Monitoring Rating Quality

Jerome S. Fons[*]
May 31, 2009

Following near-catastrophic losses on initially highly rated fixed-income securities, policy makers and market participants are grappling with the proper role and function of the credit rating industry. Attention has focused on the industry's concentrated structure, the inherent conflicts of interest and the use of credit ratings in regulations and private contracts. Ultimately, the goal of these deliberations is to improve the performance of the major rating firms and thereby better protect investors.

Congress passed the Credit Rating Agency Reform Act of 2006 in an effort to improve competition among rating firms. The implication is that greater competition will lead to better ratings. However, since no consensus has emerged around a definition of rating performance, rating firms generally do not compete based on the quality of their ratings.

This paper argues that a clear definition of rating quality, combined with an unambiguous measure of rating performance, will lead to better public policy choices and allow competitive forces to foster innovation resulting in better, more accurate ratings. No rating system can eliminate credit risk, but an accurate rating system can help to ensure that credit risk will not be borne by those unwilling or unable to bear it. Although a rating firm may claim to meet multiple objectives, we argue that this behavior will ultimately lead to compromised ratings. We assert that the proper objective for a rating system is to maximize rating accuracy as summarized by an accuracy ratio.

This is not an academic issue: Current U.S. Securities and Exchange Commission regulations covering Nationally Recognized Statistical Rating Organizations (NRSROs) call for firms seeking (and maintaining) recognition to produce "performance measurement statistics," by rating category and sector, including "historical ratings transition and default rates" within each credit rating category.[1] As discussed below, these metrics are inadequate. Moreover, self-administered performance studies are subject to conflicts of conflicts.

Following a discussion of rating uses and quality, we review alternative measures of rating performance. We propose using an indicator of rating accuracy as the preferred measure of rating quality.

About Credit Ratings

Credit ratings are opinions of future credit risk: the likelihood that an investor (or lender) will suffer losses (and the extent of any losses) as a result of a failure to honor a promise

---

[*] Special Consultant, NERA Economic Consulting and former Managing Director, Credit Policy, Moody's Investors Service.
[1] Exhibit 1, SEC Form NRSRO, available at http://www.sec.gov/about/forms/formnrsro.pdf.

to pay. Bonds are unique because they contain a contractual promise to pay.[2] A credit rating addresses the strength of that promise and often captures both the ability and willingness to pay on time and in full. Ratings are generally communicated through a simple symbol system and have become, in effect, the language of credit risk.

There are many approaches to assessing credit risk. Most are grounded in the methods of fundamental analysis, which takes into account a range of financial and operational proxies for a borrower's capacity to pay. They rely on ratios that summarize an obligor's leverage, interest coverage, liquidity profile and profitability. Other, more technical, approaches rely on statistical tools and/or market-based indicators, such as a firm's equity price or spreads on its traded bonds.

Likewise, there is a range of possible processes for assigning ratings. Today's major rating firms utilize a hybrid rating process whereby a lead analyst proposes and a committee then votes on a rating outcome. The lead analyst is responsible for monitoring and commenting on the rating, but the assignment and any subsequent changes must go through the committee process. Simpler and more direct rating processes are employed where resources may be constrained or where the benefits of a committee are not required.[3]

This paper takes an agnostic position on the proper method or process for assessing credit risk. Instead, we are concerned with the definition of rating quality and, secondarily, with the measurement of rating performance. As such, the emphasis is on the "outputs" rather than on the "inputs" to the process.

Accuracy versus Stability

When he created the modern bond rating system in 1909, John Moody clearly intended for his ratings to provide bond investors with a simple shorthand measure of relative investment quality.[4] Unfortunately, ratings have since become all things to all people. Perhaps as a consequence of their perceived independence and track record, ratings became a favorite of fiduciaries wishing to limit exposure to credit risk. A full accounting of the ways ratings are used is beyond the scope of this paper. Suffice to say, financial regulators of all stripes, legislators, fund sponsors and others have each adopted ratings-based tests or guidelines.

Because ratings are used in so many ways, a *change* in an issuer's or obligation's rating can have consequences. For example, rating triggers in loan agreements or in private financial contracts (which reference an issuer's credit rating) may result in cash (or

---

[2] For our purposes, we include under "bonds" any financial contractual obligation, including notes, commercial paper, deposits, or counterparty exposures. The terms "borrower," "issuer" and "obligor" are synonymous

[3] The major rating firms point to the committee process as an important safeguard against the conflicts of interest associated with the issuer-pays business model.

[4] See Fons (2004) for a description of the evolution of Moody's rating system.

collateral) demands if the issuer's rating is downgraded. Such demands could threaten the viability of the rated issuer.

The classic sin of a rating firm is to "reverse" a rating action. Specifically, many consider it a mistake to have a rating downgrade (upgrade) followed by a rating upgrade (downgrade) within a relatively short period of time. The most grievous error occurs when a downgrade moves the rating below investment grade and an upgrade returns the rating to investment grade. Because ratings are often used in portfolio guidelines, a rating reversal can force certain agents to sell (possibly at a loss) a bond they want to hold, only to be forced to buy it again at a higher price when upgraded. The latter purchase may be driven by a desire to correlate portfolio performance with one or more bond indices, which themselves are defined around credit ratings.

