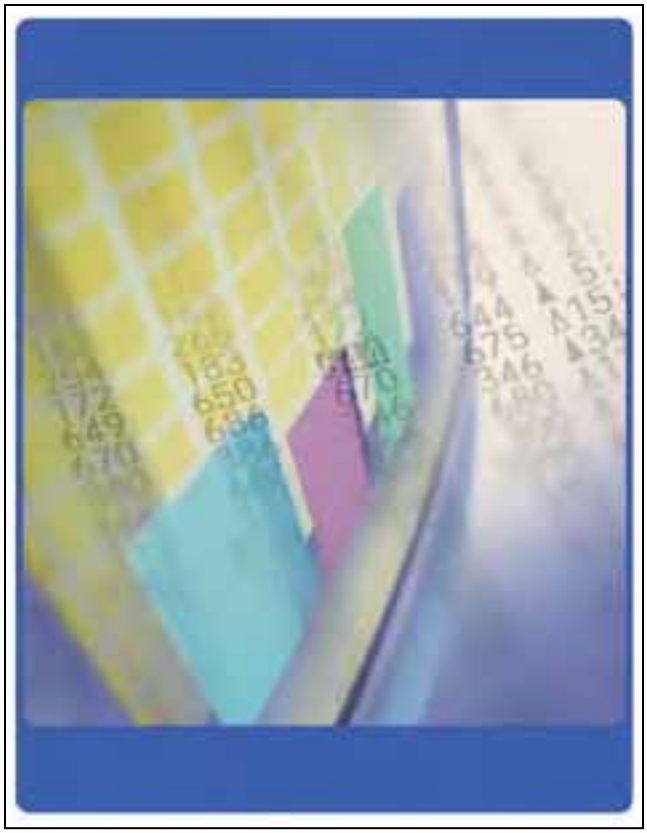
CONFERENCE SUMMARY



Forum on

Validation of Consumer Credit Risk Models

November 19, 2004

PAYMENT CARDS *Center*

FEDERAL RESERVE BANK OF PHILADELPHIA

Wharton
UNIVERSITY *of* PENNSYLVANIA

Forum on

Validation of Consumer Credit Risk Models

Sponsored by the Payment Cards Center of the Federal Reserve Bank of Philadelphia and the Wharton School's Financial Institutions Center

Peter Burns
Christopher Ody

Summary

On November 19, 2004, the Payment Cards Center of the Federal Reserve Bank of Philadelphia, in conjunction with the Wharton School's Financial Institutions Center, hosted a one-day event entitled "Forum on Validation of Consumer Credit Risk Models." This forum brought together experts from industry, academia, and the policy community to discuss challenges surrounding model validation strategies and techniques. This paper provides highlights from the forum and ensuing discussions.

The views expressed here are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. The authors wish to thank William Lang, Dennis Ash, and Joseph Mason for their special contributions to this document.

TABLE OF CONTENTS

Introduction 5

Model Validation: Challenging and Increasingly Important 7

Linking Credit Scoring and Loss Forecasting 8

Metrics for Model Validation 11

Incorporating Economic and Market Variables 14

Conclusion: Art Versus Science 16

Appendix A — Institutions Represented at the Conference 21

Appendix B — Conference Agenda 22

Introduction

On November 19, 2004, the Payment Cards Center of the Federal Reserve Bank of Philadelphia and the Wharton School's Financial Institutions Center hosted a "Forum on Validation of Consumer Credit Risk Models."[1] This one-day event brought together experts from industry, academia, and the policy community to discuss challenges surrounding model validation strategies and techniques. The discussions greatly benefited from the diverse perspectives of conference participants and the leadership provided by moderators and program speakers.[2]

Retail lenders, and particularly credit card lenders, use statistical models extensively to guide a wide range of decision processes associated with loan origination, account management, and portfolio performance analysis. The increased sophistication of modeling techniques and the broader application of models have undoubtedly played key roles in the rapid growth of the credit card industry and consumer lending in general.[3] At the same time, the widespread adoption of statistical modeling in these business processes has introduced new risk management challenges. Very simply, how do we know that our credit risk models are working as intended?

The conference discussions focused on two critical types of risk models: credit scoring models commonly used in credit underwriting and loss forecasting models used to predict losses over time at the portfolio level. These two model types differ in a number of ways, but the two modeling processes have strong theoretical links (although they are not often linked in practice).

Credit scoring models used for acquiring accounts are typically built on a static sample of accounts for which credit bureau — and often other applicant or demographic — information is available at the time of application. These data must then be combined with information about how these accounts ultimately performed in their first one to two years after acquisition. Credit scoring models are designed to predict the probability that an individual account will default or, more generally, develop a delinquency status bad enough that the bank would not have booked the account initially had it known this would happen. A number of credit scoring models only use credit bureau data to predict this probability, while others use application or demographic data in addition to credit bureau data.

Loss forecasting models predict dollar losses for a portfolio or sub-portfolio, not individual accounts. Some of the most popular loss forecasting models include cumulative loss rate models, which rely on vintage curve analysis, and Markov models, which rely on delinquency analysis of buckets. Loss forecasting models may or may not include segmentation by credit score. Economic data may be explicitly included in the model or implicitly included by using a time series covering an entire business cycle.[4]

---

[1] In May 2002, the Philadelphia Fed and the Financial Institutions Center co-hosted a multi-day conference on "Credit Risk Modeling and Decisioning." A summary of that event was published as a Special Conference Issue of the Payment Cards Center's newsletter, available on the Center's web site at: http://www.philadelphiafed.org/pcc/update/index.html.

