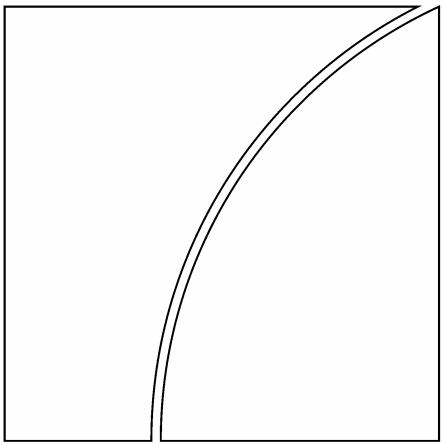
Basel Committee
on Banking Supervision

Working Paper No. 14



Studies on the Validation of Internal Rating Systems

February 2005

BANK FOR INTERNATIONAL SETTLEMENTS

The Working Papers of the Basel Committee on Banking Supervision contain analysis carried out by experts of the Basel Committee or its working groups. They may also reflect work carried out by one or more member institutions or by its Secretariat. The subjects of the Working Papers are of topical interest to supervisors and are technical in character. The views expressed in the Working Papers are those of their authors and do not represent the official views of the Basel Committee, its member institutions or the BIS.

Copies of publications are available from:
Bank for International Settlements
Information, Press & Library Services
CH-4002 Basel, Switzerland
Fax: +41 61 / 280 91 00 and +41 61 / 280 81 00
This publication is available on the BIS website (www.bis.org).

© *Bank for International Settlements 2005.*
All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.

ISSN 1561-8854

Chairman of the Validation Group

Mr Thilo Liebig	Deutsche Bundesbank, Frankfurt

Members of the Validation Group

Mr Vincent Baritsch	Financial Services Authority, London
Ms Rosalind L Bennett	Federal Deposit Insurance Corporation, Washington D.C.
Mr Martin Birn	Bank for International Settlements, Basel
Mr Stefan Blochwitz	Deutsche Bundesbank, Frankfurt
Mr Jaap W B Bos	De Nederlandsche Bank, Amsterdam
Mr Ben Carr	Financial Services Authority, London
Ms Eva Catarineu	Banco de España, Madrid
Mr Alvin Chow	Monetary Authority of Singapore, Singapore
Mr Klaus Düllmann	Deutsche Bundesbank, Frankfurt
Ms Antonella Foglia	Banca d'Italia, Rome
Mr Nils Chr Framstad	Kredittilsynet, Oslo
Mr Erik A Heitfield	Board of Governors of the Federal Reserve System, Washington D.C.
Mr Jon S Hellevik	Kredittilsynet, Oslo
Mr Mark Levonian	Office of the Comptroller of the Currency, Washington D.C.
Mr Soon Chong Lim	Monetary Authority of Singapore, Singapore
Ms Nancy Masschelein	National Bank of Belgium, Brussels
Mr Hiroaki Miyagishi	Bank of Japan, Tokyo
Mr Gregorio Moral	Banco de España, Madrid
Mr Vichett Oung	Commission Bancaire, Paris
Mr Yasushi Shiina	Financial Services Agency, Tokyo
Mr Dirk Tasche	Deutsche Bundesbank, Frankfurt
Mr Satoshi Yamashita	Financial Services Agency, Tokyo

Table of Contents

Executive summary

In June 2004, the Basel Committee on Banking Supervision issued a revised framework on International Convergence of Capital Measurement and Capital Standards (hereafter "Basel II" or the "revised Framework"). When following the "internal ratings-based" (IRB) approach to Basel II, banking institutions will be allowed to use their own internal measures for key drivers of credit risk as primary inputs to their minimum regulatory capital calculation, subject to meeting certain conditions and to explicit supervisory approval.

In light of the need under Basel II for banks and their supervisors to assess the soundness and appropriateness of internal credit risk measurement and management systems, the development of methodologies for validating external and internal rating systems is clearly an important issue. More specifically, there is a need to develop means for validating the systems used to generate the parameters (such as PD, LGD, EAD and the underlying risk ratings) that serve as inputs to the IRB approach to credit risk. In this context, validation comprises a range of approaches and tools used to assess the soundness of these elements of IRB systems.

In anticipation of the need for more knowledge regarding validation methodologies, in 2002 the Research Task Force (RTF) formed a subgroup (the Validation Group) to review and develop research on the validation of rating systems that would be useful to banks and supervisors as they consider options for implementing Basel II. The work of the Validation Group collected in this volume of studies addresses a number of topics on rating system validation, with a particular focus on empirical validation methods.

The Validation Group consists of representatives from eleven countries.[1] The main objectives of the project have been:

- to classify rating systems and their dynamic properties, and to develop a common terminology for validation purposes,
- to review validation methodologies that are currently applied in bank practice, and
- to analyse validation methodologies for the three key risk components probability of default (PD), loss given default (LGD) and exposure at default (EAD) from a theoretical perspective.

Importantly, the collection of studies presented here is not intended to represent a comprehensive survey of all available validation methods and processes. Rather, the RTF expects that this research will provide a valuable input to the work of the Basel Committee's Accord Implementation Group (AIG), national supervisors, and banks as they develop approaches for validating the risk parameters and rating systems needed to implement Basel II. In addition, it should be emphasised that these studies largely reflect the views of individual authors, and should not be viewed as representing specific Basel Committee guidance for national supervisors or financial institutions.

Although validation is foremost the responsibility of banks, both bank risk managers and bank supervisors need to develop a thorough understanding of validation methods. Supervisors will need to review banks' validation processes, and may also need to employ

[1] Participating countries are Belgium, France, Germany, Italy, Japan, Netherlands, Norway, Singapore, Spain, the U.K., and U.S.A.

validation methods in evaluating whether banks' rating systems comply with the operating standards set forth by Basel II. Some validation methods, such as benchmarking risk parameters across banks, may be more practical for supervisors to implement than banks. The focus of the research in this collection has been on validation methods in general, without regard to whether those methods are implemented by banks or their supervisors.

The key results of the studies in this volume are summarised below. When reading the summary and the original papers, it is important to understand that validation methodologies applied in the banking industry are still preliminary in nature. However, encouraged by the focus of Basel II on rating systems, it is also fair to observe that validation has received the attention of both the industry and academics, as reflected in a growing stream of literature on this topic. Many open issues remain; some are conceptual, some are caused by insufficient data, and some may be better treated as technical aspects within the implementation process. In the latter case, it is expected that they will be resolved by national supervisors or by the Accord Implementation Group.

Dynamics of rating systems

Analysis of a stylised model of rating systems indicates that the default probability assigned to each obligor rating grade and its dynamics strongly depend on the type of rating methodology and quantification techniques employed. Therefore, banks and supervisors should take into account differences in rating assignment methods and quantification approaches when applying a validation methodology.

The dynamics of default probabilities assigned to rating grades are explored by analysing the properties of stylised rating systems of the types often described as point-in-time and through-the-cycle. The impact of using idealised stressed rather than unstressed obligor-specific PDs to determine the pooled PD for a risk "bucket" (such as an internal obligor grade) is also considered. The analysis of these stylised rating systems provides some interesting insights into the impact of using the approaches outlined in the revised Framework (i.e. the historical default experience approach, the statistical model approach or the external mapping approach) for PD estimation in different rating systems.

The results of this analysis suggest that the pooled default probability assigned to each rating grade and its dynamics strongly depend on the type of rating system and the PD estimation method. The estimation from historical default rates is most meaningful when the pooled PDs are unstressed, which means that they are unbiased estimates of the likelihood of default in the following year. Furthermore, the analysis suggests that the long-run average default frequency for a through-the-cycle bucket will not provide a good approximation of that bucket's unstressed pooled PD. The reason is that the unstressed pooled PD will tend to be lower than the long-run average default frequency during cyclical peaks and higher than the long-run average default frequency during cyclical troughs.

The statistical models approach is potentially more flexible, but is only as accurate as the underlying statistical models used to estimate obligor-specific PDs.

In the case of external mapping, the analysis suggests that if there are differences in the dynamics of a bank's internal rating system and the external rating system used to quantify pooled PDs, then one might expect the mapping between internal and external grades to change from year to year. Only if a bank's approach to setting internal ratings is the same as that used in setting the external ratings can one expect the mapping between the two systems to remain stable over time.

Validation of PD, LGD and EAD

Estimation and validation methodologies for PD are significantly more advanced than those for LGD and EAD. For all three risk components, the use of statistical tests for backtesting is severely limited by data constraints. Therefore, a key issue for the near future is the building of *consistent data sets* in banks. Initiatives to pool data that have been started by private banking associations may be an important step forward in this direction, especially for smaller banks.

For the validation of PDs, we differentiate between two stages: validation of the discriminatory power of a rating system and validation of the accuracy of the PD quantification (*calibration*). Numerous methods exist for the assessment of the discriminatory power. The most common techniques are the cumulative accuracy profile (CAP) and the accuracy ratio, which condenses the information of the CAP into a single number. Portfolio-dependent confidence intervals that allow statistical inference from the accuracy ratio are given in the report.

Compared with the evaluation of the discriminatory power, methods for validating calibration are at a much earlier stage. However, stimulated by the progress of Basel II, such methods have attracted considerable interest in academic research. A major obstacle to backtesting of PDs is the scarcity of data, caused by the infrequency of default events and the impact of default correlation. Even if the final minimum requirements of the revised Framework for the length of time series for PDs (five years) are met, the explanatory power of statistical tests will still be limited. Due to correlation between defaults in a portfolio, observed default rates can systematically exceed the critical PD values if these are determined under the assumption of independence of the default events. This can happen easily for otherwise well-calibrated rating systems. As a consequence, on the one hand, all tests based on the independence assumption are rather conservative, with even well-behaved rating systems performing poorly in these tests. On the other hand, tests that take into account correlation between defaults will only allow the detection of relatively obvious cases of rating system miscalibration. Therefore, statistical tests alone will be insufficient to adequately validate an internal rating system. Nevertheless, banks should be expected to use various quantitative validation techniques, as they are still valuable tools for detecting weaknesses in rating systems.

Due to the limitations of using statistical tests to verify the accuracy of the calibration, benchmarking can be a valuable complementary tool for the validation of estimates for the risk components PD, LGD and EAD. Benchmarking involves the comparison of a bank's ratings or estimates to results from alternative sources. It is quite flexible in the sense that it gives banks and supervisors latitude to select appropriate benchmarks. An important technical issue is the design of the mapping from an individual bank's estimates to the benchmark. If benchmarking is carried out by the bank, its supervisory authority may choose to focus primarily on assessing the quality of the benchmark and the quality of the mapping. A dynamic approach to benchmarking (as described in these studies) seems to be promising, and would allow supervisors to make inferences about the characteristics of the internal rating system. Despite the usefulness of benchmarking, it should be used as a complement to, not a substitute for, statistical validation methods.

The sensitivity of capital charges to LGD and poor coverage of LGD in the literature has motivated a special emphasis in these studies on the validation of LGD. Compared to PD, much less is known about what drives LGD. Therefore, the studies concentrate more on issues that affect the estimation of LGD than on validation methods. The studies find that a qualitative assessment of the bank's LGD estimation process may be a more meaningful

validation method than the use of quantitative methods, and provides guidance on the components of such an assessment process.

In general, four methods are available for the estimation of LGDs: a workout LGD based on the discounted cash flows after default; a market LGD based on prices of traded defaulted loans; an implied market LGD that is derived from non-defaulted bond prices by means of an asset pricing model; and (in the special case of a retail portfolio) an implied historical LGD based on the experience of total losses and PD estimates. The studies in this volume focus on workout LGDs because they appear likely to be a common methodological choice of banks attempting to meet the IRB minimum requirements. Several critical issues for the estimation of workout LGDs are highlighted in the studies, including how to measure recoveries, how to allocate workout costs, and how to select an appropriate discount factor. Other important issues for estimation include consistency between the definitions of default used for PD and LGD, and the precise definition of losses (for instance whether the observed losses are censored by forcing them to be non-negative).

The obstacles that impede the validation of LGD are also present when EAD is estimated and validated. The key problem here is to determine the potential future draw-down of unused commitments. Literature on the estimation and validation of EADs is virtually non-existent and data constraints are even more severe than for LGDs, where at least one can draw some inferences from publicly available bond data.

Outlook

It appears certain that validation techniques and processes will continue to develop and evolve, as will the rating systems to which validation methodologies are applied. To some degree this evolution is driven by the requirements of Basel II. However, it also represents a continuation of longer-term trends in the financial services sector toward more rigorous and quantitative risk measurement and management methodologies, and the need to develop techniques for ensuring that those methodologies operate as intended. Ongoing developments likely will continue to attract the active attention of the research community, the industry, and supervisory authorities.

Ultimately, the task of providing supervisory guidance to institutions falls to national supervisory authorities. These authorities can reap significant benefits from active sharing of knowledge regarding concepts, techniques, and applications in such a rapidly evolving field. In this regard, the AIG recently established a subgroup on IRB validation to encourage the sharing of information and to promote consistency where appropriate in supervisory approaches to validation. In its work to date, the AIG validation subgroup has articulated several key principles that should guide IRB validation. These are:

- The bank has primary responsibility for validation.
- Validation is fundamentally about assessing the predictive ability of a bank's risk estimates and the use of ratings in credit processes.
- Validation is an iterative process.
- There is no single validation method.
- Validation should encompass both quantitative and qualitative elements.
- Validation processes and outcomes should be subject to independent review.

These principles lay the groundwork for future work of the AIG validation subgroup; the case for most if not all of these principles is strengthened by the various studies conducted by members of the RTF Validation Group and presented in this volume. It is hoped that these studies will prove valuable to both supervisors and banks in their ongoing work to refine validation concepts, techniques, and methods.

Studies on the Validation of Internal Rating Systems

I. Introduction

Rating systems are a cornerstone for the calculation of banks' regulatory capital charge in the internal ratings-based (IRB) approach of the revised Framework (Basel II) because they are the basis for the determination of a borrower's probability of default (PD). The PD and the other two risk components, loss given default (LGD) and exposure at default (EAD), are key input parameters to the regulatory capital calculation. As a consequence, validation of these three parameters and the underlying rating system is a key component of the supervisory review process.

Explicit requirements in the revised Framework underline the need to validate internal rating systems.[2] Banks must demonstrate to their supervisor that they can assess the performance of their internal ratings and their risk estimation systems consistently and meaningfully. More detailed requirements demand, for example, that realised default rates have to be within an expected range, that banks must use different quantitative validation tools and that well-articulated internal standards must exist for situations where significant deviations occur between observed values of the three risk components and their estimates.

The design of a validation methodology depends on the type of rating system. Rating systems can differ in various ways, depending on the borrower type, the materiality of the exposure, the dynamic properties of the rating methodology (e.g. point-in-time vs. through-the-cycle), and the availability of default data and external credit-quality assessments (external ratings, vendor models). As a consequence, validation is a relatively complex issue and requires a good understanding of the rating system and its properties.

The following studies summarise the work of the Validation Group. This group was formed by the Research Task Force to explore validation methodologies for rating systems from a theoretical perspective and to assess current validation practices in the banking industry.[3]

The Validation Group has explored a broad range of qualitative and quantitative validation techniques. It has considered contributions from the literature and the results from a bank survey in order to understand how validation is treated in academia as well as in the banking industry.

The validation project has progressed in three stages. The first stage began with a literature survey on validation methods and their performance in banking practice. This was important for developing a common terminology and for a classification of rating systems.

A key result of the first stage was that statistical tests are less meaningful to validate PD estimation than they are in the case of internal market risk models. Therefore, backtesting based on statistical tests is generally not powerful enough to determine if an internal rating system is acceptable. Consequently, the focus of the project was extended to benchmarking.

---

[2] See BCBS (2004), paragraphs 500–505.

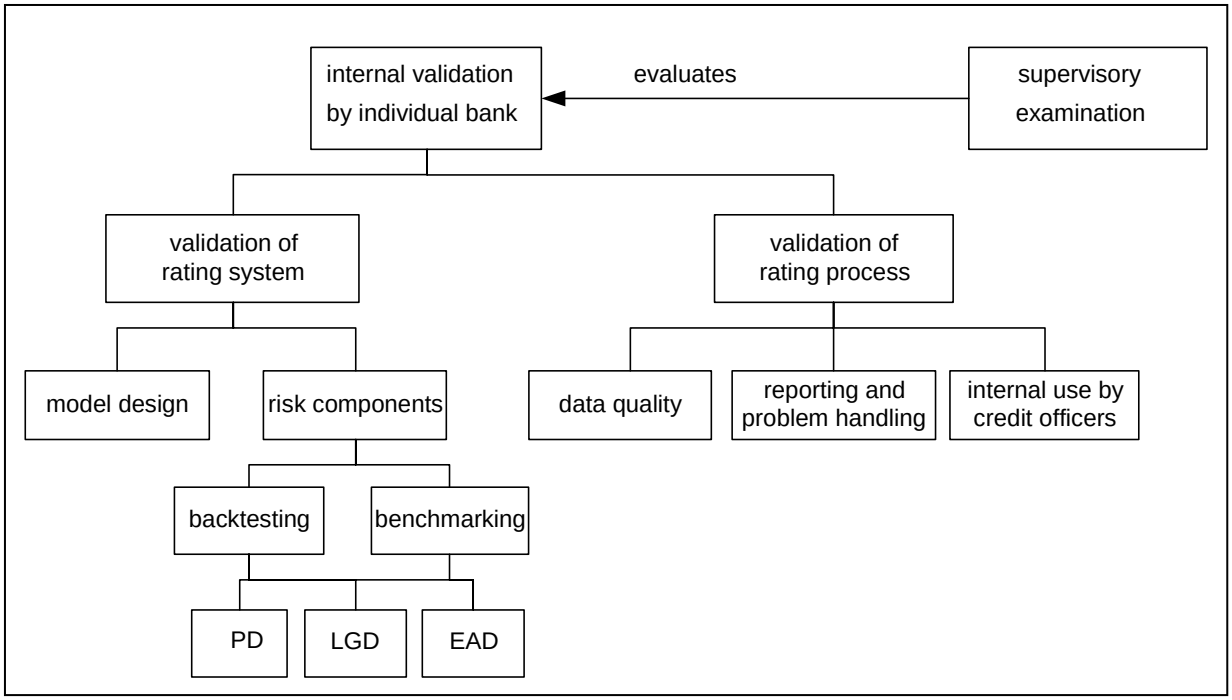
[3] Participating countries are Belgium, France, Germany, Italy, Japan, Netherlands, Norway, Singapore, Spain, the U.K., and U.S.A.

Key components of validation

An important issue at the outset of the project was to describe the key components of validation as a concept. The validation process involves the examination of the rating system and the estimation process and methods for the risk components PD, LGD and EAD. It also requires verification of the minimum requirements for the IRB approach. The application of validation methods is closely linked to the type of rating system and its underlying data basis. E.g., ratings for small business lending will typically be of a more quantitative nature, based on a rather large quantity of data. Sovereign ratings instead will typically lay more emphasis on qualitative aspects because these borrowers are more opaque and default data are scarce.

Figure 1 shows key components of a validation methodology.

Figure 1. Validation components.



Individual banks undertake validation as a means of ensuring that the output produced by internal rating systems is suitable for internal uses and to verify compliance with the use test as defined in the revised Framework. In an examination, supervisors evaluate the validation conducted by the individual bank. As a result, supervisors may use some of the same validation techniques as the banks.

Validation by a banking institution consists of two main components: validation of the rating system and the estimates of the risk components (PD, LGD, and EAD), and validation of the rating process, focusing on how the rating system is implemented.

The validation of the rating system can be further broken down into two components, the evaluation of the *rating system design* or *model design* and an assessment of the *estimates of the risk components*. In both cases, qualitative and quantitative methods can be applied.

In the case of a model-based rating system, the validation of the model *design* should include, for example, a qualitative review of the statistical model building technique, the relevance of the data used to build the model for the bank's specific business segment, the

way the risk factors that are the key inputs to the models were selected, and whether they are economically meaningful.

In the analysis of the estimates of the model parameters PD, LGD and EAD we differentiate between backtesting and benchmarking.

- *Backtesting* means the use of statistical methods to compare estimates of the three risk components to realised outcomes. This differs from the traditional backtesting of market risk models in an important way. Whereas for market risk models backtesting involves the whole model, for internal rating systems only the risk components (model inputs) are tested and the "model" is provided by the supervisor in the shape of the risk-weight functions.
- *Benchmarking* refers to a comparison of internal estimates across banks and/or with external benchmarks (e.g. external ratings, vendor models, or models developed by supervisory authorities).

In addition to an evaluation of the rating system, validation comprises an evaluation of the rating process. This involves important issues like data quality, the internal reporting, how problems are handled and how the rating system is used by the credit officers. It also entails the training of credit officers and a uniform application of the rating system across different branches. Although quantitative techniques are useful, especially for the assessment of data quality, the validation of the rating process is mainly qualitative in nature and should rely on the skills and experience of typical banking supervisors.

Structure

The studies are structured as follows: Section II provides an introduction to the terminology and a classification of rating systems. It explores how different rating concepts affect the dynamic properties of the PD estimates.

Sections III, IV and V deal with the validation of the three key risk components in the regulatory capital calculation: PD, LGD and EAD. Various quantitative validation methods for rating systems and PD estimates are discussed in Section III. The purpose is to provide a general assessment of methods that measure the discriminatory power of a rating system and the performance of the PD quantification (calibration). The technical details of the quantitative methods are described in an appendix. Section IV deals with the validation of LGD estimates. Compared with the validation of PDs, the literature on LGD validation is still scarce but growing since it is attracting more attention in academia, furthered by the Basel II proposals. The validation of EAD, the third risk component, is discussed in Section V.

Another quantitative approach that can be used as a complement to backtesting is benchmarking of internal estimates against external sources. Several benchmarking concepts are discussed in Section VI. This may be especially promising for supervisory authorities that have access to internal ratings of different banks.

II. Dynamics of rating systems

Erik A Heitfield

Introduction

This section develops a theoretical framework for comparing alternative approaches to quantifying the PD parameters reported by IRB banks. It shows that the dynamic features of reported PDs depend on how default events are predicted and on individual banks' risk-rating philosophies. Therefore, supervisors and bank risk managers will need to adapt PD validation techniques to banks' credit rating and quantification systems. They will need to understand how a bank assigns risk ratings and how it calculates default probabilities in order to accurately evaluate the accuracy of reported PDs.

Under Basel II, an IRB bank will be required to report a quantitative assessment of the probability of default for each obligor represented in its loan portfolio. The process by which PDs are assigned to obligors is clearly articulated in the revised Framework.[4] An IRB bank must first assign obligors to "risk buckets". All obligors assigned to a bucket should share the same credit quality as assessed by the bank's internal credit rating system. Once obligors have been grouped into risk buckets, the bank must calculate a "pooled PD" for each bucket. The credit-risk capital charges associated with exposures to each obligor will reflect the pooled PD for the risk bucket to which the obligor is assigned.

The revised Framework establishes minimum standards for IRB banks' internal rating processes and outlines permissible approaches to estimating pooled PDs, but it permits banks a great deal of latitude in determining how obligors are assigned to buckets and how pooled PDs for those buckets are calculated.[5] Although this flexibility allows banks to make maximum use of their own internal rating and credit data systems in quantifying PDs, it also raises important challenges for PD validation. Supervisors and bank risk managers will not be able to apply a single formulaic approach to PD validation because dynamic properties of pooled PDs depend on each bank's particular approach to rating obligors. Supervisors and risk managers will have to exercise considerable skill to verify that a bank's approach to PD quantification is consistent with its rating philosophy.

In comparing alternative approaches to quantifying PDs, it is helpful to draw a distinction between the concept of a default probability linked to an individual obligor and the pooled PD assigned to a risk bucket. The PD associated with an obligor is a measure of the likelihood that that obligor will default during a one-year assessment horizon. The pooled PD assigned to a risk bucket is a measure of the average level (i.e. the mean or the median) of the PDs of obligors within that bucket. One can therefore separate the question of how a particular bank's pooled PDs can be quantified into three related but distinct questions.

1. What properties should obligor-specific PDs possess?

2. How does the bank assign obligors to risk buckets?

3. In light of the answers to the first two questions, how can pooled PDs be derived that accurately reflect the PDs of obligors assigned to each risk bucket?

---

[4] BCBS (2004) paragraph 285 stipulates that pooled PDs should be linked to risk buckets rather than directly to obligors. Paragraphs 452–457 define the default event that PDs are intended to forecast.

[5] BCBS (2004) paragraphs 446–451 set out broad standards for the quantification of IRB risk components including PDs. Paragraphs 461–463 discuss specific requirements for assigning pooled PDs to risk buckets.

As discussed in the second part of this section an obligor-specific PD may or may not embed stress-scenario assumptions about future economic conditions. Obligor-specific PDs that incorporate current credit-quality information and do not impose stress-scenario assumptions are likely to change rapidly as prevailing economic conditions change. They will tend to fall during upturns of the business cycle and rise during economic downturns. Obligor-specific PDs that do not incorporate dynamic information on credit quality or that impose stress-scenario assumptions will tend to remain relatively stable over the business cycle. Assuming that pooled PDs accurately reflect the average level of obligor-specific PDs, the characteristics of obligor-specific PDs will have an influence on the extent to which overall credit-risk capital requirements vary over the business cycle. The revised Framework does not explicitly discuss the characteristics that obligor-specific PDs should posses, so the answer to the first question listed above may well differ from country to country depending on national supervisors' assessments of the tradeoffs between the benefits of credit-risk capital requirements that are sensitive to changing economic conditions versus the benefits of capital requirements that are relatively stable over the business cycle.

The answer to the second question will vary from bank to bank. Two canonical approaches to rating obligors are discussed in the third part of this section. Banks whose ratings are used primarily for underwriting purposes are likely to implement systems that are "through-the-cycle" (TTC). TTC ratings will tend to remain more-or-less constant as macroeconomic conditions change over time. On the other hand, banks whose ratings are used for pricing purposes or to track current portfolio risk are more likely to implement "point-in-time" (PIT) rating systems. PIT ratings will tend to adjust quickly to a changing economic environment. Between these two extreme cases lie hybrid rating systems that embody characteristics of both PIT and TTC rating philosophies. To effectively validate pooled PDs, supervisors and risk managers will need to understand the rating philosophy applied by a bank in assigning obligors to risk buckets.

Since a pooled PD is intended to measure the average level of the obligor-specific PDs for the obligors assigned to a risk bucket, the dynamic features of pooled PDs will depend on both the characteristics of obligor-specific PDs and a bank's rating philosophy. The fourth part of this section shows that some combination of obligor-specific PDs and rating philosophies will lead to pooled PDs that remain constant over an economic cycle, while others can produce pooled PDs that are either positively or negatively correlated with the business cycle. The most accurate approach to quantifying pooled PDs depends in large measure on whether and how pooled PDs are expected to vary over a business cycle. Thus, only after supervisors and risk managers have enumerated desired properties for obligor-specific PDs and are able to characterise a bank's rating philosophy will they be in a position to compare alternative approaches to PD quantification.

The revised Framework sets out three broad approaches to quantifying pooled PDs.[6] These are discussed in the fifth part of this section. The historical default experience approach to PD quantification is most appropriate in circumstances where the pooled PD associated with a risk bucket can reasonably be expected to remain stable over time. On the other hand, a statistical models approach may be more useful in circumstances where the pooled PD assigned to a risk bucket is likely to change over the business cycle or is expected to incorporate stress-scenario assumptions. The accuracy of an external mapping approach depends on whether a bank's internal rating system incorporates the same rating philosophy as the external rating system to which it is mapped.

---

[6] See BCBS (2004), paragraph 462.

Many of the conclusions drawn in this section can be derived from a stylised theoretical model of credit risk rating systems. This model is described in detail in the appendix.

Characteristics of obligor-specific default probabilities

In its purest form, a probability of default is a forward-looking forecast of the likelihood that a particular obligor will default over a fixed assessment horizon (usually one year). Not all banks have systems in place for explicitly estimating default probabilities at the obligor level, and the revised Framework does not require that IRB banks develop such systems. Rather, the revised Framework requires that IRB banks be capable of assigning aggregate **pooled PDs** to risk buckets composed of many obligors. Nonetheless, since a bucket's pooled PD is intended to measure the average PD for obligors assigned to that bucket, our analysis of the dynamic characteristics of pooled PDs begins by focusing on the characteristics of default probabilities associated with individual obligors.

An obligor-specific PD may incorporate information relevant to assessing the obligor's ability and willingness to repay its debts, as well as information about the economic environment in which the obligor operates. It is convenient to divide the information available for forecasting defaults into two categories.

- **Aggregate information** is the information observable at the time a PD is estimated that is shared in common by many obligors. This category typically includes macroeconomic variables such as exchange rates, GDP growth rates, etc.
- **Obligor-specific information** is the information that is unique to a particular obligor. Such information may be relatively static, such as an obligor's line of business, or it may be more dynamic in character, such as an obligor's leverage ratio or current revenue**.**

Often aggregate information and obligor-specific information are highly correlated. For example, it is reasonable to expect that most corporate borrowers will experience higher revenues when GDP growth is robust and lower revenues when GDP growth slows. Similarly, loan-to-value ratios for residential mortgages will tend to be negatively related to aggregate house-price appreciation rates. To the extent that aggregate and exposure-specific data are correlated, using both types of information to forecast default may be redundant. Some banks may rely heavily on information about individual exposures and place relatively little emphasis on macroeconomic variables that are correlated with these data. Other banks may prefer to make heavy use of aggregate data, while using only those exposure-specific data that could not have been predicted from available aggregate data. Either approach may be valid since aggregate and exposure-specific variables contain overlapping information.

Like all economic forecasts, obligor-specific default probabilities must either implicitly or explicitly embed assumptions about future economic conditions. These assumptions may be extrapolated from current conditions, or they may reflect conservative **stress scenarios**. A stress scenario is a collection of assumptions about future economic conditions that are unlikely to occur over an assessment horizon but would tend to induce very high credit losses if they did occur.

Obligor-specific default probabilities are sensitive to the way that they use available information and the assumptions under which they are derived. In the analysis that follows we will consider two idealised representations of obligor-specific PDs.

- An **unstressed PD** is an unbiased estimate of the likelihood that an obligor will default over the next year given all currently-available information, including static

and dynamic obligor characteristics and aggregate data. Because this PD makes use of observable macroeconomic data, it is likely to fall as macroeconomic conditions improve and rise as they deteriorate.

- A **stressed PD** measures the likelihood that an obligor will default over the next year using all available obligor information, but assuming adverse stress-scenario economic conditions. Because this PD makes use of dynamic obligor characteristics it will change as an obligor's individual characteristics change, but it will tend not to be highly correlated with the business cycle.

Figure 2. Hypothetical stressed and unstressed default probabilities for a single obligor over a business cycle.

Dashed lines show long-run average obligor PDs.

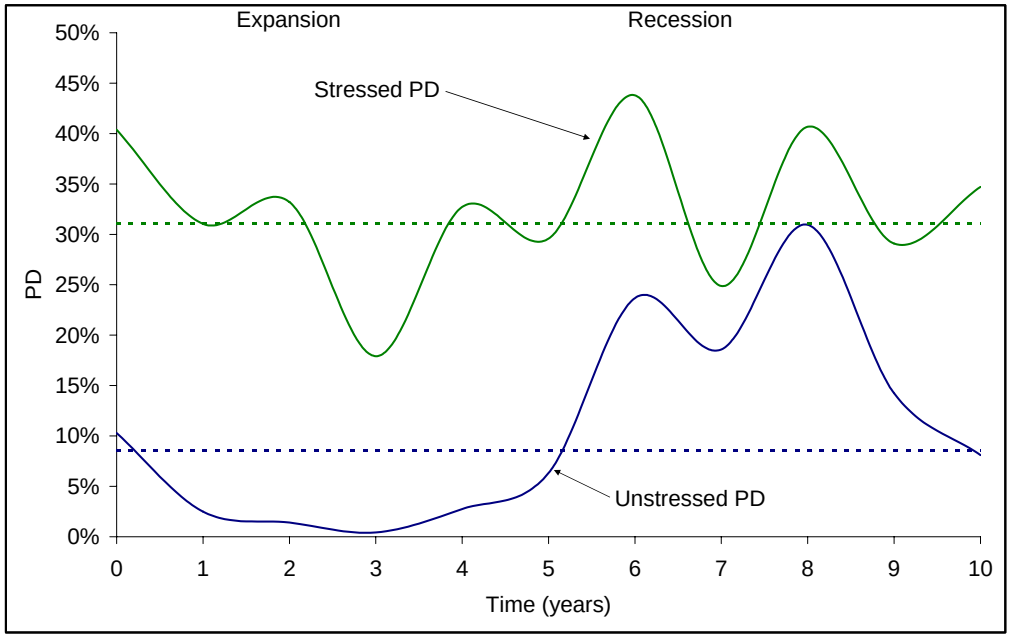


Figure 2 illustrates how these two different types of PDs might evolve over time for a single obligor. In this example, a business cycle peak occurs in years 2 and 3 and a trough occurs in years 7 and 8. Notice that the unstressed PD declines during expansion years and rises during recession years. At any particular date, most of the deviation of the unstressed PD from its long-run average is related to the business cycle. The stressed PD is "cyclically neutral" in the sense that while it moves as the obligor's particular circumstances change, it does not respond to changes in overall business conditions. Deviations between the stressed PD and its long-run average are driven by idiosyncratic factors unique to the obligor.

