

VIII. SCORING AND MODELING

Scoring and modeling, whether internally or externally developed, are used extensively in credit card lending. Scoring models summarize available, relevant information about consumers and reduce the information into a set of ordered categories (scores) that foretell an outcome. A consumer's score is a numerical snapshot of his or her estimated risk profile at that point in time. Scoring models can offer a fast, cost-efficient, and objective way to make sound lending decisions based on bank and/or industry experience. But, as with any modeling approach, scores are simplifications of complex real-world phenomena and, at best, only approximate risk.

Scoring models are used for many purposes, including, but not limited to:

- Controlling risk selection.
- Translating the risk of default into appropriate pricing.
- Managing credit losses.
- Evaluating new loan programs.
- Reducing loan approval processing time.
- Ensuring that existing credit criteria are sound and consistently applied.
- Increasing profitability.
- Improving targeting for treatments, such as account management treatments.
- Assessing the underlying risk of loans which may encourage the credit card backed securities market by equipping investors with objective measurements for analyzing the credit card loan pools.
- Refining solicitation targeting to minimize acquisition costs.

Credit scoring models (also termed scorecards in the industry) are primarily used to inform management for decision making and to provide predictive information on the potential for delinquency or default that may be used in the loan approval process and risk pricing. Further, credit risk models often use segment definitions created around credit scores because scores provide information that can be vital in deploying the most effective risk management strategies and in determining credit card loss allowances. Erroneous, misused, misunderstood, or poorly developed and managed scoring models may lead to lost revenues through poor customer selection (credit risk) or collections management. Therefore, an examiner's assessment of credit risk and credit risk management usually requires a thorough evaluation of the use and reliability of the models. The management component rating may also be influenced if governance procedures, especially over critical models, are weak. Regulatory reviews usually focus on the core components of the bank's governance practices by evaluating model oversight, examining model controls, and reviewing model validation. They also consider findings of the bank's audit program relative to these areas. For purposes of this chapter, the main focus will be scoring and scoring models. A brief discussion on validating automated valuation models (AVM) is included in the Validation section of this chapter, and loss models are discussed in the Allowances for Loan Losses chapter. Valuation modeling for **residual interests** is addressed in the Risk Management Credit Card Securitization Manual.

Scoring models are developed by analyzing statistics and picking out cardholders' **characteristics** thought to be associated with creditworthiness. There are many different ways to compress the data into scores, and there are several different outcomes that can be modeled. As such, scoring models have a wide range of sophistication, from very simple models with only a few data inputs that predict a single outcome to very complex models that have several data inputs and that predict several outcomes. Each bank may use one or more generic, semi-custom, or custom models, any of which may be developed by a scoring company or by internal staff. They may also use different scoring models for different types of credit. Each bank weighs scores differently in lending processes, selects when and where to inject the scores into the

processes, and sets cut-off scores consistent with the bank's risk appetite. Use of scoring models provides for streamlining but does not permit banks to improperly reduce documentation required for loans or to skip basic lending tenants such as collateral appraisals or valuations.

Practices regarding scoring and modeling not only pose consumer lending compliance risks but also pose safety and soundness risks. A prominent risk is the potential for model output (in this case scores) to incorrectly inform management in the decision-making process. If problematic scoring or score modeling cause management to make inappropriate lending decisions, the bank could fall prey to increased credit risk, weakened profitability, liquidity strains, and so forth. For example, a model could wrongly suggest that applicants with a score of XYZ meet the bank's risk criteria and the bank would then make loans to such applicants. If the model is wrong and scores of XYZ are of much higher risk than estimated, the bank could be left holding a sizable portfolio of accounts that carry much higher credit risk than anticipated. If delinquencies and losses are higher than modeling suggests, the bank's earnings, liquidity, and capital protection could be adversely impacted. Or, if such accounts are part of a securitization, performance of the securitization could be at risk and could put the bank's liquidity position at risk, for instance, if cash must be trapped or if the securitization goes into **early amortization**. A poorly performing securitization would also impact the fair value of the residual interests retained.

Well-run operations that use scoring models have clearly-defined strategies for use of the models. Since scoring models can have significant impacts on all ranges of a credit card account's life, from marketing to closure, charge-off, and recovery, scoring models are to be developed, implemented, tested, and maintained with extreme care. Examiners should expect management to carefully evaluate new models internally developed as well as models newly purchased from vendors. They should also determine whether management validates models periodically, including comparing actual performance to expected performance. Examiners should expect management to:

- Understand the credit scoring models thoroughly.
- Ensure each model is only used for its intended purpose, or if adapted to other purposes, appropriately test and validate it for those purposes.
- Validate each model's performance regularly.
- Review tracking reports, including the performance of overrides.
- Take appropriate action when a model's performance deteriorates.
- Ensure each model's compliance with consumer lending laws as well as other regulations and guidance.

Most likely, scoring and modeling will increasingly guide risk management, capital allocation, credit risk, and profitability analysis. The increasing impetus on scoring and modeling to be embedded in management's lending decisions and risk management processes accentuates the importance of understanding scoring model concepts and underlying risks.