[2] Speakers and moderators are listed in the program agenda at the end of this document. Copies of presentations and the program agenda are available at http://www.philadelphiafed.org/pcc/conferences/Agenda.pdf. While all of the individuals in the program made important contributions, William Lang, Dennis Ash, Shannon Kelly, and Robert Stine were especially helpful in structuring an agenda for the day.

[3] "Revolving credit" outstandings in the U.S. (largely credit card debt) grew from $100 billion to $790 billion in the 20-year period 1984-2004, as reported in the Federal Reserve Statistical Release G.19 (February 7, 2005), available at http://www.federalreserve.gov/releases/g19/hist/cc_hist_sa.txt.

Given the economic implications associated with a model's accuracy and effectiveness, issues concerning model validation are of obvious concern to the industry. Erroneous or misspecified models may lead to lost revenues through poor customer selection (credit risk) or collections management. While academics and other statisticians continue to extend and improve modeling technologies, lenders have to realistically assess the costs and benefits associated with increasing model sophistication and investing in more complex validation techniques. Hence, one of the central issues addressed during the forum was the adequacy of the attention and resources being devoted to validation activities, given these tradeoffs.

One of the central issues addressed during the forum was the adequacy of the attention and resources being devoted to validation activities.

The forum also addressed the increasing importance of validation from the regulatory perspective. Bank regulators and policymakers recognize the potential for undue risk that can arise from model misapplication or misspecification. Examining and testing model validation processes are becoming central components in supervisory examinations of banks' consumer lending businesses.

The conference format explicitly recognized these overlapping interests, and each panel was structured to include an industry, an academic, and a regulatory perspective.

The conference began with an introductory session outlining the importance of model validation and describing inherent challenges in the credit risk management process. These themes were extended in the panels that followed, dealing with validating credit scoring models and loss forecasting models. The day's final panel, entitled "Where Do We Go from Here?," attempted to draw out common threads and issues from the earlier discussions. As might be expected when such complex issues are examined, the discussions raised as many questions as answers. At the same time, the dialogue provided important insights and a better appreciation for the potential improvements that could result from greater collaboration among industry leaders, academic researchers, and regulators.

Rather than provide a chronological summary of the day's discussion, this paper highlights several key issues that emerged during the day. The paper begins with a summary of the opening presentation on the importance of model validation, which set the stage for the subsequent panels. The remainder covers three general themes that emerged from the panel discussions. These themes represent areas of particular complexity where the dialogue revealed multiple dimensions, alternative views, and, often, competing tensions. While resolving the various issues was not feasible in a single day, discussions generated important clarifications and specific suggestions for improving the model validation process.

[4] Economic data are generally not used in credit scoring models because this would require a very different sample structure. To be useful, the sample would have to include accounts with similar credit bureau and application information booked over multiple time periods, in order to reflect different economic environments. This would require a longer sample time and run the risk that the account-level data would be seriously outdated before the model was ever used. Loss forecasting models, on the other hand, are often designed specifically to include the effects of economic changes on expected loss and so use a time series of losses under varying economic circumstances, either controlling for changes in the risk profiles of the population of accounts or assuming there are none.

Model Validation: Challenging and Increasingly Important

Dennis Ash, of the Federal Reserve Bank of Philadelphia, opened the day's discussion by addressing several fundamental issues associated with validation of credit risk models. He began by describing the practical challenges that emanate from the basic modeling framework and how these factors have affected industry practices. Ash emphasized that, despite these challenges, there are a number of compelling reasons for modelers to improve validation practices. He closed with a series of questions that he encouraged participants to consider during the day's deliberations.

Ash noted that an intrinsic limitation to developing robust validation processes comes from the model construction process itself. He pointed out that scorecards (the output of the model that weighs each borrower's characteristics to compute a score) are by definition "old" when put into production and then are often used for five to 10 years without revision. By necessity, scorecards are based on historic data requiring at least a year of observation points before model construction can even begin. In essence, the model-building process results in a prediction of a future that looks like the past, which, as Ash aptly noted, is analogous to "driving a car by looking through the rear window." Furthermore, this approach simply fits patterns of correlation, which may not necessarily be related to causation, creating another level of challenge to any future validation process.

Similarly, Ash pointed out that scorecards are rarely constructed to incorporate changes in underlying economic conditions. He noted that borrower behavior tends to be quite different when interest rates are rising versus falling or in periods of economic downturns versus upturns. Performance validation, by definition, requires some quantifiable expectations about the impact of these economic factors.

That the same, often generic, scorecards are frequently used on a variety of portfolios with widely different characteristics further challenges the validation process. Different portfolios that have different terms and conditions or product features will also experience varied patterns of customer acceptance.

With these and other practical challenges facing users of credit risk models, Ash asserted that it is not surprising that banks too often pay little or no attention to model validation. Too often as well, he noted, banks ignore the most current information available in their validation processes. In an effort to recognize portfolio seasoning effects, many banks will create validation samples only from accounts booked one or two years ago. As such, they do not examine new account distributions or consider early delinquency patterns that might provide useful validation information.

Similar issues face the development and validation of loss forecasting models. Forecasts based on recent performance look at performance over the most recent outcome period, generally one year, which can then be weighted by the distributions of accounts today. This is a more accurate approach than relying on scorecard outcomes that are one to two years old and is further improved by using current weightings. Despite this, the technique does not take into account economic forecasts. More comprehensive loss predictions, which do use economic forecasts, generally use data over a complete economic cycle, which can be dated. Any forecast assumes that the future is driven by the same factors that operated in the past. Issues of causality and accuracy of data can cause degradation of the forecasts. Still, the more complete data, including economic data in addition to data on individual accounts, the longer time history, and the use of time-series analysis should make these forecasts more reliable over time.