Note that these two PD types represent ideal cases. In practice, estimated PDs may lie in-between these benchmarks. For example, some banks have proposed reporting PDs that are derived by taking moving-averages of current and lagged unstressed PDs. Such "smoothed" PDs make use of dynamic aggregate and obligor-specific information but effectively place a greater weight on information that does not change over time. Similarly, risk components that are neither entirely stressed nor entirely unstressed could be derived by imposing stress-scenario assumptions on some, but not all, available aggregate data.

Classification of rating systems

All credit rating systems assign obligors to risk buckets that are intended to distinguish among obligors of differing credit quality. However, different rating systems accomplish this task in different ways. Some banks rely almost entirely on empirical **credit scoring models**. Statistical models map obligor characteristics to credit scores and obligors with similar scores are then grouped into common risk buckets. Other systems rely more heavily on **expert judgment**. Under these systems, experienced credit officers review available information about an obligor and then apply a variety of qualitative screens to match the obligor's rating to documented rating criteria. Often rating systems make use of both statistical models and expert judgment. For example, some banks use scoring models to assign obligors to grades, but allow loan officers to modify those grades under special circumstances. Alternatively, some banks that primarily rely on expert judgment nonetheless expect loan officers to consider credit scores along with other information when assigning rating grades. Banks may use different approaches to rating different types of credit exposures. For example, some banks rely on automated scoring models to rate relatively small and homogeneous retail exposures while relying on expert judgment to rate larger corporate exposures.

Practitioners use the terms "point-in-time" or "through-the-cycle" to describe the dynamic characteristics of rating systems, but these terms often mean different things to different people. Broadly, point-in-time systems attempt to produce ratings that are responsive to changes in current business conditions while through-the-cycle systems attempt to produce ordinal rankings of obligors that tend not to change over the business cycle. Point-in-time systems tend to focus on the current conditions of an obligor, while through-the-cycle systems tend to focus on an obligor's likely performance at the trough of a business cycle or during adverse business conditions.

The analysis of PD characteristics in the second part of this section suggests that one can classify rating systems by considering the way different systems utilise available obligor-specific and aggregate information. Regardless of whether a rating system relies on expert judgment or statistical models, one can think of a risk bucket as a collection of obligors with similar PDs. Viewed in this light, differences between point-in-time and through-the-cycle rating systems reflect differences in the characteristics of obligor-specific PDs used to assign obligors to risk buckets.

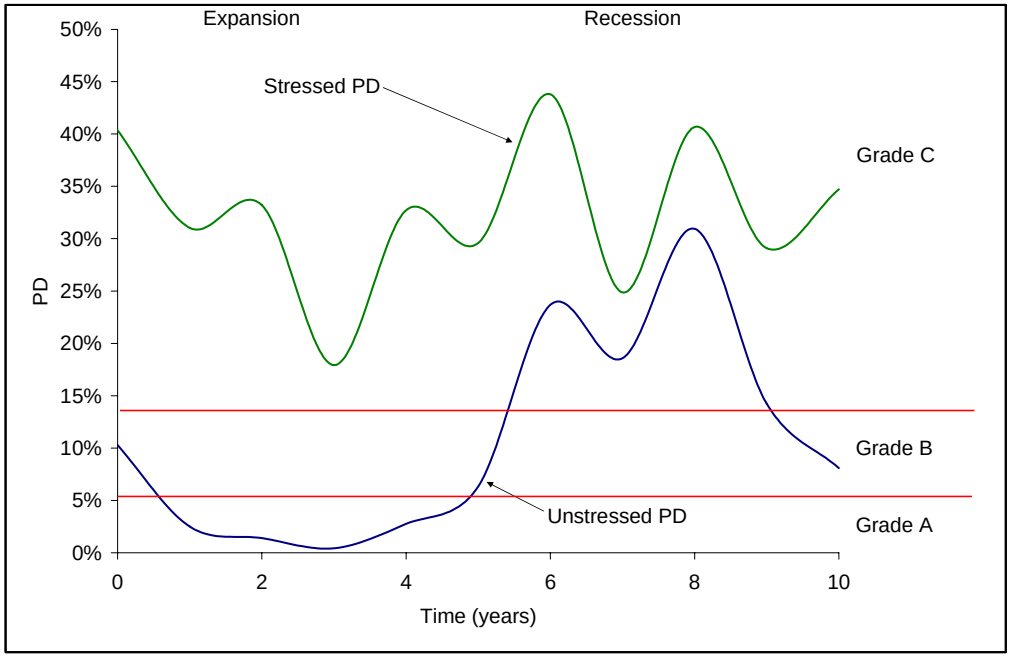
- A **point-in-time** (PIT) rating system uses all currently available obligor-specific and aggregate information to assign obligors to risk buckets. Obligors with the same PIT grade are likely to share similar unstressed PDs. An obligor's rating can be expected to change rapidly as its economic prospects change. Overall, PIT ratings will tend to fall during economic downturns and rise during economic expansions.
- A **through-the-cycle** (TTC) rating system uses static and dynamic obligor characteristics but tends not to adjust ratings in response to changes in macroeconomic conditions. Obligors with the same TTC grade are likely to share similar stressed PDs. An individual obligor's rating may change as its own dynamic characteristics change, but the distribution of ratings across obligors will not change significantly over the business cycle.

Figure 3 and Figure 4 illustrate simple examples of point-in-time and through-the cycle rating assignments based on the same obligor-specific PDs plotted in Figure 2. Notice that as the obligor's unstressed PD changes, its PIT rating changes as well. Near the peak of the business cycle the obligor receives a higher PIT rating and near the trough of the cycle it receives a lower rating. In contrast the TTC rating is tied to the obligor's unstressed PD. This PD fluctuates over time, but is not correlated with the business cycle. As a result, the obligor's TTC rating does not reflect changes in overall business conditions.

Between point-in-time and through-the-cycle rating systems lie a range of **hybrid** rating systems. These systems may exhibit characteristics of both TTC and PIT rating philosophies. For example, according to Standard and Poor's, its corporate ratings primarily reflect long-run assessments of credit quality (a TTC approach) but these ratings are allowed to vary to a limited extent as current business conditions change (typical for PIT approaches).[7] Similarly, Moody's sets ratings by assigning only "modest weight" to current business conditions (a TTC approach) but with the broad objective of ensuring that expected losses or default rates within a grade are stable over time (typical for PIT approaches).[8]

Figure 3. Example of a three-grade point-in-time rating system tied to an obligor's unstressed PD.

During the economic expansion the unstressed PD declines and the obligor receives a higher rating. During the economic recession the unstressed PD increases and the obligor receives a lower rating.



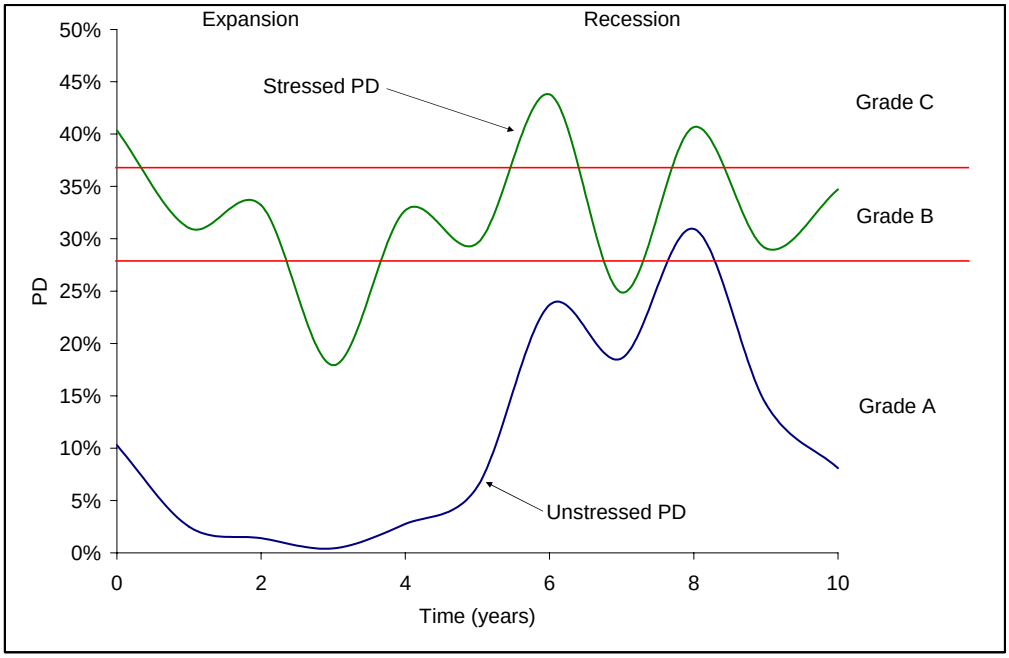
Year	0	1	2	3	4	5	6	7	8	9	10
Grade	B	A	A	A	A	B	C	C	C	B	B

[7] See Standard and Poor's (2002), p. 41–43.

[8] See Moody's Investor Service (1999), p. 6–7.

Figure 4. Example of a three-grade through-the-cycle rating system tied to an obligor's stressed PD.

While the obligor's rating changes as its stressed PD changes over time, its rating is unrelated to the business cycle.



Year	0	1	2	3	4	5	6	7	8	9	10
Grade	C	B	B	A	B	B	C	A	C	B	B

In a survey of internal rating systems at large U.S. banks, Treacy and Carey find that banks tend to focus more narrowly on current conditions in setting ratings than do public rating agencies.[9] This suggests that many bank rating systems may conform more closely to a PIT philosophy.

Dynamics of pooled PDs

This section examines the dynamic characteristics of the pooled PDs that IRB banks will be required to report. For the time being, we will abstract from the details of how these parameters are quantified, and simply assume that a bucket's pooled PD is equal to the mean of the obligor-specific PDs for obligors assigned to that bucket. This assumption is consistent with the view that pooled PDs should reflect within-bucket averages of obligor-specific PDs and allows us to focus on the dynamic features of accurately-quantified pooled PDs. The fifth part of this section will examine specific approaches to quantifying pooled PDs.

As highlighted earlier, one can think of a PIT rating system as a system designed to ensure that all obligors within a grade share roughly the same unstressed PD. This implies that the

[9] See Treacy and Carey (1998), p. 897–921.

unstressed *pooled* PD for a PIT risk bucket should remain essentially fixed over the business cycle. As business conditions improve and unstressed obligor-specific PDs tend to rise, higher-quality obligors will transition out of a PIT risk bucket (to a higher-rated bucket) and lower-quality obligors will transfer into the bucket so that the average unstressed pooled PD for the bucket remains fixed. The reverse happens when aggregate business conditions deteriorate. It is important to note that stable pooled PDs for rating grades do not imply that economic capital for the portfolio is also fixed because the distribution of obligors across grades will tend to shift over the business cycle.

Although the unstressed pooled PD for a PIT risk bucket should be stable over time, this is not true for the stressed pooled PD. Somewhat surprisingly, the stressed pooled PD for a PIT risk bucket will tend to be positively correlated with the business cycle. That is, it will tend to increase as aggregate business conditions improve and decrease as aggregate conditions deteriorate. The reason for this seemingly paradoxical result is that as overall business conditions improve, obligors with worse static characteristics (e.g. those engaged in riskier lines of business) must transition into the PIT risk bucket (from a lower rating bucket) to maintain a constant unstressed pooled PD. As a result the average value of the stressed PDs for obligors currently assigned to a PIT risk bucket will tend to rise as business conditions improve.

In the stylised analytic framework presented in the previous section, a TTC rating system is designed so that all obligors assigned to a TTC risk bucket will share roughly the same stressed PD. This implies that the stressed pooled PD for a TTC risk bucket will remain relatively stable over time. Individual obligors may transition in and out of a TTC risk bucket as their particular circumstances change, but strong cyclical patterns in such transitions should not be observed.

In contrast, the unstressed pooled PD associated with a TTC risk bucket can be expected to be negatively correlated with the business cycle. As economic conditions improve, most obligors' unstressed PDs will tend to fall, even as their stressed PDs remain stable. As a result, the unstressed PDs for obligors assigned to a TTC risk bucket will tend to fall during upswings in the business cycle and rise during economic downturns.

Table 1 summarises the dynamic characteristics of pooled PDs. The most important point to take away from this table is that without a clear articulation of the properties that obligor-specific PDs should possess and a detailed understanding of a bank's rating philosophy, supervisors and risk managers cannot know whether the pooled PDs that a bank reports should be expected to remain constant over time or should tend to be positively or negatively correlated with the business cycle. The dynamic features of accurately-quantified pooled PDs will depend on both a bank's rating philosophy and on whether obligor-specific PDs are assumed to be stressed or unstressed.

Table 1

Correlation of pooled PDs with the business cycle

Properties of obligor PDs	**Rating philosophy**	
	Point-in-time	**Through-the-cycle**
Unstressed	Stable	Negative
Stressed	Positive	Stable

Quantifying pooled PDs

The previous section examined the dynamic properties of pooled PDs under the assumption that those parameters accurately reflect the average level of obligor-specific PDs within a risk bucket. In practice, of course, IRB banks will need to estimate pooled PDs from available data. Bank supervisors and risk managers will be tasked with evaluating PD quantification methodologies to ensure that those methodologies produce accurate estimates of pooled PDs. The revised Framework outlines three broad approaches to quantifying pooled PDs.

- Under a **historical default experience approach,** the pooled PD for a risk bucket is estimated using historical data on the frequency of observed defaults among obligors assigned to that bucket.
- Under a **statistical model approach,** predictive statistical models are used to estimate a default probability for each obligor currently assigned to a bucket. The bucket's pooled PD is then calculated as the average (or the median) of the obligor-specific PDs.
- Under an **external mapping approach,** a mapping is established that links each of a bank's internal risk buckets to external rating grades. Pooled default probabilities for the external grades are calculated from external data and then assigned to the bank's internal grades via the mapping.

Each of these approaches has important strengths and weaknesses that depend on the dynamic characteristics of the pooled PDs being estimated. This section will examine each approach in turn.

Historical default experience

The **default frequency** (DF) for a risk bucket is defined as the observed default rate for the bucket over a fixed assessment horizon (usually one year). That is:

$$DF_t = \frac{D_t}{N_t},$$

where D_t is the number of defaults observed for a bucket over year t and N_t is the total number of obligors assigned to that bucket at the beginning of year t. The unstressed pooled PD for a risk bucket can be interpreted as an *ex ante* forecast of the one-year-ahead *ex post* observed default frequency for that bucket. However, because default events are generally correlated across obligors it is unlikely that in any given year a bucket's pooled PD will closely match its observed default frequency. During years when aggregate economic conditions unexpectedly improve the observed default frequency for a risk bucket will tend to fall below its dynamic unstressed pooled PD. During years when economic conditions unexpectedly deteriorate observed default frequencies will tend to lie above forecasts.

The **long-run default frequency** (LRDF) for a risk bucket is simply the average of that bucket's annual default rates taken over a number of years. In symbols:

$$LRDF = \frac{1}{T}\sum_{t=1}^{T} DF_t .$$

Over time, year-to-year differences between unstressed pooled PDs and observed default frequencies should tend to cancel out, and the LRDF for a risk bucket can be expected to converge toward the long-run average unstressed pooled PD for that bucket.

The unstressed pooled PD for a PIT risk bucket is constant over time, so the LRDF for a PIT risk bucket should closely approximate its unstressed pooled PD. This implies that given a sufficiently long history of ratings performance data, the historical default experience approach can provide an effective means of quantifying pooled PDs that reflect unstressed obligor PDs under a point-in-time ratings philosophy.

The historical default experience approach is of limited value for quantifying pooled PDs that tend to change over the business cycle. The long-run average default frequency for a TTC bucket will not provide a good approximation of that bucket's unstressed pooled PD, since that pooled PD will tend to be lower than the LRDF during cyclical peaks, and higher than the LRDF during business cycle troughs. Moreover, the historical default experience approach cannot be used to accurately quantify stressed PDs. By definition, stress-scenarios are unlikely to occur so the historical record of observed default frequencies will not generally reflect specific stress events. In general one should expect a bucket's stressed pooled PD to lie above that bucket's LRDF.

In summary, the historical default experience approach is most appropriate for quantifying unstressed pooled PDs for PIT risk buckets. It will be most accurate when long-run average default rates are calculated over a number of years. The historical default experience approach will not be effective for quantifying stressed pooled PDs, or pooled PDs that are expected to vary significantly over the business cycle.

Statistical models

The statistical models approach relies on an underlying empirical **default prediction model**. This model should take obligor-specific and aggregate information as inputs and should produce estimated obligor-specific default probabilities as outputs. Depending on the input data used, such a model could be used to generate either unstressed or stressed obligor-specific PDs. For example, unstressed PDs could be generated by using current obligor-specific and aggregate input data, while stressed PDs could be generated by substituting stress-scenario data for current aggregate input data.

Under the statistical models approach, a bucket's pooled PD is derived by taking the average of the estimated obligor-specific PDs for the obligors currently assigned to the risk bucket. This approach to quantifying pooled PDs can produce accurate estimates no matter what type of rating philosophy the bank applies. It therefore has some important advantages over the historical default experience approach. It can be used to quantify stressed pooled PDs, and it can be used to quantify pooled PDs that tend to vary significantly over the business cycle. However, it is important to recognise that the statistical models approach is only as accurate as the underlying default prediction model. If this approach is used for PD quantification, the practical challenge for bank supervisors and risk managers will lie in verifying that this model produces accurate estimates of the particular type of obligor-specific PDs under consideration.

External mapping

In some respects, the external mapping approach might appear to be the simplest of the three quantification methods outlined in the revised Framework. A bank simply establishes a mapping between its internal rating system and an external scale such as that of Moody's or S&P, calculates a pooled PD for each external grade using an external reference dataset, and then assigns the pooled PD for the external grade to its internal grade by means of the mapping. Despite its apparent simplicity, this approach poses some difficult validation challenges for supervisors and risk managers. To validate the accuracy of a bank's pooled PDs, supervisors and risk managers must first confirm the accuracy of the pooled PDs

associated with the external rating scale. They must then validate the accuracy of the bank's mapping between internal and external grades.

Quantifying pooled PDs for an external rating system poses the same estimation problems as quantifying pooled PDs for a bank's internal rating system. If a historical default experience approach is used, supervisors and risk managers must check to ensure that each bucket's pooled PD can be expected to approach its long-run default frequency over time. If a statistical models approach is used, supervisors and risk managers must validate the reliability of the underlying default prediction model. The main benefit of quantifying PDs using external ratings is that more data are likely to be available for calculating long-run default frequencies and/or estimating statistical default prediction models.

If a bank's approach to setting internal ratings matches that of the external rating agency, then the mapping between internal and external grades can reasonably be expected to remain stable over time. However, if the bank and the rating agency apply different rating philosophies the mapping may well change from year to year. For example, suppose a bank uses a PIT rating philosophy while an external rating agency uses a TTC philosophy. Since the distribution of PIT ratings shifts over the business cycle while the distribution of TTC ratings does not, one should expect the bank to apply a different mapping between internal and external grades at different points in the business cycle. Thus, validating mappings will require that supervisors and risk managers develop a detailed understanding of the rating philosophies applied by both banks and external rating agencies.

Conclusion

This section has shown that the relative merits of different approaches to quantifying pooled PDs depend on both a bank's rating philosophy (PIT or TTC) and the desired characteristics of obligor-specific PDs (stressed or unstressed). The historical default experience approach is most useful when pooled PDs are unstressed and can be expected to remain stable over time, for example when unstressed pooled PDs are assigned to point-in-time risk buckets. The statistical models approach is more flexible, but is only as accurate as the underlying statistical models used to estimate obligor-specific PDs. External mapping may allow banks to use longer data histories for estimating default prediction models or long-run default frequencies, but it requires a detailed understanding of the dynamics of both internal and external ratings.

Appendix: An illustrative model of rating system dynamics and PD quantification

Introduction

The appendix develops a stylised model of the relationship between the dynamic features of a credit rating system and the characteristics of the pooled PDs associated with that rating system. This model is not intended to be fully general; the objective here is to demonstrate how the characteristics of rating systems and pooled PDs interact. For this reason, wherever necessary, the model sacrifices realism in favour of expositional simplicity.

The appendix is organised into five sections. The first section describes a simple model of obligor default based on an underlying distance-to-default measure of credit quality. Taking this model as a point of departure, the second section defines two approaches to grouping obligors into risk buckets. A point-in-time philosophy groups obligors according to one-period-ahead predicted default frequencies. A through-the-cycle philosophy groups obligors according to stress-scenario default probabilities. The third section derives pooled PDs for risk buckets under point-in-time and through-the-cycle rating systems, and shows how the dynamic features of these pooled PDs reflect the way ratings are assigned. The fourth section uses the model to examine whether observed default frequencies can be expected to match pooled PDs. The final section summarises the results of this analysis. Throughout this appendix, Greek letters denote model parameters, capital letters denote random variables, and lower-case letters denote realisations of random variables.

Obligor default model

Default is modelled using a latent variable Z_{it}. Z_{it} is a normally distributed random variable that is unique to each obligor i and each date t. Obligor i defaults at date t if the realised value of Z_{it} lies below zero so Z_{it} can be viewed as a measure of obligor i's distance to default. Z_{it} evolves over time, and is assumed to depend on observable and unobservable risk factors and model parameters according to the formula

$$Z_{i,t+1} = \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y Y_t + U_{i,t+1}.$$

W_i is a fixed risk factor intended to capture characteristics of obligor i that do not vary over time such as industry and management quality. Y_t is a risk factor that affects the credit quality of all obligors in a bank's portfolio. It is intended to summarise that component of the macroeconomic environment that can be observed by a bank at date t. Y_t may exhibit serial correlation, but for simplicity we assume that the marginal distribution of Y_t is fixed over time. The risk factor X_{it} captures those dynamic characteristics of obligor i that are observable at date t and could not have been predicted given Y_t, so X_{it} is independent of Y_t. Taken together W_i, Y_t, and X_{it} represent the full set of information available to a bank for assessing the credit quality of obligor i at date t. The β parameters are assumed to be positive so that each risk factor is negatively related to credit quality.

The standard normal random variable $U_{i,t+1}$ reflects information that affects an obligor's default status at t+1 that cannot be observed by a bank at date t. Even after accounting for macroeconomic conditions observable at t, systematic risk will generate correlations in default outcomes across obligors. To capture this idea we assume

$$U_{i,t+1} = \omega V_{t+1} + \sqrt{1-\omega^2}E_{i,t+1}$$

where V_{t+1} is a systematic risk factor shared by all obligors, and $E_{i,t+1}$ is an idiosyncratic risk factor that is unique to obligor i. The parameter ω determines the sensitivity of default to the unobservable systematic risk factor. A value of ω near one implies that conditional on observable information at date t, defaults among obligors at date t+1 are highly correlated

events. Conversely, setting ω equal to zero implies that conditional on observable information at date t defaults at date t+1 are independent.

For simplicity, we assume that all risk factors have standard normal marginal distributions. Subject to this restriction, Y_t may depend on lagged values of Y_t and V_t. All other variables are assumed to be independent and identically distributed.

As discussed in the body of this section, one-period-ahead default probabilities can either be described as "unstressed" or "stressed". Unstressed default probabilities provide an unbiased prediction of the likelihood that an obligor will default, while stressed default probabilities predict the likelihood of default conditional on adverse stress-scenario assumptions about the prevailing macroeconomic environment. In the context of the current model, these two types of default probabilities can be defined explicitly in terms of the way observable variables are used to forecast an obligor's likelihood of default.

Let D_{it} be an indicator that is equal to one if obligor i defaults at date t and is equal to zero otherwise. An unstressed default probability uses all information available at date t to construct an unbiased measure of the likelihood that an obligor will default at date t+1. Thus, the unstressed default probability for obligor i at date t is defined as

$$\begin{aligned} PD_{it}^{U} &= E\left[D_{i,t+1} \mid W_i = w_i, X_{it} = x_{it}, Y_t = y_t\right] \\ &= \Phi\left(-\left(\alpha + \beta_W w_i + \beta_X x_{it} + \beta_Y y_i\right)\right). \end{aligned}$$

Stressed PDs incorporate adverse stress-scenario assumptions about the prevailing state of the macro economy into default forecasts. In the default model developed here the state of the macro economy at date t+1 is described by a weighted sum of the observable risk factor Y_t and the unobservable risk factor V_{t+1}. Thus a stress scenario can be defined as the condition

$$\beta_Y Y_t + \omega V_{t+1} = \psi$$

where ψ is a fixed parameter. The farther is ψ from zero (in the negative direction), the more pessimistic is the stress scenario. A stressed PD incorporates the stress-scenario assumption along with all data observable at date t. The stressed PD of obligor i at date t is given by

$$\begin{aligned} PD_{it}^{S} &= E\left[D_{i,t+1} \mid W_i = w_i, X_{it} = x_{it}, \beta_Y Y_t + \omega V_{t+1} = \psi\right] \\ &= \Phi\left(-\frac{\alpha + \beta_W w_i + \beta_X x_{it} + \psi}{\sqrt{1-\omega^2}}\right). \end{aligned}$$

Risk buckets

In assigning obligors to risk buckets, banks use exactly the same information needed to estimate default probabilities. As a result, for modelling purposes it is reasonable to assume that a bank will group obligors into risk buckets in such a way that obligors within a bucket share similar default probabilities. Viewed in this light, the two canonical rating philosophies described in the body of this section – point-in-time and through-the-cycle – reflect differences in the PDs that banks use for risk bucketing. Under a point-in-time rating system all obligors in a bucket should share similar unstressed PDs. Under a through-the-cycle rating system all obligors in a bucket should share similar stressed PDs. To obtain clear analytic results this section will analyze limiting cases in which all obligors within a point-in-

time bucket share the same stressed PD, and all obligors in a through-the-cycle risk bucket share the same unstressed PD.

A point in time risk bucket is defined as

$$\Gamma_t^{PIT} = \left\{ i \mid \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y Y_t = -\gamma_{PIT} \right\}$$

so that all obligors in the risk bucked share the unstressed default probability

$$PD_{it}^{U} \mid i \in \Gamma_{it}^{PIT} = \Phi\left(\gamma_{PIT}\right).$$

At a particular date *t*, the distribution of observed obligor-specific characteristics for obligors assigned to a point-in-time bucket will depend on the realisation of the observed macroeconomic risk factor y_t. This can be seen by noting that

$$E\left[W_i \mid i \in \Gamma_t^{PIT}, Y_t = y_t \right] = -\frac{\gamma_{PIT} + \beta_Y y_t}{\beta_W}$$

and

$$E\left[X_{it} \mid i \in \Gamma_t^{PIT}, Y_t = y_t \right] = -\frac{\gamma_{PIT} + \beta_Y y_t}{\beta_X}.$$

The mean of both the static and dynamic observable obligor-specific variables associated with obligors in a point-in-time grade tends to fall as the observable macroeconomic factor rises. Put another way, during economic downturns higher-quality obligors tend to transition to lower risk buckets, and during economic expansions lower-quality obligors tend to transition to higher risk buckets.

A through-the-cycle risk bucket is defined as

$$\Gamma_t^{TTC} = \left\{ i \mid \frac{\alpha + \beta_W W_i + \beta_X X_{it} + \psi}{\sqrt{1-\omega^2}} = -\gamma_{TTC} \right\}$$

so that all obligors in the risk bucket share the same stressed default probability

$$PD_S^{TTC} \mid \Gamma_t^{TTC} = \Phi\left(\gamma_{TTC}\right).$$

Under this philosophy the distribution of obligor-specific characteristics within a bucket tends not to change over time. Observe that

$$E_t\left[W_i \mid i \in \Gamma_t^{TTC} \right] = -\frac{\sqrt{1-\omega^2}\gamma_{PIT} + \psi}{\beta_W}$$

and

$$E_t\left[X_{it} \mid i \in \Gamma_t^{TTC} \right] = -\frac{\sqrt{1-\omega^2}\gamma_{PIT} + \psi}{\beta_X}.$$

Though individual obligors will transition between through-the-cycle risk buckets as their dynamic obligor-specific characteristics change over time, one should observe no secular pattern in transitions over the business cycle.

Pooled PDs

Under the IRB approach a bank will be expected to assign a pooled PD to each of its risk buckets. These pooled PDs should reflect the central tendency (the mean or the median) of the PDs of the individual obligors contained in each risk bucket. The stylised model of obligor default and rating systems developed in the previous two sections of this appendix can be used to derive analytic expressions or approximations for the two types of pooled PDs that can be assigned to point-in-time and through-the-cycle risk buckets. In this analysis we assume that the pooled PD assigned to a bucket is equal to the expected value (the population mean) of the PDs associated with all obligors currently assigned to that bucket. Thus, for example, the unstressed pooled PD for a through-the-cycle risk bucket can be evaluated by taking the expectation of the unstressed obligor PDs across a population of obligors who all share the same stressed PD at date *t*. Deriving analytic expressions for pooled PDs will allow us to examine how these parameters change over the business cycle.

Since all obligors assigned to a point-in-time risk bucket share the same unstressed PD, the unstressed pooled PD for a point-in-time risk bucket is simply

$$\begin{aligned} PPD^{U}_{PIT} &= E\left[PD^{U}_{it} \mid i \in \Gamma^{PIT}_{t}, Y_t = y_t \right] \\ &= E\left[E\left[D_{i,t+1} \mid W_i, X_{it}, Y_t \right] \mid \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y y_t = -\gamma_{PIT}, Y_t = y_t \right] \\ &= E\left[D_{i,t+1} \mid \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y y_t = -\gamma_{PIT}, Y_t = y_t \right] \\ &= \Phi\left(\gamma_{PIT} \right). \end{aligned}$$

Note that the expectation in the first line of the expression is taken across all possible realisations of W_i and X_{it} consistent with obligor *i* being assigned to the risk bucket whereas Y_t is fixed at the realised value y_t. The third line follows from the Law of Iterated Expectations. For a point-in-time rating bucket, the unstressed pooled PD is a constant that does not change over the business cycle. As economic conditions deteriorate and obligors' unstressed PDs tend to rise, higher-quality obligors are transferred into the bucket and lower-quality obligors are transferred out, so that on average the unstressed PD of obligors contained in the bucket remains fixed.

The stressed pooled PD for a point-in-time risk bucket at date *t* is

$$\begin{aligned} PPD^{S}_{PIT} &= E\left[PD^{S}_{it} \mid i \in \Gamma^{PIT}_{t}, Y_t = y_t \right] \\ &= E\left[E\left[D_{i,t+1} \mid W_i, X_{it}, \beta_Y Y_t + \omega V_{t+1} = \psi \right] \mid \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y y_t = -\gamma_{PIT}, Y_t = y_t \right] \\ &= E\left[D_{i,t+1} \mid \beta_Y y_t + \omega V_{t+1} = \psi, \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y y_t = -\gamma_{PIT}, Y_t = y_t \right] \\ &= \Phi\left(\frac{\beta_Y y_t + \gamma_{PIT} - \psi}{\sqrt{1-\omega^2}} \right). \end{aligned}$$

The stressed pooled PD for a point-in-time bucket will rise during economic upturns and fall during economic downturns.

To summarise, under a point-in-time rating systems pooled PDs that reflect unstressed obligor PDs can be expected to remain stable over the business cycle. Pooled PDs that

reflect stressed obligor PDs will tend to move in the same direction of the overall macroeconomic environment. They will tend to be higher during business cycle peaks and lower during business cycle troughs.