TYPES OF SCORING

Some banks use more than one type of score. This section explores scores commonly used. While most scores and models are generally established as distinct devices, a movement to integrate models and scores across an account's life cycle has become evident.

FICO Scores

Credit bureaus offer several different types of scores. Credit bureau scores are typically used for purposes which include:

- Screening pre-approved solicitations.
- Determining whether to acquire entire portfolios or segments thereof.

- Establishing cross-sales of other products.
- Making credit approval decisions.
- Assigning credit limits and risk-based pricing.
- Guiding account management functions such as line increases, authorizations, renewals, and collections.

The most commonly known and used credit bureau scores are called FICO scores. FICO scores stem from modeling pioneered by Fair, Isaac and Company (now known as Fair Isaac Corporation) (Fair Isaac), hence the label “FICO” score. Fair Isaac devised mathematical modeling to predict the credit risk of consumers based on information in the consumer’s credit report. There are three main credit bureaus in the United States that house consumers’ credit data: Equifax, TransUnion, and Experian. The credit-reporting system is voluntary, and lenders usually update consumers’ credit reports monthly with data such as, but not limited to, types of credit used, outstanding balances, and payment histories. A consumer’s bureau score can be significantly impacted by a bank’s reporting practices. For instance, some banks have not reported certain information to the bureaus. If credit limits are not reported, the score model might use the high balance (the reported highest balance ever drawn on the account) in place of the absent credit limit, potentially inflating the utilization ratio and lowering the credit score. Errors in, or incompleteness of, consumer-provided or public record information in credit reports can also impact scoring. Consumer-supplied information comes mainly from credit applications, and items of public record include items such as bankruptcies, court judgments, and liens.

Each bureau generates its own scores by running the consumer’s file through the modeling process. Although banks might not use all three bureaus equally, the scoring models are designed to be consistent across the bureaus (even though developed separately). Thus, an applicant should receive the same or a similar score from each bureau. In reality, variations (usually minor) arise due to differences in the way the bureaus collect credit information (for example, differences in the date of data collection) or due to discrepancies among information the bureaus, which could include inaccurate information. FICO scores rank-order consumers by the likelihood that they will become seriously delinquent in the 24 months following scoring. FICO scores of 660 or below may be considered illustrative of subprime lending (as set forth in the January 2001 *Expanded Guidance for Subprime Lending*), although other characteristics are normally considered in subprime lending determinations as well.

Benefits of credit bureau scoring include that it is readily available, is relatively easy to implement, can be less expensive compared to internal models, and is usually accompanied by various bureau-provided resources. Disadvantages include that scoring details are, for the most part, confidential and that it is available to every lender (no competitive differentiation).

As is the case for any type of scores generated by models, FICO scores are inherently imperfect. Nevertheless, they usually maintain effective rank ordering and can be useful tools, particularly when resource or volume limitations preclude the development of a custom score. Several types of FICO scores are in use including Classic FICO, NextGen FICO Risk, FICO Expansion, and FICO Industry Options. Collectively, the scores are called FICO scores in this manual.

There are three different Classic FICO scores, one at each of the bureaus. According to www.fairisaac.com, they are branded as Beacon scores at Equifax; FICO Risk or Classic (formerly known as EMPIRICA) scores at TransUnion; and Experian/Fair Isaac Risk Model scores at Experian. Scores range from 300 to 850, with higher scores reflecting lower credit risk.

NextGen FICO Risk scores draw their name from being touted as the “next generation” of credit bureau scores. They are branded as Pinnacle at Equifax; FICO Risk Score, NextGen (formerly PRECISION) at TransUnion; and Experian/Fair Isaac Advanced Risk Score at Experian. Compared to Classic scores, NextGen scores are reported to use more complex predictive variables, an expanded segmentation scheme, and a better differentiation between degrees of future payment performance. According to www.fairisaac.com, the score range, 150 to 950, is

widened, although odds-to-score ratios at interval score ranges remain the same. Cumulative odds may vary.

For accounts lacking sufficient credit file information to generate a Classic or NextGen FICO score, some lenders use the FICO Expansion score. The FICO Expansion score, introduced in 2004, likely draws its name from “expanding” the credit information considered in the score to beyond that collected in a standard credit report. The expanded information includes items such as payday loans, checking account usage, and utility and rental payments. The FICO Expansion score has the same range and scaling as the Classic scores.

FICO Industry Options scores draw their name from being specific to several options of industries, such as bankcard.

VantageScore

The bureaus historically used their own proprietary models (based on Fair Isaac modeling) to develop FICO scores. However, in 2006, the bureaus introduced a new scoring system under which a single methodology is used to create scores at all three bureaus. The new system is called VantageScore. Because a single methodology is used, the score for each consumer should virtually be the same across all three bureaus. Any differences are attributed to differences in data in the consumer’s files. The score will continue to incorporate typical consumer report file content but will range from 501 to 990. The scores are scaled similar to the letter grades of an academic scale (A, B, C, D, and F). Again, the higher the score, the lower the credit risk. Consumers may likely have VantageScores that are higher than their FICO scores. This is due to scaling and that phenomena alone does not indicate that a consumer is a better credit risk than he or she was under the traditional FICO score system. Further, when determining whether subprime lending exists, the new scale will need to be considered (in other words, 660 may not be a benchmark when looking at VantageScores). The industry’s rate of replacement of custom and generic scores with VantageScore remains to be seen as of the writing of this manual.