Despite these and other real challenges,

Ash argued that there are a number of compelling reasons for credit card banks and other consumer lenders to pay greater attention to model validation. Size and scale considerations are driving factors that increase the importance of carefully monitoring a model's performance. As lender portfolios become larger and more complex, scoring becomes even more embedded in decision processes, adding greater importance to monitoring a model's performance. All of these factors can have significant economic consequences.

In a highly competitive lending environment, a model's performance can have important effects on market share, perhaps even creating adverse selection problems for those who really get modeling wrong. Ash noted that implementation of Basel II requirements will quickly "raise the bar" on validation of credit risk models. Model risk in consumer lending is a factor in defining overall operational risk. Increasingly, bank examiners will be seeking evidence that scoring models are effectively differentiating pools of exposures by their credit risk characteristics and, by extension, that loss forecasting models reflect current portfolio compositions and take into account macroeconomic and other relevant exogenous factors. Validation processes, and related documentation and reporting, will need to be consistent and clearly tied to a model's purpose.

Ash noted that implementation of Basel II requirements will quickly "raise the bar" on validation of credit risk models.

Basel guidance documents provide a template for validation that should help financial institutions adopt advanced validation practices.

In closing, Ash raised a series of questions that he encouraged conference participants to consider during the day: How do we integrate model purpose and performance expectations into validation processes? How do we incorporate stress testing under different economic conditions and then establish relevant tolerance metrics in validation? What do we do when we determine that our models are not working as intended? What are appropriate monitoring standards, and how do we incorporate ad hoc analyses into standard report reviews? How can we recognize and document the role of judgment in validation processes?

Many of these questions have technical components that are generally addressed with detailed statistical considerations. The focus of this forum, however, was on the more general management principles that need to be considered in improving validation and risk management practices. These and many other issues were actively debated throughout the day. Of the various points raised, the remainder of this paper highlights three selected themes that seemed to capture a number of the key issues debated: linking credit scoring models and loss forecasting models; appropriate metrics for model validation; and the use of economic and market variables in credit scoring models.

Linking Credit Scoring and Loss Forecasting

The conference discussion focused on validation issues associated with credit scoring and loss forecasting, two common and critical risk models used in credit card banks and other consumer lending environments. However, conference participants also debated an underlying point to the discussion of validation: the extent to which these two risk models have theoretical and practical links.

Banks use credit scoring models to rank individuals based on how likely they are to default on a loan.[5] While a credit scoring model typically produces a default probability, the models are generally built to separate and rank order borrowers by risk. Thus, metrics for validation of credit scoring models typically do not rely on whether the model accurately predicts default frequency, but rather they concentrate on the model's ability to determine which borrowers are more likely to default relative to others. In contrast, validation of loss forecasting models is based on the accuracy of the models' predictions relative to those of alternative models.[6]

Banks use the scoring model's measure of relative expected performance to make a variety of decisions, such as whether to grant credit, where to set the interest rate, and how to determine the maximum borrowing limit. Bank management must dynamically adjust score cut-off criteria for granting credit as well as the criteria for setting risk-based prices and credit limits. This dynamic adjustment is generally based on an assessment of market conditions as well as on the observed absolute rate of default for a given score band.

Loss forecasting models predict aggregate dollar losses for particular portfolios over a specific period of time. A variety of methodologies can be used to predict future losses, each of which has its own technical complexities, advantages, and limitations. Banks may use more than one kind of loss forecasting model to help predict future cash flows, establish loan loss reserves, and set appropriate levels of capital.

An underlying theme during the day's discussions centered on the connection between these two risk modeling techniques. Some participants argued that the two processes are logically linked. That is, the default rate is a central component of aggregate dollar losses, and therefore, a scoring model that generates statistical measures of the likelihood of default should be a central input to loss forecasting models. Moreover, failure to exploit the connection between these modeling approaches means that lenders are not using all the relevant information available to develop more effective tools.

Failure to exploit the connection between these modeling approaches means that lenders are not using all the relevant information available to develop more effective tools.

Professor Robert Stine, of the Wharton School, observed that in his experience the two modeling functions are often conducted independently. "Banks have the credit score modelers in one office, and the loss forecasters in another office, and the two groups build their models in isolation without ever talking to each other." Stine suggested that bringing these groups together could create synergies, increase knowledge within banks, and unify different pieces of evidence involved in managerial decision-making. Others noted that this separation sometimes occurs, in part, because of differences in functional skills. Credit scoring modelers are typically statisticians housed in business units responsible for underwriting and account management, whereas in many

[5] The definition of default (or "bad") for scoring purposes is not generally the same as the definition of default a lender may use for charge-off or placing a loan on nonaccrual status.

[6] Many lenders use a "champion/challenger" approach for validating a loss forecasting model. This approach compares the current (champion) model's forecast accuracy to that of an alternative (challenger) model.

banks, loss forecasters are finance professionals working in the bank's treasury department.

In addition to pointing out institutional divisions within a firm, participants also noted technical reasons for building credit scoring and loss forecasting models independently. In particular, the absolute likelihood of default depends on factors that go beyond the characteristics of the individual borrower, and these factors are difficult to incorporate into a statistical model. For example, the likelihood of default also depends on a firm's pricing, which, in turn, depends on the pricing decisions of its competitors as well as on the overall interest rates. Moreover, industry and macroeconomic factors change dynamically, so by definition, incorporating these factors would require building far more complex, dynamic models.