At date *t*, the unstressed pooled PD assigned to a through-the-cycle risk bucket is given by

$$\begin{aligned} PPD_{TTC}^{U} &= E\left[PD_{it}^{U} \mid i \in \Gamma_t^{TTC}, Y_t = y_t\right] \\ &= E\left[E\left[D_{i,t+1} \mid W_i, X_{it}, Y_t\right] \mid \alpha + \beta_W W_i + \beta_X X_{it} + \psi = -\sqrt{1-\omega^2}\gamma_{TTC}, Y_t = y_t\right] \\ &= E\left[D_{i,t+1} \mid \alpha + \beta_W W_i + \beta_X X_{it} + \psi = -\sqrt{1-\omega^2}\gamma_{TTC}, Y_t = y_t\right] \\ &= \Phi\left(-\beta_Y y_t + \sqrt{1-\omega^2}\gamma_{TTC} + \psi\right). \end{aligned}$$

This pooled PD moves in the opposite direction as the business cycle. As economic conditions improve (Y_t increases), the unstressed pooled PD for a through-the-cycle risk bucket declines. As economic conditions deteriorate the pooled PD increases.

All obligors in a through-the-cycle risk bucket share the same stressed PD. This can be verified by confirming that

$$\begin{aligned} PPD_{TTC}^{S} &= E\left[PD_{it}^{S} \mid i \in \Gamma_t^{TTC}, Y_t = y_t\right] \\ &= E\left[E\left[D_{i,t+1} \mid W_i, X_{it}, \beta_Y Y_t + \omega V_{t+1} = \psi\right] \mid \alpha + \beta_W W_i + \beta_X X_{it} + \psi = -\sqrt{1-\omega^2}\gamma_{TTC}, Y_t = y_t\right] \\ &= E\left[D_{i,t+1} \mid \beta_Y y_t + \omega V_{t+1} = \psi, \alpha + \beta_W W_i + \beta_X X_{it} + \psi = -\sqrt{1-\omega^2}\gamma_{TTC}, Y_t = y_t\right] \\ &= \Phi\left(\gamma_{TTC}\right). \end{aligned}$$

This pooled PD remains fixed over time.

In summary, the unstressed pooled PD associated with a through-the-cycle risk bucket can be expected to move in a systematic way over the business cycle. This pooled PD should be highest at the trough of the business cycle, and lowest at the peak of a business cycle. The stressed pooled PD associated with a through-the-cycle risk bucket should remain stable throughout the business cycle.

Backtesting pooled PDs with observed default frequencies

The short-run default frequency (DF) associated with a particular risk bucket at date *t*+1 is the percentage of obligors assigned to that bucket at date *t* that default at *t*+1. Note that under this definition DF_{t+1} cannot be observed at date *t*. Hence, while default probabilities represent *ex ante* forecasts of the likelihood of default, default frequencies represent *ex post* measures of observed default rates.

When systematic risk is present observed default frequencies over a single time period are unlikely to match pooled default probabilities, even when the number of obligors assigned to a risk bucket is very large and the pooled default probability is accurately measured. This is because, in any given time period, systematic shocks will either drive up or drive down the observed default frequency relative to the forecast. This fact can be demonstrated analytically using our stylised default and rating model. In this section we assume that the number of obligors assigned to a risk bucket is always large, so that the observed default frequency for date *t*+1 is closely approximated by the expected default rate given all systematic risk factors realised by date *t*+1.

For a point-in-time risk bucket, the default frequency for date t+1 is

$$\begin{aligned} DF_{t+1}^{PIT} &= E\left[D_{i,t+1} \mid i \in \Gamma_t^{PIT}, Y_t = y_t, V_{t+1} = v_{t+1}\right] \\ &= E\left[D_{i,t+1} \mid \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y y_t = -\gamma_{PIT}, Y_t = y_t, V_{t+1} = v_{t+1}\right] \\ &= \Phi\left(\frac{\gamma_{PIT} - \omega v_{t+1}}{\sqrt{1-\omega^2}}\right). \end{aligned}$$

Comparing this default frequency with the pooled PDs for date t defined in the previous section shows that, in general, the ex post observed default frequency for a point-in-time risk bucket will not match that bucket's pooled PD. Only in the special case where no systematic risk is present (i.e. $\omega = 0$) can a PIT bucket's one-year default frequency be expected to match its unstressed pooled PD.

The default frequency for a through-the-cycle risk bucket at date t is

$$\begin{aligned} DF_{t+1}^{TTC} &= E\left[D_{i,t+1} \mid i \in \Gamma_t^{TTC}, Y_t = y_t, V_{t+1} = v_{t+1}\right] \\ &= E\left[D_{i,t+1} \mid \alpha + \beta_W W_i + \beta_X X_{it} = -\sqrt{1-\omega^2}\gamma_{TTC}, Y_t = y_t, V_{t+1} = v_{t+1}\right] \\ &= \Phi\left(\gamma_{PIT} - \frac{\beta_Y y_t + \psi + \omega v_{t+1}}{\sqrt{1-\omega^2}}\right). \end{aligned}$$

This default frequency does not match any of the pooled PDs for a through-the-cycle rating system.

The long-run default frequency is the average of observed default frequencies taken over many periods. In the context of our stylised model, the long-run default frequency for a risk bucket is simply the expected value of the observed default frequency, taken without respect to time. Thus, for a point-in-time risk bucket the long-run default frequency is

$$\begin{aligned} LRDF^{PIT} &= E\left[DF_{t+1}^{PIT}\right] \\ &= E\left[E\left[D_{i,t+1} \mid i \in \Gamma_t^{PIT}, Y_t, V_{t+1}\right] \mid i \in \Gamma_t^{PIT}\right] \\ &= E\left[D_{i,t+1} \mid \alpha + \beta_W W_i + \beta_X X_{it} + \beta_Y Y_t = -\gamma_{PIT}\right] \\ &= \Phi\left(\gamma_{PIT}\right) \end{aligned}$$

which is equal to the unstressed pooled PD for the risk bucket. The long-run default frequency for a through-the-cycle risk bucket is

$$\begin{aligned} LRDF^{TTC} &= E\left[DF_{t+1}^{TTC}\right] \\ &= E\left[E\left[D_{i,t+1} \mid i \in \Gamma_t^{TTC}, Y_t, V_{t+1}\right] \mid i \in \Gamma_t^{TTC}\right] \\ &= E\left[D_{i,t+1} \mid \alpha + \beta_W W_i + \beta_X X_{it} + \psi = -\sqrt{1-\omega^2}\gamma_{TTC}\right] \\ &= \Phi\left(\frac{\sqrt{1-\omega^2}\gamma_{TTC} + \psi}{\sqrt{1+\beta_Y^2}}\right) \end{aligned}$$

which does not correspond to any of the pooled PDs derived for a through-the-cycle rating system. However, comparing $LRGD^{TTC}$ with PPD^{U}_{TTC} reveals that the long-run default frequency will converge to the long-run average value of PPD^{U}_{TTC}.

Conclusions

Using a stylised model of obligor default and bank rating philosophies, this appendix has shown that the dynamic properties of pooled PDs depend in a predictable way on the rating philosophy applied by a bank. This analysis has three main implications for PD quantification and validation.

First, as a general proposition, supervisors should not expect the pooled PD assigned to a risk bucket to remain fixed throughout the business cycle. The dynamic properties of these parameters depend on the type of pooled PDs that a bank is required to report, and the rating philosophy applied by the bank. Under a point-in-time rating philosophy unstressed pooled PDs should remain stable over the business cycle, but stressed pooled PDs will tend to move with the business cycle. Under a through-the-cycle rating system, stressed pooled PDs will remain stable over the business cycle, but unstressed PDs will be negatively correlated with the business cycle.

Second, those pooled PDs that are not stable over the business cycle will be difficult to quantify using historical default frequency data, even if these data are available for a relatively long time period. Single-year observed default frequencies should not be expected to closely match pooled PDs. For a point-in-time rating system a bucket's long-run average default frequency should converge to its unstressed pooled PD over time.

Third, if pooled PDs are assigned to risk buckets by mapping a bank's internal risk grades to external ratings data, close attention must be paid to the ratings philosophy applied in both the internal and external rating systems. If the two rating systems do not conform to the same ratings philosophy, then mappings should not be expected to remain stable over time.

III. Rating and probability of default validation

Dirk Tasche

The statistical analysis of rating systems and score functions is based on the assumption that there are two categories of obligors of a bank:

- Obligors that will be in default at some predefined time horizon; and
- Obligors that will not be in default at this time horizon.

Usually, it is not known in advance whether an obligor belongs to the first or to the second category. Banks therefore face a **dichotomous (or binary) classification problem** as they have to assess an obligor's future status by recourse on her or his at present available characteristics only. Rating systems and score functions may be seen as classification tools in the sense of providing indications of the obligor's likely future status. The procedure of applying a classification tool to an obligor for an assessment of her or his future status is commonly called **discrimination.**

The main construction principle of rating systems can be described as "the better a grade, the smaller the proportion of defaulters and the greater the proportion of non-defaulters that are assigned this grade." Consequently, a rating system will discriminate the better, the more the defaulters' distribution on the grades and the non-defaulters' distribution on the grades differ.

The discriminatory power of a rating system thus denotes its ability to discriminate ex ante between defaulting and non-defaulting borrowers. The discriminatory power can be assessed using a number of statistical measures of discrimination, some of which are described in this section (technical details are given in an appendix). However, the absolute measure of the discriminatory power of a rating system is only of limited meaningfulness. A direct comparison of different rating systems, for example, can only be performed if statistical "noise" is taken into account. In general, the noise will be a greater issue, the smaller the size of the available default sample. For this reason, statistical tools for the comparison of rating systems are also presented. Some of the tools, in particular the Accuracy Ratio and the Receiver Operating Characteristic, explicitly take into account the size of the default sample. Moreover, the discriminatory power should be tested not only in the development dataset but also in an independent dataset (out-of-sample validation). Otherwise there is a danger that the discriminatory power may be overstated by over-fitting to the development dataset. In this case the rating system will frequently exhibit a relatively low discriminatory power on datasets that are independent of, but structurally similar to, the development dataset. Hence the rating system would have a low stability. A characteristic feature of a stable rating system is that it adequately models the causal relation between risk factors and creditworthiness. It avoids spurious dependencies derived from empirical correlations. In contrast to stable systems, unstable systems frequently show a sharply declining level of forecasting accuracy over time.

In practice, rating systems and score functions are not primarily used for yes / no-type decisions in credit granting. Rather, they form the basis for pricing credits and calculating risk premiums and capital charges. For these purposes, each rating grade or score value must be associated with a PD that gives a quantitative assessment of the likelihood with which obligors graded this way will default. Additionally, under both IRB approaches, a bank's capital requirements are determined by internal estimates of the risk parameters for each exposure. These are derived in turn from the bank's internal rating scores. The set of parameters includes the borrower's Probability of Default, as well as in some cases the expected Loss Given Default and Exposure At Default. In this connection one also speaks of the **calibration of the rating system**. As the risk parameters can be determined by the bank

itself, the quality of the calibration is an important prudential criterion for assessing rating systems.

Checking discriminatory power and checking calibration are different tasks. As the ability of discrimination depends on the difference of the defaulters' and non-defaulters' respectively distributions on the rating grades, some measures of discriminatory power summarise the differences of the probability densities of these distributions. Alternatively, the variation of the default probabilities that are assigned to the grades can be measured. In contrast, correct calibration of a rating system means that the PD estimates are accurate. Hence, for examining calibration somehow the differences of forecast PDs and realised default rates must be considered. This can be done simultaneously for all rating grades in a joint test or separately for each rating grade, depending on whether an overall assessment or an in detail examination is intended.

At first glance, the calibration of rating systems appears to be a similar problem to the back-testing of internal models for market risk. For market risk, add-ons to the capital requirements can be immediately derived from the results of the back-testing procedure. Past years' experience gives evidence of the adequacy of this methodology. However, the availability of historical data for market risk is much better than for credit risk. As market risk has to be measured on a daily basis (leading to samples of size 250 in a year), the (typically yearly) time intervals for credit risk are much coarser due to the rareness of credit events. For credit portfolios, ten data points of yearly default rates are regarded as a long time series and the current Basel II proposals consider five year series as sufficient. As a result, the reliability of estimates for credit risk parameters is not at all comparable to the reliability of back-tests of internal market risk models.

Moreover, whereas in the case of market risk there is no strong evidence that the assumption of independent (over time) observations would be violated, the analogous assumption seems to be questionable for credit losses. This holds even more for cross-sectional credit events (i.e. in the same year). As a consequence, standard independence-based tests of discriminatory power and calibration are likely to be biased when applied to credit portfolios. In particular, with respect to calibration this is a major issue to which the validation group paid considerable attention, as documented below.

The current section analyses statistical validation techniques for both the discriminatory power and the calibration (or PD-quantification) of a rating system, and assesses their usefulness for supervisory purposes. Since TTC rating systems are based on much longer time horizons than PIT rating systems, the validation methodologies set out in this section will, in practice, be more applicable to PIT rather than to TTC rating systems. An important conclusion from the group's findings is that any application of a statistical technique has to be supplemented by qualitative checks. This finding is important when considering the following description of methodologies since uncritical use of the techniques may reach misleading results. Moreover, the choice of a specific technique to be applied for validation should depend upon the nature of the portfolio under consideration. Retail portfolios or portfolios of small- and medium-sized enterprises with large records of default data are much easier to explore with statistical methods than, for example, portfolios of sovereigns or financial institutions where default data are sparse.

Discriminatory power

There are various statistical methodologies for the **assessment of discriminatory power**.[10] The following are methodologies that have been suggested in the literature or are popular in the financial industry:[11]

- Cumulative Accuracy Profile (CAP) and its summary index, the Accuracy Ratio (AR),
- Receiver Operating Characteristic (ROC) and its summary indices, the ROC measure and the Pietra coefficient,
- Bayesian error rate,
- Conditional entropy, Kullback-Leibler distance, and Conditional Information Entropy Ratio (CIER),
- Information value (divergence, stability index),
- Kendall's τ and Somers' D (for shadow ratings), and
- Brier score.

The **Cumulative Accuracy Profile** is also known as the Gini curve, Power curve or Lorenz curve. It is a visual tool whose graph can easily be drawn if two representative samples of scores for defaulted and non-defaulted borrowers are available. Concavity of the CAP is equivalent to the property that the conditional probabilities of default given the underlying scores form a decreasing function of the scores. Moreover, non-concavity indicates sub-optimal use of information in the specification of the score function. The most common summary index of the CAP is the **Accuracy Ratio** (or Gini coefficient). It is equivalent to the ROC measure so that its statistical properties can be discussed together with those of the ROC measure below. The shape of the CAP depends on the proportion of solvent and insolvent borrowers in the sample. Hence a visual comparison of CAPs across different portfolios may be misleading. Practical experience shows that the Accuracy Ratio has tendency to take values in the range of 50% and 80%. However, such observations should be interpreted with care as they seem to strongly depend on the composition of the portfolio and the numbers of defaulters in the samples.

Like the CAP, the **Receiver Operating Characteristic** (ROC) is a visual tool that can be easily constructed if two representative samples of scores for defaulted and non-defaulted borrowers are available. The construction is slightly more complex than for CAPs but, in contrast, does not require the sample composition to reflect the true proportion of defaulters and non-defaulters. As with the CAP, concavity of the ROC is equivalent to the conditional probabilities of default being a decreasing function of the underlying scores or ratings and non-concavity indicates sub-optimal use of information in the specification of the score function. One of the summary indices of ROC, the **ROC measure** (or Area Under the Curve, AUC), is a linear transformation of the Accuracy Ratio mentioned above. The statistical properties of the ROC measure are well-known as it coincides with the **Mann-Whitney** statistic. In particular, powerful tests are available for comparing the ROC measure of a rating system with that of a random rating and for comparing two or more rating systems. Also, confidence intervals for the ROC measure can be estimated with readily available

[10] Recent research results indicate that due to the concavity of the risk weight curves low discriminatory power of a rating system tends to increase capital requirements under Basel II. Hence, there is to some degree already an incentive for banks to apply rating systems with high power. This holds also under a UL-based framework.

[11] Technical details for the following methodologies are given in Appendix B if no explicit reference is given to the literature.

statistical software packages. By inspection of the formulas for the intervals, it turns out that the widths of the confidence intervals are mainly driven by the number of defaulters in the sample. The more defaulters are recorded in the sample, the narrower the interval.

The **Pietra index** is another important summary index of ROCs. Whereas the ROC measure measures the area under the ROC, the Pietra index reflects half the maximal distance of the ROC and the diagonal in the unit square (which is just the ROC of rating systems without any discriminatory power). As is the case with the ROC measure, the Pietra index also has an interpretation in terms of a well-known test statistic, the **Kolmogorov-Smirnov** statistic. As with the ROC measure, a test for checking the dissimilarity of a rating and the random rating is included in almost all standard statistical software packages.

Both the ROC measure and the Pietra index do not depend on the total portfolio probability of default. Therefore, they may be estimated on samples with non-representative default/non-default proportions. Similarly, figures for bank portfolios with different fractions of defaulters may be directly compared.

For both these indices, it is not possible to define in a meaningful way a general minimum value in order to decide if a rating system has enough discriminatory power. However, both indices are still useful indicators for the quality of a rating system.

Significance of rejection of the null hypothesis (rating system has no more power than the random rating) with the Mann-Whitney or Kolmogorov-Smirnov tests at a (say) 5% level could serve as a minimum requirement for rating systems. This would take care of statistical aspects like sample size. Lower p-values with these tests are indicators of superior discriminatory power. However, for most rating systems used in the banking industry, p-values will be nearly indistinguishable from zero. As a consequence, the applicability of the p-value as an indicator of rating quality appears to be limited.

The **Bayesian error rate** (or classification error or minimum error) specifies the minimum probability of error if the rating system or score function under consideration is used for a yes/no decision whether a borrower will default or not. The error can be estimated parametrically, e.g. assuming normal score distributions, or non-parametrically, for instance with kernel density estimation methods. If parametric estimation is applied, the distributional assumptions have to be carefully checked. Non-parametric estimation will be critical if sample sizes are small. In its general form, the error rate depends on the total portfolio probability of default. As a consequence, in many cases its magnitude is influenced much more by the probability of erroneously identifying a non-defaulter as a defaulter than by the probability of not detecting a defaulter. In practice, therefore, the error rate is often applied with a fictitious 50% probability of default. In this case, the error rate is equivalent to the Kolmogorov-Smirnov statistic and to the Pietra index.

Entropy is a concept from information theory that is related to the extent of uncertainty that is eliminated by an experiment. The observation of an obligor over time in order to decide about her or his solvency status may be interpreted as such an experiment. The uncertainty of the solvency status is highest if the applied rating system has not at all any discriminatory power or, equivalently, all the rating grades have the same PD. In this situation, the entropy concept applied to the PDs of the rating system would yield high figures since the gain in information by finally observing the obligor's status would be large. Minimisation of entropy measures like **Conditional Entropy, Kullback-Leibler distance, CIER, and information value** (see the appendix for details) is therefore a widespread criterion for constructing rating systems or score functions with high discriminatory power. However, these measures appear to be of limited use only for validation purposes as no generally applicable statistical tests for comparisons are available.

The **Brier score** is a sample estimator of the mean squared difference of the default indicator variables (i.e. one in case of default and zero in case of survival) in a portfolio and the default probability forecasts for rating categories or score values. In particular, the Brier score does not directly measure the difference of the default probability forecast and the true conditional probability of default given the scores (which is only a theoretical concept and shall be estimated by the probability forecast). Therefore, the Brier score is not a measure of calibration accuracy. Rather, the Brier score should be interpreted as the residual sum of squares that results from a non-linear regression of the default indicators on the rating or score function. As a consequence, minimising the Brier score is equivalent to maximising the variance of the default probability forecasts (weighted with the frequencies of the rating categories). Empirical results indicate that maximising the ROC measure entails maximisation of this variance. In this sense, the Brier score is a measure of discriminatory power and could be used in this sense as a part of an optimisation criterion. At present, it is not clear which testable statistical hypotheses regarding the Brier score should be considered.

The Group has found that the **Accuracy Ratio** (AR) and the **ROC measure** appear to be more meaningful than the other above-mentioned indices because of their statistical properties. For both summary statistics, it is possible to calculate confidence intervals in a simple way. The width of the confidence will depend on the particular portfolio under consideration and on the number of defaulted obligors that is available for the purpose of estimation. As a general rule, the width of the confidence interval for AR (or the ROC measure) will be the larger, and hence the quality of the estimate will be the worse, the smaller is the number of observed defaults. Consequently, these tools reflect both the quality of a rating system and the size of the samples that the rating system is built on. Therefore, they are helpful in identifying rating systems which require closer inspection by a supervisor. In particular, supervisors can reliably test if a rating model is significantly different from a model with no discriminatory power. The Brier score can be useful in the process of developing a rating system as it also indicates which of any two rating systems has the higher discriminatory power. However, due to the lack of statistical test procedures applicable to the Brier score, the usefulness of this metric for validation purposes is limited.

If not enough default observations for the development of a rating or score system are available, the construction of a **shadow rating** system could be considered. A shadow rating is intended to duplicate an external rating but can be applied to obligors for which the external rating is not available. Shadow ratings can be built when the available database contains accounting information of enough externally rated obligors. Default probabilities for the shadow rating will then be derived from statistics for the external rating. On samples of borrowers for which both the shadow and the external rating are available, the degree of concordance of the two rating systems can be measured with two rank-order statistics, **Kendall's τ** and **Somers' D**. Somers' D is a conditional version of Kendall's τ that coincides with the Accuracy Ratio in the case of a rating system with only two categories. For both these metrics, tests can be performed and confidence intervals can be calculated with some standard statistical software packages. In the case of high concordance of the shadow rating and the external rating, the shadow rating will inherit the discriminatory power and the calibration quality of the external rating if the portfolio under consideration and the rating agency's portfolio have a similar structure.

Calibration

Validation of the **calibration of a rating system** is more difficult than validation of its discriminatory power. The Group analysed statistical tools to decide about the magnitude of the difference between estimated PD and observed default rate that is still acceptable to supervisors.

When considering the statistical tools it is important to note that there are several established statistical methods for deriving PDs (Probabilities of Default) from a rating system. First, a distinction needs to be drawn between direct and indirect methods. In the case of the direct methods, such as Logit, Probit and Hazard Rate models, the rating score itself can be taken as the borrower's PD. The PD of a given rating grade is then normally calculated as the mean of the PDs of the individual borrowers assigned to each grade. Where the rating score cannot be taken as the PD (as in the case of discriminant analysis), indirect methods can be used. One simple method consists of estimating the PD for each rating grade from historical default rates. Another method is the estimation of the score distributions of defaulting borrowers, on the one hand, and non-defaulting borrowers, on the other.[12] A specific PD can subsequently be assigned to each borrower using Bayes' Formula.

In practice, a bank's PD estimates will differ from the default rates that are afterwards observed. The key question is whether the deviations are purely random or whether they occur systematically. A systematic underestimation of PDs merits a critical assessment –from the point of view of supervisors and bankers alike – since in this case the bank's computed capital requirement would not be adequate to the risk it has incurred.

This observation again underlines the need for an adequate judgement about the appropriateness of the PD estimates by the banks. In particular, the quality of the PD estimates of the following methodologies have been considered (see the appendix for more details):

- Binomial test,
- Chi-square test,
- Normal test,
- Traffic lights approach.

When independence of default events is assumed in a homogeneous portfolio, the binomial test (most powerful among all tests at fixed level) can be applied in order to test the correctness of a one period default probability forecast. Examples in the appendix show that the true type I error (i.e. the probability to reject erroneously the hypothesis of an adequate PD forecast) can be much larger than the nominal level of the test if default events are correlated. Efforts to take into account dependence in the binomial test, for example, by incorporating a one factor dependence structure and Gordy's granularity adjustment[13] in order to adjust for the finiteness of the sample, yield tests of rather moderate power, even for low levels of correlation.

The **binomial test** can be applied to one rating category at a time only. If (say) twenty categories are tested, at 5% significance level one erroneous rejection of the null hypothesis "correct forecast" has to be expected. This problem can be circumvented by applying the **chi-square** (or Hosmer-Lemeshow) test to check several rating categories simultaneously. This test is based on the assumption of independence and a normal approximation. Due to the dependence of default events that is observed in practice and the generally low frequency of default events, the chi-square test is also likely to underestimate the true Type I error.

The **normal test** is an approach to deal with the dependence problem that occurs in the case of the binomial and chi-square tests. The normal test is a multi-period test of correctness of a

---

[12] Calculating the PDs as slope of the CAP curve can be regarded as a variant of this approach.

[13] See Tasche (2003).

default probability forecast for a single rating category. It is applied under the assumption that the mean default rate does not vary too much over time and that default events in different years are independent. The normal test is motivated by the Central Limit Theorem and is based on a normal approximation of the distribution of the time-averaged default rates. Cross-sectional dependence is admissible. Simulation studies show that the quality of the normal approximation is moderate but exhibits a conservative bias. As a consequence, the true Type I error tends to be lower than the nominal level of the test, i.e. the proportion of erroneous rejections of PD forecasts will be smaller than might be expected from the formal confidence level of the test. The test seems even to be, to a certain degree, robust against a violation of the assumption that defaults are independent over time. However, the power of the test is moderate, in particular for short time series (for example five years).

For the supervisory evaluation of internal market risk models, the so-called traffic lights approach has proved to be a valuable instrument. This approach was introduced with the 1996 Marker Risk Amendment. Banks use their internal market risk models in order to forecast a certain amount of losses (Value-at-Risk) that will not be exceeded by the realised losses with a high probability of 99%. Depending on the number of observed exceedances, the so-called multiplication factor that is applied to the Value-at-Risk estimate is increased. There is a *green zone* of exceedances where no increment to the multiplication factor is necessary. In the *yellow zone,* the increment is effectively proportional to the number of exceedances, whereas in the *red zone* the maximum value for the increment has to be applied.

The concept of a traffic lights approach can be transferred to the validation of PD estimates. However, it is unlikely that direct consequences for the capital requirements of a bank can be derived from this approach. A recently proposed version of a **traffic lights approach** is – in contrast to the normal test – completely independent of any assumption of constant or nearly constant PDs over time.[14] It can be considered as a multi-period back-testing tool for a single rating category that is based on the assumption of cross-sectional and inter-temporal independence of default events. The distribution of the number of defaults in one year is approximated with a normal distribution. Based on the quantiles of this normal distribution, the number of defaults is mapped to one of the four traffic light colours: green, yellow, orange, and red. This mapping results in a multinomial distribution of the numbers of colours when observed over time. Inference on the adequacy of default probability forecasts this way becomes feasible. By construction of the tool with a normal approximation that neglects potential cross-sectional and inter-temporal correlations, higher than expected frequencies of type I errors (i.e. erroneous rejections of default probability forecasts) may occur. As a consequence, this traffic lights approach is conservative in the sense of yielding relatively more false alerts than not detecting bad calibrations. Simulation results indicate that the traffic lights approach is not too conservative since the frequency of false alerts can be kept under control. Furthermore the simulation study suggests that the type II errors (i.e. the probabilities of accepting biased estimates as correct) are not higher than those of the normal test.

In conclusion, at present no really powerful tests of adequate calibration are currently available. Due to the correlation effects that have to be respected there even seems to be no way to develop such tests. Existing tests are rather conservative – such as the binomial test and the chi-square test – or will only detect the most obvious cases of miscalibration as in the case of the normal test. As long as validation of default probabilities per rating category is required, the traffic lights testing procedure appears to be a promising tool because it can be

[14] See Blochwitz, Hohl and Wehn (2003).

applied in nearly every situation that might occur in practice. Nevertheless, it should be emphasised that there is no methodology to fit all situations that might occur in the validation process. Depending on the specific circumstances, the composition of a mixture of different techniques will be the most appropriate way to tackle the validation exercise.

Open issues

As demonstrated above, the Accuracy Ratio and the Area Under the Curve are powerful tools for the validation of discriminatory power. In particular, tested and proven methods that in addition take care of sufficient sample size can be used for statistical inference on the discriminatory power of a rating system.

Since the influence of correlation between default events should not be neglected, the situation is less satisfactory when the calibration of a rating system has to be validated than in the case of discriminatory power. In particular, further research seems warranted concerning the following issues:

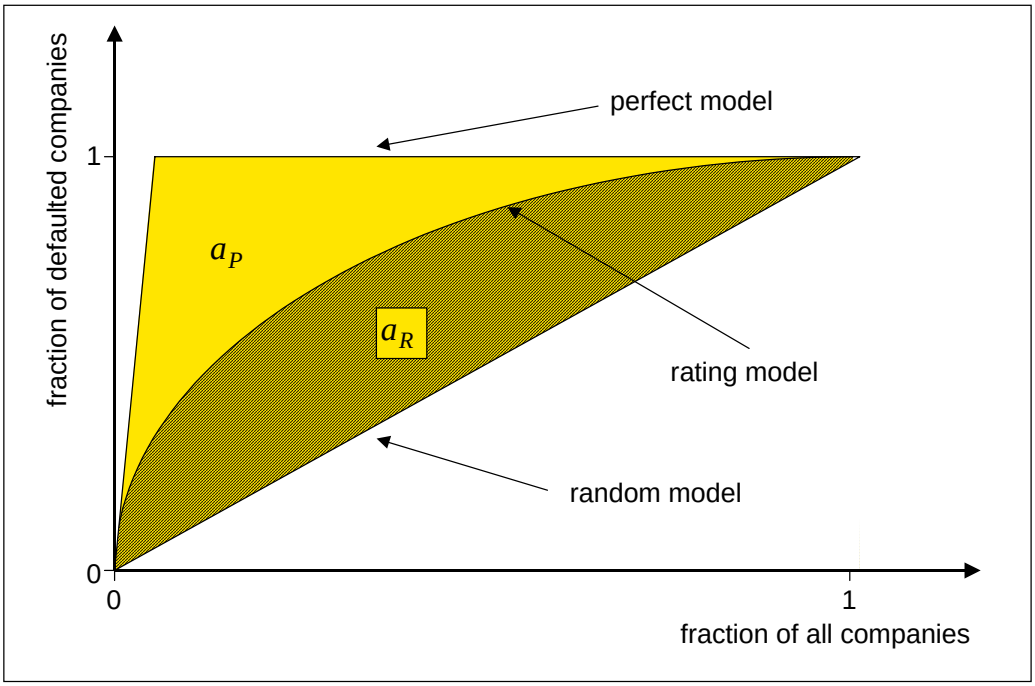
- More evidence should be provided on the appropriateness of the assumptions (cross-sectional independence and inter-temporal independence of default events, normal approximation) that are the basis of the traffic lights approach by Blochwitz et al.
- Additionally, the construction of calibration tests that incorporate dependence modelling appears desirable.
- The independence-based simultaneous Hosmer-Lemeshow test of PDs in several rating grades should be supplemented with further simultaneous tests.
- Are there useful alternative paths (e.g. different assumptions on the underlying model) to the construction of traffic lights approaches to PD calibration?
- Can any recommendations be given on the minimum size of data samples, in particular concerning a lower limit of default observations, that are to be used for discrimination or calibration purposes? If so, what would be a suitable methodology for determining these recommendations?

Appendix: Statistical measures of discriminatory power and calibration quality

Cumulative Accuracy Profiles and Accuracy Ratio

Consider an arbitrary rating model that produces a rating score. The score under consideration could be a rating score like Altman's Z-score or a score obtained from a Logit-model or from any other approach. A high rating score is usually an indicator of a low default probability. To obtain the CAP curve, all debtors are first ordered by their respective scores, from riskiest to safest, i.e. from the debtor with the lowest score to the debtor with the highest score. For a given fraction *x* of the total number of debtors the CAP curve is constructed by calculating the percentage *d(x)* of the defaulters whose rating scores are equal to or lower than the maximum score of fraction *x*. This is done for *x* ranging from 0% to 100%. Figure 5 illustrates CAP curves.

Figure 5. Cumulative Accuracy Profiles.



A perfect rating model will assign the lowest scores to the defaulters. In this case the CAP is increasing linearly and than staying at one. For a random model without any discriminative power, the fraction *x* of all debtors with the lowest rating scores will contain *x* percent of all defaulters. Real rating systems will be somewhere in between these two extremes. The quality of a rating system is measured by the Accuracy Ratio *AR*. It is defined as the ratio of the area a_R between the CAP of the rating model being validated and the CAP of the random model, and the area a_P between the CAP of the perfect rating model and the CAP of the random model, i.e.