Other Scores

In addition to or instead of generic credit bureau scores, many banks use other types of scores. Brief discussions on a variety of these scores follow, in alphabetical order. The bureaus and other vendors offer models for many of these types of scoring.

Application Scoring:

Application scoring involves assigning point values to predictive variables on an application before making credit approval decisions. Typical application data include items like length of employment, length of time at current residence, rent or own residence, and income level. Points for the variables are summed to arrive at an application score. Application scores can help determine the credit’s terms and conditions.

Attrition Scoring:

Attrition scores attempt to identify consumers that are most likely to close their accounts, allow their accounts to go dormant, or sharply reduce their outstanding balance. Identification of such accounts may allow management to take proactive measures to cost-effectively retain the accounts and build balances on the accounts.

Bankruptcy Scoring:

Bankruptcy scores attempt to identify borrowers most likely to declare bankruptcy. HORIZON (by Fair Isaac) is a common credit bureau bankruptcy score.

Behavior Scoring:

Behavior scoring involves assigning point values to internally-derived information such as payment behavior, usage pattern, and delinquency history. Behavior scores are intended to embody the cardholder's history with the bank. Their use assists management with evaluating credit risk and correspondingly making account management decisions for the existing accounts. As with credit bureau scores, there are a number of scorecards from which behavior scores are calculated. These scorecards are designed to capture unique characteristics of products such as private label, affinity, and co-branded cards.

Behavior scoring systems are often periodically supplemented with credit bureau scores to predict which accounts will become delinquent. Using a combination allows management to evaluate the composite level of risk and thus vary account management strategies accordingly.

Adaptive control systems (ACS) commonly use behavior scoring. ACS bring consumer behavior and other attributes into play for decisions in key management disciplines (for instance, line management, collections, and authorizations) so as to reduce credit losses and increase promotional opportunities. ACS include software packages that assist management in developing and analyzing various strategies taking into account the population and economic environment. They are a combination of software actionable analytics and optimization techniques and use risk/reward logic. ACS recognize that accounts can go in several directions. They consider the possible outcomes of the options and determine the "best" move to make. With ACS, challenger strategies can be tested on a portion of the accounts while retaining the existing strategy (champion strategy) on the remainder. Continual testing of alternative strategies can help the bank achieve better profits and control losses. Many large banks use TRIAD (developed by Fair Isaac) or a similar ACS, but smaller banks may lack the capital or the infrastructure to implement such a process.

Collection Scoring:

Collection scoring systems rank accounts by the likelihood of taking delivery of payments due. They are used to determine collection strategies, collection queue assignments, dialer queue assignments, collection agency placement, and so forth. Collection scores are normally used in the middle to late stages of delinquency.

Fraud Detection Scoring:

Fraud detection scores attempt to identify accounts with potential fraudulent activity. Fraud continues to be pervasive in the credit card lending industry and detection of potential fraudulent activity can help identify and control losses as well as assist management in developing fraud prevention controls.

Payment Projection Scoring:

Payment projection scoring models use internal data to rank accounts, normally by the relative percentage of the balance that is likely to be repaid. Some models only forecast the relative percentage, while others rank the likelihood a cardholder will pay a moderate to high level of the account balance. The scores are normally used in the early to middle stages of delinquency.

Recovery Scoring:

Recovery scoring models rank order the amount of recovery that is expected after charge-off. They aid management in deploying the necessary resources where collection is most likely and help with agency placement and sale decisions.

Response Scoring:

Response scoring models are used to manage acquisition costs. By identifying the consumers that are most likely to respond, a bank is able to tailor its marketing campaigns so as to target its marketing toward those consumers that are most likely to respond and to steer away from spending marketing dollars on consumers that are least likely to respond.

Revenue Scoring:

Revenue scoring models rank order the potential revenue expected to be generated on new accounts during the first 12-month period. The models use predictive indicators such as usage ratios, the level of revolving balances, and other card-usage patterns. Revenue scoring allows management to focus marketing initiatives on what are expected to be the most profitable accounts. Used in conjunction with credit bureau scores in screening applicants, they allow management to evaluate the revenue potential as well as the risk ranking of prospects. Consequently, management is better able to identify its target market and tailor its solicitations to that market.

Revenue scoring is also used to manage existing accounts according to revenue potential. Strategies can be formulated recognizing the risk, revenue, and frequency of cardholder use. From this information, management is better able to reward low-risk, product-loyal consumers by reducing APRs or waiving fees. Conversely, management is apt to raise APRs and fees for consumers who exhibit higher risk or that evidence little product loyalty.