Different people make different distinctions between credit scoring models and loss forecasting models.

Indeed, some conference participants suggested that attempting to incorporate industry and macroeconomic factors into credit scoring models is inherently too complex and would ultimately lead to substantial error. In light of these complexities, some practitioners argued that by concentrating on producing a relative risk ranking of borrowers, lenders can effectively capture fairly stable relationships between borrower-specific information and the relative risk of default.

Intuitively, it would seem that changes in economic or market conditions would change the absolute likelihood that people will repay their loans. However, it was argued that most "good risks" will remain less likely to default than "bad risks," regardless of economic or market conditions. Thus, one would expect rank ordering to be more stable in changing conditions than the absolute rate of default. In this view, instead of trying to build statistical scoring models that give absolute risk in varying conditions, it is better to build relatively stable rank-ordering models and then rely on managerial judgment to change cutoffs for credit scores and make other business decisions to account for different conditions.

While acknowledging that there are substantial difficulties in making greater use of scoring models in loss prediction, Nick Souleles, of the Wharton School, contended that some of these difficulties are surmountable and that there might also be substantial gains in tackling them. As noted earlier, different people make different distinctions between credit scoring models and loss forecasting models. One distinction concerns what is being measured: credit scoring models predict default, whereas loss models usually predict expected losses. Another distinction concerns the "cardinality" of the results: credit scoring models typically produce only a rank ordering of risk, whereas loss models predict dollar losses.

Souleles argued that both of these distinctions are somewhat artificial and that, in principle, the two models should share common foundations. For example, it is possible to rank order consumers by expected losses or profitability and conversely to produce cardinal probabilities of default. Indeed, while earlier generations of scoring models were based on discriminant analyses that simply tried to separate "bad" and "good" accounts, many current scoring models are based on logistic and related models, which formally provide (and assume) cardinal probabilities of default. Hence, when people say they use scoring models only to rank order risk, they are, in practice, ignoring the additional information available in the underlying model. As argued earlier, this is done for

robustness. In Souleles' view, though, this suggests that the underlying models are not stable enough and that it might be better to deal with robustness and model instability directly.

With respect to "cardinality," his view is that lenders cannot avoid making cardinal decisions, so they might as well systematize their decisions as best as possible. While in the past credit scoring models were often used simply to decide whether or not to extend a loan, today very few decisions are so binary. For instance, on booking a credit card account, a lender must decide on the credit limit and the interest rate, both of which are continuous variables, and the appropriate interest rate should generally depend on the (cardinal) expected probability of default.

While in the past credit scoring models were often used simply to decide whether or not to extend a loan, today very few decisions are so binary.

Representatives of the regulatory community also noted that in the Basel II framework, risk ranking and forecasting are linked by requiring a portfolio to be segmented into homogeneous pools of risk, a job for which scoring is a prime tool, and then requiring various risk parameters to be estimated for each pool: the probability of default, the loss given default, and the exposure at default. These risk parameters, in turn, determine the minimum capital requirements for that pool. The capital requirements can then be added across pools to get the total capital requirement. Basel risk parameters and capital requirements are not necessarily the same as a bank's internal estimates of loss and economic capital, but the link between the Basel process and internal risk models may provide an impetus to banks to more effectively incorporate scoring into their loss forecasts.

In the face of current limitations to credit scoring models, banks have generally chosen to approach loss forecasting from a variety of directions that do not involve exploiting the potential connection with credit scoring models. While participants had varying views as to the efficacy of various approaches that would bring these two modeling techniques closer together, they generally agreed that industry and academic researchers are moving in the direction of greater linkage and that implementation of Basel II will likely spur these developments. Furthermore, as the accuracy of prediction in credit scoring models improves, there will be a greater incentive to exploit the connection with loss forecasting. More broadly, credit scoring models that generate more reliable point estimates of the rate of default could serve explicitly as inputs into a variety of other decision-making models, such as lifetime value models or pricing models. Academics, regulators, and those in the financial services industry all have good reason to actively follow these developments.

Metrics for Model Validation

During the discussion on model validation, the issue of appropriate metrics was another prominent theme. Recognizing that there is no common yardstick by which credit scoring and loss forecasting models can be measured, the conference panelists offered a framework for thinking about how model purpose, model use, and expectations for results play into the evaluation of credit scoring and loss forecasting models. Despite widespread agreement about the importance of clearly articulating models' purpose, use, and expected results, opinion diverged on the merits of using such standard statistical tests as the Gini coefficient and the K-S statistic. In the end, as with other discussion topics, forum participants broadly

acknowledged that developing effective processes and exercising sound judgment were equally as important as the particular statistical measurement technique used.

Dennis Glennon, of the Office of the Comptroller of the Currency, provided a helpful description of the relationships between the fundamental uses of credit scoring and loss forecasting models and the tools used to evaluate their performance.

In defining credit scoring models as essentially a classification tool, he argued that they be evaluated simply based on how well they separate "good" and "bad" credits over time. One common approach is to consider some measure of divergence between "goods" and "bads." An effective classification tool should result in accepting a high proportion of "goods" consistent with expectations. The K-S statistic and the Gini coefficient are common measures of a model's ability to separate risk. A second, related consideration is to evaluate whether the scoring model rank orders well over time. Instability in ordering would suggest that the model is not capturing the underlying and relatively constant information about how risky different credits are.

Instability in ordering would suggest that the model is not capturing the underlying and relatively constant information about how risky different credits are.