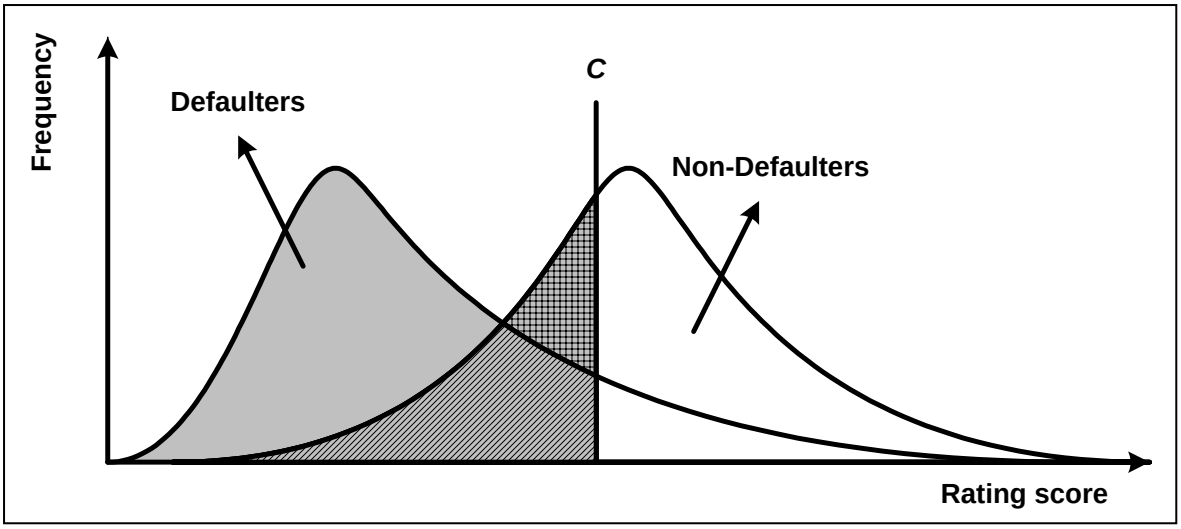
$$AR = \frac{a_R}{a_P}.$$

Thus, the rating method is the better the closer *AR* is to one.

Receiver Operating Characteristic and the area under the Receiver Operating Characteristic curve

The construction of a ROC curve is illustrated in Figure 2 which shows possible distributions of rating scores for defaulting and non-defaulting debtors. For a perfect rating model the left distribution and the right distribution in Figure 2 would be separate. For real rating systems, perfect discrimination in general is not possible. Both distributions will overlap as illustrated in Figure 6.

Figure 6. Distribution of rating scores for defaulting and non-defaulting debtors.



Assume someone has to find out from the rating scores which debtors will survive during the next period and which debtors will default. One possibility for the decision-maker would be to introduce a cut-off value *C* as in Figure 6, and to classify each debtor with a rating score lower than *C* as a potential defaulter and each debtor with a rating score higher than *C* as a non-defaulter. Then four decision results would be possible. If the rating score is below the cut-off value *C* and the debtor defaults subsequently, the decision was correct. Otherwise the decision-maker wrongly classified a non-defaulter as a defaulter. If the rating score is above the cut-off value and the debtor does not default, the classification was correct. Otherwise a defaulter was incorrectly assigned to the non-defaulters' group.

We define the hit rate *HR(C)* as

$$HR(C) = \frac{H(C)}{N_D},$$

where *H(C)* is the number of defaulters predicted correctly with the cut-off value *C,* and N_D is the total number of defaulters in the sample. This means that the hit rate is the fraction of defaulters that was classified correctly for a given cut-off value *C*. The false alarm rate *FAR(C)* is defined as

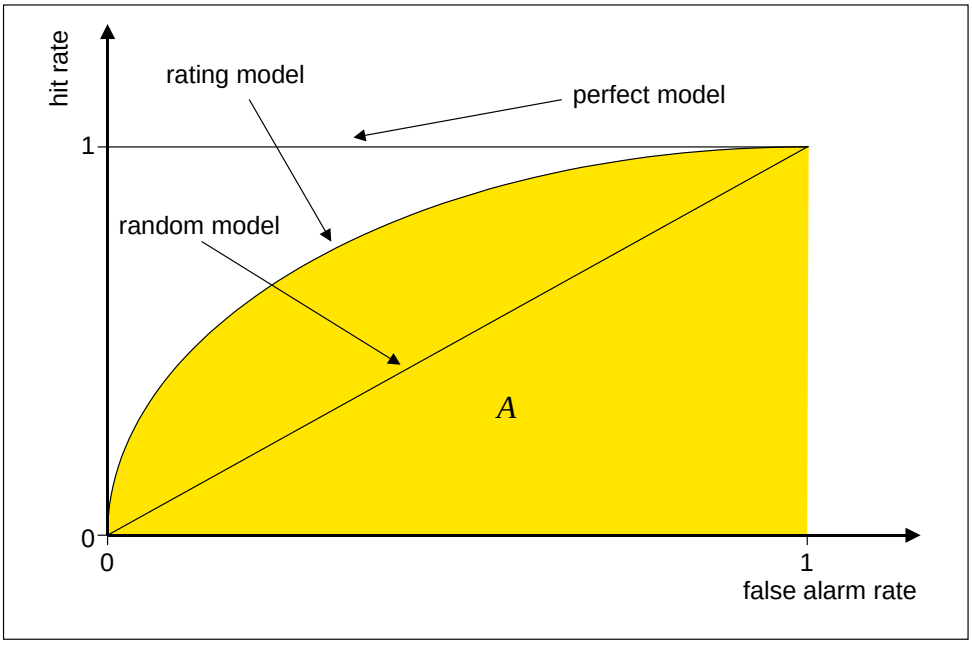
$$FAR(C) = \frac{F(C)}{N_{ND}},$$

where *F(C)* is the number of false alarms, i.e. the number of non-defaulters that were classified incorrectly as defaulters by using the cut-off value *C*. The total number of non-

defaulters in the sample is denoted by N_{ND}. In Figure 6, *HR(C)* is the area to the left of the cut-off value *C* under the score distribution of the defaulters (coloured plus hatched area), while *FAR(C)* is the area to the left of *C* under the score distribution of the non-defaulters (chequered plus hatched area).

The ROC curve is constructed as follows. For all cut-off values *C* that are contained in the range of the rating scores the quantities *HR(C)* and *FAR(C)* are computed. The ROC curve is a plot of *HR(C)* versus *FAR(C)*. This is illustrated in Figure 7.

Figure 7. Receiver Operating Characteristic Curves.



Theoretically, we could get a continuum of pairs *HR(C)* and *FAR(C)*. In practice, of course, we only have a finite sample of points on the ROC curve. The entire ROC curve is found by linearly interpolating this set of points. The points *(0,0)* and *(1,1)* are contained in every ROC curve, because for *C<min(rating score)* we have *H(C)=F(C)=0* and for *C>max(rating score)* we obtain $H(C) = N_D$ and $F(C) = N_{ND}$.

A rating model's performance is the better the steeper the ROC curve is at the left end and the closer the ROC curve's position is to the point (0,1). Similarly, the model is the better the larger the area under the ROC curve is. We denote this area by *A*. By means of a change of variable, it can be calculated as

$$A = \int_0^1 HR(FAR)d(FAR) .$$

The area *A* is 0.5 for a random model without discriminative power and it is 1.0 for a perfect model. It is between 0.5 and 1.0 for any reasonable rating model in practice.

Connection between Accuracy Ratio and Area under the Curve

It turns out that the area A under the ROC curve and the CAP accuracy ratio AR are connected by means of the linear transformation[15]

$$AR = 2A - 1.$$

As a consequence, neither A nor AR depend on the proportion of defaulters in the sample that is used for their estimation.

Confidence Intervals for the AR and the area under the ROC

We discuss a simple method of calculating confidence intervals for *A,* the area under the ROC curve. The same reasoning applies to the Accuracy Ratio by means of the relation proven above. We start with a probabilistic interpretation of *A*.

Consider the following experiment. Two debtors are drawn at random, the first one from the distribution of defaulters, the second one from the distribution of non-defaulters. The scores of the defaulter and the non-defaulter determined in this way can be interpreted as realisations of the two independent continuous random variables S_D and S_{ND}. Assume someone has to decide which of the debtors is the defaulter. A rational decision-maker might suppose that the defaulter is the debtor with the lower rating score. The probability that he is right is equal to $P(S_D < S_{ND})$. A simple calculation shows that this probability is exactly equal to the area under the ROC curve *A*.

As we already noted, we can interpret *HR(C)* as the probability that S_D is below the cut-off value *C,* $P(S_D < C)$, and *FAR(C)* as the probability that S_{ND} is below *C,* $P(S_{ND} < C)$. Applying this to the definition of *A* we obtain

$$A = \int_0^1 HR(FAR)dFAR = \int_0^1 P(S_D < C)dP(S_{ND} < C) = \int_{-\infty}^{+\infty} P(S_D < C) f_{S_{ND}}(C)dC = P(S_D < S_{ND}),$$

where the probability density function of S_{ND} is denoted by $f_{S_{ND}}$. Hence, the area under the ROC curve is equal to the probability that S_D produces a smaller rating score than S_{ND}.

This interpretation relates to the U-test of Mann-Whitney. If we draw a defaulter with score s_D from S_D and a non-defaulter with score s_{ND} from S_{ND} and define $u_{D,ND}$ as

$$u_{D,ND} = \begin{cases} 1, if & s_D < s_{ND} \\ 0, if & s_D \geq s_{ND} \end{cases},$$

then the test statistic $\hat{U}$ of Mann-Whitney is defined as

$$\hat{U} = \frac{1}{N_D \cdot N_{ND}} \sum_{(D,ND)} u_{D,ND},$$

[15] See, e.g., Engelmann, Hayden and Tasche (2003).

where the sum is over all pairs of defaulters and non-defaulters *(D,ND)* in the sample.

Observe that $\hat{U}$ is an unbiased estimator for $P\left(S_D < S_{ND}\right)$, i.e.

$$A = E\left(\hat{U}\right) = P\left(S_D < S_{ND}\right).$$

Furthermore, we find that the area $\hat{A}$ below the ROC curve calculated from the empirical data is equal to $\hat{U}$. For the variance $\sigma^2_{\hat{U}}$ of $\hat{U}$ we find the unbiased estimator $\hat{\sigma}^2_{\hat{U}}$ as

$$\hat{\sigma}^2_{\hat{U}} = \frac{1}{4(N_D - 1)(N_{ND} - 1)}\left[1 + (N_D - 1)\hat{P}_{D,D,ND} + (N_{ND} - 1)\hat{P}_{ND,ND,D} - 4\left(N_D + N_{ND} - 1\right)(\hat{U} - \frac{1}{2})^2\right],$$

where $\hat{P}_{D,D,ND}$ and $\hat{P}_{ND,ND,D}$ are estimators for the expressions $P_{D,D,ND}$ and $P_{ND,ND,D}$ which are defined as

$$P_{D,D,ND} = P\left(S_{D,1}, S_{D,2} < S_{ND}\right) + P\left(S_{ND} < S_{D,1}, S_{D,2}\right) - P\left(S_{D,1} < S_{ND} < S_{D,2}\right) - P\left(S_{D,2} < S_{ND} < S_{D,1}\right),$$

$$P_{ND,ND,D} = P\left(S_{ND,1}, S_{ND,2} < S_D\right) + P\left(S_D < S_{ND,1}, S_{ND,2}\right) - P\left(S_{ND,1} < S_D < S_{ND,2}\right) - P\left(S_{ND,2} < S_D < S_{ND,1}\right).$$

The quantities $S_{D,1}$, $S_{D,2}$ are independent observations randomly sampled from S_D and $S_{ND,1}$, $S_{ND,2}$ are independent observations randomly sampled from S_{ND}. This unbiased estimator $\hat{\sigma}^2_{\hat{U}}$ is implemented in standard statistical software packages.

For $N_D, N_{ND} \to \infty$ it is known that $\left(A - \hat{U}\right) / \hat{\sigma}_{\hat{U}}$ is asymptotically normally distributed with mean zero and standard deviation one. This allows the calculation of confidence intervals at confidence level α for $\hat{U}$ using the relation

$$P\left(\hat{U} - \hat{\sigma}_{\hat{U}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right) \le A \le \hat{U} + \hat{\sigma}_{\hat{U}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right) \approx \alpha,$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Empirical analysis we carried out indicates that the number of defaults should be at least around 50 in order to guarantee that the above formula is a reasonable approximation. We note that there is no clear rule for which values of $\hat{U}$ the asymptotic normality of $\hat{U}$ is a valid approximation, because $\hat{U}$ can solely take values in the interval [0,1]. If $\hat{U}$ is only a few (two or three) standard deviations away from one it is clear, that the normal approximation cannot be justified. However, this problem is not likely to occur for real rating systems. A value of $\hat{U}$ close to one means that the score distributions of defaulters and non-defaulters are almost separated, i.e. the rating model is almost perfect. No real rating model is known to have that discriminative power.

Under the assumption that the normal approximation to the distribution of the estimate $\hat{U}$ for the area A under the ROC curve works, one obtains the following representation I_α of the confidence interval at level α for the true value of A:

$$I_\alpha = \left[\hat{U} - \hat{\sigma}_{\hat{U}} \Phi^{-1}\left(\frac{1+\alpha}{2}\right), \hat{U} + \hat{\sigma}_{\hat{U}} \Phi^{-1}\left(\frac{1+\alpha}{2}\right) \right].$$

In other words, for any value a between $\hat{U} - \hat{\sigma}_{\hat{U}} \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$ and $\hat{U} - \hat{\sigma}_{\hat{U}} \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$ the hypothesis that $a = A$ cannot be rejected at level α. Observe that the width of this confidence interval in particular depends on the confidence level α as well as on the total numbers of defaulters N_D and non-defaulters N_{ND} in the sample. The expression, given above for the variance $\hat{\sigma}^2_{\hat{U}}$ of the estimator $\hat{U}$, admits an easy upper bound[16] which can be used for getting a feeling of the influence of the confidence level and the sample size on the width of the confidence interval:

$$\hat{\sigma}^2_{\hat{U}} \leq \frac{A(1-A)}{\min\left(N_D, N_{ND}\right)}.$$

By means of this inequality, the following table with confidence interval widths as a function of the confidence level α and the number of defaults[17] N_D in the sample can be calculated.

[16] See Equation (7) of Bamber (1975).

[17] We assume that the number of defaulters in the sample is dominated by the number of surviving obligors.

Table 2

Upper bound of width of the confidence interval for the area A under the ROC curve in dependence of confidence level α and number of defaults N_D.

The calculations were based on the assumption that the true value of *A* is 0.75. With a sample of 500 defaulted obligors, one can be sure at 95% confidence level that the difference of the upper and the lower confidence bounds will not be greater than 4.4% = 0.0329 / 0.75 of the true value of *A*.

N_D	α			
	90%	**95%**	**99%**	**99.5%**
10	0.1951	0.2324	0.3055	0.3329
25	0.1234	0.1470	0.1932	0.2105
50	0.0872	0.1039	0.1366	0.1489
100	0.0617	0.0735	0.0966	0.1053
250	0.0390	0.0465	0.0611	0.0666
500	0.0276	0.0329	0.0432	0.0471
1000	0.0195	0.0232	0.0305	0.0333
2500	0.0123	0.0147	0.0193	0.0211
5000	0.0087	0.0104	0.0137	0.0149
10000	0.0062	0.0073	0.0097	0.0105

The Pietra Index

Geometrically, the **Pietra Index** can be defined as the maximum area a triangle can obtain that is inscribed between the ROC curve and the diagonal of the unit square (cf. Figure 3). With this definition, the Pietra Index can be easily estimated from sample data[18]. Equivalently, the Pietra Index can be seen as half the maximum distance of ROC curve and diagonal. In case of a concave ROC curve that will in most cases be the result of an optimisation procedure, the interpretation of the Pietra Index as a distance leads to the representation

$$\text{Pietra Index} = \sqrt{2}/4 \max_C |HR(C) - FAR(C)|.$$

The sup-term on the right-hand side of this equation is just the well-known Kolmogorov-Smirnov test statistic of the distribution functions *HR* and *FAR* that have been defined above.

[18] See Lee (1999), p. 455–471.

Bayesian error rate

Denote with p_D the rate of defaulters in the portfolio and define the hit rate *HR* and the false alarm rate *FAR* as above. In case of a concave ROC curve the Bayesian error rate then can be calculated via

$$\text{Error rate} = \min_C(p_D(1-HR(C)) + (1-p_D)FAR(C)).$$

If p_D equals 50% then the Error rate can be expressed as

$$\text{Error rate} = 1/2 - 1/2 \max_c |HR(C) - FAR(C)|.$$

As a consequence, the error rate is then equivalent to the Pietra Index and the Kolmogorov-Smirnov statistic. Of course, at first glance the assumption $p_D = 50\%$ does not appear to be very reasonable. However, for technical reasons it might sometimes be necessary to develop a score function on a sample which is not representative in terms of the proportion of defaulters and non-defaulters. In this situation, the error rate with probability of default would be a good choice since it does not depend on the true proportion of defaulters and non-defaulters.

The error rate can be estimated with parametric methods, for instance under the assumption of normally distributed scores, or non-parametrically with the Kolmogorov-Smirnov statistic in case of $p_D = 50\%$ or with kernel estimation methods. If the estimation is carried out with parametric methods, the underlying assumptions have to be checked with appropriate statistical tests.

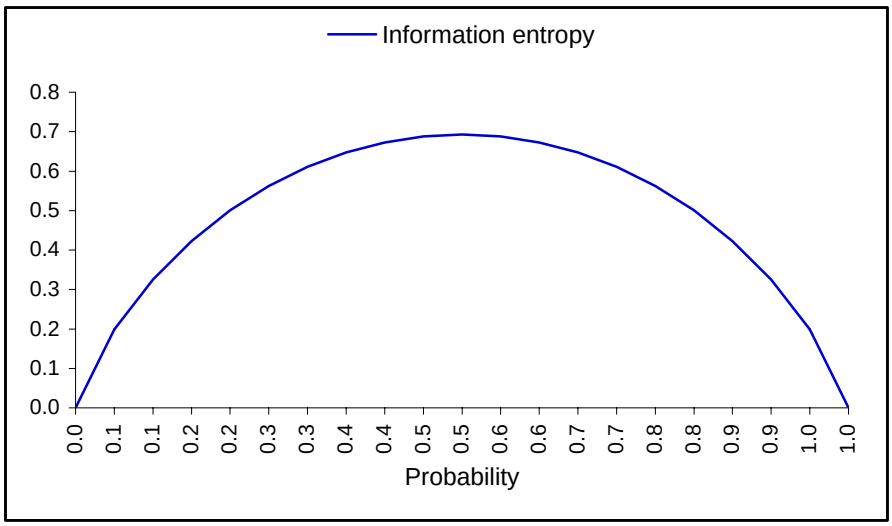
Entropy measures

We define the **Information Entropy** $H(p)$ of an event with probability *p* as

$$H(p) = -\left(p\log(p) + (1-p)\log(1-p)\right).$$

The Information Entropy is visualised as a function of the default probability *p* in Figure 8.

Figure 8. Information Entropy as a function of default probability.



We see that the information entropy takes its maximum at *p=1/2*, the state which reveals the greatest uncertainty. If *p* equals zero or one, either the event under consideration itself or its complementary event will occur with certainty and thus not reveal any information at all.

Consider as rating system that, applied to an obligor, produces a random score *S.* If D denotes the event "obligor defaults" and $\bar{D}$ denotes the complementary event "obligor does not default", we can apply the information entropy H to the $P(D \mid S)$, the conditional probability of default given the rating score *S.* The result of this operation can be considered a conditional information entropy of the default event,

$$H(P(D \mid S)) = -\Big(P(D \mid S)\log P(D \mid S) + P(\bar{D} \mid S)\log P(\bar{D} \mid S)\Big),$$

and as such is a random variable whose expectation can be calculated. This expectation is called **Conditional Entropy** of the default event (with respect to the rating score *S*), and can formally be written as

$$H_S = -E\Big[P(D \mid S)\log P(D \mid S) + P(\bar{D} \mid S)\log P(\bar{D} \mid S)\Big].$$

It can be shown that the Conditional Entropy of the default event is at most as large as the unconditional Information Entropy of the default event, i.e.

$$H_S \quad \leq \quad H(p).$$

Hence the average amount of information that can be revealed if the rating score is known is always smaller than the information in case of no a-priori knowledge. The difference of $H(p)$ and H_S should be as large as possible because in this case the gain of information by application of the rating scores would be a maximum. Formally, this difference is called **Kullback-Leibler Distance***,* i.e.

$$\text{Kullback-Leibler Distance} \;= H(p) - H_S.$$

Obviously, the range of values the Kullback-Leibler Distance can take depends on the unconditional probability of default. In order to arrive at a common scale for any underlying population, the Kullback-Leibler Distance is normalised with the unconditional Information Entropy. This operation leads to the notion of **Conditional Information Entropy Ratio (CIER)** with the representation

$$CIER_S = \frac{H(p) - H_S}{H(p)}.$$

The value of *CIER* will be the closer to one the more information about the default event is contained in the rating scores *S.*

Whereas the CIER looks at the gain in information about the default event that is reached with the rating scores under consideration, another entropy-based measure of discriminatory power, the **Information Value**, is rather a measure of the difference between the defaulter score distribution and the non-defaulter score distribution. In this sense it is similar to the Pietra Index. For the sake of brevity, we consider the case of continuous score distributions with density f_D for the defaulters and density $f_{\bar{D}}$ for the non-defaulters. The Information Value *IV* is defined as the sum of the relative entropy of the non-defaulter distribution with respect

to the defaulter distribution and the relative entropy of the defaulter distribution with respect to the non-defaulter distribution, i.e.

$$IV = E\left[\log \frac{f_D(S)}{f_{\bar{D}}(S)} \mid D\right] + E\left[\log \frac{f_{\bar{D}}(S)}{f_D(S)} \mid \bar{D}\right].$$

Alternatively, the Information Value can be written as

$$IV = \int \left(f_D(s) - f_{\bar{D}}(s)\right) \log \frac{f_D(s)}{f_{\bar{D}}(s)}\, ds.$$

This last representation shows that the Information Value takes only non-negative values. However, there is no theoretical upper bound to its range. As the Information Values measures the difference of two distributions and the difference between the defaulter and the non-defaulter distributions should be as large as possible in order to obtain a rating system with high discriminatory power, high Information Values indicate powerful rating systems.

Kendall's τ and Somers' D

Kendall's τ and Somers' *D* are so-called rank order statistics, and as such measure the degree of comonotonic dependence of two random variables. The notion of comonotonic dependence generalises linear dependence that is expressed via (linear) correlation. In particular, any pair of random variables with correlation 1 (i.e. any linearly dependent pair of random variables) is comonotonically dependent. But in addition, as soon as one of the variables can be expressed as any kind of increasing transformation of the other, the two variables are comonotonic. In the actuarial literature, comonotonic dependence is considered the strongest form of dependence of random variables.

If (X,Y) is a pair of random variables, Kendall's τ is defined by

$$\tau_{XY} = P(X_1 < X_2, Y_1 < Y_2) + P(X_1 > X_2, Y_1 > Y_2) - P(X_1 < X_2, Y_1 > Y_2) - P(X_1 > X_2, Y_1 < Y_2),$$

where (X_1,Y_1) and (X_2,Y_2) are independent copies of (X,Y). Hence τ_{XY} can be seen as the difference between two probabilities, namely the probability that the larger of the two *X*-values is associated with the larger of the two *Y*-values and the probability that the larger *X*-value is associated with the smaller *Y*-value. In case of continuous random variables, Kendall's τ takes on the value 1 if the variables are comonotonic. Somers' *D* is defined as the difference of the conditional probabilities that correspond to the probabilities in the definition of Kendall's τ, given that the two *Y*-values are not equal. Formally, the definition of Somers' *D* can be expressed as

$$D_{XY} = \frac{\tau_{XY}}{\tau_{YY}}.$$

Note that $\tau_{XY} = \tau_{YX}$, but in general $D_{XY} \neq D_{YX}$. Denote, similarly to the section on the connection of ROC and CAP curves, by *S* the random variable that describes the score of an obligor that is chosen at random from the whole portfolio. Let *J* denote the (complementary) default indicator with value 0 if the obligor under consideration defaults and 1 otherwise. By a short calculation one then obtains that

$$D_{S\,J} = P(S_D < S_{ND}) - P(S_D > S_{ND}) = AR,$$

where as above S_D and S_{ND} denote scores of defaulters and non-defaulters respectively and AR stands for Accuracy Ratio as it was defined in the context of CAP curves. Thus, Somers' *D* may be interpreted as a generalisation of the Accuracy Ratio. Since *J* in a certain sense is nothing but a perfect rating – which is unfortunately known only at a time when it is not of interest any longer – this observation on the connection between Somers' *D* and the Accuracy Ratio suggests to choose Somers' *D* as a criterion for the concordance of the scores *S* and any other rating *R*. Hence, a mathematical formulation of the problem of how to find a best possible shadow rating system *S* to imitate some given (e.g. externally) rating system *R* would be to solve the optimisation problem

$$\max_S D_{SR}.$$

The solution $S^{\#}$ of this problem should yield a value $D_{S^{\#}R}$ of Somers' *D* which is as close as possible to one, the theoretically highest value of *D*. By means of some standard statistical software packages, confidence intervals for estimates of Kendall's τ and Somers' *D* and estimates for differences of theses indices can be determined. This way, statistical inference becomes feasible on the quality of a shadow rating or the relative performance of shadow ratings with respect to the same reference rating.

Brier score

The Brier score is a method for the evaluation of the quality of the forecast of a probability. It has its origins in the field of weather forecasts. But it is straightforward to apply this concept to rating models.

Let $p_0, p_1, ..., p_K$ be the estimated default probabilities of the debtors in the K rating classes of a rating system. The Brier score is defined as[19]

$$B = \frac{1}{n}\sum_{j=1}^{n}(p_j - \theta_j)^2 ,$$

In the above formula *n* is the amount of rated debtors, p_j is the forecasted default probability of debtor *j*, and θ_j is defined as

$$\theta_j = \begin{cases} 1, & \text{if obligor } j \text{ defaults} \\ 0, & \text{else} \end{cases}.$$

From the above definition it follows that the Brier score is always between zero and one. The closer the Brier score is to zero the better is the forecast of default probabilities. A disadvantage of the Brier score is its performance for small default probabilities. In this case the "trivial forecast" will have a rather small Brier score. By trivial forecast we mean that all debtors are assigned the default frequency of the overall sample. In this case the expected Brier score is equal to the variance of the default indicator, i.e.

[19] See Brier (1950).

$$\bar{B} = (1-p)p, \ ,$$

where p is the default frequency of the overall sample. For $p \to 0$ the Brier score converges to zero. The only possibility of applying this score in a meaningful way is to compute the Brier score relative to the trivial score since the absolute values are very close together for cases with few defaults.

Binomial test

The binomial test is a natural possibility for the validation of *PD* estimates banks have to provide for each rating category of their internal rating systems. Its construction relies on an assumption of the default events in the rating category under consideration being independent.

The binomial test works as follows:

null hypothesis H0:	the *PD* of a rating category is correct
alternative hypothesis H1:	the *PD* of a rating category is underestimated

Given a confidence level q (e.g. 99%) the null hypothesis is rejected if the number of defaulters k in this rating category is greater than or equal to a critical value k^* which is defined as

$$k^* = \min\left\{k \mid \sum_{i=k}^{n} \binom{n}{i} PD^i \left(1-PD\right)^{n-i} \le 1-q\right\},$$

where n is the number of debtors in the rating category. The critical value k^* can be approximated by an application of the central limit theorem to the above formula. This approximation results in

$$k^* = \Phi^{-1}(q)\sqrt{nPD\left(1-PD\right)} + nPD,$$

where Φ^{-1} denotes the inverse function of the standard normal distribution. If it is preferred to express it in terms of an observed default rate p^* that is allowed at maximum

$$p^* \approx \Phi^{-1}(q)\sqrt{\frac{PD\left(1-PD\right)}{n}} + PD.$$

Therefore both concepts are roughly equivalent. In the sequel we concentrate on the binomial test. However, all results apply also to the normal approximation to the binomial test.

The resulting test is very simple and intuitive. However, it is based on the assumption that defaults are independent events. From empirical studies it is known that this is not the case. Defaults are correlated with small coefficients of correlation. Typical values for default correlation are around 0.5% to 3%. These numbers seem rather small. Applying the binomial test under the assumption of correlated defaults makes the mathematical framework more complicated. It is not possible to state a simple formula for the distribution of the default rates as above. To analyse the influence of default correlation on k^* we carry out two analyses.

The first one is based on exact numerical calculation involving numerical integration, the second one on an analytical approximation method.

In both settings we assume a one period model. Similar to the one-factor approach underlying the risk-weight functions of the IRB approach of Basel II, at the end of the period the value R_i of the assets of debtor i depends on a systematic factor X common to all debtors and a factor ε_i that is specific to the debtor. We assume further that

$$R_i \sim N(0,1),$$

$$X \sim N(0,1),$$

$$\varepsilon_i \sim N(0,1),$$

$$Cov(X, \varepsilon_i) = 0,$$

$$Cov(\varepsilon_{i,}\varepsilon_j) = 0, \quad i \neq j,$$

$$Corr(R_i, R_j) = \rho,$$

where *N(0,1)* is the standard normal distribution. The parameter ρ is called "asset correlation". Under these assumptions R_i can be modelled as

$$R_i = \sqrt{1-\rho}\,\varepsilon_i + \sqrt{\rho}\,X,$$

In this setting a debtor defaults if the value of his assets is below some threshold γ_i at the end of the period, i.e.

$$R_i \leq \gamma_i\,.$$

We define the default indicator Y_i as

$$Y_i = \begin{cases} 1, & R_i \leq \gamma_i, \\ 0, & \text{else.} \end{cases}$$

Assuming that all debtors have the same probability of default, i.e. all debtors are in the same rating category and the same threshold $\gamma := \gamma_i$ applies to all debtors, and that the asset correlation is the same for all pairs of debtors, we obtain the following properties for the default indicator Y_i

$$E(Y_i) = PD = \Phi^{-1}(\gamma),$$

$$Var(Y_i) = PD(1-PD),$$

$$Corr(Y_i, Y_j) = \frac{\Phi_2(\gamma, \gamma; \rho) - PD^2}{PD(1-PD)} =: \delta,$$

where $\Phi_2(\cdot,\cdot;\rho)$ denotes the bivariate standard normal distribution function with correlation ρ. The correlation δ is called "default correlation".

Denote by D_n the observed number of defaults among the n obligors in the rating category under consideration. Assuming that the default number distribution follows a Basel-II-like structure as described above, the distribution of D_n can be calculated by means of

$$P(D_n \le k) = \int_{-\infty}^{\infty} \sum_{l=0}^{k} \binom{n}{l} \Phi\left(\frac{\gamma - \sqrt{\rho}\, x}{\sqrt{1-\rho}}\right)^l \left(1 - \Phi\left(\frac{\gamma - \sqrt{\rho}\, x}{\sqrt{1-\rho}}\right)\right)^{n-l} \phi(x)\, dx,$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denotes the standard normal density. This formula for $P(D_n \le k)$ can be evaluated by numerical integration. Hence the exact correlation-dependent critical values k^* can be found by iterated calculations of $P(D_n \le k)$.

Additionally, we consider an easy (Basel-II-like) approximation to the distribution of D_n by means of the Vasicek distribution. Define the observed default rate L_n as

$$L_n = \frac{D_n}{n},$$

where, as above, D_n is the observed number of defaults and n is the total number of debtors in the rating category. Since defaults conditional on a realisation x of X are independent it follows from the law of large numbers that conditional on X

$$L_n \xrightarrow{n\to\infty} P(Y_i = 1 \mid X) = \Phi\left(\frac{\gamma - \sqrt{\rho}\, X}{\sqrt{1-\rho}}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. For "large" numbers of n we obtain using the above result

$$\mathrm{P}\left(D_n \le k\right) = \Pr\left(L_n \le \frac{k}{n}\right) \approx \mathrm{P}\left(\Phi\left(\frac{\gamma - \sqrt{\rho} X}{\sqrt{1-\rho}}\right) \le \frac{k}{n}\right) = \Phi\left(\frac{\sqrt{1-\rho}\, \Phi^{-1}\left(\frac{k}{n}\right) - \gamma}{\sqrt{\rho}}\right).$$

Setting the above probability equal to the confidence level q and solving for k we can obtain an approximation for the critical value k^*.