DUAL-SCORING MATRIX

A dual-scoring matrix is a system which uses one score on one axis and another score on its other axis. Examiners should normally expect to see dual scoring in more complex credit card operations. Any scoring system may interface with another, but a commonly employed dual-scoring matrix uses application and credit bureau scores. The use of two scores allows management to more effectively segment applicants. Each score has a cut-off level (as discussed later in this chapter). Applicants that either pass or fail both cut-off scores are either accepted or rejected, respectively. A gray area arises when an applicant passes one cut-off but fails the other. These situations afford management a greater opportunity to maximize approvals or minimize losses by including potentially good credit risk or by excluding potentially bad credit risk that may have gone undetected in a single-scoring system. Taking advantage of this opportunity requires a thorough tracking system so that management can determine the historical loss rates for the score combinations in the gray area. Cut-off scores can then be adjusted so that the best scoring combinations are approved and so that applicants who would be approved under a single-score system, yet still pose unacceptable risks, can be identified and excluded.

CREDIT SCORING MODEL DEVELOPMENT

Scoring can be done with generic models, semi-custom models, or custom models. When properly designed, models are usually more reliable than subjective or judgmental methods. However, development and implementation of scoring models and review of these models present inherent challenges. These models will never be perfectly right and are only good if users understand them completely. Further, errors in model construction can lead to inaccurate scoring and consequently to booking riskier accounts than intended and/or to a failure to properly

identify and address heightened credit risk within the loan portfolio. Errors in construction can range from basic formula errors to sample-bias to use of inappropriate predictive variables.

A scoring model evaluates an applicant's creditworthiness by bundling key attributes of the applicant and aspects of the transaction into a score and determines, alone or in conjunction with an evaluation of additional information, whether an applicant is deemed creditworthy. In brief, to develop a model, the modeler selects a sample of consumer accounts (either internally or externally) and analyzes it statistically to identify predictive variables (independent variables) that relate to creditworthiness. The model outcome (dependent variable) is the presumed effect of, or response to, a change in the independent variables.

The sample selected to build the model is one of the most important aspects of the developmental effort. A large enough sample is needed to make the model statistically valid. The sample must also be characteristic of the population to which the scorecard will be applied. For example, as stated in the March 1, 1999 *Interagency Guidance on Subprime Lending* (Subprime Lending Guidance), if the bank elects to use credit scoring (including application scoring) for approvals or pricing in a subprime lending program, the scoring model should be based on a development population that captures the behavioral and other characteristics of the subprime population targeted. Because of the significant variance in characteristics between subprime and prime populations, banks offering subprime products should not rely on models developed solely for products offered to prime borrowers.

Both a large number of good and bad accounts are necessary to maximize the model's effectiveness. There are no hard and fast rules, but the sample selected normally includes at least 1,000 good, 1,000 bad, and about 750 rejected applicants. Often, the sample contains a much higher volume of accounts. The definition of good and bad accounts (the dependent variable) differs among banks, especially between prime and subprime issuers. Furthermore, definitions of bad for scoring purposes are not necessarily the same as definitions of bad used by banks for charge-off or nonaccrual consideration. For prime portfolios, good accounts tend to be defined as accounts with sufficient credit history and little or no delinquency. Bad accounts for prime portfolios are normally distinguished by adverse public records, delinquency of 90 days or more, accounts with a history of delinquency, and accounts charged-off. Rejected applicants are applicants that management refused to accept because of their risk parameters. Certain inferences are made to break down the rejected applicants into good and bad accounts. This procedure, known as **reject inferencing**, makes certain assumptions on how rejected applicants would have performed had they been accepted and attempts to mitigate any accept-only bias of the sample. The process is used as it would be cost-prohibitive and potentially detrimental to make loans to consumers who would otherwise be rejected just for the sake of improving models.

After a representative sample has been assembled, the accounts are analyzed to determine the characteristics and attributes common to each group. The characteristics may be based on data sources such as the consumer's credit report, the consumer's application, and the bank's records. Characteristics are the questions asked on the application or performance categories of the credit bureau report. Attributes are the answers given to questions on the application or entries on the credit bureau report. For example, if education is a characteristic, college degree or high school diploma illustrate possible attributes.

The characteristics, which may number in the hundreds, are refined into a much smaller group of predictive variables, which are those items thought to best indicate whether a new applicant will eventually fall into the good or bad performance category. Ideally, the predictive variables also maintain a stable relationship with the performance measurement over-time. Commonly used predictive variables include, but are not limited to, prior credit performance, current level of indebtedness, amount of time credit has been in use, pursuit of new credit, time at present address, time with current employer, type of residence, and occupation. Examiners should expect that management has excluded factors lacking predictive value or that by law cannot be used in the credit decision-making process (such as race).

Once the predictive variables have been selected, points are assigned to the attributes of those variables. Each attribute is awarded points, and determining the number of points to award each attribute may be the most difficult element of the process. There are several methods for calculating and assigning points, all using a form of multivariate statistics. A scoring table is constructed, for which characteristics are on one axis and attributes are on the other axis. Points are awarded to each cell of the matrix. The consumer's characteristics and attributes are compared with the scoring table, or scorecard, and are awarded points according to where they fall within the table. The points are tallied to arrive at the overall score. Whether a high score means low or high risk depends on the model's construction.

Once designed and prior to implementation, the model is evaluated for integrity, reliability, and accuracy by a party independent of its design. This process is referred to as validation. A sample from the development sample may be held-out and scored with the new model. Performance is then monitored, and a model that demonstrates separation and rank ordering on the hold-out sample is considered valid. Validations for independent samples are also usually conducted prior to release of the model and post-implementation.