As a result, there are powerful interests who advocate stable ratings. In some cases, they may be willing to tolerate "wrong" ratings in an effort to minimize market disruption. It is possible that the increased use of mark-to-market accounting by fixed income fund managers is partly to blame for the fixation on stable ratings. Rating changes can have an impact on bond prices, though most concede that ratings generally lag price movements.[5]

When a rating is downgraded, there is a possibility that the bond market will react negatively, which could adversely affect the performance of a mark-to-market portfolio, and, in turn, the reported capital position of a financial institution. Buy-and-hold investors are generally less affected by rating movements and therefore should favor accuracy over stability.

The fact remains, however, that *stable ratings cannot be accurate ratings*. Efforts to stabilize ratings have cost the ratings industry its reputation and have contributed to untold bondholder pain. It is likely that the historic stability of ratings facilitated their use in portfolio governance. These uses backfire when ratings change frequently. The industry has thus created a trap where ratings become self-fulfilling, and ultimately, harmful to end-users. A focus on rating accuracy will lead to higher rating volatility and may help break the self-fulfilling ratings trap by discouraging rating use in investment guidelines.

Market participants have confronted this issue for more than a decade in slightly different terms as part of the debate surrounding point-in-time (PIT) versus through-the-cycle (TTC) rating systems. The former are generally produced by model-based methods. The latter are characteristic of the major rating agencies. The consensus view is that PIT ratings are the preferred measure of credit for capital regulatory purposes.[6]

Although inherently unobservable before the fact, an obligor's (or obligation's) credit risk is not a static attribute. It naturally changes through time due to both idiosyncratic

---

[5] See Hite and Warga (1997) and Hull, Predescu and White (2004) for evidence that market prices (or spreads) often precede rating changes.
[6] See, for example, Gordy and Howells (2004).

and systemic developments. The elements contributing to a borrower's credit profile are not carved in stone. Profit margins, earnings stability and underlying asset values can and do fluctuate. Often, these fluctuations are reflected through changes in bond prices. A highly accurate rating system will reflect these changes, rather than emphasizing rating stability.

Consider the table below, reprinted from Figure 13 as found in Moody's (2009) which compares the performance of two different "rating" systems, Moody's traditional bond ratings and ratings implied from bond market spreads, for the period January 1999 through March 2009. In essence, bond-implied ratings place those bonds (or issuers) with the smallest spreads (or yield differentials from comparable default-free securities, such as US treasury obligations) at the highest rating grades and those with the widest spreads at the lowest.

As described more fully in the Measuring Rating Performance section below, an accuracy ratio is a statistic indicating the degree to which a rating system separates in advance (in this case, either 1-year or 5-years) those issuers (or obligations) who default from those who do not. A perfectly accurate rating system will produce an accuracy ratio of 100%. A very poor system will score close to 0%.

Also shown in the table is the average rating (three years) prior to default, a summary measure of rating performance. Here, the lower the rating, the better the warning of impending distress. Finally, the table shows two measures of rating volatility, the rating action rate (the percent of issuers experiencing a rating change within the past year) and the large rating action rate (the percent of issuers experiencing a rating change of at least three rating "notches" within the past year).

For each of the accuracy measures shown, bond-implied ratings dominate Moody's ratings. That is, bond-implied (or "market") ratings better separate defaulters from non-defaulters and provide an earlier warning of impending credit risk.

Comparisons with Bond Market Implied Ratings: January 1999 through March 2009

	Historical Average	
	Moody's Ratings	Bond-Implied Ratings
1-Year Accuracy Ratio	81.20%	88.30%
5-Year Accuracy Ratio	67.80%	72.70%
Average Rating Prior to Default	B2	Caa1
Rating Action Rate	20.10%	95.00%
Large Rating Action Rate	4.20%	55.70%

Also revealed by these statistics is the fact that Moody's ratings are much more stable than market ratings. Roughly one in five issuers saw its Moody's rating change in a given year, on average. By comparison, nineteen out of twenty issuers experienced a change in their market-implied rating. This mainly reflects the inherent volatility in bond spreads. Yet no one seriously considers movements in spreads to be "mistakes." Rather, such changes simply reflect changes in investor sentiment.

The reported stability for agency ratings can be attributed both to institutional factors (such as the rating committee process) and to deliberate measures (such as the use of rating outlooks and the credit watch list).[7] If anything, these measures draw information content away from the published rating and toward these ancillary signals.

We turn next to a summary of tools that have been developed to measure rating performance.

Measuring Rating Performance

This section provides an overview of the techniques that have been developed to assess the performance of rating systems. Before doing so, we discuss a number of features that differentiate rating systems.

Attributes of Ratings

Ratings can be assigned to both issuers and to their obligations. *Ex post*, or after the fact, an issuer either pays an obligation as promised, or it does not, in which case we say that a default has occurred. Despite its seemingly black-and-white nature, things are more complicated than that. First off, there are many ways a default can occur, including missing payments, filing for protection from creditors or completing a distressed debt exchange.[8]

In a world of perfect foresight, there would be just two rating categories: "good" and "bad". Issuers who pay their obligations as promised would be rated good and those that do not would be rated bad. Lacking a crystal ball, the best we can do is form an opinion about the likelihood of a default (over a fixed horizon). Ratings are attractive because they convey this likelihood using a simple grading system.