We are now able to analyse the effect of introducing default correlation to the binomial test and can judge how critical the assumption of independence is in the binomial test. Note that in the independent case applying the approximation in order to calculate the critical value just yields the expected number of defaults plus one. Therefore, the use of the approximation formula is not sensible if the assumption of independence appears to be justified.

Example 1: PD = 1.0%, q = 99%, N = 100

Asset correlation	Default correlation	Exact k^*	Approximate k^*
0.0%	0.0%	5	2
5.0%	0.41%	6	4
10.0%	0.94%	7	5
15.0%	1.60%	8	7
20.0%	2.41%	10	8

Example 2: PD = 0.5%, q = 99%, N = 1000

Asset correlation	Default correlation	Exact k^*	Approximate k^*
0.0%	0.0%	11	6
5.0%	0.25%	20	18
10.0%	0.58%	29	27
15.0%	1.03%	37	35
20.0%	1.60%	45	44

Example 3: PD = 1.0%, q = 99%, N = 1000

Asset correlation	Default correlation	Exact k^*	Approximate k^*
0.0%	0.0%	19	11
5.0%	0.41%	35	32
10.0%	0.94%	49	47
15.0%	1.60%	63	62
20.0%	2.41%	77	76

Example 4: PD = 5.0%, q = 99%, N = 1000

Asset correlation	Default correlation	Exact k^*	Approximate k^*
0.0%	0.0%	68	51
5.0%	1.20%	128	125
10.0%	2.55%	172	169
15.0%	4.08%	212	210
20.0%	5.78%	252	250

Example 5: PD = 1.0%, q = 99%, N = 10000			
Asset correlation	**Default correlation**	**Exact k^***	**Approximate k^***
0.0%	0.0%	125	101
5.0%	0.41%	322	320
10.0%	0.94%	470	468
15.0%	1.60%	613	611
20.0%	2.41%	755	753

From the above examples we find that the assumption of independence in the binomial test is not robust for the higher percentiles. Even small violations lead to dramatic changes in k^*. Since default correlation is not zero in reality this means that large deviations of estimated *PD* and observed default rate are not unlikely. The existence of default correlation makes it impossible even for banks with a large number of debtors to estimate a *PD* which is close to the observed default rate with a probability close to one. The reason for this is that the law of large numbers does not work because of the existence of correlation (defaults are only independent conditional on a realisation of the systematic factor X, unconditional defaults are not independent). The distribution of future default rates does not collapse to one point even for a very large number of debtors in a rating category. The consequence of this statistical phenomenon is that the validation of a *PD* estimate is in fact not the validation of the correctness of a single number but the validation of the mean of a distribution. As a further consequence, in order to construct cross-sectional tests of the adequacy of PD estimates that take into account correlation effects, assumptions on the underlying dependence structure have to be stated.

However, as the critical values of PD tests that incorporate correlations tend to be much greater than the critical values of the (independence-based) binomial test, this latter test will be conservative in the sense of too early rejections of the null hypothesis H0. In other words, when the binomial test is applied in a context of correlation, the true size of the type I error (unjustified rejection of H0) will be higher than the nominal level of the test indicates. In order to get a feeling for the magnitude of this effect we present the true sizes of the type I error as coloured regions in the "pool size-PD" space (Figure 5). For a level of confidence in the uncorrelated case $q_{\text{uncorrelated}}$ the true size of the type I error $q_{\text{correlated}}$ in the single factor model setting is given by

$$q_{\text{correlated}} = \Phi\left(\frac{\sqrt{1-\rho}\,\Phi^{-1}\left(\Phi^{-1}\left(q_{\text{uncorrelated}}\right)\sqrt{\frac{PD\left(1-PD\right)}{n}}+PD\right)-\Phi^{-1}\left(PD\right)}{\sqrt{\rho}}\right).$$

Setting $q_{\text{uncorrelated}}$ to 99.9% (i.e. $\Phi^{-1}\left(q_{\text{uncorrelated}}\right)=3.09$) and assuming the correlation ρ as set for the QIS 3, the picture in Figure 5 is the result. Surprisingly, even for the quite high QIS 3

correlations the true size of type I error $q_{\text{correlated}}$ does not fall below 80% for realistic PDs as well as for realistic pool-sizes (see Figure 9).

Figure 9. True size of type I error at nominal level 99.9% as function of pool size and PD.

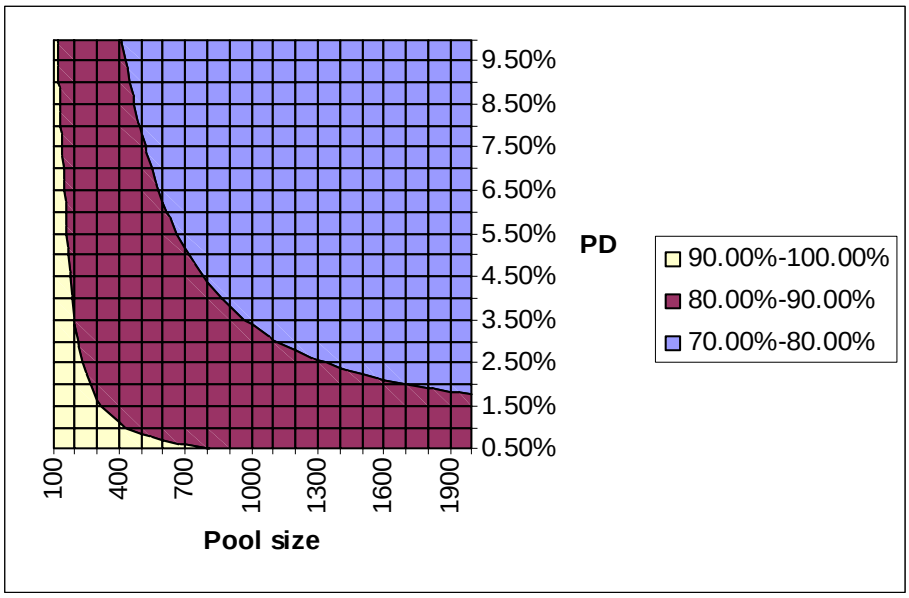


Figure 9 suggests that the binomial test might serve as kind of an early warning system. A rejection of H0 ("PD estimate adequate") by the binomial test would trigger a closer inspection of the underlying rating system, but – due to the high size of the type I error – would not entail immediate consequences for the approval of the rating system as a whole.

Chi-square (or Hosmer-Lemeshow) test

Let $p_0,\ldots,p_k$ denote the forecasted default probabilities of debtors in the rating categories $0,1,\ldots,k$. Define the statistic

$$T_k = \sum_{i=0}^{k} \frac{(n_i\, p_i - \theta_i)^2}{n_i\, p_i\, (1-p_i)}$$

with $n_i =$ number of debtors with rating i and $\theta_i =$ number of defaulted debtors with rating i. By the central limit theorem, when $n_i \to \infty$ simultaneously for all i, the distribution of T_k will converge in distribution towards a χ^2_{k+1}-distribution if all the p_i are the true default probabilities. Note however, that this convergence is subject to an assumption of independent default events within categories and between categories.

The p-value of a χ^2_{k+1}-test could serve as a measure of the accuracy of the estimated default probabilities: the closer the p-value is to zero, the worse the estimation is. However, there is a further caveat: if the p_i are very small the rate of convergence to the χ^2_{k+1}-distribution may be low. But note that relying on the p-value makes possible a direct comparison of forecasts with different numbers of rating categories.

Simulation study on the performance of the normal and traffic lights tests

Model for the simulation study

Goal of the study is to generate close-to-reality time series of annual default rates and to apply both the normal and the traffic lights test methodologies to them. Closeness to reality in this case means that the rates in different years can be stochastically dependent and that the same holds for the default events within one year. Essentially, the model can be considered as an extension of the Vasicek model which was used in deriving the Basel II risk-weight functions into the time dimension.

- Assume that a fixed portfolio is being observed in years $t = 1,\ldots,T$.
- At time t the number of obligors in the portfolio is the a priori known deterministic number N_t.
- The change in the general economic conditions from year $t-1$ to year t is expressed by the random variable S_t. Small values of S_t reflect poor economic conditions, large values stand for good conditions.
- The joint distribution of S is normal with standardised marginal distributions and correlation matrix $\Sigma = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1T} \\ r_{21} & 1 & r_{23} & \cdots & r_{2T} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ r_{T1} & \cdots & \cdots & r_{TT-1} & 1 \end{pmatrix}$. Defining $r_{st} = \vartheta^{|s-t|}$ for some appropriately chosen $\vartheta \in [0,1]$ is common practice in panel analysis and is also the approach which is followed in this simulation study.
- The unconditional default probability in year t is p_t.
- Conditional on S, the numbers of default D_t are independent and binomially distributed with sizes N_t and conditional default probabilities $p_t(S) = \Phi\left(\frac{p_t - \sqrt{\rho_t}\, S_t}{\sqrt{1-\rho_t}}\right)$. The ρ_t are interpreted as the correlations of the changes in the obligors' asset values from year $t-1$ to year t. The annual percentage default rates d_t will consequently be calculated as $d_t = D_t / N_t$.

Normal test

The construction of the normal test of PDs is based on the following observation:

If $X_1, X_2, X_3 \ldots$ are independent random variables with (not necessarily equal) means $\mu_1, \mu_2, \mu_3, \ldots$ and common variance $\sigma^2 > 0$ then by the central limit theorem the distribution of the standardised sum

$$\frac{\sum_{t=1}^{T}(X_t - \mu_t)}{\sqrt{T}\,\sigma}$$

will converge to the standard normal distribution for T tending towards ∞.

In most cases of practical interest, the rate of convergence is quite high. Therefore, even for small values of T (e.g. $T = 5$) approximating the standardised sum with the standard normal distribution seems reasonable.

In order to apply the normal approximation to the case of PD forecasts $PD_1, PD_2, \ldots, PD_T$ and observed percentage default rates $d_1, d_2, \ldots, d_T$, an estimator τ^2 of the assumed common variance σ^2 must be specified. The obvious choice of τ^2 is

$$\tau_0^2 = \frac{1}{T-1}\sum_{t=1}^{T}(d_t - PD_t)^2.$$

This estimator will be unbiased if the forecast PDs exactly match the true PDs. However, τ_0^2 will be upwardly biased as soon as some of the forecasts differ from the corresponding true PDs. The bias can be considerably reduced by choosing

$$\tau^2 = \frac{1}{T-1}\left(\sum_{t=1}^{T}(d_t - PD_t)^2 - \frac{1}{T}\left(\sum_{t=1}^{T}(d_t - PD_t)\right)^2\right).$$

Under the hypothesis of exact forecasts, τ^2 is unbiased. In case of mismatches, it is also upwardly biased, but to a less extent than τ_0^2.

The normal test of the hypothesis *"None of the true probabilities of default in the years* $t = 1, \ldots, T$ *is greater than its corresponding forecast* PD_t*"* goes as follows:

Reject the hypothesis at confidence level α *if*

$$\frac{\sum_{t=1}^{T}(d_t - PD_t)}{\sqrt{T}\,\tau} > z_\alpha,$$

where z_α *is calculated as the standard normal* α*-quantile (e.g.* $z_{0.99} \approx 2.33$*). If the critical value* z_α *is not exceeded by the test statistic, accept the hypothesis at level* α.

Traffic lights test (according to Blochwitz, Hohl, and Wehn)

The traffic lights approach to PD testing by Blochwitz, Hohl, and Wehn is based on the following observation:

If default events are assumed to be independent and, additionally, independence in time is taken as given, under the Null hypothesis a multinomial distribution with well-defined probabilities of the outcomes (identified with the traffic light colours) can be chosen as test statistic.

Denote by $N_1, \ldots, N_T$ the sizes of the portfolio under consideration in the years $t = 1, \ldots, T$, and by $D_1, \ldots, D_T$ the corresponding numbers of defaults. If the default events in year t are independent and that all the obligors in the portfolio have the same probability of default p_t,

the number D_t of defaults in year t is binomially distributed with probability parameter p_t and size parameter N_t. As a consequence, by the central limit the distribution of the standardised default rate

$$R_t = \frac{D_t - N_t p_t}{\sqrt{N_t p_t (1-p_t)}}$$

can be approximately described by the standard normal distribution as long as $N_t p_t$ is not too small. Define probabilities[20] q_g, q_y, q_o, q_r (corresponding to the colours green, yellow, orange, and red) with $q_g + q_y + q_o + q_r = 1$ and the mapping $C(x)$ by

$$C(x) = \begin{cases} g, & x \le \Phi^{-1}(q_g), \\ y, & \Phi^{-1}(q_g) < x \le \Phi^{-1}(q_y), \\ o, & \Phi^{-1}(q_y) < x \le \Phi^{-1}(q_o), \\ r, & \Phi^{-1}(q_o) < x, \end{cases}$$

where Φ^{-1} denotes the inverse function of the standard normal distribution function. With this definition, under the assumption of independence of the annual numbers of default, the vector $A = (A_g, A_y, A_o, A_r)$ with A_c counting the appearances of colour c in the sequence $C(R_1), \ldots, C(R_T)$ will be approximately multinomially distributed with

$$P\left[A = \left(k_g, k_y, k_o, k_r\right)\right] = \frac{T!}{k_g!\, k_y!\, k_o!\, k_r!} q_g^{k_g}\, q_y^{k_y}\, q_o^{k_o} q_r^{k_r},$$

for every quadruple $\left(k_g, k_y, k_o, k_r\right)$ of non-negative integers such that $k_g + k_y + k_o + k_r = T$. In order to construct critical regions for tests of the underlying probabilities of defaults, for the case of $T \le 9$ the statistic

$$V = 1000 A_g + 100 A_y + 10 A_o + A_r$$

turned out to be appropriate. With this notation, the traffic lights test of the hypothesis *"None of the true probabilities of default in the years $t = 1, \ldots, T$ is greater than its corresponding forecast PD_t"* can be specified as follows:

Reject the hypothesis at confidence level α if

$$V \le v_\alpha,$$

[20] For the purpose of the simulation study, the probabilities were chosen as $q_g = 0.5$, $q_y = 0.3$, $q_o = 0.15$, and $q_r = 0.05$.

where v_α is calculated as the greatest number v with the property that $P[V \le v] < 1-\alpha$. If the critical value v_α is exceeded by the test statistic, accept the hypothesis at level α.

Subject and results of the simulation study

Both the normal test as well as the traffic lights test were derived by asymptotic considerations – with regard to the length of the time series of the observed default rates in case of the normal test and with regard to the portfolio size in the case of the traffic lights test. As a consequence, even in the case of complete independence in time and in the portfolio it is not clear that the type I errors[21] observed with the tests will be dominated by the nominal error levels. Of course, the compliance with the nominal error level is much more an issue in the case of dependencies of the annual default rates or of the default events in the portfolio. When compliance with the nominal error level for the type I error is confirmed, the question has to be examined which test is the more powerful, i.e. for which test the type II errors[22] are lower.

In order to clarify these points, simulation scenarios for $T = 5$ years of default experience as described in Table 3 (for type I errors) and

Table 4 (for type II errors) were generated. Tables Table 5 and

Table 6 list the parameter settings that were used for the implementation[23] of the scenarios. For all simulation runs, a constant-over time portfolio size of 1,000 obligors was fixed. Finally, Table 7 and
Table 8 report the observed error rates in the cases of the type I error and the type II error respectively.

With regard to the type I errors, according to Table 7 both test methodologies seem to be essentially in compliance with the nominal error levels. At high error levels (10% and 5%), the normal test fits the levels better than the traffic lights test does. For low error levels (2.5% and less), the order of compliance is just reversed with the traffic lights performing better. Both test methodologies face in some scenarios relatively bad performance at the very low levels. Serious outliers are observed at 5% level for the traffic lights test as, in the dependence scenarios with larger PDs (DC_LC and DV_LV), type I errors of more than 10% occur.

In general, according to
Table 8 the traffic lights test appears to be more powerful than the normal test. In case of low PD forecasts to be checked, compared to the case of larger PDs for both test methodologies power is very low. However, whereas in most scenarios differences in power are not dramatic, the traffic lights test sees a heavy collapse of performance in the "independence with varying larger PDs" (I_LV) scenario where at levels 2.5% and 1% the normal test is more than 10% better.

To sum up, both test methodologies seem to be reliable with respect to compliance with the nominal error levels, even in case of dependencies that were not taken into account in their designs. The traffic lights test is more powerful than the normal, and should therefore be

[21] I.e. the probabilities of erroneously rejecting the hypothesis.

[22] I.e. the probabilities of not rejecting the hypothesis if specific alternatives are true.

[23] Every scenario was investigated with 25,000 simulation runs.

preferred to the normal test. However, the normal test appears to be slightly more robust than the traffic lights test with respect to violations of the assumptions underlying its design. This observation might favour simultaneous applications of the tests.

Table 3

Scenarios for type I error simulations

Scenario	Description
I_SC	Independence of default events and annual default rates, small and constant unconditional PDs.
I_LC	Independence of default events and annual default rates, larger and constant unconditional PDs.
DC_SC	Time dependence, constant asset correlations, small and constant unconditional PDs.
DC_LC	Time dependence, constant asset correlations, larger and constant unconditional PDs.
I_SV	Independence of default events and annual default rates, small and varying unconditional PDs.
I_LV	Independence of default events and annual default rates, larger and varying unconditional PDs.
DV_SV	Time dependence, varying asset correlations, small and varying unconditional PDs.
DV_LV	Time dependence, varying asset correlations, larger and varying unconditional PDs.

Table 4

Scenarios for type II error simulations

Scenario	Description
I_SV	Independence of default events and annual default rates, small and varying unconditional PDs.
I_LV	Independence of default events and annual default rates, larger and varying unconditional PDs.
DV_SV	Time dependence, varying asset correlations, small and varying unconditional PDs.
DV_LV	Time dependence, varying asset correlations, larger and varying unconditional PDs.

Table 5

Parameter settings for type I error simulations (see Section 1)

Scenario	Correlation in time (ϑ)	Asset correlations	Forecast of PDs (in %)	True PDs (in %)
I_SC	0	0; 0; 0; 0; 0	0.3; 0.3; 0.3; 0.3; 0.3	0.3; 0.3; 0.3; 0.3; 0.3
I_LC	0	0; 0; 0; 0; 0	3.0; 3.0; 3.0; 3.0; 3.0	3.0; 3.0; 3.0; 3.0; 3.0
DC_SC	0.2	0.05; 0.05; 0.05; 0.05; 0.05	0.3; 0.3; 0.3; 0.3; 0.3	0.3; 0.3; 0.3; 0.3; 0.3
DC_LC	0.2	0.05; 0.05; 0.05; 0.05; 0.05	3.0; 3.0; 3.0; 3.0; 3.0	3.0; 3.0; 3.0; 3.0; 3.0
I_SV	0	0; 0; 0; 0; 0	0.1; 0.2; 0.3; 0.4; 0.6	0.1; 0.2; 0.3; 0.4; 0.6
I_LV	0	0; 0; 0; 0; 0	1.0; 2.0; 3.0; 4.0; 6.0	1.0; 2.0; 3.0; 4.0; 6.0
DV_SV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	0.1; 0.2; 0.3; 0.4; 0.6	0.1; 0.2; 0.3; 0.4; 0.6
DV_LV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	1.0; 2.0; 3.0; 4.0; 6.0	1.0; 2.0; 3.0; 4.0; 6.0

Table 6

Parameter settings for type II error simulations (see Section 1)

Scenario	Correlation in time (ϑ)	Asset correlations	Forecast of PDs (in %)	True PDs (in %)
I_SV	0	0; 0; 0; 0; 0	0.1; 0.2; 0.3; 0.4; 0.6	0.15; 0.25; 0.35; 0.45; 0.65
I_LV	0	0; 0; 0; 0; 0	1.0; 2.0; 3.0; 4.0; 6.0	1.5; 2.5; 3.5; 4.5; 6.5
DV_SV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	0.1; 0.2; 0.3; 0.4; 0.6	0.15; 0.25; 0.35; 0.45; 0.65
DV_LV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	1.0; 2.0; 3.0; 4.0; 6.0	1.5; 2.5; 3.5; 4.5; 6.5

Table 7

Type I errors (normal = with normal test, traffic = with traffic lights test)

Nominal level	0.1	0.05	0.025	0.01	0.005	0.001
I_SC, normal	0.109	0.059	0.045	0.027	0.020	0.014
I_SC, traffic	0.135	0.085	0.043	0.011	0.007	0.001
I_LC, normal	0.130	0.081	0.055	0.037	0.028	0.016
I_LC, traffic	0.104	0.062	0.030	0.013	0.005	0.001
DC_SC, normal	0.092	0.049	0.030	0.017	0.013	0.007
DC_SC, traffic	0.124	0.076	0.029	0.018	0.016	0.008
DC_LC, normal	0.116	0.070	0.044	0.026	0.019	0.010
DC_LC, traffic	0.136	0.113	0.026	0.024	0.023	0.018
I_SV, normal	0.111	0.059	0.043	0.024	0.017	0.012
I_SV, traffic	0.132	0.088	0.043	0.013	0.005	0.001
I_LV, normal	0.128	0.077	0.051	0.032	0.024	0.014
I_LV, traffic	0.096	0.060	0.029	0.012	0.004	0.001
DV_SV, normal	0.083	0.037	0.021	0.010	0.007	0.003
DV_SV, traffic	0.115	0.071	0.027	0.017	0.015	0.007
DV_LV, normal	0.113	0.062	0.036	0.019	0.013	0.005
DV_LV, traffic	0.126	0.108	0.023	0.022	0.022	0.017

Table 8

Type II errors (normal = with normal test, traffic = with traffic lights test)

Nominal level	0.1	0.05	0.025	0.01	0.005	0.001
I_SV, normal	0.736	0.836	0.875	0.922	0.944	0.964
I_SV, traffic	0.685	0.782	0.874	0.946	0.972	0.990
I_LV, normal	0.252	0.366	0.467	0.575	0.643	0.754
I_LV, traffic	0.259	0.374	0.600	0.688	0.760	0.871
DV_SV, normal	0.862	0.927	0.956	0.977	0.984	0.992
DV_SV, traffic	0.811	0.868	0.950	0.965	0.969	0.983
DV_LV, normal	0.775	0.858	0.908	0.946	0.961	0.979
DV_LV, traffic	0.733	0.760	0.933	0.935	0.936	0.955

IV. Loss given default validation

Rosalind L Bennett, Eva Catarineu and Gregorio Moral

Introduction

LGD is an important element of the IRB approach to capital measurement. LGD is particularly important because the minimum regulatory capital charge is highly sensitive to the LGD that is reported by a financial institution. Under the advanced approach, financial institutions are allowed to use internally derived measures of LGD rather than a standard LGD given by the revised Framework. Thus, the validation of internal measures of LGD is crucial to the validation of the appropriateness of the capital measures.

There are several key issues to address when discussing LGD. First, is how LGD is defined and measured which, in turn, results in a need to establish what we understand by "loss" and "default". The first part of this section introduces four existing methods to measure LGDs depending on the source, type of facilities and definition of loss they use. The first method is derived from the observation of market prices on defaulted bonds or marketable loans soon after default (market LGDs). The second is based on the discount of future cash flows resulting from the workout process from the date of default to the end of the recovery process (workout LGDs). In the third, LGD is derived from non-defaulted risky bond prices by means of an asset pricing model (implied market LGDs). Finally, for retail portfolios, LGD can be inferred from the experience of total losses and PD estimates (implied historical LGDs).

The second part of this section, devoted to general aspects of LGD estimation, discusses the reference data sets (RDS) that are needed to estimate the risk parameters (PDs, LGDs, EADs) and the pooling of internal and external data.

The third part of the section looks specifically at workout LGDs. This part outlines the different components in the process of computing the workout loss of a defaulted facility when using the method of discounting cash flows. First, it focuses on the issues arising from the computation of workout LGDs: issues related to the different types of recoveries (cash and non-cash), direct and indirect costs of recovery, defining when the recovery process is over and the treatment of repossessions. Then, this part discusses alternative interest rates for discounting the cash flows. This issue is of primary importance in the calculation of workout LGDs and the impact of using alternative discount rates on LGD estimation is illustrated in numerical examples. Finally, the treatment of facilities for which there is incomplete information is considered.

The fourth part of this paper looks at LGD validation and serves as a useful illustration of problems that arise in the validation process. The final part presents the conclusions. In addition the appendix provides a survey of a number of empirical studies on LGD estimation. The intention is to use common empirical findings to provide some additional guidelines that may be helpful in evaluating a bank's LGD estimation.

Definition and measurement of LGD

Before attempting to estimate or validate LGD one must first understand what is meant by the term LGD. This section looks at the definition of LGD, and the definitions of default and loss on which LGD is based, and presents a classification of the procedures that, in principle, can be used to estimate LGDs.

Definition of LGD

In general, LGD is the loss, expressed as a percentage of the EAD, on a credit facility if the credit defaults.[24] We can further refine this definition to distinguish between the ex-post measures of LGD on defaulted facilities and ex-ante measure on non-defaulted facilities.

LGD for a non-defaulted facility can be defined as the ex-ante estimate of loss conditional on the default, expressed as a percentage of the EAD. The LGD associated with a non-defaulted facility can be viewed as a random variable. Frequently, we are interested in having one figure for the value of the LGD which is typically based on an estimate for the expectation of this random variable, i.e. expected LGD.

LGD for a defaulted facility is the ex-post loss expressed as a percentage of the exposure at the time of default. If there is complete information on all of the losses related to a facility, and a method to calculate losses has been chosen, we can directly calculate realised LGD. If there is not complete information on the losses related to a defaulted facility, for example if the facility is in the process of workout, LGD is a random variable. We can calculate an estimate of LGD for these defaulted facilities by using complete information from a sample of similar facilities.

A reference data set (RDS), which includes realised LGDs on defaulted facilities, can be used to estimate LGD on non-defaulted facilities. There are different methods that can be used to assign an LGD to non-defaulted facilities. These can be classified as subjective and objective methods according to the type of input used.

- Subjective methods are based on expert judgment. Banks use these methods in portfolios for which there are no defaults or in the early stages of use of their internal models.
- Objective methods use numerical data containing information on LGD as the main input. Additionally, it is possible to subdivide objective methods into explicit and implicit methods.
  - In explicit methods, the LGD is estimated for each facility using a reference data set (RDS) of defaulted facilities. The first step is to determine the realised LGD of each facility included in the RDS. The second step consists of assigning an LGD to each non-defaulted facility (using a model).[25] As we will show in following sections, the loss can be computed using either market values (explicit market LGD) or discounted cash-flows derived from the recovery process (workout LGD).
  - Implicit methods are not based on the realised LGD on defaulted facilities contained in a RDS. Instead, LGD is derived using a measure of total losses and PD estimates.
    - The implied market LGD methods derive LGD from risky bond prices using an asset pricing model. However, rather than using defaulted facilities like the explicit market LGD, they use a reference data set that includes non-defaulted facilities and credit spreads instead of realised LGD, as basic inputs. The credit

[24] Under the advanced approach, "LGD must be measured as the loss given default as a percentage of the EAD." See BCBS (2004), paragraph 297.

[25] The easiest procedure is to use the sample mean of the realised LGDs in the reference data set, though there are more sophisticated procedures.

spreads on risky bonds reflect, among other things, the expected loss on the bonds. Recent models illustrate how to decompose this measure of expected loss into the PD and the LGD. These methods can be very useful in those portfolios with very few defaults such as loans to banks, sovereign loans and very large corporate loans. It is important to note that in certain portfolios, information on defaults is scarce and an empirically grounded estimate requires the use of information from non-defaulted facilities.

- The revised Framework considers an implicit method for obtaining LGDs for retail portfolios.[26] This estimate of LGD uses the PD estimates and the experience of total losses in the portfolio to derive the implied LGD. In the following analysis, this method is called the implied historical LGD.

Table 9 summarises the different objective methods of calculating realised LGDs and assigning LGDs to non-defaulted facilities.

Table 9

Classification of the objective methods to obtain LGDs

Source	Measure	Type of facilities in the RDS		Most applicable to
		Defaulted facilities	Non-defaulted facilities	
Market values	Price differences	Market LGD		Large corporate, sovereigns, banks
	Credit spreads		Implied market LGD	Large corporate, sovereigns, banks
Recovery and cost experience	Discounted cash flows	Workout LGD		Retail, SMEs, large corporate
	Historical total losses and estimated PD	Implied historical LGD		Retail

The aim of this paper is to present a general framework for the validation of LGDs from a regulatory point of view. It is clear that purely subjective methods for estimating LGD are not valid from a regulatory standpoint. The revised Framework specifies that the "LGD estimates must be grounded in historical recovery rates."[27] Regarding the objective methods, there is some debate whether the revised Framework requirements exclude any of the four general methods for estimating LGDs, in particular the implied market method since it uses information from non-defaulted facilities. Furthermore, asset pricing models used under the implied market method are in early stages of development and certain capital issues remain

[26] See BCBS (2004), paragraph 465.

[27] See BCBS (2004), paragraph 470.

controversial (for example, what part of the spread is attributable to credit risk and what part for other factors). On the other hand, the validation of implied historical LGDs relies essentially on the validation of the PDs used in this method. For these reasons, in the following sections we focus on the explicit methods and primarily on workout LGD.

Definition of default

There is a broad spectrum of possible definitions of default, which can be generally classified as either objective or subjective. An objective definition is used whenever the classification of a defaulted facility depends only on observable characteristics that are beyond the control of the bank. For instance, an objective definition of default could be based on a minimum number of days past due and a minimum amount that is past due. On the other hand, banks frequently use subjective definitions of default that, for instance, depend on some internal classification of the facilities that are in trouble, based on risk managers appraisals or decisions made by the banks themselves, such as starting a legal process.

The definition of default employed in Basel II, is based on two sets of conditions (at least one of the conditions must be met): first that "the bank considers that the obligor is unlikely to pay [in full]", and second, that "the obligor is past due more than 90 days on any material credit obligation".[28] The first is a subjective condition. The second is an objective condition, where the minimum number of days past due that triggers default is 90 days, and the minimum past due threshold amount is the level of materiality fixed by the bank.

The most natural definition of default depends on the portfolio and the product. In retail portfolios, for instance, banks often use an object definition of default with a time threshold of 90 days, though other time thresholds are also used (for example, 45 day or 180 day thresholds). In corporate portfolios, the definition of default is typically based on subjective conditions and is more complex.

In practice, bank databases frequently do not include all historical defaulted facilities but only those facilities that ultimately resulted in a loss. It may be the case that using the Basel II definition of default may result in instances of default that do not ultimately result in a loss. The use of databases that only include those facilities that resulted in loss when calculating LGD is equivalent to the use of a more stringent definition of default which may differ substantially from the Basel II definition of default.

Consequently, it is important to identify which definition of default is being used because:

- the definition of default directly influences LGD,
- for regulatory purposes the LGD estimates must be consistent with the Basel II definition of default,
- the default definition used for the LGD computations will have to be consistent with the one used when estimating PDs to obtain sensible values for economic capital and expected loss, and
- it makes little sense to perform direct benchmarking exercises on LGDs (among banks, portfolios or at different moments of time) if different definitions of default have been employed.

[28] See BCBS (2004), paragraph 452.

Furthermore, it is useful to develop methods that establish a link between LGD estimates which use different default definitions.[29]

Definition of loss

Before beginning any LGD estimation procedure, it is also necessary to define loss. Many of the implications related to the definition of default discussed in the previous section are also applicable here. It may make little sense to compare LGD estimations among banks or portfolios that are using different definitions of loss.

First, it is important to notice that economic loss, as defined in the revised Framework is not the same as accounting loss: "The definition of loss used in estimating LGD is economic loss. […] This must include material discount effects and material direct and indirect costs associated with collecting on the exposure."[30] Economic loss can be determined using different methods that we have previously classified as either explicit or implicit methods. Focusing on the explicit methods, when associating an economic loss to every element included in the reference data set, two different approaches can be used: market LGD and workout LGD.[31]

Market LGD depends on the market price of a defaulted facility soon after the date of default (typically around 30 days). Most rating agency studies on recoveries use this approach.[32] This method is useful since prices reflect the investor's assessment of the discounted value of recoveries. However, if markets are illiquid or they are driven by shocks unrelated to expected recoveries, this measure may not be appropriate. These concerns are particularly relevant for relatively new loan markets.