Validation has long been fundamental to a successful score modeling process, and evaluating a bank's model validation process has long been a central component of the examination. The Subprime Lending Guidance requires management to review and update models for subprime lending to ensure that assumptions remain valid. Validation is also an integral part of the proposed rulemaking for the revised Basel capital accord.

BASEL CONSIDERATIONS REGARDING CREDIT SCORING

A brief discussion on the new Basel capital accord is housed in the Capital chapter. Under the proposed rulemaking, banks that use an Internal Ratings Based (IRB) approach would use internal estimates of certain risk parameters as key inputs when determining their capital requirements. The IRB approach requires banks to assign each retail exposure to a segment or pool with homogeneous risk characteristics. These characteristics are often referred to as primary risk drivers and may include credit scores.

A bank must be able to demonstrate a strong relationship between the IRB risk drivers (such as scores) and comparable measures used for credit risk management. Thus, even if a bank uses custom scores for underwriting or account management, generic bureau scores could possibly be used for IRB segmentation purposes if the bank can demonstrate a strong correlation between these measures. A bank using credit scores as segmentation criterion would have to validate the choice of the score (bureau, custom, and so forth) as well as demonstrate that the scoring system has adequate controls.

Examiners will expect that all aspects of the risk segmentation system, including credit scoring, are subject to thorough, independent, and well-documented validation. Validation for the risk segmentation system is ultimately tied to validation of the bank's quantification of IRB risk parameters. Examiners will also expect that the IRB validation process include:

- Evaluating the developmental evidence or logic of the system.
- Ongoing monitoring of system implementation and reasonableness (verification and benchmarking).
- Comparing realized outcomes with predictions (back-testing).

VALIDATION

Examiners should determine whether management provides for appropriate, ongoing validation of scoring models, whether used as part of an IRB framework, for credit risk management, or for

other purposes. Validation is a process that tests the scoring system's ability to rank order as designed and essentially answers whether the model is accurate and working properly. Model validation does not only increase confidence in the reliability of a model but also promotes improvements and a clearer understanding of a model's strengths and weaknesses among management and user groups. Model validation can be costly, particularly for smaller banks. But, using un-validated models to manage risks is a poor business practice that can be even more costly as well as lead to safety and soundness concerns. Risks from not validating are elevated when a bank bases its credit card lending decisions on the scoring model alone (and does not consider other factors in the decision-making process), when the model is otherwise vital, or when the model is complex.

Examiners do not validate models; rather, validation is the responsibility of bank management. Examiners do, however, test the effectiveness of the bank's validation function by selectively reviewing aspects of the bank's validation work. Examiners could also identify concerns with a model's performance as a by-product of the credit risk review or other examination procedures.

Examiners should evaluate the bank's validation framework, including written validation policies, to determine if it is proper. Key elements of a sound validation policy generally include:

- **Competent and Independent Review** - The review should be as independent as practicable. The reviewer can be an auditor with technical skills, a consultant, or an internal party. In practice, model validation requires not only technical expertise but also considerable subjective business judgment.
- **Defined Responsibilities** - The responsibility for model validation should be formalized and defined just as the responsibility for model construction should be formalized and defined.
- **Documentation** - Validation cannot be properly performed if a sufficient paper trail of the model's design is not available. Weak documentation can be particularly damaging to the bank if the modeler leaves and the replacement is left with little to reference. Model documentation should summarize the general procedures used and the reasons for choosing those procedures, describe model applications and limitations, identify key personnel and milestone dates in the model's construction, and describe validation procedures and results. Technical complexity does not excuse modelers from the responsibility of providing clear and informative descriptions of the model to management.
- **Ongoing validation** - Validation should occur both pre- and post-implementation. Models should be subject to controls so that coding cannot be altered, except by approved parties. Most models are normally altered in response to changes in the environment or to incorporate improvements in understanding of the subject. Model alterations that are inappropriate can result in dodging risk limits or disguising losses.
- **Auditor involvement** - Examiners should expect that the bank's audit program ensures that validation policies and procedures are being followed.

A clear understanding of the scoring model's intended use is critical to properly assessing a model's performance. But, regardless of the intended use, the three key components of a validation process, as mentioned in the prior section, apply: evaluation of the conceptual soundness of the model; ongoing monitoring that includes verification and benchmarking; and outcomes analysis.

Evaluating conceptual soundness involves assessing the quality of the model's construction and design. Examiners should determine whether management reviews documentation and empirical evidence supporting the methods used and the variables selected in the model's design. Modelers adopt methods, decide on characteristics, and make adjustments. Each of these actions requires judgment, and validation should ensure that judgments are well-informed. Examiners should expect management to review developmental evidence for new models and when a material change is made to an existing model.