Glennon noted that, by contrast, loss forecasting models are essentially predictive tools that require metrics that evaluate "goodness-of-fit" and "accuracy." "Goodness-of-fit," he explained, measures how much of the variation in losses can be explained by changes in the independent variables. In regression analysis, this is most commonly measured as the R-squared of the regression. By contrast, a loss forecasting model's "accuracy" is best determined by how close predictions of losses are to those actually realized. Commonly used metrics to test predictive accuracy include the mean-squared error and the mean-absolute error.

Glennon's general conclusion was that validation methodologies should be closely associated with how the model is used. For example, in cases where a bank has a business need to use the estimated probability of default produced by a scoring model, validation criteria should include evaluations of the model's goodness-of-fit and accuracy. However, if a bank only uses the rank-ordering properties of the score, validation should concentrate on the model's ability to separate risk over time.

Although participants agreed that models should be evaluated based on purpose and defined by expectations, there was less agreement about whether commonly used statistical tests are appropriate to the needs of model-based consumer lenders, such as credit card companies. Professor David Hand, of London's Imperial College, argued that the standard metrics for validating credit scoring models are, indeed, inadequate and potentially misleading.

Hand started with the observation that credit scoring models are used to assign applicants to one of a discrete number of possible actions by the bank. For example, in deciding whether to accept an applicant for a credit card, a bank accepts applicants above a certain score and rejects those below it. When the bank makes the accept/reject decision, it doesn't matter how much the person is above or below the cutoff. Therefore, the distribution of applicants' scores is irrelevant to the model's performance at assigning applicants to actions. Hand pointed out that the model's only observable

measure of performance is the number of "bad" applicants accepted. Nevertheless, the commonly used statistical tests of a model's performance, such as the K-S statistic or Gini coefficient, measure the model's ability to rank risk throughout the entire sample without giving any special weight to performance near the accept/reject region. More generally, Hand argued that banks should not use metrics that rely on continuous distributions to evaluate models used for assigning applicants to discrete actions.

Hand further suggested that standard statistics for evaluating the risk separation properties of scoring models were often not well aligned with the use of those models. In particular, he presented research on the measures one should use when evaluating a model that establishes a cut-off score for granting or denying credit. Hand's model shows that alternative measures that concentrate on ranking performance of marginal borrowers (those borrowers near the potential score cutoff) produce better results than standard validation criteria that measure how the model ranks performance for the entire sample.

Hand's model shows that alternative measures that concentrate on ranking performance of marginal borrowers (i.e., those borrowers near the potential score cutoff) produce better results than standard validation criteria that measure how the model ranks performance for the entire sample.

Keith Krieger, of JPMorgan Chase, noted that Hand's argument holds only for the K-S statistic when banks choose a cutoff different from the point of maximum divergence. Michael Mout, of Capital One, also noted that banks do not always develop and evaluate models for a use as specific as accepting or rejecting applicants. For example, a scoring model might be used to provide a bank with information for testing new products to borrowers who are below the cutoff for existing products. Mout also argued that the consistent use of an agreed-upon metric is important, noting that a consistent metric is essential for comparing models during development, across portfolios, and over time. Thus, he concluded that there could be difficulty in tying a metric too closely to a cut-off criterion that was dynamically changing.

While the discussion raised questions about whether Hand's approach was applicable in all situations, there was agreement on Hand's more general point that evaluating a model's performance depends critically on a clear understanding of the model's intended use.

Nick Souleles also pointed out the importance of establishing a clear yardstick for a model's purpose. Moreover, he argued that the appropriate yardstick for lending models should be the maximization of a bank's risk-adjusted lifetime returns from its loans or accounts rather than accurate estimates of the probability of default or expected losses.

He also noted that at the portfolio level, the return on a portfolio of loans depends on more than the risk characteristics of an individual loan or segment. The covariance in returns across loans is an additional, crucial parameter. To illustrate the importance of covariance in returns, suppose that the average probability of default as measured by credit scores is the same in Michigan and in Alaska. However, suppose that the timing is such that default rates in Alaska have a low covariance with

the national default rate, while the default rates in Michigan are highly correlated with the national default rate. In this case, loans to Alaskans will reduce the volatility of the portfolio, holding all else fixed. While this example is simply illustrative, not a policy recommendation, the point is that most lenders would value lower volatility for the same average default rate.

Souleles presented recent research showing that it is possible to formally model which consumers are likely to be more cyclical than others. Further, he pointed out that this sort of cyclicality can potentially break the rank ordering of risk implicitly assumed by many credit scorers, since, in a downturn, the risk from cyclical consumers will deteriorate faster than that from noncyclical consumers.

Forum participants also concurred that models must be validated relative to clearly understood expectations. Rather than establishing some arbitrary statistical criteria for a model's performance, the central question for validation is whether the model is working as intended and producing results that are at least as good as alternative approaches. A clear understanding and documentation of expected performance is a necessary and fundamental basis on which all validation approaches must be built. On a pragmatic level, validation must assist management in determining whether the benefits of potential improvements to the model are worth the added costs of developing and implementing new models.

There was considerable discussion as to whether expectations for a model's performance solely required establishing objective statistical criteria or whether judgment was a necessary component. Some practitioners noted that a model's performance depends on multiple factors. For example, a model's performance is likely to be better in stable economic environments than unstable ones. Some forum participants argued that any evaluation of a model's performance needs to take into account these complex factors and that model developers could not solely rely on a statistical measure to assess a model's performance. At least one participant noted that the discussion on tools for a model's validation highlights just how much "art" remains in what initially appears to be a scientific and strictly numerical decision.