Another measure of LGD, the workout LGD, values loss using information from the workout. The loss associated with a defaulted facility is calculated by discounting the cash flows, including costs, resulting from the workout from the date of default to the end of the recovery process. The loss is then measured as a percentage of the exposure at default.[33] The timing of cash flows and both the method and rate of discount are crucial in this approach. There are four main issues that arise when using the workout approach to compute the loss of a defaulted facility. First, it is important to use the appropriate discount rate. Second, there are different possibilities about how to treat zero or negative LGD observations in the reference data. Third, the measurement and allocation of costs associated with workout can be

---

29 Moral and Garcia (2002) present a methodology that establishes a relationship between LGD estimates that are based on two different definitions of default, one being broader than the other. They establish this relationship for instances where the mathematical expectation is used as the LGD estimate. In their analysis, they use an estimate of the conditional probability to convert the expected LGD under one definition of default to the expected LGD under a different definition of default.

30 See BCBS (2004), paragraph 460.

31 In general, the RDS can include defaulted and/or not defaulted facilities.

32 Rating agencies use par values just before the default as EAD value to compute the loss. Note that LGD is simply 100 percent minus the market price rate expressed as a percent. Alternatively, the market LGD can be computed by comparing the price of a defaulted facility soon after the date of default to the market price just before the default date. The issue of what is the appropriate measure of EAD is complicated, and although important and interesting, the details of the issue will not be covered in this section.

33 This discussion will not include difficulties associated with the estimation and validation of the exposure at default. Particular difficulties arise from estimating the amount of exposure, or the credit conversion factors (CCF), associated with commitments to defaulting borrowers.

complicated. Fourth, it is not clear how to define the completion of a workout. These four issues are discussed below.

Some issues related to LGD estimation

Before moving onto more specific aspects of workout LGDs, it is worth briefly discussing three issues related to LGD estimation: what facilities have to be included in the reference data sets (RDS) used in the estimation process; when a recovery process is over; and how estimates of LGD that use a specific definition of default can be transformed into estimates under other definitions of default.

For a certain portfolio, an internal or external reference data set (RDS) is required to estimate the risk parameters (PDs, LGDs and EADs) that are needed for internal uses and the computation of capital requirements in the IRB approach. Ideally, these RDSs should:

- cover at least a complete business cycle,
- contain all the defaults produced within the considered time frame,
- include all the relevant information to estimate the risk parameters, and
- include data on the relevant drivers of loss.

In practice, banks use RDSs that include internal and/or external data that may cover different time frames, use different definitions of default and, in some cases, contain a biased sample of all the defaults produced within the timeframe. Thus, it is necessary to check for consistency within the RDSs. Otherwise, the final estimates of LGD will be inaccurate or biased.

To estimate LGDs, defaulted facilities that are still in the recovery process (also known as incomplete workouts) are frequently excluded from the reference data set. However, banks may find it useful to include incomplete workouts in certain cases if they can associate some loss estimates with those facilities.

In the case of workout LGDs, banks must define when a workout is finished.[34] Sometimes banks employ a recovery threshold (for example, whether the remaining non-recovered value is lower than 5% of the EAD), or a given time threshold (such as one year from the date of default). If the definition results in the exclusion of many defaulted facilities from the LGD estimates, the treatment of incomplete workouts must be revised.

Another issue that arises in the estimation of LGD is the problem of dealing with different definitions of default. For instance, some have proposed that when LGD is zero or negative the observation should be eliminated from the reference data set, also referred to as truncating the dataset.[35] However, if zero or negative observations are eliminated, the definition of default changes to a more stringent definition of default. If the reference data set includes a large number of observations where realised losses are negative, it is important to revisit the definition of default and loss.

---

[34] For a market LGD (corporate bond or a marketable loan), well established criteria indicate the number of days after which the recovery process is over. However, in the case of a workout LGD it is more complicated to set up the final date of a workout.

[35] A closely related concept, censoring, is discussed in what follows.

In certain portfolios, for example a portfolio of large corporate loans, it is not possible for individual banks to obtain a suitable sample of losses from defaulted facilities to estimate LGD. Therefore, institutions may turn to external data sources. If external estimates were available for LGD using a particular definition of default, then it is necessary to adjust for the difference in the definition of default. Therefore, it is useful to develop methods that establish a link between LGD estimates that use different definitions of default.[36]

Workout LGD

This section looks at the process of computing the workout loss of a defaulted facility, and discusses issues related to the measurement of the various components of the workout LGD including recoveries, costs and the discount rate.

Components of workout LGD

There are three main components for computing a workout loss: the recoveries (cash or non-cash), the costs (direct and indirect) and the discount factor that will be fundamental to express all cash-flows in terms of monetary units at the date of default. If all the cash flows associated with a defaulted facility from the date of default to the end of the recovery process are known (i.e. we have complete information) then the realised LGD, measured as percentage of the EAD at the time of default, is given by:

$$Realised\ LGD = \left[1 - \frac{\sum_i R_i(r) - \sum_j P_j(r)}{EAD}\right] \qquad (1)$$

where R_i is each of the i discounted recoveries of the defaulted facility, P_j is each of the j discounted payments or costs during the recovery period and r represents a discount rate.

When loss is calculated by setting all negative observations of loss to zero, as shown in equation (2), it is referred to as censoring the data.

$$Realised\ LGD = Max\left[1 - \frac{\sum_i R_i(r) - \sum_j P_j(r)}{EAD}, 0\right] \qquad (2)$$

Censoring the data does not change the definition of default. In contrast, when defaults with zero or negative LGD are removed from the dataset, referred to as truncating the data, the definition of default has changed.

If the data has not been censored or truncated, then the realisations of LGD could be negative (representing an economic gain on the asset).[37] This would be more likely if the definition of default was broader, such as 30-days past due. In principle, the constraint of realised LGD being greater or equal to zero, for regulatory purposes, can be imposed for prudential reasons. However, banks do not necessarily impose this condition in their estimates for non-regulatory use (for instance, pricing). In practice, when applying equation

---

[36] See Moral and García (2002) for an example of such a method.

[37] Typically, this occurs when the opportunity cost implicit in the discount rate is lower than the income from interest and fees on the defaulted loan. Censoring the data is consistent with not recognising future margin income.

(1) or (2) banks will have to make decisions about several issues that are described in following sections.

Recoveries

It is not always straightforward to obtain recoveries derived from a defaulted facility since they are not always included in the historical databases maintained by banks. Recoveries from a workout process can be cash recoveries or non-cash recoveries. Cash recoveries are relatively easy to track and incorporate into the LGD calculations. Conversely, non-cash recoveries, like repossessions or restructurings, are more difficult to track and are typically treated on a case-by-case basis.

In the case of non-cash recoveries resulting from repossessions, there are two possible treatments. The first treatment is to consider the recovery process complete at the time of the repossession. The second treatment considers the recovery process complete only when the repossessed good has been sold to a third party.

In general, it is convenient to consider the recovery process complete on the date of repossession for several reasons. First, management of the assets and the processes after the repossession date may change. Second, often there is a subsidiary involved in these processes, which makes it harder to identify future cash flows and costs related to the workout process. Third, the credit risk of a defaulted facility lasts until the repossession date and then it becomes another sort of risk (for example market risk). Finally, for certain types of facilities such as mortgages, the time lag between the dates of repossession and sale to a third party can be lengthy. If the workout was not considered complete until the asset was sold to a third party, there would be a substantial reduction in the number of elements with complete information in the reference data set.

However, use of the first method raises the issue of how to value the repossessed good. There are different approaches to converting the non-cash recovery associated with the repossession into an artificial cash recovery. One possibility is to apply a haircut coefficient to the book value of the repossessed good. The haircut coefficient can be calibrated using historical experience combined with current conditions.

Costs

Another element of the workout loss formula is the total discounted value of workout costs that, in theory, should include both direct and indirect costs of the workout process of the asset, and whose measurement can be a very difficult task. Direct costs are those associated with a particular asset (for example, a fee for an appraisal of collateral). Indirect costs are necessary to carry out the recovery process but are not associated with individual facilities (for instance, overhead associated with the office space for the workout department).

The assumptions about how indirect costs are allocated to individual assets will affect the final estimate of workout LGD. In practice, it is not easy to assign direct costs to defaulted facilities, and it is even more difficult to allocate indirect costs.[38] One possible approach to overcoming these difficulties is to identify the key recovery costs for each product, to model

[38] Both direct and indirect costs vary over time due to internal and external changes. For example, a decrease in a transfer tax on real estate implies a change in the direct costs of repossession for mortgages.

them using a sample of facilities for which the true costs are known and to allocate costs of recoveries out of the sample using the model.[39] Despite these difficulties, a bank that chooses to calculate realised LGD using the workout method must include all of the costs, including indirect costs.

Discount rates

To calculate the economic loss of a defaulted facility, using the observed recoveries and costs, it is necessary to discount them (back to the default date) using some discount rate. The impact of the chosen discount rate on the LGD estimates is particularly important in portfolios where the recovery period is long and has a low risk level. There is an important debate about how to choose the appropriate discount rate. Theoretically, the appropriate discount rate is the risk-appropriate rate. Practically, the difficulty is to infer the risk-appropriate rate from observable variables when there are no markets for these facilities. In the LGD literature and in practice, different rates have been proposed as suitable. These discount rates can be classified in two broad groups: historical and current rates.

Historical discount rates are fixed for each defaulted facility. All of the cash-flows associated with a defaulted facility are discounted using a rate determined at a particular date in the life of the defaulted facility. Alternatively, at the date of default a discount rate curve can be constructed with rates for each date over the expected life of the workout and the cash flows can be discounted using the curve. Typically, the discount rate is defined as either the contractual rate fixed at the origination date, the risk-free rate plus a spread at the default date for the average recovery period, a suitable rate for an asset of similar risk at the default date, or a zero-coupon yield plus a spread at the default date.

Current discount rates are fixed on each date in which LGD is being estimated. All the cash-flows associated with a defaulted facility are discounted by using a rate, or a curve, that is determined at the current date. These rates can be either average rates computed at the moment when the loss is being calculated (such as the average risk-free rate plus a spread during the last business cycle or the average rate of similar risky assets over the last business cycle) or spot rates plus a spread existing at that moment. The use of current rates allows the consideration of all available information and facilitates the comparison between LGD estimates from different portfolios.

The interest rate that the bank chooses for discounting the cash flows may have significant effects on the final estimates of workout LGDs. Moral and Oroz (2002) illustrate the effect of changes in the discount rate on the estimate of the workout LGD using data from a portfolio of Spanish mortgages. They compare three cases. First, they use an interest rate for each defaulted facility, linked to the date of default. Second, they use the same discount rate for every cash flow associated with every facility (ranging from 2% to 6%). Finally, they discount all defaulted facilities taking into account the existing yield curve at the date when the LGD calculation was performed.

The results illustrate the importance of the discount rate. In their example, an increase in the discount rate of one percentage point (in absolute terms, i.e., from 5% to 6%) implies an increase in the LGD estimate of around 8%. Furthermore, their results show that the variability of LGD resulting from using current rates is significant. With an observation period of 900 days, they compute the 900 LGD estimates with the same reference data set and

[39] An example of this methodology can be found in Moral and García (2002) in their empirical work using a data set of residential mortgages in Spain.

using different rates associated with the existing yield curve in each day of the observation period. They observe that the maximum percentage difference is around 20%. In their example, the use of both historical rates and a 5% rate produce conservative LGD estimates compared to the method based on spot rates during the observation period. This result is explained by the low levels of interest rates during the observation period. The authors also acknowledge that, in general, it is difficult to estimate the appropriate risk rate for discounting defaulted facilities, but for this case of residential mortgages with a very low historical volatility of losses, it is reasonable to take a rate close to the risk-free rate. Current practices, however, are very diverse.[40]

Validation of LGD

The LGD validation process involves the examination of all the elements that are needed to produce LGD estimates. This includes all of the assumptions made to construct a reference data set, calculate realised LGD, and generate LGD estimates from the reference data set. Validation also requires verification that the minimum regulatory requirements are met. Figure 10 below presents an example of the validation process. The same type of process can be used to verify that LGD estimates used for internal capital purposes are appropriate as well. At each step in the figure, both the assumptions made and the calculation must be validated.

Consider a grade that contains facilities with the same or similar type of collateral, product, industry, purpose, etc.[41] Given a facility of this grade, for which an LGD has to be estimated, the first step is to obtain a reference data set (RDS) of defaulted facilities.[42] This RDS must include seven years of data for corporate but for retail five years are enough (ideally a complete economic cycle),[43] use a definition of default consistent with the one used to estimate PDs (ideally the same definition),[44] and include internal and/or external facilities similar to the given facility.[45] The reference data set should also include data on the relevant drivers for the loss estimates.

---

[40] In some countries, like Spain, some banks use constant interest rates, commonly 5%, although other banks use different rates for each defaulted facility related to the date of default. In other countries, like the United States, the fixed rate that banks use may be much higher.

[41] See BCBS (2004), paragraph 399.

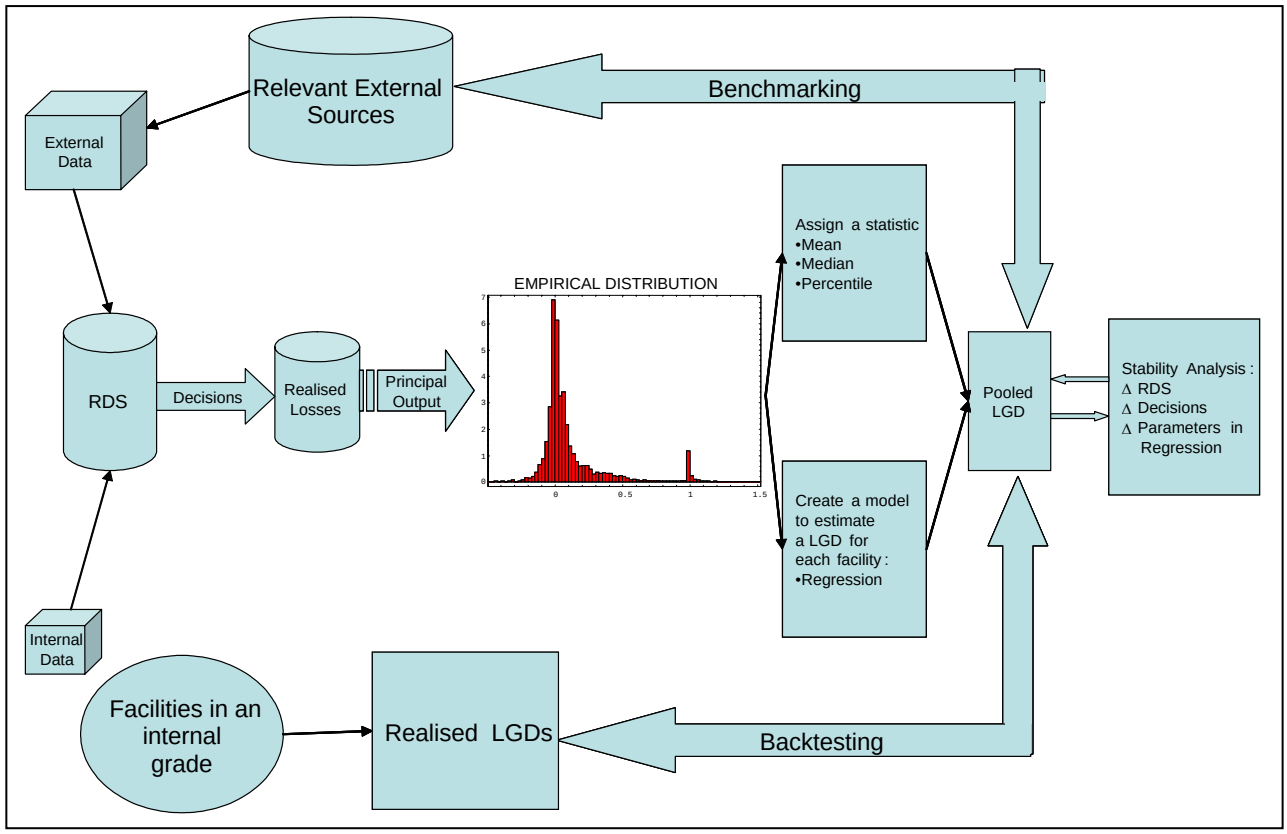
[42] It is assumed that there are enough data in the RDS.

[43] See BCBS (2004), paragraphs 472 and 473.

[44] See BCBS (2004), paragraph 456.

[45] See BCBS (2004), paragraphs 448 and 450.

Figure 10. An example of the validation process.



In practice the main validation questions are:

- Are there selection biases in the RDS?
- Are "bad years" included in the time frame of the RDS?[46]
- How has the similarity (between the elements in the RDS and the elements in the internal LGD grade) been tested?
- What tests have been performed in order to prove the consistency between the definitions of default used in this RDS and the one used to estimate PDs?
- How to treat portfolios that exhibit cyclical variability in realised LGDs?

The revised Framework requires the estimate of LGD to reflect economic downturn conditions where necessary.[47] This LGD cannot be less than the long-run default-weighted average which must be based on the average economic loss of all observed defaults within the data source.[48]

[46] "Bad years" are years in which defaults were relatively frequent as well as years with high realised LGDs.

[47] See BCBS (2004), paragraph 468.

[48] The requirement for the use of a default-weighted LGD is the result of the existence of "bad years" in the RDS.

After constructing the RDS, the second step is to assign an economic loss to each defaulted facility included in the RDS.[49] In the case of the workout LGD, this stage involves assumptions (such as the discount rate, allocation of direct and indirect costs, treatment of non-cash recoveries and incomplete recoveries), as we have shown in the previous sections. For incomplete recoveries the bank must have an "expected loss […] given current economic circumstances and facility status"[50] that can be assigned to these facilities.

On the other hand, the use of the market LGD, if possible, is straightforward because there are neither allocations of costs, nor problems with explicit rates of discount. The primary relevant problems are the liquidity of the market and the comparability of the instruments to the portfolio of the bank. Frequently, the more similar the instruments are to the portfolio of the bank the less liquid is the market and, thus, the information from the market becomes less reliable.

At the end of this stage an empirical distribution of realised LGDs is obtained. An analysis of the empirical distribution can help detect problems related to data outliers, changes in segmentation, and temporal homogeneity of the facilities included in the RDS.

After that, the simplest method to assign an LGD estimate to a particular facility (and all the elements in the considered internal grade), is to choose a statistic of the empirical distribution. If the realised LGDs exhibit cyclicality, banks will need to incorporate it into their estimate of LGD. One example is to use an average of loss severities observed during periods of high credit losses.[51] Another example is to use an appropriate percentile of the distribution instead of the mean as a more conservative LGD estimate.

If the RDS is large enough, empirical percentiles can be used for building a confidence interval for the LGD estimate. In other cases, a more complex method must be used, for example using bootstrapping techniques or assuming a parametric distribution for the LGD estimate.

Finally, a more sophisticated procedure consists of a model that refines the LGD estimate taking into account other variables existing in the RDS.[52] For example, if there are quantitative variables (such as interest rates at the date of default, loan to value ratios, GDP at the moment of default, etc) correlated with the realised LGDs, a regression model can be estimated. Regression models are particularly susceptible to over-fitting and thus it is particularly important to perform both "out of time" and "out of sample" tests in order to assess their true predictive power. This type of model must be carefully constructed to produce the "long-run default-weighted average LGD" required under the revised Framework rather than "point in time" LGD estimates.[53]

In any case, it is necessary to perform:

- **Stability analysis**: Analyse how changes in the RDS and changes in the assumptions made for determining the losses and/or parameters of the model impact the LGD estimate. It is especially important to analyse the volatility of the

[49] See BCBS (2004), paragraphs 460 and 468.

[50] See BCBS (2004), paragraph 471.

[51] See BCBS (2004), paragraph 468.

[52] As noted above, the RDS should include data on the relevant drivers of loss.

[53] See BCBS (2004), paragraph 468.

LGD estimates when the RDS timeframe changes. According to the revised Framework a "bank must take into account the potential for the LGD of the facility to be higher than the default-weighted average during period when credit losses are substantially higher than average".[54]

- **Comparisons between the LGD estimates and "relevant external data sources"**.[55] The main problems that arise when using external data sources are related to: different default definition; biases in the external data sample; and different measures of losses (market losses, workout losses, implied market losses, etc).
- **Comparisons between realised LGD of new defaulted facilities and their LGD estimate.**[56] One must be careful not to directly compare the realised losses which are "point in time" to those produced from a model intended to provide a long-run average.

Conclusions

LGD is an important element of the advanced internal ratings-based (AIRB) capital requirements. Under the AIRB, minimum capital requirements are highly sensitive to the LGD that is reported by an institution. Therefore, developing LGD estimation methods and establishing validation procedures for the LGD estimates are crucial for the AIRB framework.

Consistency between the definition of default used in the reference data set (RDS) associated with the LGD and the definition of default used when estimating PDs is particularly important. In this section, objective methods for computing LGDs were classified into explicit and implicit methods. Explicit methods include explicit market LGD and workout LGD, which use information exclusively from defaulted facilities. Implicit methods include implied historical LGD and implied market LGD, and use information from non-defaulted facilities. There is some debate as to whether revised Framework excludes any of the four methods of LGD, especially those that use information from non-defaulted facilities. Therefore, the remainder of the section focused on explicit methods, with particular attention to the workout method of calculating LGD.

There are several general issues involved in the definition of the loss associated with a defaulted facility and the estimation of LGD. The decision whether to "censor" the data (forcing the non-negativity of loss) or to use original data and the requirements on the RDS used to estimate the risk parameters are both important. When it comes to focusing on the details of constructing a workout LGD, the most important decisions include how to measure recoveries and to allocate costs; and how to choose the appropriate discount rates.

The survey of empirical research on LGD (see the appendix) reveals that most studies use data from U.S. corporate bonds rather than information from loan portfolios. The studies provide some indication of the range of LGD estimates and the types of observed empirical distributions from different data sources. Furthermore, the survey of empirical studies provides information on the most common drivers of LGD.

---

[54] See BCBS (2004), paragraph 468.

[55] See BCBS (2004), paragraph 502.

[56] See BCBS (2004), paragraph 501.

The development of LGD estimation methods is still in its early stages. Most banks do not have sophisticated models of LGD that use internal data sets and take into account risk drivers. This section demonstrates that the process for estimating workout LGDs is complex even when banks use unsophisticated models. At the current stage of development in the estimation of LGDs, validation should focus on the examination of the elements that are needed to produce LGD estimates for facilities, to verify that the minimum requirements are met, and to assess the sufficiency of other requirements imposed by banks themselves.

Areas for further research

The main open issues in the area of LGD validation are:

- *Decisions made in order to calculate realised LGDs.* This section demonstrates the importance of several decisions that have to be made when estimating LGDs. Among others, these include whether to censor the realised losses (imposing the non-negativity of the losses), choosing the interest rates used in the discounting, and deciding when the recovery process is over.
- *Further research needed on estimation methods.* This section has highlighted several problems inherent in measuring LGD even when using relatively simple estimation methods such as averages. Further research should be conducted on more complicated estimation methods, such as regression techniques including several risk drivers.
- *Benchmarking and backtesting*. This section demonstrates that benchmarking exercises on LGD are especially difficult due to the high number of factors that affect these estimates. More research is needed to establish comparison procedures between different LGD estimates. Furthermore, the report is not explicit on how to carry out backtesting (i.e. comparisons between realised LGD and LGD estimates). For instance, a direct comparison between long-run default-weighted LGD estimates and “point-in-time” LGD observations in a specific year may not be meaningful. Further research in this field is needed.

Appendix: Summary of empirical studies of Loss Given Default

Survey of empirical research

Given the lack of data, validation of LGD will prove to be difficult. Surveying a number of empirical studies and evaluating the common findings provides some initial guidelines that may be helpful in evaluating a bank's LGD estimation.

To that end, Table 11 summarises a survey of empirical research on LGD. This research provides useful information about LGD based on publicly available data. A major limitation of most of these studies is that publicly available data on LGD are typically for U.S. corporate bonds. Intuitively, one would expect that the LGD on loans would be lower than that on bonds since loans are typically senior to bonds. In addition, problem loans are more actively managed than problem bonds. Troubled loans are often restructured in such a way that the borrower gets relief by putting up more collateral, thus lowering the LGD.

There is some information on LGD for loans, but with few exceptions the data are for syndicated loans in the U.S. Since syndicated loans are large corporate loans in which banks and other institutions participate, the results from this market are not applicable to many segments in a loan portfolio such as retail credits.

Despite these weaknesses, it is important to review the empirical results to have some intuition for the characteristics of LGD. There are three aspects of the studies summarised in Table 11 that are important for validation. First, many of the studies report descriptive statistics for LGD in their sample. The descriptive statistics provide a check on whether the statistics that a bank generates internally are reasonable. Second, some of the studies evaluate the empirical distributions of LGD in their sample. This is an important component for validation since the information contained in simple descriptive statistics depends on the distribution of the LGD. For example, if the distribution is normal then the average and standard deviation are informative. However, if the distribution is highly skewed the average and standard deviation will be misleading. Third, the studies typically analysed the main determinants of the LGD. This will help provide some guidance for evaluating empirical models of LGD. If particular drivers have been statistically significant in these studies, then internal bank models should at least investigate whether the same drivers may be help explain LGD in their portfolio.

The first aspect of the LGD results is the descriptive statistics from these studies, as summarised in Table 10. The average LGD of bonds reported in these studies ranges from 58% to 78%. An average LGD across all priority levels and collateral classes is misleading. LGD depends on the priority of the credit in bankruptcy and whether the credit is secured or unsecured. Some studies broke the LGD data into secured and unsecured and senior and junior debt. LGD for senior secured bonds ranged from 42% to 47%; for senior subordinated bonds LGD ranged from 58% to 66% and for junior subordinated bonds the LGD ranged from 61% to 69%.

The descriptive statistics on loans were segregated by secured and unsecured and by business line. The average LGD for senior secured loans in these studies ranged from 13% to 38%. Two studies that reported average LGD for senior unsecured loans of 21% and 48%. The reported average LGD for commercial loans ranged from 31% to 40%. One study reported an average LGD for consumer loans of 27%.

Table 10

Summary of descriptive statistics for LGD from surveyed empirical studies

Study	Average	Median (if available)
All bonds		
Acharya et al (2004)	58.04%	62.00%
Altman et al (1996)	58.30%	
Hamilton et al (2003)	62.80%	70.00%
Altman et al (2001)	64.15%	59.95%
O'Shea et al (2001)	78.00%	
Senior secured bonds		
Altman et al (1996)	42.11%	
Hu and Perraudin (2002)	47.00%	
Altman et al (2001)	47.03%	42.58%
Senior subordinated bonds		
Roche et al (1998)	58.00%	
Altman et al (1996)	65.62%	
Junior subordinated bonds		
Roche et al (1998)	61.00%	
Altman et al (1996)	68.66%	
Senior secured loans		
Carty et al (1998)	13.00%	0.00%
Roche et al (1998)	18.00%	
Carty and Lieberman (1996)	21.00%	8.00%
Carty and Lieberman (1996)	29.00%	23.00%
Gupton et al (2000)	30.50%	
O'Shea et al (2001)	37.00%	17.00%
Hamilton et al (2003)	38.40%	33.00%
Senior unsecured loans		
Carty et al (1998)	21.00%	10.00%
Gupton et al (2000)	47.90%	
Commercial loans		
Eales and Bosworth (1998)	31.00%	21.00%
Hurt and Felsovalyi (1998)	31.80%	
Asarnow and Edwards (1995)	34.79%	21.00%
Araten (2004)	39.80%	
Consumer Loans		
Eales and Bosworth (1998)	27.00%	20.00%

The distribution of LGD is the second important aspect of these studies. If the distribution is not normal, then the average LGD is misleading. Several studies analysed the distribution of the LGD and found that the distributions were highly dispersed. Most of the studies found the distribution of LGD was unimodal and skewed towards low LGD. A few studies found that the beta distribution fit the data better than a normal distribution. Two of the studies of the LGD on loans found that the distribution of LGD was bimodal. None of these studies concluded the distribution of LGD is approximated by a normal distribution. Therefore, a simple average LGD is most likely misleading.

Many of the empirical studies went beyond simple averages attempting to isolate significant drivers of LGD. The most important of the determinants, as shown above, was the security and priority of the claims. Another important determinant is the credit cycle. If the economy is in a period of high defaults, LGD is higher than in periods of low defaults. Once this effect has been taken into account, more general macroeconomic factors typically do not matter.

Another important determinant is industry and the liquidity of collateral. Credits that are backed by liquid collateral such as cash and accounts receivable experienced lower LGD than credits that were backed by less liquid collateral such as property, plant and equipment. Industries that are more likely to use less liquid collateral will have higher LGD.

The size of the borrower does not seem to affect LGD. The amount of other outstanding obligations does tend to increase LGD, especially for unsecured loans. Furthermore, if the creditor has external sources of support, LGD is lower. The size of the loan has an ambiguous effect on LGD; some studies found a positive effect, some studies a negative effect and some no effect.

This review of empirical studies on LGD is not meant to provide hard and fast benchmarks for validation. Nor is the list of drivers of LGD in any way a comprehensive list of potential drivers of LGD. Rather, this review is meant only to serve as a starting point for validation.

Table 11

Summary of empirical studies of Loss Given Default

	Sample characteristics		
Bibliographic citation	**Period**	**Facilities included**	**Methodology and findings**
Acharya, Viral V., Sreedhar T. Bharath and Anand Srinivasan (2004), Understanding the Recovery Rates on Defaulted Securities. Unpublished manuscript, April.	1982–1999	• Bank loans and corporate bonds. • Standard and Poor's Credit Pro database.	• Market LGD[a]. • Average LGD bonds 58.04%, median 62%, standard deviation 25.34%. • Average LGD on bank loans is 22 percentage points lower than on senior secured bonds. • Substantial variation in LGD over time. • Industry distress increases LGD by 10 percentage points relative to a healthy industry. • Once account for industry distress, macroeconomic conditions do not affect LGD.

Altman, Edward I., Brooks Brady, Andrea Resti, and Andrea Sironi (2003), The Link between Default and Recovery Rates: Implications for Credit Risk Models and Pro-cyclicality. NYU Stern School Solomon Working Paper, March.	1982–2001	• 1,300 defaulted corporate bonds.	• Market LGD[a]. • Measured just after default. • Weighted average recovery rates for all corporate bond defaults in the U.S. (weights are market value of the defaulting debt issue). • Average LGD 1982–2001of all bonds 62.8%. • Annual averages. • Explain variation in annual recovery rates. • OLS regression techniques. • Negative correlation between weighted average annual recoveries and annual default rates across all seniority and collateral levels.
Altman, Edward I. and Vellore M. Kishore (1996), Almost Everything You Wanted to Know About Recoveries on Defaulted Bonds. *Financial Analysts Journal* 52 (6), p.57–64.	1978–1995	• 696 defaulted bond issues. • By seniority. • By industrial class.	• Market LGD[a]. • Overall average LGD is 58.3%. • Average LGD for senior unsecured debt 42%; senior unsecured 52%; senior subordinate 66% and junior subordinate 69%. • Statistically different LGD by industry class even when adjusted for seniority. • Initial rating of investment grade versus junk-bond category has no effect on recovery when adjusted for seniority. • Time between origination and default has no effect on recovery. • No statistical relationship between size and recovery.
Altman, Edward I., Andrea Resti and Andrea Sironi (2001), Analyzing and Explaining Default Recovery Rates. A Report Submitted to the International Swaps and Derivatives Association, December.	1982–2000	• Defaults on Publicly traded corporate bonds. • 1,000 bonds.	• Market LGD[a]. • Measure price on defaulted bonds as close to default date as possible. • Weighted average annual recovery rates (weighted by market value of debt issue).

			• Weighted average LGD on all seniorities is 64.15%; median is 59.95%; based on 1,289 observations. • Weighted average LGD on senior secured debt is 47.03%; median is 42.58%; based on 134 observations. • Predict annual recovery rates rather than recovery rates on individual facilities using OLS regression. • Explanatory variables: o Default rates. o Dollar amount of the bond outstanding at time of default. o Dollar amount of high yield bonds outstanding in the year. o Defaulted Bond Index (NYU-Salomon Center) which is a market weighted indicator of the average performance of defaulted publicly traded bonds. • Annual GDP growth. • Change in GDP growth from previous year. • Annual return on S&P 500. • Change in annual return on S&P 500. • The multivariate model for 1987–2000 had the following significant explanatory variables and effect on LGD: the bond default rate (+) and its change (+), the amount of high yield bonds outstanding (+) and the Defaulted Debt Index (-). • The multivariate model for 1982–2000: all of the above except the Defaulted Debt Index. • Macroeconomic factors were not significant in a multivariate context.