The purpose of the second component of validation, ongoing monitoring, is to confirm that the model was implemented appropriately and continues to perform as intended. Process verification and benchmarking are its key elements. Process verification includes making sure that data are accurate and complete; that models are being used, monitored, and updated as designed; and that appropriate action is taken if deficiencies exist. Benchmarking uses alternative data sources or risk assessment approaches to draw inferences about the correctness of model outputs before outcomes are actually known. The time needed to generate a sufficient number of representative accounts (good and bad) to evaluate the effectiveness of the model post-implementation will vary depending on the product-type or customer group. Consequently, benchmarking becomes an important tool in the validation process because it provides an earlier-read of model performance than is available from back-testing.

The third component of validation, outcomes analysis, compares the bank's forecasts of model outputs with actual outcomes. It should include back-testing, which is the comparison of the outcomes forecasted by the models with actual outcomes during a sample period not used in model development (out-of-sample testing).

Benchmarking and back-testing differ in that when differences are observed between the model output estimates and the benchmark, it does not necessarily indicate that the model is in error. Rather, the benchmark is an alternative prediction, and the difference may be due to different data or methods. When reviewing the bank's benchmarking exercises, examiners should find out whether management investigates the source of the differences and determines whether the extent of the differences is appropriate.

Examiners can compare the delinquency rate at each score interval as a simple test of overall performance of the scoring system. If the system is performing adequately, a correlation between the scores and delinquency rates (that is, delinquency rates increase as projected risk (as reflected in the scores) increases) should be evident. Examiners may also want to review the results of various tests that management may be using. For example, divergence statistics and the population stability index are sometimes used. Divergence statistics measure the distance between the average score of satisfactory accounts and average score of unsatisfactory accounts. The greater the distance, the more effective the scoring system is at segregating good and bad accounts. If the difference is small, a new or redeveloped scoring system may be warranted. The population stability index compares divergence with the original development sample and helps identify and measure erosion in the model's predictive power. Other advanced statistical tools include Chi square, Kolmogorov-Smirnov (K-S) tests, and Gini coefficients. While examiners generally do not need to know the specifics of all of these types of tests, they should be aware that these tests are common in the industry and should expect management to be able to explain the validation tools used. Management's development of effective processes and exercise of sound judgment are just as important as the measurement technique used.

Incorporation of combinations of model expertise and skill levels in the validation process is not uncommon. For example, internal staff could be used to verify the integrity of data inputs while a third party could be used to validate model theory and code. Examiners should determine what management's procedures are for ensuring that vendors' validation procedures are appropriate and meet the bank's standards. Management is ultimately responsible for ensuring the validation processes used, whether internal or external, are appropriate and adequate.

While scoring models developed in-house are becoming more prevalent, banks continue to purchase a number of models from vendors and the bureaus. Vendors are sometimes unwilling to share key formulas, assumptions, and/or program coding. In these cases, the vendor typically supplies the bank with validation reports performed by independent parties. The independent party's work can only be relied on if the information provided is sufficient to determine the adequacy of the scope, the proper conveyance of findings to the vendor, and the adequacy of the vendor's response thereto. Examiners assessing risks of modeling activities should pay

particular attention to situations in which management has exclusively relied on a vendor's general acceptance by others in the industry as sufficient evidence of reliability and has not conducted its own comprehensive review of the vendor and its practices.

Examiners should evaluate management's processes for re-tooling or re-developing models that exhibit eroding performance. If evidence reliably shows that the behavior shift is small and likely to be of short duration, a policy shift or change to the model may not be warranted. But, if evidence suggests that the behavior shift is material and is likely to be long-term, there are several approaches management may consider to limit losses, depending on the ability to identify the most likely reason(s) for the performance shift. It can adjust its underwriting policy to narrow the market to a group believed to perform better than the population in general. This usually involves making changes to the bank's business strategy and, thus, is rather limited as a short-term risk management tool. Banks may also develop or purchase scoring models based on more recent information about the current population. In this case, the bank must weigh the costs of developing or purchasing a model against that of carrying an increased number of bad accounts booked by the existing model. One of the most common, and often the easiest, adjustments is to manage the cut-off score to maintain a targeted loss rate consistent with profit objectives.

CUT-OFF SCORE

Each bank develops its own policies and risk tolerances for its credit card lending programs. Setting cut-off scores is one way banks implement those risk tolerances. A cut-off score is the point below which credit will not be extended and at or above which credit will be extended (assuming a higher score equates to better creditworthiness). A bank might have more than one cut-off score, with each tailored to a specific population. The ability to customize cut-off scores allows management to maximize the approval rate without sacrificing asset quality. Some banks have cut-off bands, which define a range of scores for which the consumer would undergo additional judgmental review.

Selecting a cut-off score involves determining the optimum balance between approval and loss rates. Management evaluates how much additional revenue will be added if the approval rate is increased and what the cost associated with the incremental increase in the bad rate will be. They also often give consideration to marketing expenses and customer service expenses. How management chooses to balance the competing goals determines the cut-off score. Odds charts are often involved in setting cut-off scores and are discussed in the next section.

As time passes, cut-off scores and models become less predictive because of economic changes, demographic shifts, and entry into new markets. Examiners should assess management's practices for reviewing cut-off scores and models, including resulting acceptance and loss rates. By monitoring the rates, management can appropriately adjust the cut-off score to change either acceptance rates or loss rates, depending on the strategic goals. For example, management could grow the portfolio by lowering the cut-off score (when lower scores equate to higher risk), taking on an elevated degree of credit risk and accepting increased loss rates. These dynamics of the scoring environment highlight the need for thorough tracking and **calibration** procedures.

