in larger proportions of safe borrowers in the LLD than in the underlying portfolios. Although the logistic regression results imply that rating disagreements are somewhat more likely for such safe borrowers, the capital allocation consequences of such disagreements are smaller in absolute terms, and thus results in this paper may present a rosier picture than would results for total commercial portfolios.
- Rating comparisons are possible only for multi-lender borrowers, and such borrowers tend to take out larger loans. Because rating disagreements are somewhat less likely for larger borrowers, the frequency and size of disagreements across lenders if they cross-rated their total portfolios probably would be larger than implied by this study's results.
- As noted, the effects of differences across lenders in quality of mappings to agency grades, in quantifications of grades, and in capital allocation systems may lead to effective differences in pricing hurdles and capital allocations that are not captured in a study such as this one, which focuses only on ratings.

Future studies are likely to be able to address such concerns, but benchmarking exercises are useful even when generalizations must be made with caution, if only to give lenders ideas about how best to focus their limited risk-management resources.

Appendix A. Some Details of Data and Computations

A1 Mapped ratings and fractional differences in grades

Each lender's internal rating system is different, typically involving a different scale. As each lender joins the LLD group, LPC works with the lender to map its ratings to a common ten-grade scale, and it is the mapped grades that are recorded in the LLD. Thus, the statistics reported in this study describe not solely the consistency of actual internal rating assignments, but also the consistency with which the various internal scales are mapped to the common scale. Because the only practical way to compare ratings assigned by different internal systems is by mapping to a common scale or by converted internal ratings to PDs (another form of common scale), and because LPC tries hard to make mappings consistent, in my opinion, the mapped grades are useful for purposes of comparison.

Mapping does not eliminate the fact that the underlying internal scales are not exactly aligned. In the LLD, this is reflected in the existence of fractional grade assignments. For example, a lender's internal 'E' might be mapped to a '5.5' on LPC's standard scale rather than to a 5 or a 6. Thus, rating comparisons often yield differences that are fractions of grades. Two lenders with exactly the same number of grades on their internal scales, but whose scales do not line up perfectly, may each assign a borrower to the internal grade that is most similar to the grade assigned by the other lender, but because the two scales map slightly differently to the common scale a fractional measured difference in ratings results. Moreover, fractional differences may be common for lenders that agree about borrower quality but that have very different numbers of grades on their internal scales, especially where one lender has fewer than ten internal grades. For ease of interpretation, when presenting statistics about grade disagreements, I round down differences of half a grade or less. For example, where the difference between two ratings is higher than 0.5 grades but less than or equal to 1.5 grades, I consider it to be a one-grade effective difference. However, for PDs and capital allocations, even fractional differences in mapped grades can have significant practical implications for loan pricing and capital allocations, and thus I preserve fractional grades in performing calculations.[18]

A2 Effective dates of observations

Each record in the LLD database contains an "effective" date which is an "as-of" date for the record's data. Internal ratings appearing in the LLD were the official values in lenders' database systems on the associated effective dates. However, the dates are quite diverse, even for a given lender and year, and thus comparing ratings across lenders is complicated by their disparate timing.[19] For example, if lender A has a record for a facility for borrower 1 dated 6/1/95, and lender B has a record for the same borrower dated 6/30/95, if the grades differ, is that indicative of a difference of opinions across the lenders, or simply that both lenders changed their grade for the borrower in mid-June of 1995? Discussions with LPC lead me to believe that a difference of opinion is usually implied, but perhaps not always.

I address this problem by using three methods to compute cross-lender comparisons. In the "common-month" and "common-quarter" methods, I assume that any records with effective dates in the same month or quarter are effectively the same in their timing, and that any differences in ratings that are observed are due to genuine differences of opinions about grade assignments. The example given in the preceding paragraph would qualify as a comparable case under the monthly and quarterly methods. These methods probably tend to overstate the rates of genuine rating differences, and to a greater extent the longer the duration of a common period. The common-quarter method is the one used for results presented in the text.[20]

In the "timespan" method, I identify sets of records for the same lender, borrower and facility where the rating is the same from one effective date to the next, and I assume that the rating stayed constant at the recorded value during the timespan from the earliest

[18] To get a sense of the effects of fractional mappings on results, one might generate results only for lenders whose internal grades map to integer LPC grades. Unfortunately, only six of the LLD lenders never report a fractional grade, and only 42 comparables exist among these lenders using the common-quarter method described below. The frequency of disagreements of zero, one and two grades among the six lenders are about the same as the full-sample frequencies of disagreements of 0.5, 1.5 and 2.5 grade, consistent with rounding down at the half-grade mark, but of course the sample is very small.

[19] I suspect the effective dates are usually not the date of the last formal internal review of the internal rating, but rather are the date of the last change to any variable in the data record.

[20] Results are similar for a common-year method, which yields 4033 comparable observations.

to the latest effective date appearing in the set. I then form comparisons by finding records for the same borrower at other lenders that have effective dates inside the timespan. The previous example would not qualify as a comparable under the timespan method, but if lender A also had a record for borrower 1's facility dated 6/1/96 and the rating was unchanged from a year earlier, then a single comparable would result (because lender B's record's effective date of 6/30/95 is inside the timespan 6/1/95-6/1/96). Because this method requires that one lender's rating stay constant for some period of time, it may tend to understate rating inconsistencies because borrowers with volatile credit quality may tend to be excluded.