Araten, Michel, Michael Jacobs Jr., and Peeyush Varshney (2004), Measuring LGD on Commercial Loans: An 18-Year Internal Study. *RMA Journal* 86 (8), p. 96–103.	1982–1999	• 3,761 large corporate loans originated by JP Morgan.	• Workout LGD[a]. • Mean LGD 39.8%, standard deviation 35.4%. • Range from -10% to 173%. • Mode LGD approximately 5%. • Broke down LGD by type of collateral and found LGD lowest for loans collateralised by accounts receivable. • Found a positive correlation between LGD and the default rate using annual data from 1986–1999.
Asarnow, Elliot and David Edwards (1995), Measuring Loss on Defaulted Bank Loans: A 24-Year Study. *Journal of Commercial Bank Lending* 77 (7), p. 11–23.	1970–1993	• Citibank loans. • U.S. borrowers. • 831 C&I loans (senior, secured and unsecured). • 89 structured secured loans. • Major corporate and upper middle market loans.	• Workout LGD[b]. • LGD calculated for entire workout period. • Define defaulted C&I loans as any loan ever classified as doubtful or non-accrual. • Resolution occurs when the outstanding balance is zero. • Outstanding balance at default is widely distributed from $1 million to $190 million. • LGD expressed as a percent of outstanding balance at default. • Use the contractual lending rate as the discount rate. • Didn't have lending rates on the individual loans so used the annual average of the interest rate on domestic C&I loans as reported in Citibank financial statements. • Simple average LGD for C&I loans 34.79% and for structured loans is 12.75%. • Median LGD for C&I loans 21%. • LGD for C&I loans relatively stable over time. • LGD for C&I loans has a bimodal distribution. • No statistical difference in LGD by size of loan at default and recovery.

Carty, Lea V., David T. Hamilton, Sean C. Keenan, Adam Moss, Michael Mulvaney, Tom Marshella, and M.G. Subhas (1998), Bankrupt Bank Loan Recoveries. Special Comments. Moody's Investors Service, June.	1986–1997	• 200 bank loans. • Industrial, retail and consumer products are 58% of the sample. • Includes 136 senior secured bank loans with details on collateral. • 98 loans with market prices from 1989–1996.	• Market LGD[a] (Workout LGD[b] used for new or amended debt instruments). • Market LGDs are from Bloomberg, IDC, Citibank, Goldman Sachs, BDS securities, Lehman Brothers, Merrill Lynch, Loan Pricing Corporation, borrower's financial statements and libraries of domestic stock exchanges. • Cash flows for the workout LGD came from bankruptcy documents. • Discount rate for the workout LGD is the contractual lending rate (or, if unavailable, an estimate of the loan rate). • Average LGD for senior secured bank loans is 13%; median LGD 0%; standard deviation 23%. • Average LGD for senior unsecured bank loans is 21%; median 10%; standard deviation 27%. • Dispersion in LGD is high. • Distribution of LGD is skewed towards low LGD. • Lowest LGD was less than 0 (a gain) and the highest LGD was 92.6%. • Average LGD was lower for pre-packaged Chapter 11 rather than Chapter 11 bankruptcies. • Broke down the 136 bank loans by collateral type and found the most liquid collateral (accounts receivable, cash and inventory) produce lower average LGD of 10.23% (versus 14.57% for property, plant and equipment and 26.45% for stock of subsidiaries). • Market LGD average for 98 loans from 1989–1996 is 30%; median is 25%; standard deviation 21%.

Carty, Lea V. and Lieberman (1996), Defaulted Bank Loan Recoveries. Special Comment. Moody's Investors Services, November.	1989–1996	• U.S. traded loans (most directly comparable to syndicated loans). • 58 borrowers, one loan per borrower.	• Market LGD[a]. • Senior secured syndicated bank loans average LGD is 29% and median is 23%. • Wide range of recovery rates on loans and the distribution is skewed towards low LGD. • Definition of default is default on publicly held bonds; therefore each loan must have a corresponding publicly traded bond. • Use bid prices on the loans to calculate recovery; use an average of 5 dealer bids. • Take a price that is between 2 and 8 weeks of default.
	1990–1996 (except 1 loan prior to 1990)	• Loan Pricing Corporation's Loan Loss Database. • 229 senior secured loans that were non-accrual and had completed their workout. • Small to mid-size borrowers.	• Workout LGD[b]. • Calculated LGD by summing the present value of interest payments, principal payments and post-default draw downs on the loan. • Contractual lending rate not available so use Loan Pricing Corporations market-based model of the spread over LIBOR paid by borrowers as the discount rate. • Distribution more skewed towards low LGD than the traded loans above. • Average LGD 21%; median 8%. • LGD rates on loan secured by current assets are lower than those secured by property, plant and equipment. • Average LGD does not vary with the asset size of the borrower.
Ciochetti, Brian A. (1997), Loss Characteristics of Commercial Mortgage Foreclosures. *Real Estate Finance* 14 (1) 53–69.	1986–1995	• 2,013 commercial mortgages issued by insurance companies.	• Average LGD over whole sample is 30.60%. • Annual average LGD ranged from 19.6% in 1987 to 38.2% in 1996.
Eales, Robert and Edmund Bosworth, (1998), Severity of Loss in the Event of Default in Small Business and Large	1992–1995	• Westpac Banking Corp. (Australia).	• Workout LGD[b]. • All loans internally worked out at the bank.

Consumer Loans. *Journal of Lending & Credit Risk Management* 80 (9), p. 58–65.		• Small business loans. • Larger consumer loans. • 94% secured loans. • 5,782 customers that defaulted on one or more loan. • Loans with less than US$ 6.7 million outstanding. • All fully worked out. • Customer level, not loan level.	• Discount rate is the cost of equity capital estimated from a Capital Asset Pricing Model (CAPM). • If the original loan rate had been used the reported LGDs would have been about 10% lower (e.g. multiplied by 0.9). • Cash flows include: o Further loan disbursements (-). o Internal loan workout costs (-). o External workout costs (-). o Principal repayments (+). o Interest payments (+). o Proceeds of the realisation of security (+). o Recoveries made after account closure (+). • Data covers a period of healthy economy and stable inflation. • LGD is sensitive to the size of the customer's debt with the bank; on average smaller loans have higher LGD. • Average LGD for business loans 31%; median LGD for business loans 22%. • Average LGD for consumer loans 27%; median 20%. • Distribution of LGD for the whole sample has a long tail and is bimodal. • Distribution of LGD for secured loans is unimodal and skewed towards low LGD. • Distribution of LGD for unsecured loans is bimodal. • LGD can be greater than 100%. • LGD cannot be zero because workout expenses are included. • Average LGD for secured loans is lower than for unsecured loans.

Frye, Jon (2003), LGD in High Default Years. Federal Reserve Bank of Chicago, unpublished manuscript.	1983–2001	• Moody's Default Risk Service Database. • U.S. non-financial issuers. • 859 bonds and loans (88 senior secured loans).	• Market LGD[a]. • Separate into "bad years" (the high default years 1990, 1991, 2000, 2001) and "good years" (low default years 1983–1989 and 1992–1999). • Unsecured or subordinated seniorities tend to have the greatest LGDs in both the good and bad years. • For most debt types, LGD is higher in bad years than in good years.
Frye, Jon (2000), Depressing Recoveries. Federal Reserve Bank of Chicago Emerging Issues Series, October. Also see abridged version in *Risk* 13 (11), p. 108–111.	1983-1997	• Moody's Default Risk Service database. • U.S. corporate bonds, sovereign bonds. • Excludes bonds backed by a second entity. • Must have a default price. • Non-financial issuers.	• Market LGD[a]. • Recovery is the market price one month after default. • Default is usually a missed payment. • Assumes that the market valuation is accurate on average in a given year. • Moody's rating scale changed twice in 1982–2000 so only look at 1983–1997 where the rating scale was unchanged. • The first loan recovery in the data is in 1996 and there are only recoveries from 15 loans in total so eliminate the loans from the sample. • High default years of 1990 and 1991 exhibit lower recovery rates (higher LGD) than other years. • The distribution of LGD differs in high versus low default years. • In an economic downturn, bond recoveries might decline 20 to 25 basis points from their normal year average.

Gupton, Greg M., Daniel Gates and Lea V. Carty (2000), Bank Loan Loss Given Default. Special Comment. Moody's Investors Services, November.	1989–2000	• U.S. traded loans (most directly comparable to syndicated loans). • 121 defaults. • Includes 119 senior secured loans; 33 senior unsecured loans; 29 unspecified loans.	• Market LGD[a]. • Default price measured one month after default. • Define default on loan to occur when there is a default on public debt; therefore data includes only loans where the borrower also has public debt. • Wide distribution of recovery rates. • Average LGD for senior secured loans is 30.5% and for senior unsecured loans is 47.9%. • Distribution of LGD is skewed towards high LGD. • Bank loans had a lower LGD than the corresponding public debt. • 80 of the 121 firms have completed the bankruptcy process. • The presence of multiple loans increases the LGD for senior unsecured loans; has no effect on LGD for senior secured loans. • A better Moody's rating at default lead to a lower LGD. • No evidence that industries have different LGDs.

Gupton, Greg M. and Roger M. Stein (2002), LossCalc™: Moody's Model for Predicting Loss Given Default (LGD). Special Comment. Moody's Investors Services, February.	1981–2002	• 1,800 defaulted loans, bonds and preferred stock. • U.S. debt obligations only. • Both senior secured and senior unsecured loans. • Also includes Corporate mortgages and industrial revenue bonds. • Over 900 defaulted public and private firms. • Issue size is US$ 680,000 to US$ 2 billion; median size US$ 100 million.	• Market LGD[a]. • Default price one-month after default. • Beta distribution fits the recovery data better than a normal distribution. • There are a small number of LGD less than zero (gains). • LossCalc™ predicts immediate LGD and one-year horizon LGD. • Methodology is to map the beta distribution of the LGDs to a normal distribution and then perform OLS regression. • Historical averages of LGD by debt type (loan, bond, preferred stock) are an explanatory factor for facility level LGD. • Historical averages of LGD by seniority (secured, senior unsecured, subordinate etc) are an explanatory factor for LGD. • Except for financial firms or secured debt, firm leverage is an explanatory factor for LGD. • Moving average recoveries for 12 broad industries are an explanatory factor for LGD. • One-year PDs from RiskCalc are an explanatory factor for LGD. • Moody's Bankrupt Bond Index is an explanatory factor for LGD. • Average default rates for speculative grade bonds from 12-months prior to facility default are an explanatory factor only for immediate LGD. • Changes in the index of leading economic indicators are an explanatory factor for LGD.

Hamilton, David T., Praveen Varma, Sharon Ou and Richard Cantor (2003), Default and Recovery Rates of Corporate Bond Issuers: A Statistical Review of Moody's Ratings Performance 1920–2002. Special Comment, Moody's Investors Service.	1982–2002	• 2,678 bond and loan defaults. • Includes 310 senior secured bank loan defaults.	• Market LGD[a]. • Default price measures one month after default. • Distribution of recovery rates is a beta distribution skewed towards high recoveries (low LGD). • Average LGD for all bonds is 62.8%; median LGD for all bonds is 70%. • Average LGD for senior secured bank loans is 38.4%; median LGD for senior secured bank loans 33%. • LGD in all debt instruments increased in 2001 and 2002. • Average LGDs vary by industry. • LGD and default rates are positively correlated.

Hu, Yen-Ting and William Perraudin (2002), The Dependence of Recovery Rates and Defaults. CEPR Working Paper, February.	1971–2000	• Moody's Corporate Bond Default Database. • Long-term bond defaults (no bank loans). • 958 observations. • Includes 910 observations from U.S. issuers.	• Market LGD[a]. • Definition of default is a missing or delayed payment, failing for bankruptcy or distressed exchange where exchange is meant to help borrower avoid default. • Average LGD for all bonds that default in one quarter. • Distribution of LGD is unimodal and somewhat skewed. • Average LGD for senior secured bonds is 47%. • OLS regressions and inverse Gaussian regressions with recoveries as dependent variable. • Explanatory variables: o Industry dummies. o Domicile dummies (emerging markets, non-U.S. OECD and off-shore banking centre). o Seniority dummies. o External support dummy. • Transportation, banking and thrifts, sovereign and all others had higher LGD and utilities had lower LGD than the reference U.S. senior secured industrial bond. • Senior subordinated and subordinated bonds had lower LGD than senior secured. • External backing lowered the LGD.

Hurt, Lew and Akos Felsovalyi (1998), Measuring Loss on Latin American Defaulted Loans: A 27-Year Study of 27 Countries. Citibank/Portfolio Strategies.	1970–1996	• 1,149 defaults on C&I bank loans in 27 countries in Latin America. • Excludes sovereign loans. • Only loans larger than US$ 100,000.	• Workout LGD[b]. • Includes only loans that have been fully worked out. • Default is defined as loans that are doubtful or non-accrual. • Average length of time to resolve the loans was 19.7 months. • Average LGD on Latin American loans is 31.8%; standard deviation of 28.8%. • Nine of the 1,149 loans had a LGD over 100%. • LGD had a skewed distribution; large number of loans with small LGD (0% to 15%). • Larger loans had higher LGD. • Large loans in Latin America often involve economic groups that are frequently family-owned and recovery is more difficult; smaller loans are typically secured by trade receivables. • Average LGD is relatively stable across years. • Sovereign events such as large devaluation or default on external debt did not effect the average LGD. • Average LGD on Latin American loans is 31.8%; standard deviation of 28.8%.

Moral, Gregorio and Raúl García-Baena (2002), LGD Estimates in a Mortgage Portfolio. Banco de España, Estabilidad Financiera, Nº3.	1996–2001	• Spanish mortgages. • 1,532 defaulted mortgages applying a 90-day past due definition of default. • 687 mortgages taken to court.	• Workout LGD[b]. • LGD cannot be negative. • Separates the recovery period into two different parts. The first part goes from the date of default to the final recovery (final cash-flow or repossession). The second part goes from the repossession date to the sale date of the repossessed property. • Haircut coefficient applied to non cash recoveries 90%. • Cost allocation using a simple model (legal and repossession costs). • Use the sample mean as estimate for LGD under the mortgages taken to court definition of default. LGD under the 90-day past due definition of default is obtained by multiplying by an estimate of the conditional probability. • Analyses the evolution of the estimated conditional probability used to transform the LGD estimate. • Uses a fixed discount rate of 5% to discount the cash-flows. • Average LGD under the 90-day past due definition of default is 12.65%, Median LGD=11.55%. • LGD under the mortgages going to court definition of default is 28.20%, Median=25.75%. • Confidence interval for LGD using the 90-day past due definition =[9.7%, 15.7%]. • Empirical distributions of LGD are highly skewed. • Obtains the LGD estimates as a function of some parameters (legal costs, repossession costs, haircut coefficient for cash-flows associated to repossessions). • Studies the sensitivity of LGD estimates under sample changes using bootstrapping and jackknife methodologies. • Describes a possible methodology to utilise the information provided by workouts that have not been completed.

Moral, Gregorio and María Oroz (2002), Interest Rates and LGD Estimates. Unpublished manuscript.	1993–2000	• Spanish residential mortgages originated in a specialised bank. • 3,887 defaulted mortgages applying a 90-days past due definition. • 464 mortgages taken to court.	• Workout LGD[b]. • LGD cannot be negative. • Separates the recovery period into two different parts. The first part goes from the data of default to the final recovery (final cash-flow or repossession). The second part goes from the repossession date to the sale date of the repossessed property. • Estimates a haircut coefficient of 92.2% to apply to non-cash recoveries. • Uses the sample mean as an estimate of LGD under two different definitions of default: 90-days past due and mortgages taken to court. • Converts the LGD under one definition to LGD under the other definition using a conditional probability measure. • Uses a fixed interest rate of 5% to discount cash flows. • Average LGD using the 90-day past due definition of default is 1.17%, median is 0. • Average LGD using the taken to court definition of default is 14.4%. • Evaluate the impact of different discount rate assumptions on the LGD estimates. • Fixed discount rate for all loans. • Fixed discount rate for each loan based on historical interest rate on date of default. • Discount rate using the yield curve from the date of computation. • Conclude that the LGD estimate is sensitive to changes in the discount rate. • Both historical rates and a 5% fixed rate provide conservative estimates compared with the method using the current yield curve.

			• Conclude that when about 90% of the 90-day past due default have zero losses the risk-free interest rate can be used for the discount rate.
O'Shea, Steven, Sharon Bonelli and Robert Grossman (2001), Bank Loan and Bond Recovery Study: 1997–2000. Loan Products Special Report. Fitch Structured Finance, March 19.	1997–2000	• 35 companies with both senior secured bank loans and subordinated debt. • 35 loans. • 18 industry sectors. • Bid-side prices for loans and bonds obtained from Loan Pricing Corporation, Bloomberg, Interactive Data Corp, Advantage Data Corp., broker/dealer quote sheets, individual commercial banks and institutional investors.	• Market LGD[a]. • Measured prices one-month following bankruptcy. • 18 of the 35 loans had LGD of 0–19%. • Distribution of bank loan LGDs skewed towards low LGD; distribution of bond LGDs skewed towards high LGD. • Average LGD for loans 37%; median LGD for loans 17%. • Average LGD for all classes of bonds 88%. • Average LGD for industries were lower for industries with hard or liquid assets or strong franchise value. • Companies with less bank debt had lower LGDs than companies with more bank debt. • Average number of days in bankruptcy was 326. • Average LGD was higher for firms that were in bankruptcy longer.

Roche, James, William Brennan, Derek McGirt and Mariarosa Verde (1998), Bank Loan Ratings in Bank Loans: Secondary Market and Portfolio Management, edited by Frank J. Fabozzi. Frank J. Fabozzi Associates.	1991–1997	• 60 widely syndicated secured bank loans (aggregating $25 billion). • Source: Loan Pricing Corporation. • Retail sector was 18 of the 60 companies.	• Market LGD[a]. • Measured market LGD using the last price listed for each security. • The authors believe that these prices are a better indication of ultimate realisation value than the price one month subsequent to bankruptcy. • Loss rate on secured bank loans 18%, senior subordinated debt 58%; subordinated debt 61%. • LGD correlated to industry factors. • Average period that a loan was in distress was 19 months. • Size of the borrower (measured in sales) has no significant influence on the LGD. • Negative correlation between LGD and stock prices (measured by the Dow Jones Industrial Average).
Van de Castle, K. and D. Keisman (1999), Recovering Your Money: Insights into Losses from Defaults. Standard and Poor's Credit Week, June 16.	1987–1997	• 829 debt instruments from Standard and Poor's Credit Loss Database.	• Market LGD[a]. • Average LGD for bank loans 15.5% with a standard deviation of 24.9%. • Results from regression analysis indicate the type of debt, collateral type and percent of debt below the tranches are determinants of LGD.

(a) Market LGD is calculated from observed market prices on defaulted or marketable loans soon after the actual default event. (Schuermann (2003), p. 6).

(b) Workout LGD is calculated from the discounted cash flows resulting from the workout and the exposure at default. The cash flows include collections and workout expenses and are properly discounted. (Schuermann (2003), p. 6.)

V. Exposure at default validation[57]

Jaap W B Bos

Under the Advanced Internal Ratings Based approach, banks are allowed to use their own internal estimates of expected exposure at default (EAD) for each facility. Conceptually, EAD consists of two parts, the amount currently drawn and an estimate of future draw downs of available but untapped credit. The estimates of potential future draw downs are known as credit conversion factors (CCFs). Since the CCF is the only random or unknown portion of EAD, estimating EAD amounts to estimating these CCFs. Compared to PDs and LGDs, relatively little is known about EAD. Therefore, the following section concentrates more on issues that effect the estimation of EAD than on validation methods. At this stage, it seems that validation by a qualitative assessment of the bank's estimation process may be more meaningful than the use of quantitative methods.

When estimating EAD it is important to recognise that EAD, even more so than PD and LGD depends on how the relationship between bank and client evolves in adverse circumstances, when the client may decide to draw unused commitments. What this means is that the realised EAD is to a large extent influenced by earlier decisions and commitments by the institution.

EAD is generally believed to depend on the type of the loan and the type of the borrower. For example EADs for credit cards are likely to be different from EADs for corporate credit. At present, literature on these issues as well as data sources are scarce.

A bank can take the following characteristics of a credit into account in its EAD estimation.

Fixed or floating rate

- EADs of floating rate credits are more difficult to predict and their volatility depends also on the volatility of the underlying benchmark rate (e.g. Libor).

Revolving or non-revolving

- Revolving credits may have normal utilisation rates that are different from normal utilisation for non-revolving credits.

Covenants

- Empirical findings indicate that the draw-down of a credit line at the time of default tends to decrease with the quality of the borrower's credit rating at the time the commitment was granted. The argument behind this observation is that a bank is more likely to require covenants for borrowers with lower credit quality which restrict future draw-downs in cases where the credit quality has declined. Some covenants can lower EAD, but may come at the cost of higher PDs.

[57] This section makes use of the analysis in Asarnow and Marker (1995) and especially Araten and Jacobs (2001).

Restructuring

- If an obligor experiences payment difficulties or is in default, credit restructuring may result in stricter covenants and make the obligor less likely to use the unused portion of a commitment.

Obligor-specific characteristics

- Taking into account obligor-specific characteristics, such as the history with the institution and customer profitability, is likely to improve EAD estimations. However, the scarcity of default data on high-quality borrowers is one reason why there is no clear-cut evidence as to how these characteristics influence EAD. For example, there is some evidence that EAD decreases with credit quality, though this may be in part the result of stricter limits and covenants for non-investment grade borrowers.[58]

In addition, EAD may depend on a number of other factors. First, the longer the time to maturity, the larger is the probability that the credit quality will decrease, as the obligor has both an increased opportunity and perhaps an increased need to draw down the remaining credit line. Second, the more the borrower has access to alternative sources and forms of credit, the lower the EAD is expected to be. Third, the current usage of the commitment is also likely to affect the EAD, though it is as yet uncertain what this effect will be. Banks should investigate the explanatory power of any additional factors that are perceived to influence EAD. Inclusion of such variables, if data are available and reliable, may help isolate and reduce measurement errors in EAD, and explain EAD volatility.

In summary, much remains unknown with respect to possible measurement errors of EAD. The inclusion of additional explanatory variables may be of help, assuming these are available and reliable. In doing so, most effort needs to be guided towards explaining EAD volatility, as this may turn out to be the biggest obstacle to reliable EAD prediction.

[58] See Araten and Jacobs (2001).

VI. Benchmarking

Vichett Oung

Definition

In the context of validation, benchmarking can be defined as a comparison of internal ratings and estimates with externally observable (whether public or non public) information. For IRB systems, an example of public benchmarks includes the ratings given by ratings agencies such as S&P or Moody's. In the case of the major rating agencies, detailed information about the firms they rate usually available, which make testing and analysis of their rating systems more feasible (though many would still view them as "black boxes"). Other examples of public but harder-to-analyse benchmarks include Moody's KMV EDFs, which are usually well disclosed but difficult to test as the technology used is proprietary. Non-public benchmarks are typically supervisory benchmarks that are usually not disclosed.

The most straightforward benchmarking is usually carried out for PD estimates because they are obligor-specific and therefore it is relatively easy to define a set of borrowers which are benchmarked. Moreover, PDs are expressed on a universally understood zero-to-one interval scale. It is more difficult to use benchmarking for EAD and LGD estimates because they are exposure-specific, though such practice is growing. Benchmarking of the ratings that often underlie estimates is even more difficult because of the need to map different rating scales to a common scale.

More generally in the following discussion, although emphasis is put on PD validation, we will regard IRB systems benchmarking as a whole (PD, LGD and EAD), thus summarising the notion of benchmarking to the idea of mapping to an external rating system. Possible ways to address EAD benchmarking are discussed in the appendix.

Objectives of benchmarking

At the current stage of knowledge, unambiguous and complete statistical tests enabling a formal acceptance or rejection of an IRB system do not appear to be available (see Sections III, IV and V). Difficulties mainly relate to the effect of default correlation, data constraints, and the definition of meaningful and robust target criteria for validating IRB systems. In this respect, recourse to benchmarking is often viewed as a complement to formal statistical backtesting of internal rating systems. As a matter of fact, benchmarking does appear in many aspects to be part of the whole process of producing internally generated estimates at banks' IRB systems. For example, banks frequently use external and independent references to calibrate their own IRB system in terms of PD. However, a bank's internal rating should reflect its internal risk management practices, and should not be a mere replication of an external benchmark model.

In principle, it is possible to differentiate between two ways of carrying out benchmarking for a certain set of borrowers or exposures:

1. Comparison of the internal estimates of risk components (e.g. PD) across a panel. For example, banks or supervisors may wish to compare PD estimates on corporates with respect to a peer group. The main purpose is to assess the correlation of the estimates or conversely the identification of potential "outliers" (e.g. variance analysis or robust regression) but not to determine if these estimates are accurate or not.

2. Comparison of internal estimates with an external and independent benchmark, for example, a rating provided by a supervisory authority or rating agency. Here the external benchmark is implicitly given a special credibility, and deviations from this

benchmark provide a reason to review the internal estimates. In this approach, the benchmark is used to calibrate and/or validate internal estimates. Given difficulties with identifying absolute benchmarks, one should be critical when using benchmarks.

In either case, benchmarking appears to be part of validation but may to some extent be more flexible, as it allows banks and supervisors to decide what benchmark is most appropriate and to enforce decision rules on how the IRB system should behave. In this respect, benchmarking replaces a purely formal validation process (which is not always available) with a more empirical and operational approach.

The first approach is of particular interest to supervisors and can be pursued in any country where banks adopt an IRB approach. Banks can be approached directly to provide, for example, the PD estimates for a common set of borrowers. However, although simple to implement, this approach raises several difficulties.

- A major technical problem is often **the identification of common borrowers across banks**. This can be alternatively viewed as the construction of a peer group. Depending on the information sources available, tax codes, identification codes from public or private credit registers or manual selection may solve this technical problem.
- Once these common borrowers have been identified, comparing the different ratings across banks would require their **mapping on a master scale**. This issue is similar to mapping to a benchmark (see below).
- Once PDs for a peer group have been collected, benchmarking may support further analysis: a widely used approach is to use benchmarking to identify outliers. In this respect, **benchmarking can also be viewed as part of non parametric tests to detect potential and systematic bias in a bank's methodology**; As a matter of fact, identifying and analysing "outliers" may prove difficult in effect, as differences in estimates may just stem from differences in methodologies. For example, PD estimates may differ because of a different definition of default. Thus, benchmarking might be regarded more as a variance analysis which can still be useful as it provides a qualitative indicator of potential differences in technologies used within the peer group. Such differences need of course to be further analysed but, in this respect, supervisors should be careful not to provide at least *ex ante* wrong incentives by imposing their own explicit or implicit benchmark.

Pursuing the second approach, involving an external benchmark, raises two major concerns:

- The **selection of the benchmark**: selecting an appropriate benchmark may not be such an obvious exercise as it requires some prior knowledge or some inference of the features of the underlying model. Choosing a benchmark PD, for example, depends upon whether the PD analysed is stressed or unstressed, and dynamic or static (see Section II on rating system classification).
- The **mapping to the benchmark**: the mapping refers to the one-to-one relation that can be inferred between the unobserved model and its benchmark. Ideally, in the case of a perfectly matching benchmark, this relationship will be perfectly one-to-one, but this will not be true the general case. As such, formalising this relationship may be quite difficult.

As a matter of fact, whether used with a relative or absolute objective, a comparison with an external benchmark may in practice still appear to be rather subjective regarding the last two aspects, i.e. what benchmark to use and how to use it, the difficulty reflecting the fact that both issues are related. Most of the benchmarks used are, for example, PDs from rating

agencies or commercial vendor models, often regardless of their true adequacy. As a consequence, mapping procedures may often be simple and misleading.

Yet, benefits of using benchmarking as a complement to validation could be greater if a more objective approach to constructing decision rules for banks and supervisors was used. Moreover, if benchmarks and mapping methodologies were available, validation could be less costly and burdensome. Supervisory evaluation could then focus on assessing the quality of the benchmark and the quality of the mapping. In addition, it could allow useful inference on the underlying characteristics of the IRB system with respect to the benchmark. The following sections look at the need to formalise the selection of benchmarks and the mapping procedure.

Moreover, there is no clear distinction between the range of PD values that a grade spans and the fact that the distribution of PD values of obligors assigned to the grade may not be uniform. For example, even though a grade might be properly defined to span the PD interval [.01,.02), and the average PD estimated for the grade might be .015, the true average PD for the obligors assigned to the grade at any given time might be higher or lower depending on the distribution of individual obligor PDs. Moreover, changes in the distribution over time might change the true average PD.

Selection of benchmarks

In practice, credit risk modelling in banks follows a rather "bottom up" approach. Notwithstanding reasons related to the empirical calibration of models, the segmentation of credits by portfolios (bank, sovereign, corporate, etc.) is often justified by an operational reality; banks develop different risk management and commercial techniques depending on the business activity. This segmentation actually reflects the fact that credit risk is governed by economic factors which are specific to the type of portfolio.

In this respect, the segmentation of portfolios according to specific economic characteristics entails that the underlying (local) default models are also (locally) different. This means for example, that the factors governing risks, say on banks, are not necessarily the same, or do not necessarily follow the same dynamics as those governing, say corporate.

It appears then that in a bottom-up approach, which is most likely to be used by banks, a portfolio's segmentation would necessitate (locally) specific default models which are deemed to be (locally) quite different. This observation has two major implications:

- With respect to the selection of benchmarks, considering IRB benchmarking as a mapping to an external rating system entails that the selection of an appropriate benchmark would rest upon the assessment of its qualities in adequately representing the expected economic characteristics of the portfolio studied. For example, many banks use rating agencies grades as benchmarks not only for their corporate portfolios but also more extensively for their SME and SME retail portfolio. The issue here is that **selecting a benchmark is not independent from the characteristics of the underlying economic model being analysed which would need to be inferred first**. In the given example, one should question whether benchmarking PD estimates of SME portfolios on say, S&P ratings, is consistent (in terms of granularity, calibration, discriminative power, etc.). For example, some banks would use the same benchmark (e.g. S&P), to classify risks for their corporate, SME and SME retail portfolios. One may therefore question whether the granularity of a rating system benchmarked on S&P ratings (about 20 classes) is consistent for describing SME and retail SME risks. It appears likely that this granularity would be too excessive for SME risks, thus entailing the possibility of

non-significant or non-discriminative risk buckets. Conversely, the granularity is expected to be higher for retail SME, thus entailing a likely too low discriminative power.

- With respect to the aggregation on a master scale, an issue may also arise on the consistency of the aggregation of specific underlying default models on a master default model, on the one hand, and eventually the consistency of the master default model obtained.

Overall, inconsistencies in the rating criteria, dynamic properties, and granularity of two rating systems make benchmarking exercises involving such disparate systems operationally and conceptually difficult, which reduces the value of such exercises. Thus, consistency in benchmarks is desirable. The need for more knowledge about which kinds of inconsistency matter in practical terms is still an open issue. These last aspects are discussed in the following section.

Mapping to the benchmark or to a master scale

Benchmarking generally requires a mapping procedure, i.e. rules relating unambiguously one rating system to the other, to be defined. Unfortunately, mapping procedures often appear to be rather simple or crude. Most of the time, this mapping rests on empirical comparisons of **average** PDs as a basis for grouping and matching risk buckets on a master scale. On this particular point, a distinction should be made between the range of PD values that a grade spans and the fact that the distribution of PD values of obligors assigned to the grade may not necessarily be uniform. For example, even though a grade might be properly defined to span the PD interval [.01,.02), and the average PD estimated for the grade might be .015, the true average PD for the obligors assigned to the grade at any given time might be higher or lower depending on the distribution of individual obligor PDs. Moreover, changes in the distribution over time might change the true average PD. This should be borne in mind when addressing the mapping process which is usually based on average PD.