There are two types of mistakes a rating firm can make: Rating too high an issuer that subsequently defaults and rating too low an issuer that does not subsequently default. Some would argue that the first error is more costly than the second, but there are costs to both errors.[9] In statistical decision theory, these are commonly referred to as Type I and Type II errors. Depending on how we frame the hypothesis at hand, one can use the term

---

[7] As a result, the major rating systems show *path dependent* behavior. This is because analysts and committees are generally not free to place a rating where it best reflects credit risk, but instead must take account of recent rating history. Consequently, it is likely that such ratings are most accurate when an issuer is first rated.

[8] The focus on default/non-default may in fact be too narrow. Certain issuers and obligations may benefit from support from a third party (i.e., by being rescued). Some investors (particularly those who mark-to-market) may face losses as concerns mount about the availability or form of support. One could argue that a rating system should signal the risk of such losses, even where no default technically occurred.

[9] For example, a borrower rated too low may have to pay a higher interest rate than is warranted.

Type I error where the rating is too high and Type II where the rating is too low. A proper performance measurement system will take into account both error types.

We list here some of the key attributes of traditional rating systems.

- *Probability of Default versus Expected Loss.* As distressed debt investors are keenly aware, not all defaults are equal. For example, it is entirely possible that a default could occur where bondholders receive one hundred cents on the dollar, with interest. It is also possible for bondholders to receive only a few pennies on the dollar owed. In the latter case, the *loss given default* would be quite high. A rating system that focuses only on whether or not an issuer will default need not address loss given default. This is because loss severity is observed only at the security or bond level. One could estimate an issuer's average severity of loss across all defaulted obligations, but it would not apply to any particular debt. On the other hand, it is natural for a system designed to rate individual obligations to incorporate an estimate of loss severity.[10] Such a rating system targets *expected loss*, rather than default. As a rough approximation, we can define expected loss as the product of the probability of default and loss given default.
- *Risk Measure.* Rating systems have emerged that attempt to signal overall creditworthiness, or *financial strength*, as opposed to default risk or expected loss. These often, though not always, employ a separate rating scale that can be "mapped" into a traditional default/loss rating. Such ratings indicate the likelihood that an issuer will need to be rescued; if no support materializes, a default will occur. These are often provided for banks and insurance companies.
- *The Rating Scale.* As discussed above, the simplest rating system would have two categories: good and bad. For decades, the major rating firms managed well with ten categories. The current standard is 21 rating grades, or 22, if you count "default" as a category. While this may reflect the limits of human ability to discriminate levels of credit risk, it is most likely an historical artifact. Statistical and/or market-based estimates can result in a continuous "score" or rating. Perhaps the most useful system would be expressed in basis points.
- *Cardinal versus Ordinal.* Traditional rating systems are designed to ordinally rank credit risk. That is at any point in time, the strongest issuer is rated at the top of the scale and the weakest at the bottom. The rating firm sometimes manages the distribution of ratings over time to ensure that each category is sufficiently populated. Conversely, a cardinal rating system maps into a fixed probability of default schedule (or expected loss schedule). No attempt is made to manage the distribution, except by the initial mapping cut-offs. By definition, a cardinal system is also ordinal, but an ordinal system is not necessarily cardinal.
- *Time Horizon.* Ratings are opinions about future events, but just how far into the future? It is worth noting that fundamental credit analysis does not typically employ advanced forecasting methods. In an ideal world, a bond's rating horizon would match the maturity of the underlying bond. But most issuers have

---

[10] This is typically achieved through "notching" junior obligations below a firm's senior debt.

outstanding bonds with varying maturities. As a compromise, the major rating firms claim that their long-term ratings have predictive power for up to five years hence.[11] Most model-based rating systems are designed to predict risk up to a one-year horizon.

In order of historical development, we present here an overview of widely used tools for assessing the performance of credit ratings.

1. Default Rates

The first, market-wide effort to document the power of rating systems was the product of the Corporate Bond Research Project, sponsored by the FDIC and overseen by the National Bureau of Economic Research as part of the New Deal Era's Works Project Administration. The study represented one of the first applications of computers to financial analysis and culminated in a three-volume set of books. For our purposes, the key results of this comprehensive study are contained in Hickman (1958). In examining the performance of all domestically issued and held "straight" corporate bonds offered between 1900 and 1943, Hickman calculated one-year – as well as "life-span" – realized yields, default rates and loss rates, based on initial rating and other characteristics.

In performing these calculations, Hickman conducts an ageing analysis by forming groupings of bonds, based on the issue date and other characteristics (such as rating grade), and following these through time. This approach allows one to identify patterns in defaults or losses. This is the fundamental concept behind a default rate.

However, as one observer has noted, there are at least 7000 different ways to calculate a default (or loss) rate. A more important point is that the choice of calculation should be guided by the purpose or intended use. As noted above, Hickman chose individual domestically issued bonds as his unit of study. One could argue that the focus should be on companies, not bonds, since it is the companies that default on bonds and a company can have many bond issues outstanding, each possibly issued at a different date.

Although not yet invented at the time of Hickman's study, structured finance obligations are not strictly considered to be corporate bonds. As a result, the major rating firms exclude these from their key performance studies, even though they performed very badly during the recent crisis. Investors also suffered losses on rated preferred stock, a type of fixed income security, but these too are excluded. Performance data on rated bank loans, commercial paper, municipal obligations, and financial strength ratings for financial institutions tends to be spotty, if available at all.