Rather than establishing some arbitrary statistical criteria for a model's performance, the central question for validation is whether the model is working as intended and producing results that are at least as good as alternative approaches.

While there was general agreement that the validation process is part science and part art, some participants argued for the need to establish clear quantitative criteria as part of the validation process. Such criteria need not be the sole measure of model performance, but they are necessary for establishing scientific rigor and discipline in the validation process. Although participants did not reach consensus on this topic, they generally recognized that experts must learn to balance evidence from a variety of metrics when building and evaluating models.

Incorporating Economic and Market Variables

Throughout the conference, participants discussed the advantages and disadvantages of including additional market and economic variables in both credit scoring and loss forecasting

models. In her presentation, Dina Anderson, of TransUnion, illustrated that credit scoring models are limited because they do not account for macroeconomic variables or, more generally, any factors influencing loan repayment that are outside of an individual's control. Anderson described an individual who loses her job during a recession and goes late on credit card payments until she finds a new job. If the job loss is simply due to bad luck, she will not be any riskier after getting a new job than she was before. "In reality," Anderson noted, "the likelihood that the customer is 'good' remains the same." However, because she was delinquent, credit scoring models will move her into a higher risk pool, despite the fact that her underlying risk is unchanged. Therefore, the model is not appropriately reflecting the risk probability over time because of causal factors that it does not include.

During his presentation, Souleles also addressed issues of model stability. He began by noting that model instability is an issue for both scoring and loss models. Models are calibrated using historical data, so if relevant unmodeled conditions change, the model can have trouble forecasting out of sample. Souleles pointed out that one useful response is to try to incorporate more of the relevant conditions into the model, in particular, macroeconomic conditions. Time-series analysis of macro variables, such as the unemployment rate, requires long sample periods, presumably covering at least one business cycle. Until recently, sample periods that were long enough were hard to come by, but he suggested that the 2001 recession provided new data that could be useful in predicting the effects of future increases in unemployment.

Moreover, even with shorter sample periods, he believes that it is still possible to use cross-sectional variation in, say, unemployment rates across counties, to model the effects of unemployment. Souleles showed results from his study of this subject, which found that increases in unemployment rates, declines in house prices, and health shocks (e.g., the loss of health insurance) increase default rates.[7] Such macro variables help predict default even after controlling for standard credit scores. While the scores still provide most of the predictive "lift," the macro variables provide enough additional lift to warrant their inclusion. Knowing this, lenders often respond informally, for example, by adjusting their score "cutoffs" (for at least binary decisions). Souleles argued that it would be better to formally include the macro variables in the model, in addition to the usual credit variables.

Souleles argued that it would be better to formally include the macro variables in the model, in addition to the usual credit variables.

Souleles pointed out that it is relatively easy to control for macro variables in reduced form, without building a complete structural model of the economy. While some in the audience argued that controlling for macro variables introduces too much subjectivity, Souleles responded that limiting oneself to the variables that happen to be available at the credit bureau is no less subjective. Nonetheless, Souleles warned that, in the absence of a structural model, one must remember that future recessions might be different from past recessions. He showed data from the period 1995-97, during which the bankruptcy rate significantly increased, even when controlling for credit scores and macroeconomic conditions (which were im-

[7] "An Empirical Analysis of Personal Bankruptcy and Delinquency," (with D. Gross), *Review of Financial Studies*, 15(1), Spring 2002.

proving at the time). Lenders will always have to back up their models with judgment. Still, he concluded that one should try to quantify that which can be quantified and use the experience of recent recessions to increase a model's accuracy (as compared to the alternative of ignoring that experience altogether).

Joseph Breeden, of Strategic Analytics, also emphasized that banks should quantify the expected effects of scenarios on future losses. Whether explicitly or implicitly, all loss forecasts are based on predictions regarding the vintage life-cycle, changing credit quality, seasonality, management action, the macroeconomic environment, and the competitive environment, which together form a scenario. By overtly including these factors, management can determine how much of the difference between actual and expected losses is a result of the model and how much is a result of the scenario. Even if a macroeconomic forecast is inaccurate, by explicitly including it, banks can examine outcomes over a range of possible future conditions. Breeden suggested that banks could even solve the model backwards, determining what would need to happen to the economy for a portfolio's performance to fulfill management's expectations. As in other areas of the discussion, this topic elicited a number of important insights for further research.

Breeden suggested that banks could even solve the model backwards, determining what would need to happen to the economy for a portfolio's performance to fulfill management's expectations.

Conclusion: Art Versus Science

In a speech in early December 2004, Federal Reserve Governor Susan Schmidt Bies noted that "although the importance of quantitative aspects of risk management may be quite apparent – at least to practitioners of the art – the importance of the qualitative aspects may be less so. In practice, though, these qualitative aspects are no less important to the successful operation of a business."[8] Later in her talk she added, "Some qualitative factors – such as experience and judgment – affect what one does with model results. It is important that we not let models make the decisions, that we keep in mind that they are just tools, because in many cases it is management experience – aided by models to be sure – that helps to limit losses." In a related sense, a good bit of the conference discussions focused on the role of judgment in the validation of credit risk models. By noting this balance of technical and judgmental factors, participants recognized the importance of both "art" and "science" in credit risk modeling.