Results from the common-month and timespan methods are presented in Tables A1 through A3, with results from the common quarter method reproduced for comparison. Results using the timespan method differ somewhat from those using the common-period method, with the former yielding greater frequencies of agreement or of modest differences in PD and K. This is consistent with the timespan method tending to sample-select borrowers with more stable ratings. Such borrowers are also more likely to be investment-grade borrowers (studies of rating transition rates find smaller likelihoods of rating changes for such borrowers). Differences of a grade or so have less absolute economic significance for such borrowers than for below-investment-grade borrowers.

The largest number of comparable cases is generated by the timespan method (5029) but observations for that method are at the facility rather than the borrower level. In the common-month and common-quarter methods, facility ratings are aggregated to the borrower level, as described below.

A3 Facility versus borrower grades, and facility records

The internal rating systems of some LLD lenders include only a facility grade and no borrower grade.[21] That is, the lender's own databases record only grades that mix an evaluation of the borrower's probability of default with an evaluation of the likely effect of the structure of the particular facility on loss given default (LGD). Where one lender's facility rating is compared to another lender's borrower rating, some differences are to be expected, especially where the first lender's facility differs in seniority from that of the

[21] The identity of such lenders is not recorded in the LLD.

second. Moreover, the fact that data are recorded in the LLD at the facility rather than the borrower level means that calculations done at the facility level will tend to overweight comparisons for borrowers with multiple facilities.

Unfortunately, there is no perfect solution to these problems. In the common-month and common-quarter methods, I aggregate facility records for each lender-borrower combination having effective dates within the month or quarter, and I use the value of the worst rating appearing in the aggregated records. This solves the overweighting problem, and may be an adequate solution to the facility-rating problem where a borrower's facilities include one that is senior and unsecured, but some borrowers have only a single, secured facility at a given lender. In the timespan method, aggregation to the borrower level is very difficult because of the lack of a common unit of time across comparable pairs of facilities. Thus, all timespan-method analysis is at the facility level. The main results are not very sensitive to which method is used, which offers some reassurance that issues related to facility versus borrower ratings and differing effective dates are not a major complication. Moreover, there are not too many cases where facility ratings disagree for a given borrower and lender, and when these are dropped results change very little.

Table A1. Frequency of Rating Disagreements Across Lenders for the Same Borrower (percent)

Number of grades of difference*	Common-month method N = 1933		Common-quarter method N = 2588		Timespan method N = 5029	
	Marginal frequency	Cumulative frequency	Marginal frequency	Cumulative frequency	Marginal frequency	Cumulative frequency
0	45	45	44	44	47	47
1	38	83	39	83	38	85
2	12.3	95.3	12.1	95.1	10.8	95.8
3	3.1	98.4	3.3	98.4	2.7	98.5
4 or more	1.6	100	1.6	100	1.5	100

* '0' means disagreements of half a grade or less, '1' disagreements of more than half a grade but less than or equal to 1.5 grades, etc. The three methods refer to different ways of determining that two lenders served borrowers at the same time in the absence of exact rating-action dates.

Table A2. Frequency of Implied Probability of Default (PD) Differences Across Lenders for the Same Borrower (percent)

PD Difference (basis points)	Common-month method N = 1933		Common-quarter method N = 2588		Timespan method N = 5029	
	Marginal frequency	Cumulative frequency	Marginal frequency	Cumulative frequency	Marginal frequency	Cumulative frequency
0	23.8	23.8	23.8	23.8	29.7	29.7
>0, ≤10	24.1	47.9	23.2	47.0	25.3	55.0
>10, ≤25	20.3	68.2	20.6	67.6	19.1	74.1
>25, ≤50	8.9	77.1	9.1	76.7	7.8	81.9
>50, ≤100	4.2	81.3	4.4	81.1	3.5	85.4
>100, ≤200	9.0	90.3	9.2	90.3	6.7	92.1
>200, ≤500	8.8	99.1	8.7	99.0	7.2	99.3
>500	0.9	100	1.0	100	0.7	100

Table A3. Frequency of Implied Total Capital Allocation Differences Across Lenders for the Same Borrower (percent)

Capital difference (percentage points)	Common-month method N = 1933		Common-quarter method N = 2588		Timespan method N = 5029	
	Marginal frequency	Cumulative frequency	Marginal frequency	Cumulative frequency	Marginal frequency	Cumulative frequency
0	23.8	23.8	23.8	23.8	29.7	29.7
>0, ≤0.5	20.4	44.2	19.4	43.2	22.3	52.0
>0.5, ≤1	7.4	51.6	7.3	50.5	5.5	57.5
>1, ≤2	19.8	70.4	19.2	69.7	7.8	75.3
>2, ≤4	11.6	82.0	12.1	81.8	10.4	85.7
>4, ≤8	10.6	92.6	10.7	92.5	8.1	93.8
>8, ≤15	6.8	99.4	6.9	99.4	5.8	99.6
>15	0.6	100	0.6	100	0.4	100