Hickman grouped his sample bonds by year of origination. One could instead form "cohorts" of bonds (or issuers) that were outstanding at the start of each year, regardless

---

[11] Most rating firms also offer short-term ratings (which are simply mapped off their long-term ratings) and apply these to securities with original maturities of less than one year.

7

of the year of origination. The latter approach may make more sense if the unit of study is the issuer, rather the bond. Hickman also weighted his results by the (non-inflation adjusted) dollar amount of each offering. Aggregate default rates calculated this way place a greater emphasis on the performance of large companies (or bond offerings) and more recent cohorts. An alternative would be to focus on the count of companies.

Then there is the thorny issue of how to calculate a multi-period default rate. If the cohorts are formed annually (as opposed to say, monthly or quarterly), a one-year default rate is fairly simple to produce: First, we form a cohort of issuers (or bonds) holding a given rating at the start of the period (in this case the beginning of a year). We follow this cohort for one year and observe which default and which do not. The numerator of the default statistic is the number (or amount) of issuers (or bonds) that default during the year and the denominator is the number (or amount) of issuers (or bonds) outstanding at the start of the year. The reported one-year default rate is typically a simple (or weighted by cohort size) average across many years.[12]

It is possible to simply extend this approach to multiple periods. That is, rather than focus on one year, the numerator can contain those issuers (or bonds) that default within two, three or even twenty years hence. An average of these, calculated for non-overlapping periods, might prove helpful if one had a very long sample period. But the large rating firms use a more recent sample period, say, starting 1970 or later.[13] This leaves too small a sample set. Instead, typical practice is to compute one-year marginal, or "forward," default rates and roll these up into a multiple period default rate.[14]
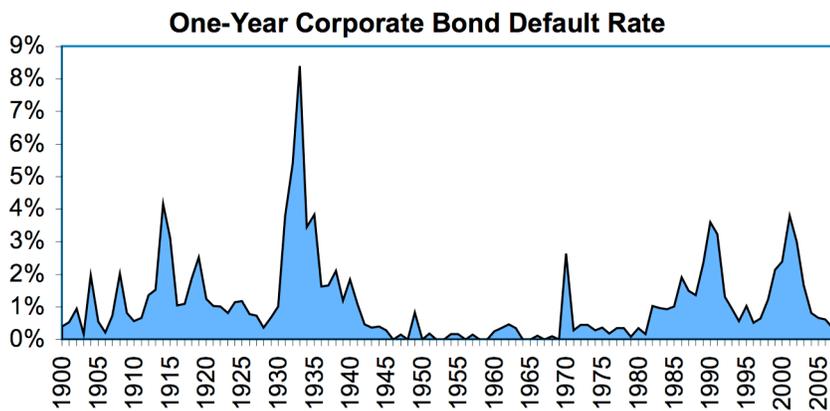
Source: Hickman, Moody's

---

[12] Because not all defaults occur at year-end, this in turn represents an average of many shorter-term default rates.

[13] In the mid-1990s, Moody's extended its default database back to 1920 using firm-level rating histories taken from old rating manuals. Although the quality of this data is not commensurate with more recent information, it is nevertheless interesting.

[14] A marginal (or forward) one-year default rate has as a numerator the number of defaults occurring in a forward, one-year period and as a denominator the number of issuers (or obligations) *able* to default during that year.

When a multi-period default rate is composed of multiple marginal one-year default rates, one must account for the fact that, as the cohort ages, the denominator can decline as a result of prior defaults and other retirements.[15] There are many ways to account for this. By not taking these "censored" rating histories into account, one will arrive at understated default rates. Likewise, there are many ways to average the experience of overlapping observations. Suffice to say, there is no single best way to compute an average 10- or 20-year default rate.

The formula below illustrates how multi-period default rates are calculated using Moody's approach, which is based on an issuer default rate.[16] Define d_t as the marginal default rate for issuers with a given rating, say Baa1. This is the one-year default rate t years after formation of a cohort of Baa1 issuers. The variable $x(t)$ is the number of issuers from the original cohort that defaulted in year t. The number of issuers rated Baa1 outstanding at the start of the tth year is $n(t)$.

$$d_t = \frac{x(t)}{n(t) - x(t-1) - \frac{1}{2}[w(t) + w(t-1)]}$$

The calculation in the denominator is designed to remove those firms not able to default in the period either because they defaulted in the prior year or because their rating was withdrawn. Since withdrawals can happen throughout the year, a simple average of current and prior withdrawals is used to smooth for this. The Baa1 multi-year *cumulative* default rate D_T for a horizon of T years is then defined as:

$$D_T = d_1 + (1 - d_1)d_2 + (1 - d_1)(1 - d_2)d_3 + ... + \prod_{t=1}^{T-1}(1 - d_t)d_T$$

Because bondholders often care about potential losses (in addition to defaults), it is natural to extend the analysis shown above to *loss rates*.[17] There are at least two ways to do this. One way is to calculate losses at the individual bond issue level and use the sum across these as the value for $x(t)$ in the default rate calculations shown above. The alternative method is to calculate a single, average loss given default rate across all bond issues and multiply this figure by each d_i term shown in the formula above for the cumulative default rate. The result will be a cumulative loss rate.