At the most basic level, the construction of any statistical credit scoring and loss forecasting model requires some element of judgment, wherein the statisticians themselves decide whether to formally model the full array of (often endogenous) processes underlying repayment and default. The discussion relating to incorporating macroeconomic data into model design reflects one such issue, as Souleles noted, that even without a formal structural model of

[8] Susan S. Bies, "It's Not Just about the Models: Recognizing the Importance of Qualitative Factors in an Effective Risk-Management Process," The International Center for Business Information's Risk Management Conference, Geneva, Switzerland, December 7, 2004. Speech online at: http://www.federalreserve.gov/boarddocs/speeches/2004/20041207/default.htm

the macroeconomy, measurements of available reduced-form parameters often improve model fit.

The art, of course, lies in choosing the parameters to include and in calibrating a meaningful model. Those choices, in turn, rely on a clearly stated and documented understanding of the model's intended purpose and use. Models used to rank order credit scores have different inherent limitations than those used to generate accurate predictions. Furthermore, models used for binary classifications (accept/reject) face different limitations than those used for multiple joint decisions (accept/reject, interest rate, and credit line). Models incorporating changes in economic or industry performance may face limitations not yet known. Nonetheless, we can be sure that as competitive pressures and technical advances continue, implementation of new model validation techniques will rise in importance.

Consistency is a critical factor, and judgmental input must be controlled and managed with the same precision used with other model inputs.

The industry typically refers to such judgment as "overrides": Management decides to take action notwithstanding the model's results. While most participants agreed that managerial judgment, aided by credit scoring and loss forecasting models, can lead to better account management, that judgment needs to be implemented carefully. Consistency is a critical factor, and judgmental input must be controlled and managed with the same precision used with other model inputs. When judgmental inputs are inconsistent and subject to frequent changes, the model becomes less important to the credit scoring and loss forecasting management process. If the model is routinely overridden, the model becomes superfluous and should be either abandoned or revised. As one individual observed, the perceived need for constant change and re-calibration is likely a sign that the model is no longer functioning as intended and needs to be replaced. Judgmental factors may therefore add noise or accuracy (or both) to actual credit and loss outcomes. Hence, when models are augmented by managerial judgment, results from the modeling and subsequent validation processes can become seriously compromised. Therefore, while there was broad agreement that model performance must allow for judgmental factors, a number of participants argued that incorporating judgmental factors increases the need for rigorous testing and validation.

Validation, and more generally risk management, is an entire process that requires an interplay between effective managerial judgment and statistical expertise. It is not simply establishing a set of statistical benchmarks. Ronald Cathcart, of CIBC, aptly summarized the benefits and drawbacks of incorporating judgmental factors in the construction, use, and validation of credit scoring and loss forecasting models when he emphasized the need for consistency in the use of managerial processes throughout the model's life. Cathcart defined eight common steps or stages generally found in credit risk modeling beginning with "problem definition" to "maintenance and monitoring."[9] As he described these eight steps, he noted that judgmental factors are incorporated throughout the model's life and all steps require distinct validation approaches to ensure consistency throughout the entire process.

[9] The eight steps as defined by Cathcart are included in his PowerPoint presentation available on the Center's web site at: http://www.philadelphiafed.org/pcc/conferences/Ronald_Cathcart.pdf.

Cathcart also emphasized the importance of documentation, a point echoed by others in the discussion. While this may seem obvious, a number of participants from the regulatory community noted that the lack of documentation of judgmental processes is an all too common deficiency found in bank exams. Very simply, internal risk managers and bank examiners have a common need to understand how judgment is being employed and how well outcomes matched expectations or previous performance. While lenders should have clearly established expectations of how a model will perform and how it should inform management decisions, they should also have criteria that elicit managerial review to determine whether a model has come to the end of its useful life.

As a result, documentation is expected to become an ever more critical factor in the Basel II world. As model risk becomes a bigger factor in overall risk considerations, model validation becomes paramount. Underpinning the Basel II framework is the regulatory acceptance of individual banks' approaches to model-based decisioning. Lenders must be able to demonstrate to their regulators how their models are performing against expectations and how risk exposures fit within defined bands of acceptability. In essence, Basel II raises the bar for validation processes. As noted in Basel Retail Guidance, "A bank must establish policies for all aspects of validation. A bank must comprehensively validate risk segmentation and quantification at least annually, document the results, and report its findings to senior management."[10]

Lenders must be able to demonstrate to their regulators how their models are performing against expectations and how risk exposures fit within defined bands of acceptability.

Models are quickly becoming a critical area of potential innovation and competitive advantage. While participants generally accepted this premise, several argued that a reliance on demonstrated validation outcomes will lead to the elimination of judgment in the lending process. As articulated by several members of the regulatory community, this is clearly not the intention or direction they will be pursuing. The application of judgmental factors is recognized as a critical element of the risk management process. How such factors are applied and how expectations for performance will be affected now, however, need to be well documented.

In the end, it was generally agreed that while credit scoring and loss forecasting models and their statistical validation appear to be a well-grounded quantitative science that is becoming an important focus of regulatory compliance, they remain inextricably intertwined with the art of management.

---

[10] Internal Ratings-Based Systems for Retail Credit Risk for Regulatory Credit; 69 *Federal Register*, pp. 62,748 ff, October 27, 2004.