VALIDATION CHARTS AND CALIBRATION

Most scores are rank-order measurements that, by themselves, are generally not indicative of the likelihood or magnitude of an event or outcome. Rather, they summarize a plethora of consumer data and essentially do little except rank order the consumer's risk against the risk of other consumers. But, in addition to this rank-ordering, scores must give accurate outcome (usually default) probabilities to be the most useful. Calibration is the process by which a model's output (in this case scores) is converted into the actual rate of the outcome (default) and includes adjusting or modifying for the difference between the expected rate based on the historical

database and the actual rate observed. The process is aimed at converting or modifying the model's output into a probability based on the expected odds for the historical population and adjusting for the relevant population. Often, it is thought of as the process of determining and fine tuning the grades or gradation of a quantitative measuring system by comparing them with a set standard or starting point. Frequently the standard used might be a bureau's validation chart.

In general, validation charts (also commonly known as odds charts) reflect the estimate of the percentage of borrowers in a defined population who will evidence a certain trait or outcome, such as delinquency, loss, or bankruptcy. Examiners normally expect management to develop its own odds chart(s) when it has sufficient historical data. When properly developed, customized odds charts are more predictive than odds charts that are available from the bureaus. Validation charts available from the bureaus display the odds of poor performance (such as delinquency, loss, or bankruptcy) observed at a given bureau score. Each set of charts available from the bureaus is specific to a model, an industry, and an application (where application refers to how the scores will be used). For example, the bureaus have validation charts available for the bankcard industry and for subprime lending. The bureaus' validation charts can be helpful as a starting point for management in setting risk strategies but do not precisely predict the actual odds that each bank will experience. Rather, a bank's particular market will have different characteristics and, thus, different odds. The risk ranking based on bureau score will generally hold, but the actual odds of going bad that each score represents will vary between banks and portfolios. Thus, management must provide for sufficient calibration processes. For example, if the bureau odds chart indicates that 1 out of every 20 consumers with a credit score of XYZ will be a bad account and the bank is realizing 5 out of every 20 consumers with a credit score of XYZ is a bad account, calibration most likely is needed.

Calibration most often adjusts or refines an odds chart when significant variation exists from the general forecast. But, there are other instances for which the scores and scaling could be adjusted, or calibrated. For example, calibration might be used to make all scores positive. For example, if a model's scores are (52), (6), and 15, an entity could add 52 points, so the scores would be 0, 46, and 67. Also, calibration might be used to compress the scale (for example, if every 31 points doubles the odds of bad, a bank could calibrate the scale such that the bad odds are doubled every 20 points). Calibrations might also be done to make users feel comfortable (for example, if an existing cut off score is XYZ based on an internal model that predicts that one percent of accounts with a score of XYZ will be bad, then calibration could be used to ensure that accounts that are scored XYZ would continue to tie to the likelihood that one percent will be bad. In this way, the bank would not have to change the cut-off score to keep getting the same caliber of customers). Examiners should ascertain whether recent calibrations are well-documented and have been properly executed.

OVERRIDES

Overrides are discussed in the Underwriting and Loan Approval Process chapter. Exceptions outside of management's credit scoring parameters are called overrides and may be high-side or low-side. When management overrides the cut-off score, they introduce information into the ultimate credit decision that is not considered in the scoring system. If the scoring system is effectively predicting loss rates for a designated population and the system reflects management's risk parameters, examiners should expect that management use overrides with considerable caution. Excessive overrides may negate the benefit for an automated scoring system. A high volume of overrides is equivalent to having no cut-off score and jeopardizes management's ability to measure the success of the credit scoring system. Once a bank approves credits that fail to meet the scoring system's criteria, it has broken its odds and may be taking on higher levels of risk than acceptable for the bank's risk appetite and/or capabilities to control. However, business reasons may justify a temporary increase in override rates. For example, when transitioning to a new system, override rates might rise until a reasonable level of confidence in the new approach is achieved.

CREDIT SCORING MODEL LIMITATIONS

Determining whether scoring models are managed by people who understand the models' strengths and weaknesses is an integral part of the examination process. Users lacking a complete understanding of how the models are made, how they should be used, or how they interface with the bank's lending policies and procedures can expose the bank to risks, as discussed throughout this chapter. Scoring is only useful if its limitations are properly understood, and examiners should draw a conclusion about whether an understanding of the model's limitations by management is evident.

One limitation is that scoring model output is only as good as the input that is used. If data going into the scoring model is inaccurate (for instance, if information on the consumer's credit bureau report is erroneous), the model's output (score) will be erroneous. Depending on how the erroneous information is weighted in the scoring formula, the impact on the score could be substantial. Moreover, if management does not select and properly weight the best predictive variables, the model's output will likely be less effective than had the most predictive variables been used and properly weighted. Management must make sure that the variables used in the models are appropriate, predictive, and properly weighted to arrive at the best credit decision and that data inputs are complete and accurate.