Default rates, grouped by rating category and time horizon, are interesting historical statistics.[18] They help us understand how an investor, with an exposure to an issuer or bond of a given rating, may have fared on average over a range of holding periods. They

---

[15] For example, a rating may be withdrawn or a firm may merge with another.
[16] Please see Moody's (2006) for a fuller presentation, along with default rate results.
[17] The loss rate defined here is simply losses as a percent of par amount owed. By contrast, Hickman defines the loss rate as the difference between the yield promised at bond offering and the yield realized through the date of "extinguishment."
[18] While intended to convey information about *mean* defaults, the approaches described here do not provide useful data about the *distribution* of defaults around the mean.

are not, however, terribly useful indicators of a rating system's performance. The best they can tell us is that the default experience of lower rating categories is greater than that for higher categories.[19] For example, we depict below historical default rates for the two hypothetical rating systems, X and Y.

Rating system X is characterized by exponentially increasing default risk and fairly depicts patterns found in real-world agency ratings. Rating system Y shows default risk slowly tapering off at lower rating categories. Both systems order risk correctly, on average, in that credit risk increases as ratings fall. But which system best protects investors?

---
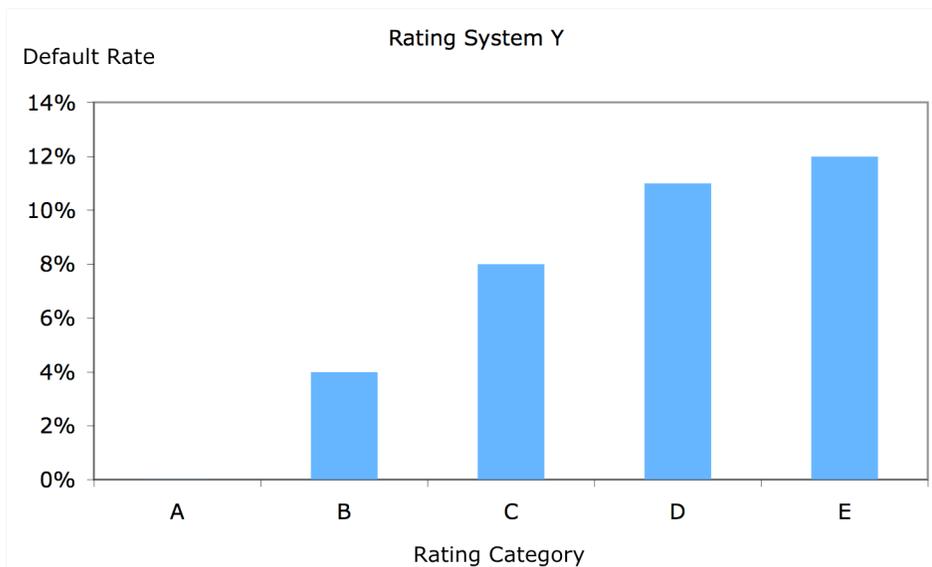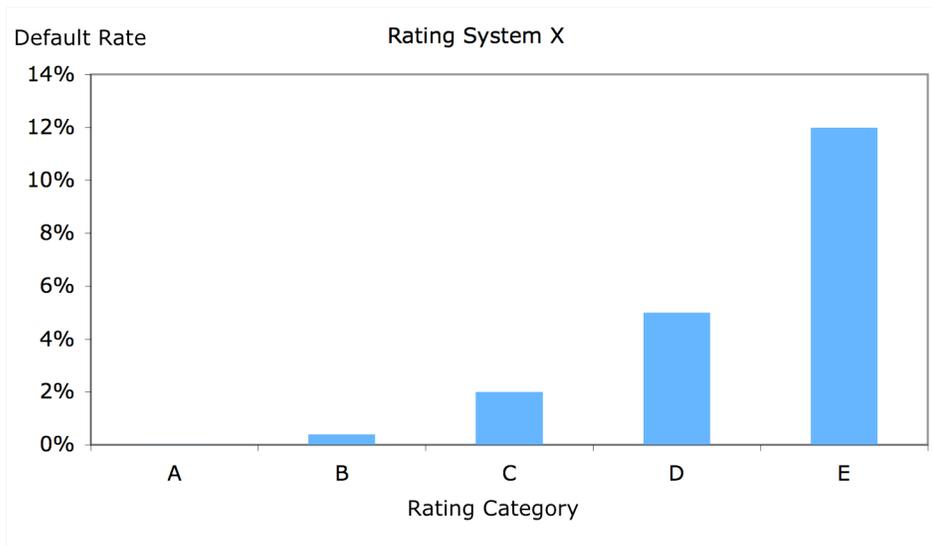
[19] For more detail, the interested reader is referred to the default study publications issued by each of the major rating firms.

Default rates can be helpful if the rating system is intended to be a cardinal system. In that case, judging rating performance would turn on determining whether actual default experience for each rating category closely matches an "idealized" default rate. One could construct a metric that summarized the deviations from the ideal. However, the major rating firms have positioned their ratings as ordinal rankings. To the extent lower rating categories exhibit higher default rates, the rating system meets a broad-level ordinal test. Their biggest shortcoming is that default rates do not tell us how well a rating system separates defaulters from non-defaulters. Moreover, default rates do not balance Type I rating errors against Type II errors.