APPENDICES

APPENDIX A
Institutions Represented at the Conference

American General Corporation
Argus Information and Advisory Services
Bank of America
Bridgeforce
Capital One
CIBC
CIT
Citigroup
Cornell University
Daimler Chrysler
Drexel University
Equifax
Ernst & Young
Experian-Scorex
Fair Isaac & Co., Inc.
Federal Deposit Insurance Corporation
Federal Reserve Bank of Atlanta
Federal Reserve Bank of Philadelphia
Federal Reserve Bank of Richmond
Federal Reserve Board of Governors
GE Consumer Finance
Household Credit Card Services
Imperial College London
Innovalytics, LLC
JPMorgan Chase
KeyBank
KPMG
LoanPerformance, Inc.
MBNA
Merrill Lynch
Office of the Comptroller of the Currency
Penn Mutual Life Insurance Company
PNC Bank
Strategic Analytics
TransUnion
U.S. Department of Justice
US Bank Corp.
Wells Fargo
Wharton School

APPENDIX B
Conference Agenda

8:30 am **Registration and Coffee**

9:00 am **Welcome and Introduction**
Carol Leisenring
Co-Director, The Wharton School's Financial Institutions Center
Peter Burns
Vice President & Director, Payment Cards Center
Federal Reserve Bank of Philadelphia

9:15 am **What Is the Challenge and Why Is It Important?**
Dennis Ash, *Federal Reserve Bank of Philadelphia*

- What do we mean by model validation?
- Why focus on credit scoring and loss forecasting models?
- What are the risks of not getting it right? And what are the opportunities for those that can do better?

9:45 am **Break**

10:15 am **Validating Credit Scoring Models**
Moderator: Christopher Henderson, *MBNA America Bank*
Panelists: David Hand, *Imperial College London*
Dina Anderson, *TransUnion*
Michael Mout, *Capital One*

- How often do we need to validate and what does this timing depend on?
- Will one measure do?
- What do we do when the future is different from the past because of changes in the economy, changes due to portfolio acquisitions, changes in product terms, etc.?

12:00 pm **Informal Lunch**

Conference Agenda

1:00 pm **Validating Loss Forecasting Models**

Moderator: Joseph Breeden, *Strategic Analytics*
Panelists: Dennis Glennon, *Office of the Comptroller of the Currency*
Nick Souleles, *The Wharton School*
Ron Cathcart, *Canadian Imperial Bank of Commerce*

- How are loss forecasting models different from credit scoring models?
- What techniques (roll rate, vintage analysis, scoring-based approaches, etc.) are best used for forecasting dollar losses?
- How do we best validate loss forecasting models and how is this different from or similar to validation of credit scoring models?

2:45 pm **Break**

3:00 pm **Where Do We Go From Here?**

Moderator: William Lang, *Federal Reserve Bank of Philadelphia*
Panelists: Robert Stine, *The Wharton School*
Erik Larsen, *Office of the Comptroller of the Currency*
Sumit Agarwal, *Bank of America*
Huchen Fei, *JPMorgan Chase*

- What should we most care about going forward?
- What are the gaps in our understanding?
- What things do we need to work on: to run the business, to provide effective oversight, and to resolve theoretical questions?

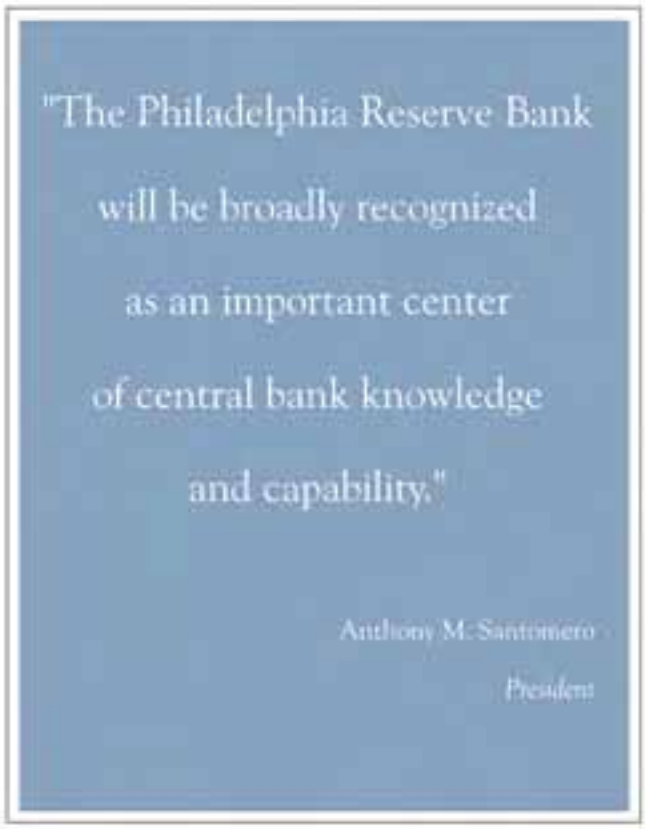
The Wharton Financial Institutions Center
2307 Steinberg Hall-Dietrich Hall
3620 Locust Walk
Philadelphia, PA 19104

http://fic.wharton.upenn.edu/fic/

Payment Cards Center
Federal Reserve Bank of Philadelphia
Ten Independence Mall
Philadelphia, PA 19106

http://www.philadelphiafed.org/pcc/

Federal Reserve Bank of Philadelphia



PAYMENT CARDS
Center

FEDERAL RESERVE BANK OF PHILADELPHIA

Ten Independence Mall
Philadelphia, PA 19106-1574
215-574-7110
215-574-7101 (fax)
www.philadelphiafed.org/pcc

Peter Burns
Vice President and Director

Stan Sienkiewicz
Manager

The Payment Cards Center was established to serve as a source of knowledge and expertise on this important segment of the financial system, which includes credit cards, debit cards, smart cards, stored-value cards, and similar payment vehicles. Consumers' and businesses' evolving use of various types of payment cards to effect transactions in the economy has potential implications for the structure of the financial system, for the way that monetary policy affects the economy, and for the efficiency of the payments system.