The effectiveness of the model output (scores) can also be constrained by factors such as changing economic conditions and business environments. Examiners should identify whether management monitors warning signs of market deterioration, such as increases in personal bankruptcies, which may affect the accuracy of model assumptions. Robust models are typically more resilient to these types of changes.

Models, even if good at risk-ranking an overall market segment, can be limited if they do not reflect the bank's population. A model is typically developed for a certain target population and may be difficult to adapt to other populations. In most cases, a credit scoring model should only be used for the product, range of loan size, and market that it was developed for. When a bank tries to adapt the model to a different population, performance of that population may likely deviate from expectation. When a bank implements or adapts a model to a new market or population for which it was not designed, examiners should determine whether management performs an analysis similar in scope to the one used to validate the model at implementation.

Credit scoring is good at predicting the probability of default but generally not at predicting the magnitude of losses. (Normally, other models, such as loss models, focus on predicting the level (magnitude) of risk.) Generic credit scoring models in particular most likely rank order the risk appropriately but generally do not accurately predict the level of the risk. Thus, banks that use generic models should not assume that their loss rates will be the same as those reflected in industry odds charts. How accounts ultimately perform depends on a number of factors, including account management techniques used, the size of line granted, and so forth.

Scorecards could be considered, by their very nature, to be antiquated when they are put into production. They are based on lengthy historic data and take time to develop. Moreover, models are calibrated using historical data, so if relevant un-modeled conditions change, the model can have trouble forecasting out of sample.

Along similar lines, during times of strong economic growth, models may be ill-prepared to predict borrower performance in recessionary conditions, particularly if the historic period observed did not include recessionary conditions. There are several behaviors that could impact the model's effectiveness in recessionary times. One is that consumers might prioritize their payments to pay off secured debt rather than unsecured debt. In hard times, this could leave a bank that is holding the consumer's unsecured credit card debt as one of the last to get paid, if paid at all.

The effectiveness of scoring models can also be limited by human involvement. For example, when models are augmented by managerial judgment (for instance, in the case of overrides), results from the model and subsequent validation processes can become seriously compromised. In addition, unsupported overconfidence in the models could lead some banks to move up or down market to make larger or more risky loans, respectively. Without proper model validation, such movements could result in the bank taking on more credit risk than it can control.

AUTOMATED VALUATION MODELS

Automated valuation models (AVMs) are sometimes used to support evaluations or appraisals. Examiners should look at management's periodic validation of AVMs for mitigating the potential valuation uncertainty in the model and should confirm whether its documentation covers analyses, assumptions, and conclusions. Validation includes back-testing a representative sample of the valuations against market data on actual sales (where sufficient information is available) and should cover properties representative of the geographic area and property type for which the tool is used. Many vendors provide a "confidence score" which usually relates to the accuracy of the value provided. Confidence scores come in many formats and are calculated based on differing systems. Examiners should determine whether management understands how the models work as well as what the confidence scores mean and should confirm whether management has identified confidence levels appropriate for the risk in given transactions.

SUMMARY OF EXAMINATION GOALS – SCORING AND MODELING

The examiner's role is to evaluate scoring, model usage, and model governance practices relative to the bank's complexity and the overall importance of scoring and modeling to the bank's credit card lending activities. The role includes:

- Identifying the types of scoring systems used in the credit card lending programs and whether the models are generic, custom, or vendor-supplied. A model inventory is normally available for review.
- Determining how management uses scores in its decision-making processes and whether each model's use is consistent with the intended purpose.
- Assessing whether designated staff possess the necessary expertise.
- Determining whether management thoroughly understands the models used.
- Reviewing cut-off scores and odds charts to assess the level of risk being taken.
- Testing the effectiveness of the bank's validation function by selectively reviewing various aspects of the bank's validation work for key models.
 - Evaluating the scope of validation work performed.
 - Reviewing reports summarizing validation findings and any additional workpapers necessary to understand findings.
 - Evaluating management's response to the reports, including remediation plans and timeframes.
 - Assessing the qualifications of staff or vendors performing the validation.
- Assessing the bank's calibration procedures, including documentation thereof.
- Determining whether credit bureau, behavior, and/or other scores enhance account management and collection practices.
- Assessing override policies and practices.
 - Review the number/volume and types of overrides.
 - Verify that override reports are reviewed by management and that performance is adequately tracked.
 - Determine the impact, if any, of overrides on asset quality.
- Assessing whether the bank's audit program appropriately considers models and oversight thereof.
- Identifying instances in which management has taken action when performance of the scoring model deteriorated and determine if the action was appropriate, effective,

- and timely.
- Determining if management is prepared to take future action if the scoring model's performance deteriorates.
- Determining if there are any models under development.
 - Identify potential impacts on the bank from implementation of the forthcoming models.
 - Understand what prompted the model development.
 - Ascertain the planned implementation date of the model.
- For models developed by third parties, assessing whether the systems are supervised and maintained in accordance with vendor-provided specifications and recommendations.

Examiners normally select models for review in connection with the examination when model use is vital or increasing. Focus may also be placed on models new or acquired since the prior examination. Quantitative or information technology (IT) specialists are sometimes needed for some complex models, but examiners normally can perform most model reviews.