2. Average Rating Before Default

An ideal rating system will provide sufficient warning that a default is imminent. The average rating before default is an indicator of a rating system's ability to provide ample warning. Constructing this statistic is fairly easy. Each rating category is converted to a number, although there is generally no accounting for differences in scale or risk between the rating categories. According to Moody's (2009), the "rating of the defaulted issuer is sampled every month for 36 months prior to default as well as immediately prior to default. These 37 observations are averaged together to create a single representative number for each defaulted issuer. These representative numbers are then averaged together to create the reported average rating."[20]

The average rating before default focuses only on those issuers that defaulted during the sample period. For better or worse, it equally weighs each issuer and rating category. And the statistic addresses Type I error only. A rating system could maximize this statistic by simply rating *all* issuers very low. Thus, this is not a particularly helpful measure.

3. Rating Migration Statistics

In the early1990s, researchers began to devise ways to summarize the likelihood that an issuer (or bond) would see its rating change. The idea is fairly simple: Examine a cohort of issuers or bonds with the same rating, follow it for a period (say, one year), and calculate the percentage falling into each possible "state." The states are the various rating categories, along with a default state and a withdrawn state. A certain percentage will remain at the original rating, a presumably smaller percentage will be rated higher and others rated lower. The percentage in the default state will correspond to the one-year default rate described above.

---

[20] Moody's (2009), page 11.

The results of this exercise are displayed in a table, such as the one shown below.

		A	B	C	D	E	Default	Withdrawn
					Rating at End of Period			
	A	**89%**	4%	3%	1%	0%	0%	3%
Rating at	**B**	2%	**82%**	5%	4%	2%	1%	4%
Start of	**C**	2%	4%	**78%**	5%	4%	2%	5%
Period	**D**	1%	3%	4%	**75%**	5%	6%	6%
	E	1%	2%	3%	4%	**70%**	12%	8%

Migration tables, sometimes called transition matrices, illustrate a number of interesting features of a rating system. In this example, issuers or bonds holding higher ratings have a greater tendency to remain at their start-of-period ratings, and therefore exhibit less rating volatility. The table provides a summary of historical rating activity and, as long as history repeats itself, sheds light on the likelihood that an issuer or obligation with a given rating will move to a different category. To the extent rating movements are correlated with bond prices, such information can help one anticipate expected returns.

Rating transition matrices are often employed in modeling the future credit profile of a portfolio of issuers or obligations. Using fairly restrictive assumptions, a one-period transition matrix is multiplied by itself to yield a two-period rating distribution.   Raising the matrix to the Nth power provides an estimate of the rating distribution N periods from now.

Aside from rank ordering the default column, a rating migration table does not provide a particularly helpful measure of rating performance. Of course, if absolute rating stability is considered a good thing, one would like to see the diagonal (bolded items in the table above) as close as possible to 100%. But it does not indicate if a rating system provides timely warning of default risk.

4. Measures of Rating Activity

Rating agencies have for years reported various statistics summarizing rating activity. The implication is that movements in ratings point to fundamental shifts in aggregate credit quality. This is particularly true for the *upgrade/downgrade ratio*.[21] The upgrade/downgrade ratio is simply the number of issuers (or bonds) upgraded during a fixed period, divided by the number of issuers (or bonds) downgraded over the same period. The ratio can theoretically range from zero to infinity. A value greater than one indicates that upgrades exceed downgrades. A value below one indicates that downgrades exceed upgrades. Although highly volatile, for most of its reported history,

---

[21] Confusingly, this is also reported as the downgrade/upgrade ratio.

the ratio has tended to be less than one.[22] For much of the past decade, this ratio has been less than one as reported by Moody's.

The upgrade/downgrade rate has a number of shortcomings. For one thing, it is meaningless when there are no downgrades. More importantly, it is not weighted by rating activity: The ratio would read the same value if we multiply the numerator and denominator by a scale factor of 1000. And it does not tell us how much aggregate rating levels are changing.

Another popular measure is the *rating action rate*. This is simply the number of rating changes occurring over a given period, divided by the number of ratings outstanding at the start of the period. As reported in the table shown in the *Accuracy versus Stability* section above, a figure of around 20% per year is typical for agency ratings. That is, roughly eight out of ten issuers see no change in their ratings during a typical year.

First introduced in Carty and Fons (1993), *rating drift* summarizes any shifts in the aggregate (i.e., average) credit rating over a given period. It is a weighted measure of the number of notches the average rating has moved due to upgrades and downgrades. When upgrades exceed downgrades, in terms of the overall number of rating grades moved, rating drift will be positive. If the reverse is true, drift will be negative. This is not a measure of rating performance, *per se*, but rather an indicator of rating volatility.

5. The Accuracy Ratio

An *accuracy ratio* is a statistic used to summarize the ability of a rating system to separate *ex ante* (i.e., before the fact) those issuers that subsequently default from those that do not default. The grouping (or cohort) choices are identical to those outlined in the default rate section above.

The basic theory underpinning the accuracy ratio has its roots in the Lorenz Curve, a tool economists developed to graphically represent income inequality. Rather than plotting incomes, we plot the cumulative accuracy of a rating system, using a cumulative accuracy profile (CAP).[23] KMV (now part of Moody's Analytics) was among the first to propose applying this concept to the assessment of rating performance.[24] It was later expanded and refined by Sobehart, *et. al.* (2000) and proposed as a ratings performance benchmark by Cantor and Mann (2003).