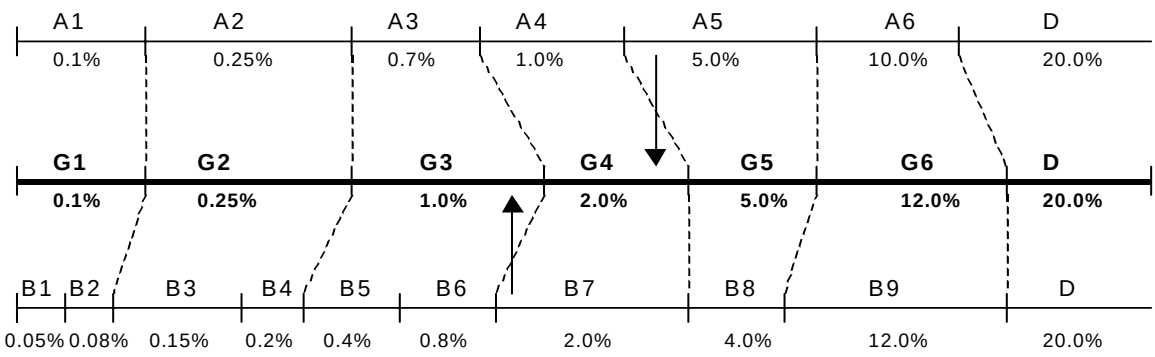
An illustration of an empirical mapping is given in the following example:

Rating scale A

Grade	A1	A2	A3	A4	A5	A6	D
PD (1 y)	0.1%	0.25%	0.7%	1%	5%	10%	20%

Rating scale B

Grade	B1	B2	B3	B4	B5	B6	B7	B8	B9	D
PD (1 y)	0.05%	0.08%	0.15%	0.2%	0.4%	0.8%	2%	4%	12%	20%



While simple to implement, such an approach may not seem satisfactory on theoretical grounds and with respect to validation purposes for two reasons. First, the PDs compared are not necessarily homogenous. They depend on the definition of default used, the horizon (TTC or PIT), and the conditioning (stressed, unstressed). **These properties are linked to the underlying default model and would need to be inferred in the first place** as suggested by the need to classify rating systems according to their dynamic properties (see Section II).

Second, even in the case where the PDs compared are homogenous, this approach does not take into account the granularities of each rating system (see above) which are proxies of the true distribution of values of the underlying default model. The problem stems from the fact that the distribution of obligors on a bucket is not observed, only an average PD. Merging buckets on the ground of average PD implies an assumption that merging the corresponding distribution does not alter the resulting average PD. This may be true, but is not in the general case. The previous example illustrates the potential errors: part of the borrowers rated B7 (PD 2%) are rated G3 (PD 1%); similarly part of the borrowers rated A5 (PD 5%) are reassigned in bucket G4 (PD 2%). The problem is the more serious with risky borrowers. Defaulted borrowers on rating scale A are reassigned a better risk bucket (G6) on the master scale. Regarding PD, the problem of mapping could be considered as optimising the granularity of the master scale in order to minimise the loss of information. With respect to validation, special attention should therefore be given to the appropriate granularity of the master scale and benchmark used. As mentioned before, this would need some inference of the economic characteristics of the benchmark or master model.

For example, consider a bank's portfolio specialised in corporate and SME lending. Assume the risks on the SME portfolio are discriminated using a 10-grade rating scale, while the risks on the corporate portfolio are discriminated using a 20-grade rating scale. On both portfolios, the bank uses specific default models to derive its PD estimates. In theory if the bank wishes to know its total risks on the whole portfolio, it would need to somehow aggregate the two sub portfolios. One way to do this is to calculate separately capital requirements on each sub portfolio and to aggregate them. This approach raises consistency issues between models (see below). A more usual approach is to build a master scale and its corresponding master default model. If the master scale has only 15 grades, then information will be lost on the risk distribution of the underlying corporate risks. If 20 grades are used, then non significant or

redundant risk buckets may be added for the description of the underlying SME risks. Overall the discriminative power of the master scale is likely to be affected.

Default models equivalence and mapping

A further issue behind mapping consistency and benchmarking in general relates to whether IRB parameters can be validated exogenously, regardless of the default model implied. Preliminary findings of these studies suggest that this assumption may be questioned.

As a matter of fact, IRB parameters are generally internally derived at banks from their own specific models. Banks then usually "validate" their estimates by mapping them to external estimates such as agencies rating or any other external data. In doing so they implicitly assume that the default model underlying the benchmark is somehow "equivalent" to their own internal default model. While this may be an acceptable assumption for C&I default models which indeed are by construction mapped onto benchmark default models (S&P, Moody's KMV, etc.) for which there is some sort of in-built equivalence with the ASFR model, it may be more of a problem for SMEs and retail default models which may greatly differ (see above).

In this respect, benchmarking could only be accepted as proper "indirect" validation if an actual equivalence between the internal default model and the benchmark default model could be demonstrated. Such instruments are still missing and could be subject to further research. Meanwhile, the general consistency between the bank's internal default model and its benchmark should be targeted. This problematic arises in particular when mapping (local) default models onto a master model.

Rating dynamics and benchmarking

Most of the validation process focuses on a static approach, whereby the features of an IRB system (calibration, performance) are assessed against a given set of criteria. The above discussion on mapping stressed the need to infer the stochastic behaviour of rating transitions within an IRB system. Qualifying the behaviour of a rating system and its components over time would therefore require a more dynamic approach to benchmarking which would enable *unveiling* and *classifying* rating systems according to their attributes and *dynamic properties* (see Section II).

Most of the validation process focuses so far on a static approach, whereby the features of an IRB system (calibration, performance) are assessed against a given set of exogenous criteria. In contrast, the above discussion regarding benchmarking stresses the need to shift from a completely exogenous approach which is in many respects still incomplete and not satisfactory, to a more complete approach which in its ideal form would encompass default model equivalence. While this objective may yet be out of reach, a more pragmatic way of ensuring default model consistency is **backtesting**, which can be viewed as an **ex post benchmarking**. Thus consistency between default models, as a pre-requisite to benchmarking, should also be viewed ***over time***, **meaning that similar (rating) dynamics should also be expected.**

An example of such dynamic backtesting is provided in appendix 2, which opens possible ways of application. Using an actual bank portfolio composition observed over time through a benchmark (supervisory) IRB system, its rating dynamics are constrained to follow the benchmark IRB system dynamics in a state space model, which is then used to predict a counterfactual portfolio. This counterfactual portfolio is then regressed on the actual portfolio,

under the null hypothesis that both portfolios follow the same dynamics. In the case studied this hypothesis is rejected.

Conclusions

Benchmarking is an important part of the validation process and in many cases appears an important empirical complement to a more rigorous and formal approach. To some extent, this alternative is more flexible as it gives room to banks and supervisors to decide what benchmark is most closely related to their IRB system. Still, benchmarking remains in many respects very subjective and would need to be more formalised. Conversely, if benchmarks and mapping methodologies were available, validation could be less costly and less burdensome. Supervisory evaluation could then focus on assessing the quality of the benchmark and the quality of the mapping. However, such assessment would need to focus on a more dynamic approach to benchmarking that could allow useful inference on the underlying characteristics of the IRB system.

Areas for further research

Some of the issues on benchmarking discussed above still need further investigation. The main issues of concern that have been identified at this stage are:

- *Peer group definition and utilisation*. Are there minimum requirements for using peer groups for benchmarking?
- *Incentives issues.* How to avoid potential selection bias or gaming in choosing a benchmark? How to make benchmarking incentive compatible? How to prevent potential herding behaviour?
- Mapping validation and default models consistency?

The main conclusions from the work of the Validation Group are summarised in the Executive Summary. A key result is that validation is a central prerequisite for meaningful capital requirements and a consistent application of the revised Framework across banks.

Appendix 1: A state space model for benchmarking

A state-space model is characterised by a system of equations whereby the general form is the following:

$$\begin{cases} Y_t = A_t Z_t + D_t + \varepsilon_t \\ Z_t = T_t Z_{t-1} + C_t + \eta_t \end{cases}$$

The first equation of this system is called the "measurement equation" and relates a time series (t=1,..,T) of observed vectors of n elements to a time series of vectors (m *x 1*) Z_t, called "state vectors". A_t is a matrix of *n x m* of parameters, D_t a *n x 1* vector of parameters, and ε_t a *n x 1* vector of independent disturbances with zero expectations and H_t covariance matrix. In a univariate model, n=1, and the measurement equation becomes:

$$y_t = A_t' Z_t + d_t + \varepsilon_t \quad \text{with} \quad Var(\varepsilon_t) = h_t \quad E(\varepsilon_t) = 0 \quad t = 1,...,T$$

In general, Z_t is not observable, but is supposed to be generated by a first order Markov process which is described by the "transition equation" of the system:

$$Z_t = T_t Z_{t-1} + C_t + \eta_t \quad t = 1,...,T$$

T_t is a *m x m* matrix called the "transition matrix" of the system, C_t is a *m x 1* vector also called the *input* of the system, and η_t is m x 1 vector of independent disturbances with zero expectations and covariance matrix Q_t.

Besides, the initial state must be known ($E(Z_0)=a_0$, $Var(Z_0)=P_0$). In the standard state-space form, the disturbances ε_t and η_t are generally assumed to be not correlated, either with the initial state vector at any time. The expression of a problem under the state-space form opens the way to the use of a number of powerful instruments such as the Kalman covariance filter.

The Kalman covariance filter is a recursive algorithm which determines the optimal estimator of the state vector at a date t, from all information available at this date. This information comprises present and past observations at time t, and the system matrices known at time t. This estimate is given by the expectation of Z_t conditional to the past and present information Y_0, Y_1, .., Y_t, i.e. $Z_{t/t} = E\left[Z_t \middle| Y^t = Y_0,..,Y_t \right]$. The corresponding problem is known as a filtering problem. The Kalman filter is also used for smoothing purposes (improving the estimation by considering future observations with respect to date t) and for predictions.

Let

- $\Sigma_{t/-1} = V\left(Z_t \middle| Y^{t-1} \right)$ the covariance matrix of Z conditional to the past of Y at date *t,*
- $F_{t/-1} = V\left(Y_t \middle| Y^{t-1} \right)$ the covariance matrix of the prediction error for Y conditional to the past of at date *t.*

It can be demonstrated then that the optimal estimator for the state vector is at each date t by the following recurrence equations:

- "Updating" equations for the observations: they enable to calculate the state vector a date t with the information obtained at this date from the filter

$$\hat{Z}_{t/t} = \hat{Z}_{t/t-1} + K_t\left(Y_t - C_t - T_t\hat{Z}_{t/t-1}\right)$$

with

$K_t = \Sigma_{t/t-1}T_t'\left(Q_t + T_t\Sigma_{t/t-1}T_t'\right)^{-1}$ gain of the filter

$Y_t - C_t - T_t\hat{Z}_{t/t-1}$ prediction error

$\Sigma_{t/t} = \left(I - K_tT_t\right)\Sigma_{t/t-1}$ covariance matrix of the prediction error

- "Prediction" equations: they enable the calculation of new information, i.e. at a future date, using past information

$$\hat{Z}_{t/t+1} = D_t + T_{t+1}\hat{Z}_{t/t}$$
$$\Sigma_{t+1/t} = Q_t + T_{t+1}\Sigma_{t/t}T_{t+1}'$$

Another reason for the importance of the Kalman filter is that when the disturbances and the state vector are normally distributed, it enables the calculation of the likelihood function of the model, via the decomposition of the prediction error. It enables then the estimation of any unknown parameters within the model and the realisation of statistical tests.

For a given value of the model's parameters of the form θ_k, the likelihood function can be expressed as:

$$f_T\left(Y^T;\theta_k\right) = \prod_{t=1}^{T}\hat{f}_t\left(Y_t / Y^{t-1};\theta_k\right)$$

with $\hat{f}_t\left(Y_t / Y^{t-1};\theta_k\right)$, the conditional normal density function of Y_t / Y^{t-1} which is given by the Kalman filter. The value of the Log-likelihood function for θ_k is $L_T\left(\theta_k\right)$ and can be calculated:

$$L_T(\theta_k) = -\frac{T}{2}Log(2\pi) - 0.5\cdot\sum_{t=1}^{T}Log\det(F_{t/t-1}(\theta_k)) - 0.5\cdot\sum_{t=1}^{T}\left[y_t - \hat{y}_{t/t-1}\left(\theta_k\right)\right]\left(F_{t/t-1}\right)^{-1}\left[y_t - \hat{y}_{t/t-1}\left(\theta_k\right)\right]'$$

One can then estimate the parameter by maximising the likelihood

$$\hat{\theta}_T = Arg\max_{\theta} L_T\left(\theta\right).$$

Appendix 2: Application to implementation: a dynamic benchmarking model[59]

A model of capital requirements dynamics

The proposed model of the dynamics of IRB capital requirements under Basel II stems from the reverse engineering of the formation of capital requirements. IRB capital requirements are assumed here to be an *ex ante* and exogenous constraint on the bank's portfolio. Future interesting research may derive the targeted or "optimal" risk weight from more sophisticated equilibrium economic models but this is beyond the scope of this paper. This capital constraint then expresses that at any time, the targeted average risk weight of the portfolio $\overline{RW_t}$, as derived by the foundation IRB approach, is equal to the weighted sum of risk by buckets within the portfolio, or in other words, that capital requirements set *ex ante*, are binding on the portfolio composition. Under a dynamic form, the capital constraint appears then compatible with the following process

$$\overline{RW_t} = RW_t' \cdot Z_t + e_t \quad e_t \sim N(0,h).$$

We name this relation, the *measurement equation*, with e_t a Gaussian white noise representing the measurement error, or innovation, RW_t' the line vector of buckets risk weights as determined in the IRB Foundation model, and Z_t the row vector representing the portfolio composition by risk bucket.

Besides, if one accepts that the transition matrix governs the dynamics of the portfolio composition pending on economic evolution, one can infer that the composition of the portfolio at the future date *t+1* can be obtained from the initial portfolio à date *t* with the transition matrix valid at this date (evolution of the quality of the initial portfolio) and the new credits distributed over the period. The dynamics of the portfolio appear thus compatible with the following process

$$Z_{t+1} = \Pi_t' \cdot Z_t + \nu_t + \eta_t.$$

We name this relation, *the transition equation*, with η_t the row vector of disturbances considered here to be Gaussian white noises. The production of new credits is captured here by the row vector ν_t, also called the *input* of the model. To simplify, we assume that ν_t is stationary and thus represents here the invariant part of the portfolio stemming from the average credit activity. Because of data limitations, we must also impose additional conditions in order to restrict the number of parameters to estimate. In the extent the risk classes transitions are assumed to be Markovian, the disturbance associated with the transition equation are also assumed to be independent and identically distributed. Some studies show that the hypothesis of Markovian transitions may be questioned on the long term[60], but that it would still remain valid on the short term[61], which is our case with quarterly transitions. Relaxing the Markovian hypothesis will be the subject for future further research, especially with the inclusion of richer model specifications such as the use of non Markovian transition matrices based on Aalen-Johansen matrices.

---

[59] This application is being developed at the Commission Bancaire and Banque de France following research on the subject proposed by Bardos, Foulcher and Oung (2000).

[60] See Lando and Skødeberg (2002).

[61] See Bangia, Diebold, Kronimus, Schagen and Schuermann (2002).

In total, assuming ex ante capital constraints, the dynamics of the bank's portfolio within an IRB framework appears thus reasonably described by a "measurement equation – transition equation" system, which typically characterises the class of models of the so called "state-space" form Harvey.[62]

$$\begin{cases} \overline{RWA}_t = RW' \cdot Z_t^* + e_t & \text{measurement equation (capital constraint)} \\ Z^*{}_t = \Pi_t' \cdot Z_{t-1}^* + \nu + \eta_t & \text{transition equation (portfolio dynamics)} \end{cases}$$

$$E(e_t) = 0$$

$$E(\eta_t) = 0$$

model parameters and hypothesis

$$V(e_t) = h$$

$$V(\eta_t) = Q = q \cdot I_n$$

In this class of models, the "state vector", is a N x 1 vector representing the portfolio composition a time t. It is not observed but can be predicted with the use of a Kalman filter. RW is a N x 1 vector of supervisory risk weights which to some extent may be viewed as a proxy for the bank's risk aversion. RWA is a scalar that represents the capital requirement that should be observed on the portfolio at a given time.

The predicted portfolio Z^* is considered as a counterfactual portfolio insofar it satisfies the capital constraint as well as the portfolio dynamics which are imposed by the transition matrix. This benchmark portfolio theoretically represents the portfolio that should be observed, taking into account the capital constraint and of the risk evolution entailed by economic fluctuations, without any other portfolio adjustment. One can then compare this counterfactual portfolio to the benchmark portfolio actually observed and infer the existence of over or under reactions in the adjustments thus revealed.

The estimation of the counterfactual portfolio and of the model parameters are obtained with successive iterations of:

- the Kalman filter for a given value of the model's parameters for all observations known at date t = 1,...,T;
- the computation of the model log-likelihood $L_T(\theta)$ for a given value of the model's parameters θ and the corresponding filtered results for the sate vector at all date t = 1,...,T.

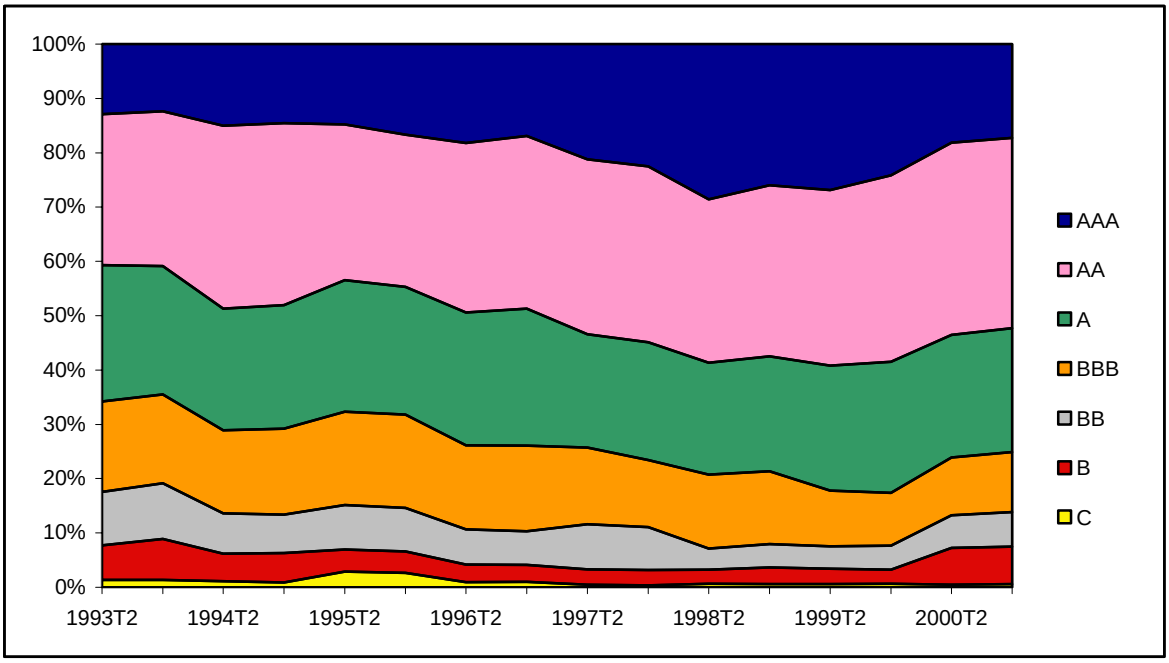
The whole procedure is reiterated until convergence of the model's parameters $\hat{\theta}_T = \underset{\theta}{\text{Argmax}}\ L_T(\theta)$. Initial values are calibrated from empirical observations obtained from the Banque de France central credit register (CCR). The capital constraint is here exogenous and set equal to the risk weights previously obtained from the FIRB model. The portfolio dynamics is given by the average (quarterly) transition matrix over the period. This matrix can be obtained from annual empirical matrices under the hypothesis of Markovian transitions. Assumptions on the new inflows of credits γ needs some analysis of banks credit cycle and could also be obtained from CCR or alternative regulatory call reports and research.

[62] See Harvey (1989).

To compare the actual portfolio to the counterfactual portfolio, the composition of the former is regressed over the composition of the latter under the null hypothesis that both portfolios have similar dynamics. In this respect, if there is no significant portfolio adjustment, one should expect the intercept to be null and the sensitivity to be significantly close to one. Specific tests could be performed to test the corresponding null hypothesis. Rejection of the null hypothesis would then be interpreted as a rejection of the model's assumptions about the dynamics of the bank's rating system.

An example of application: Inferring rating systems dynamic properties using dynamic benchmarking

Figure 11. A bank's portfolio evolution within an IRB system.



In this application, an observed rating system is used as benchmark for supervisory purposes. Assume that this rating system is a PIT one but a TTC benchmark rating system could also be used.

Figure 11 shows the observed (ex post) evolution of the composition of a bank's corporate portfolio according to this benchmark rating system. For convenience, S&P type labels have been used to represent the risk buckets: one can observe that this composition significantly varies over time. In particular, the rebalancing of bucket exposures in favour of low risky classes is noticeable during the boom period of 1998 to 1999, whereas a decline seems to follow during 2000 with the economic slowdown. However, it seems premature at this stage to assert whether this phenomenon was mainly the result of mechanic rating migrations between risk buckets, or the result of a dynamic portfolio strategy from the banks.

Table 12

Average quarterly transition matrix

	C	B	BB	BBB	A	AA	AAA
C	0.8102	0.0988	0.0187	0.0126	0.0075	0.0029	0.0013
B	0.1112	0.7242	0.1054	0.0235	0.0102	0.0062	0.0017
BB	0.0059	0.1129	0.7236	0.1247	0.0125	0.0065	0.0017
BBB	0.0036	0.0045	0.0709	0.782	0.1204	0.0088	0.0039
A	0.0003	0.0041	0.0085	0.072	0.7933	0.1148	0.0050
AA	0.0001	0.0011	0.0023	0.0071	0.0638	0.8484	0.0766
AAA	0.0000	0.0005	0.0006	0.002	0.0063	0.0648	0.9257

We now want to make inference about the dynamic behaviour of another IRB system which is proprietary to the bank and not easily observed by the supervisor unless at some high cost. A less costly way to make such inference is to benchmark the observed portfolios resulting from this proprietary IRB system over time with the counterfactual portfolio of our benchmarking model described above. We assume then that the dynamics of the banks portfolio is governed by the transition matrix of the IRB system under review and we constrain these dynamics by the same capital requirements as a proxy for the bank's risk aversion.

Figure 12. GDP growth and FIRB capital requirements on corporate portfolio.

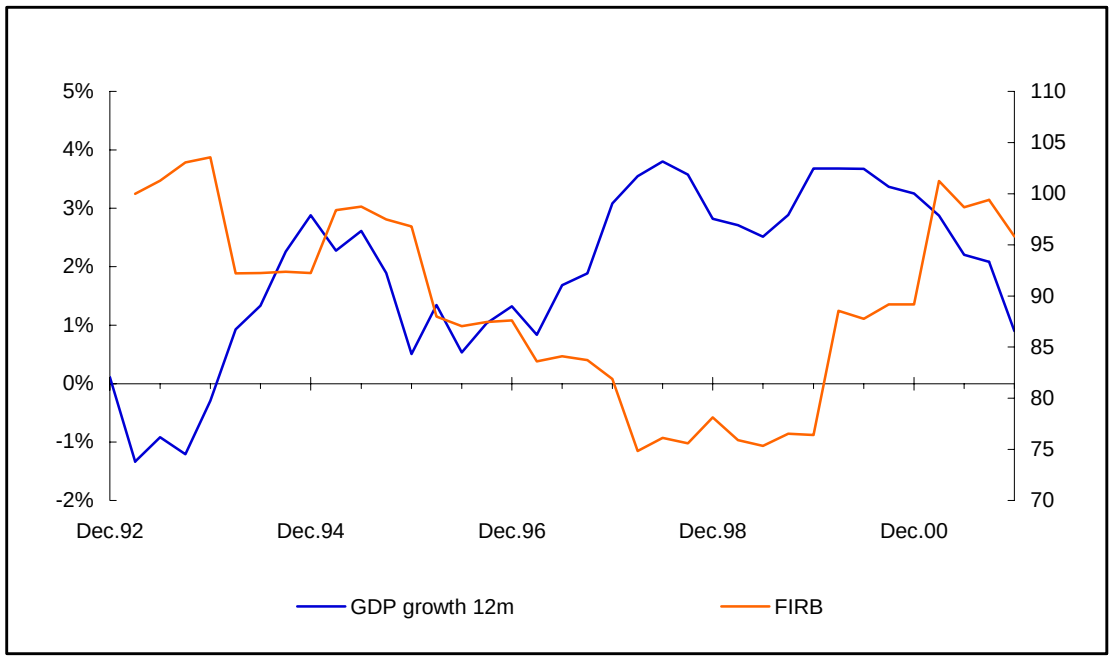


Figure 12 shows the FIRB capital requirements (expressed as an index) which were simulated *ex post* on the portfolios previously obtained. The results obtained on the average

corporate portfolio suggest cyclical capital requirements seemingly consistent with economic activity, which we assume particularly well measured by GDP growth as regards industrial business. As a matter of fact, capital requirement seem to peak at the end of 1993, a year of strong recession in France, improve gradually until end 1994, before receding again during the first half of 1995. This evolution seems consistent with the economic activity observed at the same time. From then onwards, capital requirements seem to lag behind the cycle: capital requirements start falling from the second half of 1995, thus anticipating a recovery which only occurs from the beginning of 1996, and increase strongly over 1999, thus anticipating an economic slump, which actually occurred in late 2000.

Figure 13. Composition of the optimal portfolio though time and under capital constraint (average quarterly transition matrix for the period).

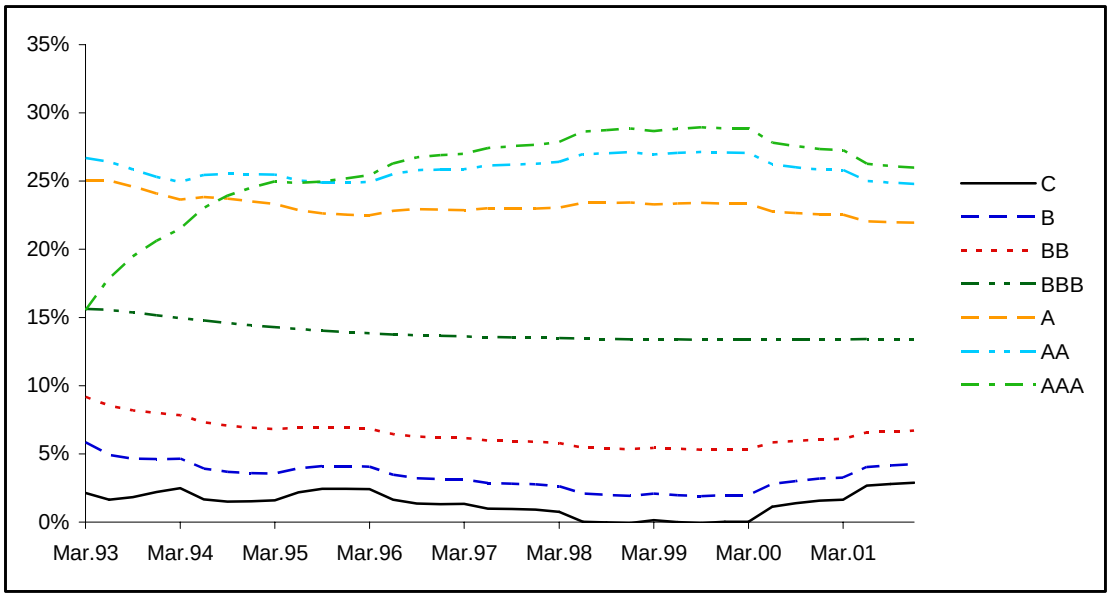


Figure 13 shows the counterfactual portfolio predicted by our benchmarking model. We note that the distribution of the optimal portfolio's composition, under capital constraint, changes in a consistent manner with economic fluctuations, with a shift of the distribution in favour of good classes in periods of good economic conditions, and a shift towards riskier classes in periods of poor economic conditions. This positive observation seems to support the idea that the optimal portfolio does capture the dynamics of risks evolution.

Table 13

Correlation of the observed portfolio with the counterfactual portfolio (average transition matrix of the period)

	TTR	TR	R	N	F	TF	TTF
Alpha	***0.0046**	0.0106	***-0.0298**	***-0.1690**	-0.0056	0.0629	-0.1037
t statistic	2.56	0.73	-2.17	-2.75	-0.07	0.48	-2.21
Beta	***0.4282**	***1.1960**	***1.4743**	***2.2283**	***1.0030**	****0.9604**	***1.1529**
t statistic	3.93	2.84	7.04	5.04	2.86	1.91	6.38
R^2	0.2918	0.1680	0.5812	0.4109	0.1705	0.0704	0.5313
RMSE	0.0059	0.0248	0.0121	0.0182	0.0149	0.0230	0.0348
F statistic	15.42	8.07	49.57	25.41	8.19	3.65	40.68
Prob > F	0.0004	0.0076	<.0001	<.0001	0.0072	0.0645	<.0001

*, **, ***, significant at the 1%, 5% and 10% confidence
model $Z^i_t = \alpha + \beta.Z^i_t{}^* + \varepsilon_t$

To compare the actual portfolio to the counterfactual portfolio, we now regress the composition of the former over the composition of the latter under the null hypothesis that both rating systems have similar dynamics. In this respect, if there is no significant portfolio adjustment, we should expect the intercept to be null and the sensitivity should be significantly close to one. The results obtained and presented in Table 13 demonstrate the opposite. One cannot systematically accept the null hypothesis for the intercept, on the one hand, and nearly all sensitivities measured are significantly different from one. A sensitivity greater (less) than one means that the actual portfolio overreacts (under reacts) with respect to the counterfactual portfolio used as a benchmark. The results obtained also suggest that this sensitivity would be particularly strong on median risks (class N), whereas it would be much lower for extreme risks, noticeably very high risks (TTR), excepted for very low risks which remain slightly over-weighted. This concentration on very low risks may probably be related to the potential size effect of corporate loans. Such results would support the conclusion that the IRB system under review may be more pro-cyclical than the benchmark rating system used.

References

Araten, Michael and Michael Jacobs Jr (2001), Loan Equivalents for Revolving Credits and Advised Lines. *The RMA Journal*, May, p. 34–39.

Asarnow, Elliot and James Marker (1995), Historical Performance of the U.S. Corporate Loan Market: 1988–1993. *The Journal of Commercial Lending*, 10 (2), p. 13–32.

Bangia, Anil, Francis X Diebold, André Kronimus, Christian Schagen and Til Schuermann (2002), Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking and Finance* 26 (2–3), p 445–474.

Bamber, Donald (1975), The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph. *Journal of Mathematical Psychology* 12, p. 387–415.

Bardos M, S Foulcher and Vichett Oung (2003), Exigences de capital et cycle économiques: une étude empirique sur les données françaises. Bulletin de la Commission bancaire no. 28, April.

Basel Committee on Banking Supervision (BCBS) (2004), International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Bank for International Settlements, June.

Blochwitz, Stefan, Stefan Hohl and Carsten Wehn (2003), Reconsidering Ratings. Unpublished Working Paper.

Brier, G W (1950), Verification of forecast expressed in terms of probability.

Engelmann, Bernd, Evelyn Hayden, and Dirk Tasche, (2003), Testing Rating Accuracy. *Risk*, January 2003, p. 82–86.

Harvey, Andrew C (1989), Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press: Cambridge.

Lando, David and Torben M Skødeberg (2002), Analyzing Rating Transitions and Rating Drift with Continuous Observations. *Journal of Banking and Finance* 26, p. 423–444.

Lee, Wen-Chung (1999), Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine* 18, p. 455–471.

Moody's Investor Service (1999), The Evolving Meaning of Moody's Bond Ratings, Moody's: New York, p. 6–7.

Moral, Gregorio and Raul García (2002), Estimación de la severidad de una cartera de préstamos hipotecarios. Banco de España. Estabilidad Financiera, no 3. (Estimation of LGD for a mortgage portfolio. Financial Stability Review. No English translation available yet.)

Moral, Gregorio and Maria Oroz (2002), Interest rates and LGD estimates. Unpublished manuscript.

Schuermann, Til (2003), What do we know about Loss-Given Default? Federal Reserve Bank of New York. Unpublished manuscript.

Standard and Poor's (2002), Factoring Cyclicality into Corporate Ratings. 2002 Corporate Rating Criteria. McGraw-Hill: New York, p. 41–43.

Tasche, Dirk (2003), A traffic lights approach to PD validation, Working paper.

Treacy, William F and Mark S Carey (1998), Credit Risk Rating at Large US Banks. *Federal Reserve Bulletin* 84 (11), November, p. 897–921.