Below is a sample CAP plot for a rating system. The horizontal axis depicts equal-width percentile "buckets" of issuers, sorted by rating, from lowest to highest. The vertical axis shows the cumulative share of *ex post* defaulters, expressed as a percent of all defaults, occurring within each bucket during the period. The jagged line depicts the performance

---

[22] See Okashima and Fridson (2000) for an example of this ratio's predictive power.
[23] In signal detection theory, these are called receiver operating characteristic (ROC) curves.
[24] McQuown (1993) highlights "power curves" which are similar in principle to a CAP.

of the rating system under consideration. In this example, those issuers rated *ex ante* in the lowest 20th percentile – perhaps corresponding to a rating of B and lower – accounted for 85% of issuers who subsequently defaulted. The histogram below the CAP chart shows the distribution of rated defaults that generated this CAP curve.
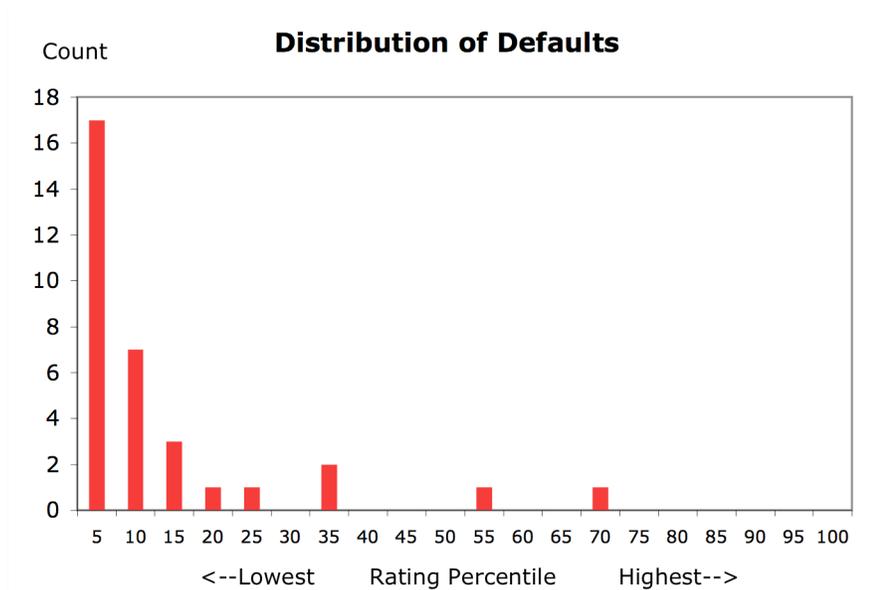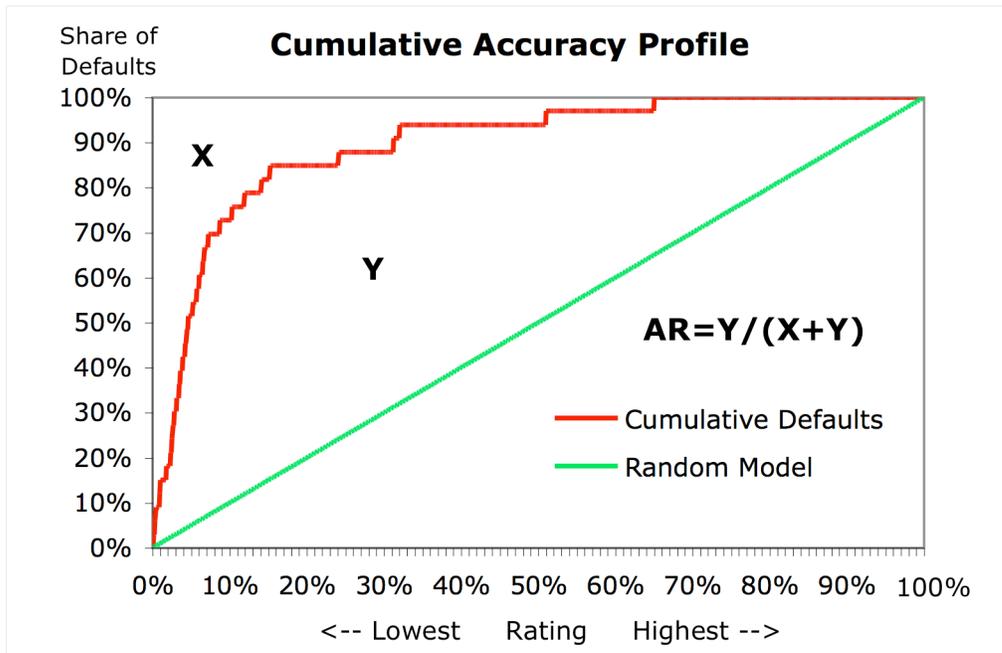
With a little reflection, it can be seen that the closer the CAP is to the upper left corner, the more "accurate" the rating system. The best possible rating system will have all defaults occurring among those issuers with the lowest *ex ante* rating. The diagonal line

represents a lower bound to accuracy, as this would be the CAP for a completely random rating system where, for example, the lowest 50% of rated issuers accounted for exactly 50% of all defaults.

The accuracy ratio (AR) is expressed in terms of this random rating system and is calculated as the area between the subject rating system's CAP and the random rating system's CAP (labeled area Y in the chart above), divided by the area above and to the left of the random CAP (area X plus area Y). A perfect AR score would be 100%, a random system's AR would be 0%.[25]

Mathematically, the AR can be computed as:

$$AR = \frac{2}{N} \sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{i} x_j}{D} - \frac{i}{N} \right),$$

where N is the number of rating partitions, of equal percentile width, as a percent of the total rated sample (sorted from 1=riskiest to N=safest), D is the total number of defaults and x_i is the number of defaults occurring among issuers in the percentile partition i, where $i = 1, 2, 3,…,N$. The interior summand cumulates defaults for each successive partition. In order to compare the accuracy of any two rating systems, the rated populations must be exactly the same.

Averaging accuracy ratios across multiple periods can be done in several ways. Just as with a multi-period default rate, one can create a cohort and follow it for several years and base the accuracy ratio on the ability of a rating system to rank order issuers that default several years out. As a rule, the AR falls as the horizon lengthens.

When summarizing the results of several one-year accuracy ratios, one can simply compute a series of accuracy ratios, either for overlapping or non-overlapping samples, and average these together. This would reflect the ability of a rating system to provide an ordinal ranking of credit risk. Alternatively, cohorts across time periods can be "pooled" into one very large data set and then sorted by rating. This method is preferred if the rating system has cardinal accuracy as the objective.

Although it is just one way to summarize a CAP, an accuracy ratio provides an important measure of the "power" of a credit risk rating system to discriminate between those issuers who subsequently default and those who do not. Of the metrics discussed in this paper, it is the only measure to do this. Importantly, it equally weighs Type I and Type II errors. And, as with each of the measures described in this paper, it is best to have a large sample size. For most real-world rating systems, accuracy is highly correlated with

---

[25] Some prefer to calculate the AR by adjusting the "perfect" CAP for the number of defaults. See Cantor and Mann (2003), footnote 7, for the pros and cons of doing this.

overall credit distress: it falls when defaults are high and rises when defaults are subdued.

Summary Recommendations

This paper discusses rating system objectives and provides an overview of the metrics employed to judge the performance of a rating system. We argue that the trade-off between stable ratings and accurate ratings is not manageable. Consequently, efforts to meet multiple rating objectives should be abandoned. In order to restore investor confidence, *rating quality must be synonymous with rating accuracy.* In particular, actual or artificial barriers to timely ratings must be eliminated. When an incorrect rating is maintained (for whatever reason), investors relying on that rating – particularly those contemplating a bond purchase – will be misled and may needlessly suffer losses.

It is up to policymakers and market participants to insist on accurate ratings above all else. Rating firms should commit to rating accuracy as a primary objective and abandon rating stability as a goal. Moreover, rating firms should be judged on their ability to provide timely and accurate ratings. We recommend the accuracy ratio described above as the only measure for judging rating performance.

An independent body should be responsible for compiling rating performance reports. There is too much internal pressure to fudge results and too many choices as to methods for this responsibility to rest within a rating firm. A firm compiling its own performance record has every incentive to report in a way that makes it look good.

The independent body would be the arbiter of default or loss, and each rating firm would supply to it an electronic feed of rating histories. Apples to apples comparisons could be broadly distributed and updated through time. Such comparisons will give policy makers, investors and other users of ratings confidence that the industry is fulfilling its proper role.

References

Lea V. Carty and Jerome S. Fons (1993), "Measuring Changes in Corporate Credit Quality," *Moody's Special Comment*, November 1993.

Richard Cantor and Christopher Mann (2003), "Measuring The Performance Of Corporate Bond Ratings," *Moody's Special Comment*, April 2003

Jerome S. Fons (2004), "Tracing the Origins of 'Investment Grade'", *Moody's Special Comment*, January 2004.

Michael B. Gordy and Bradley Howells (2004),"Procyclicality in Basel II: Can We Treat the Disease Without Killing the Patient?," Board of Governors of the Federal Reserve System, May 2004.

W. Braddock Hickman (1958), *Corporate Bond Quality and Investor Experience*, Princeton University Press, 1958.

Galen Hite and Arthur Warga (1997), "The Effect of Bond-Rating Changes on Bond Price Performance," *Financial Analysts Journal*, May/June 1997.

John Hull, Mirela Predescu, and Alan White (2004), "The Relationship Between Credit Default Swap Spreads, Bond Yields, and Credit Rating Announcements," *Journal of Banking & Finance*, November 2004.

McQuown, John Andrew (1993), "A Comment On Market vs. Accounting Based Measures of Default Risk", KMV Corporation, September 1993.

Moody's (2006), "Default and Recovery Rates of Corporate Bond Issuers, 1920-2005," *Moody's Special Comment*, revised March 2006.

Moody's (2009), "The Performance of Moody's Corporate Debt Ratings, March 2009 Quarterly Update," *Moody's Special Comment*, May 2009.

Okashima, Kathryn and Martin S. Fridson (2000), "Downgrade/Upgrade Ratio Leads Default Rate," *The Journal of Fixed Income*, September 2000.

Jorge R. Sobehart, Sean C. Keenan and Roger M. Stein (2000), "Benchmarking Quantitative Default Risk Models: A Validation Methodology," *Moody's Rating Methodology*, March 2000.